0.1 Overview

In modeling constraint-based phonotactics, there are three broad kinds of decisions to be made: which framework to use, which parameters (constraints) to use, and how to tune (rank or weight) the parameters. This experiment will weigh in on the question of framework choice, using the function they use to combine the effects of violations to distinguish among them.

Several constraint-based frameworks for instantiating phonological grammars have been proposed. Optimality Theory (**?**) is a framework with strictly ranked constraints. This is in contrast with its predecessor Harmonic Grammar (**????**), which assigns weights to constraints. Linear Optimality Theory (**?**) is similar to Harmonic Grammar, but only allows constraints to assign penalties, not rewards; Harmonic Grammar, in contrast, can in principle allow both positively and negatively weighted constraints. Maximum Entropy (**?**) is a version of Harmonic Grammar that converts the harmony scores given by Harmonic Grammar into probabilities. **?** have used Maximum Entropy to assign probabilities to all the surface forms of a language, rather than the surface forms of a given input candidate.

One feature of constraint-based frameworks is that they recognize units of violation that are independent of the words they appear in. In other words, two different words can be said to contain the same violation, and one word can be said to contain multiple different violations. Thus, these frameworks all have some method of combining units of violation into a grammaticality score for the entire word. However, they differ in how they define the combination operation. As a result, they make different predictions about the relative ungrammaticalities of words whose violations are in a subset-superset relationship. This experiment will seek to distinguish between frameworks on the basis of such groups of words.

0.2 Optimality Theory

Optimality Theory (OT) predicts that adding mild violations to a word with a severe violation has no effect, so that the function from number of violations to grammaticality is flat for any given first violation as long as it remains one of the worst violations in the word. In other words, OT's function for combining the penalties of multiple violations is to return the maximum of those penalties as the penalty for the whole word (multiplied by the number of times that penalty is incurred).

Figure 0.2 illustrates by crossing a strong violation, [mr] in the onset, with a weak violation, [o:sp] in the rhyme. Each occurrence of the strong violation counts as a penalty of 2, while each occurrence of the weak violation counts as a penalty of 1. The penalty of each word is simply the maximum of the penalties of the violations in the word, so that *mroasp* is not any worse than *mrote*.

However, predating OT, **?** found that speakers have an above chance probability of preferring a word with one violation to a word with that same violation and a less severe one, suggesting that even the milder violations affect the grammaticality of a word. This contradicts OT's prediction that an additional violation that is lesser than the first violation will not affect the ungrammaticality of the word. Additionally, **?**

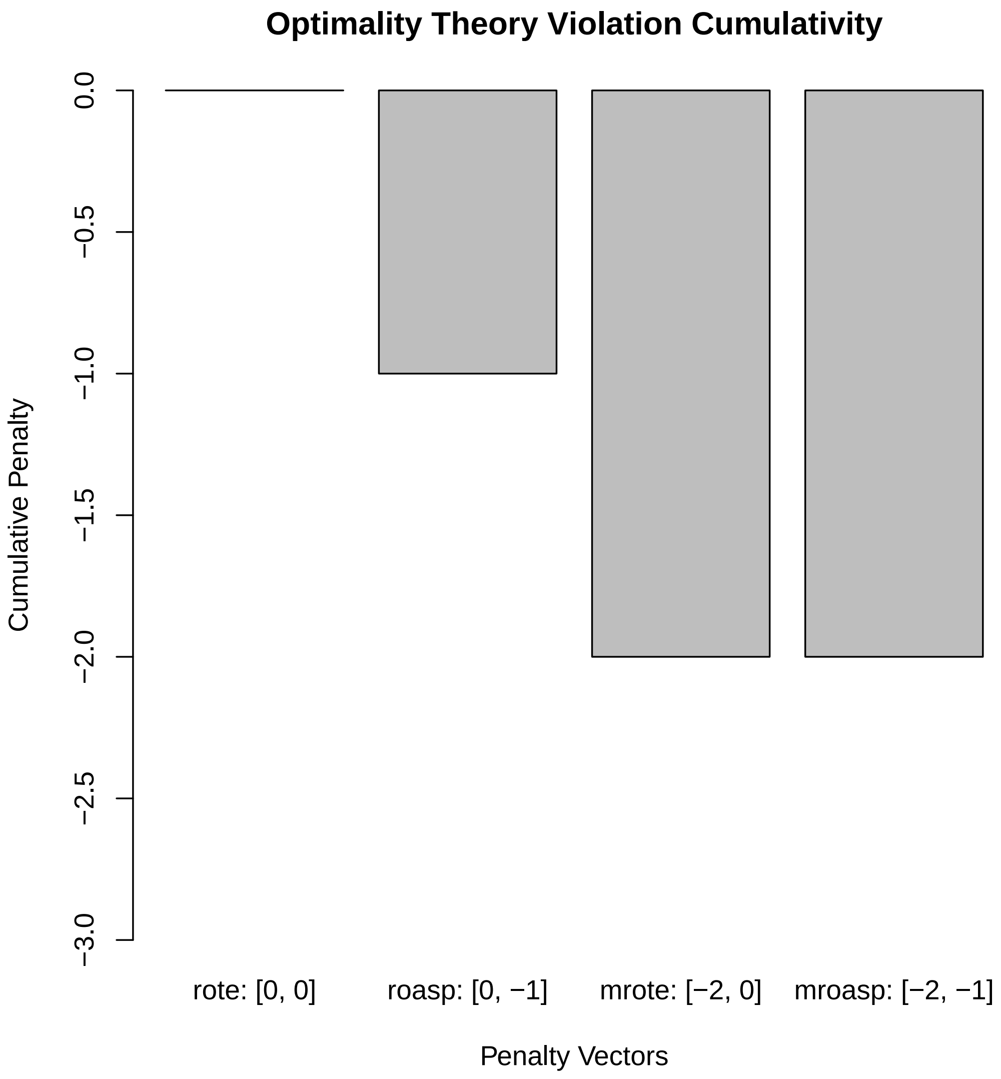

Figure 1. Cumulativity of violations in Optimality Theory.

found that a word like *mrupation*, with one severe violation followed by a common English sequence, was preferred to a word like *spleitisak*, with several minor violations. This is in contrast with OT's prediction that the strong violation *mr* matters more than any number of lesser violations. **?** found that in some cases, multiple violations of a constraint produce lower acceptability than a single violation in syntax, as well.

? designed experiments to directly test the question of cumulativity of violations, addressing potential alternative explanations for these two results, and found that models that take into account all violations of a word, not just its worst violation, fit the data significantly better. Albright used two types of words, those with phonotactic violations in the onset and those with not only phonotactic violations in the onset but also milder violations in the rime. In a variety of analyses, he fitted models that rate words by their worst violation only, and ones that rate words by the sum of all their violations. The models that take into account all violations in the word were more strongly correlated with experimental findings. This study showed that cumulative models reflect speaker judgments better than noncumulative models, but did not distinguish among various cumulative models.

I conclude that OT's strategy of combining violations by finding their maximum is not empirically supported in the domain of phonotactics, and I turn to Harmonic Grammar and Maximum Entropy.

0.3 Linear vs. Exponential Combination

? found evidence that all violations in a word contribute to the word's ungrammaticality, but it did not compare various models that work this way against each other. The shape of the curve relating number of violations to phonotactic judgments bears on the question of which framework we should use to model phonotactic well-formedness. As **?** points out, Harmonic Grammar predicts a well-restricted set of cumulativity effects, unlike Optimality Theory with Local Constraint Conjunction (**?**). But the weighted constraints of Harmonic Grammar can be combined in a linear fashion, producing the framework commonly associated with the name, or exponentiated and normalized, as in Maximum Entropy. These approaches predict differently shaped curves.

(1) Harmonic Grammar: The harmony $\mathcal{H}$ of a word x is the dot product of the violation vector v, representing violations of x on each constraint in the constraint set C, with the constraint weight vector w.

$$\mathcal{H}(x) = \sum_{i \in C} v_i w_i$$

(2) Maximum Entropy: The probability p of a word x is the exponentiated negative harmony of the word, normalized relative to the candidate set X.

$$p(x_i) = \frac{\exp(-\mathcal{H}(x_i))}{\sum_{j \in X} \exp(-\mathcal{H}(x_j))}$$

Consider a constraint with a weight of two and candidates A, B, C, and D that violate it zero, one, two, and three times respectively. In linear Harmonic Grammar, the candidates have the harmony scores 0, -2, -4, and -6; they decrease by two each time, in a linear pattern. In Maximum Entropy, if we assume these candidates exhaust the possibilities, they have the probabilities 0.865, 0.117, 0.015, and 0.002; candidate A has the majority of the probability because it is the best choice available, and each additional violation decreases the probability by a smaller amount than the last.

0.4 Experiment 1

This experiment uses evidence from the accumulation of violations to distinguish between Harmonic Grammar and Maximum Entropy as models of phonotactic knowledge. I predict that, in accordance with the Maximum Entropy model, a violation in the presence of other violations will have a smaller effect on the grammaticality of the word than it would have in isolation, as the grammaticality approaches a floor.

0.4.1 Method

In order to test the prediction made by the Maximum Entropy model, I gathered acceptability data on words much like the ones used in the examples above: words with no obvious violations, words that are the same except with the addition of a violation in the onset, words the same as the first group except with a violation in the coda, and words with both the onset and coda violation.

0.4.1.1 Participants

One hundred participants were recruited from Mechanical Turk. They were each paid $0.75 on the assumption that the experiment took 5 minutes to complete. They were all located in the United States and claimed to be over 18 years old. In order to maintain a level of consistency in the participant pool, I ran the experiment only on weekdays between the hours of noon and 5pm Eastern time, which corresponds to regular workday hours in the four continental US timezones.

Participants were excluded if they were not native speakers of English and if their data was suspect. Native status was determined by answers to two demographics questions: one asking their native language and one asking the language they use at home. Participants were only included if English was given in response to both of these questions. Other languages given in addition to English were not considered reason for exclusion. Two participants were excluded on the basis of native language.

The quality of the data was assessed in a variety of ways.

First, there were twelve filler trials. These had the same form as test trials, a yes/no question about the acceptability of a nonce word in English. However, six words were constructed to be more English-like than even the best test words, and six were constructed to be less English-like than even the worst test words. Participants were excluded if they accepted good fillers equally as often as bad fillers,

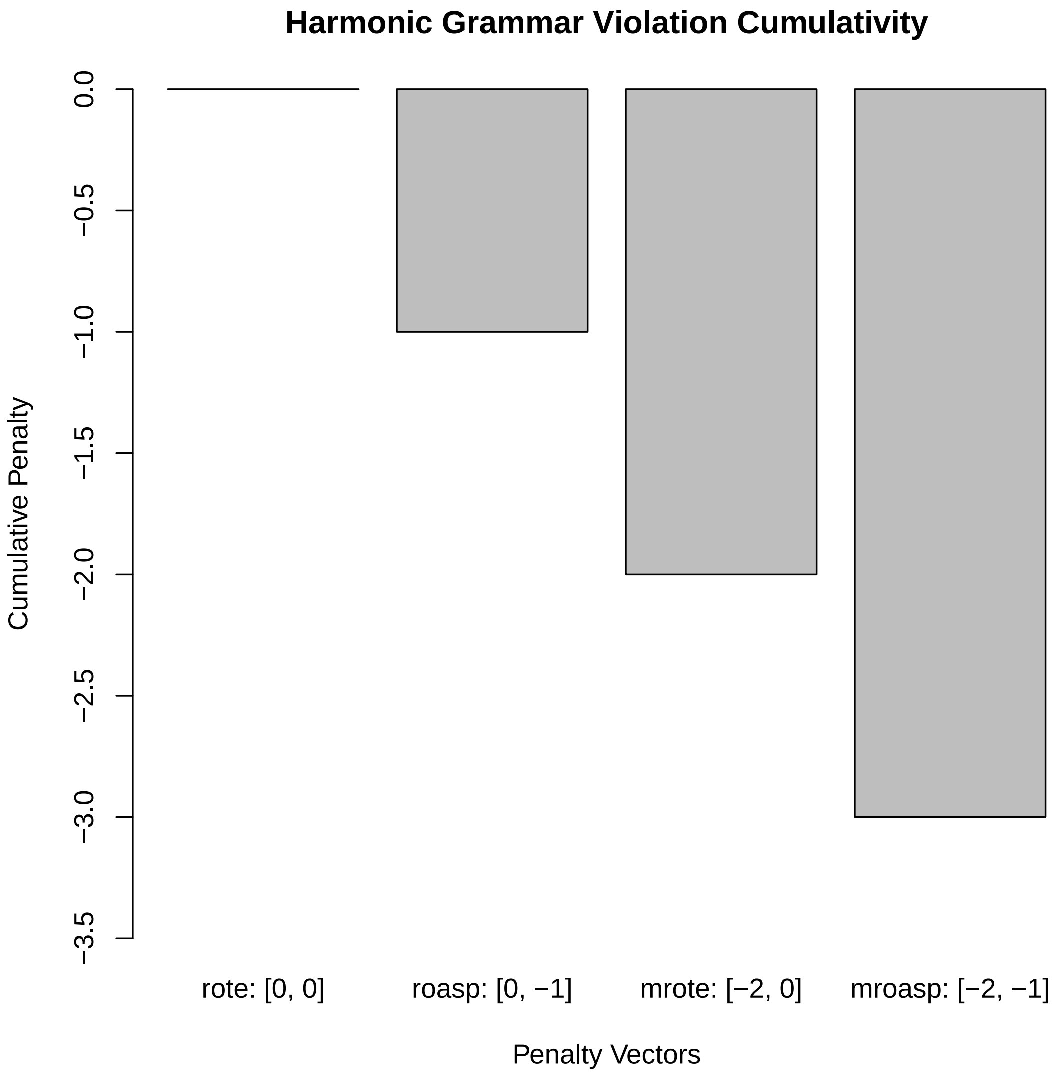

Figure 2. Cumulativity of violations in Harmonic Grammar.

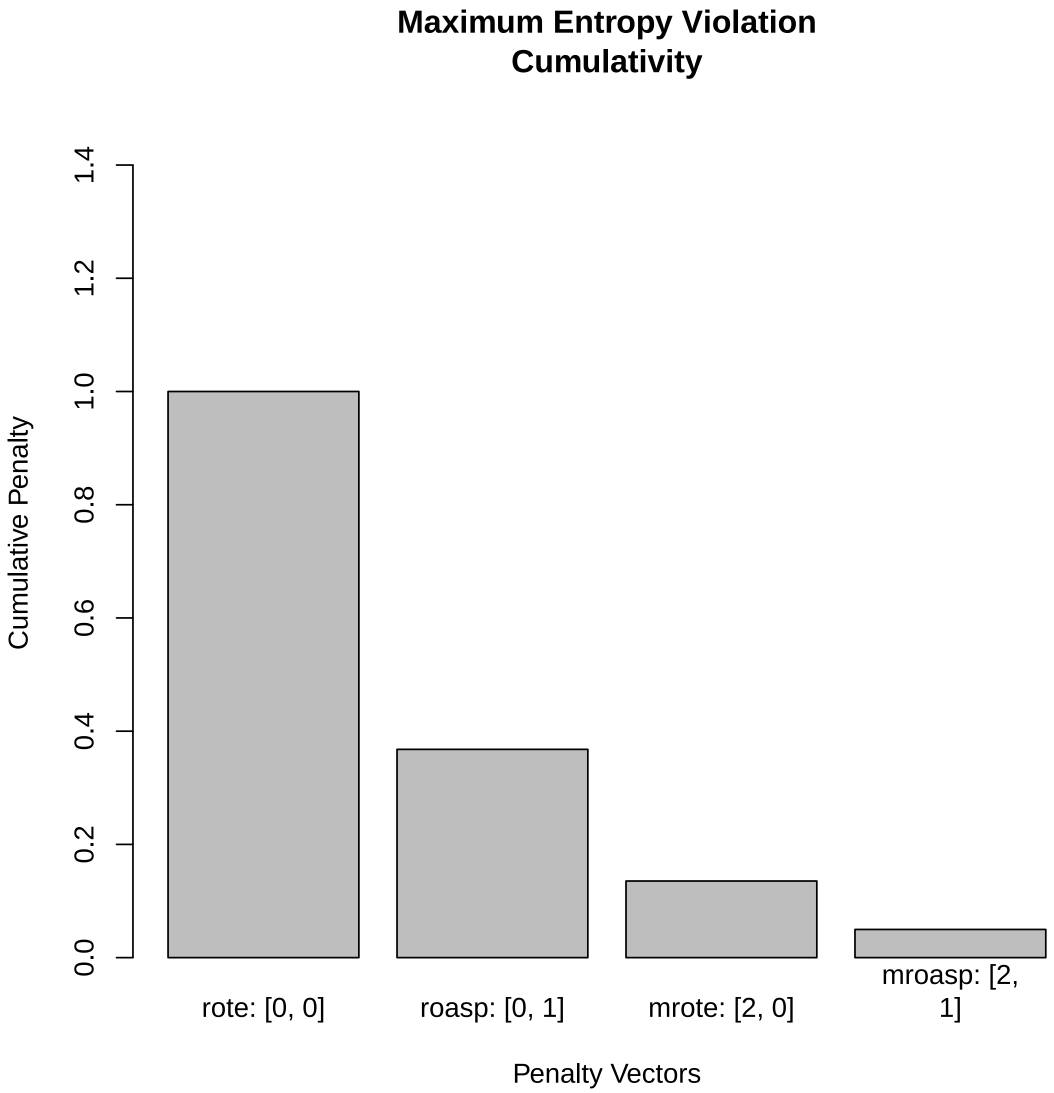

Figure 3. Cumulativity of violations in Maximum Entropy grammar.

indicating that they may not have been paying attention or answering carefully. Four participants were excluded for this reason. No participants accepted bad fillers more often than good fillers.

Second, any participant who consistently chose whichever option was on a particular side of the screen was excluded. "Consistently" was defined as more than 90% of test questions. No participants fell into this category.

The third exclusion criterion was answer speed. In pilot data, the fastest reaction times were over 300ms, so times under 50ms suggest that a computer program is clicking through the experiment automatically. No participants were found to have this behavior.

After these exclusions, the data from 94 participants was used in the analysis.

0.4.1.2 Materials

The words were presented orthographically. The benefit of visual presentation is that participants are much less likely to fail to perceive the violations. There is a large body of evidence that speakers misperceive certain sound combinations that would be severe phonotactic violations in their native language. The result is that their behavioral data does not reflect the presence of the violation (**??????**).If some violations were not perceived, the results of the experiment would be compromised, and indeed, some of the violations used in this study, such as [tl], are known to be among those misperceived by English speakers (**??**).The downside of visual presentation is that we are testing orthotactics more directly than phonotactics. The materials were designed to minimize the ambiguity in the relationship between spelling and sound in order to mitigate this problem as much as possible.

Each of 24 test items appeared in four conditions. The four conditions are created by crossing two factors: presence of an onset violation and presence of a rime violation.

For each item, there is a unique good onset, bad onset, good coda, and bad coda. There is also a vowel assigned to the item. The four conditions of the item are created by taking all combinations of one onset, the vowel, and one coda.

The good onsets are all biconsonantal or triconsonantal and attested in native English words. The bad onsets are all biconsonantal, though some are represented by three letters, such as *thl*. The bad onsets are not attested in native English words, although some are found in foreign proper nouns, such as *Sri Lanka* and *Vladimir*.

The good codas consist of one consonant, sometimes spelled with two letters, such as *ss* or *ck*. The bad codas are biconsonantal. These codas are not attested in native English words, and are believed to violate English phonotactics. Many of them have sonority profiles not allowed in English, but extremely bad sonority profiles — those with a sonorant followed by an obstruent — were avoided to keep participants from mentally inserting a schwa into the coda cluster, breaking it into two syllables.

The vowels are taken from the set {*a*, *e*, *i*, *o*, *u*, *oo*}, in order to have a uniform distribution of orthographically distinct vowels across the items.

Items were created randomly from these components, and then altered to avoid any actual English words.

(3) Example test item

a. Good onset, good coda (GG): plag
b. Bad onset, good coda (BG): tlag
c. Good onset, bad coda (GB): plavb
d. Bad onset, bad coda (BB): tlavb

A Latin square design was applied to the test items so that each participant only saw one word from each item set.

Due to the Latin square, each participant saw six words from each condition. Accordingly, twelve fillers were added to the set of materials: six "good" fillers and six "bad" fillers. The good fillers were words expected to be even more English-like than the GG words, as they have real English suffixes added to them. **?** suggests that this addition will increase acceptability. The bad fillers were words expected to be even less English-like than the BB words, because they had longer violation-containing clusters. These fillers serve two purposes: to reduce the chances of ceiling and floor effects in the test words by giving participants examples of more extreme acceptabilities, and to show when participants are answering randomly or inattentively. Participants who are attentive should accept good fillers more often than they accept bad fillers.

The full set of materials is provided in Appendix I.

0.4.1.3 Procedure

The experiment was built using Speriment and run using psiTurk (**?**) to interact with Mechanical Turk.

The task is a two-alternative forced choice task between the responses "Yes" and "No" as answers to the question "Based on how it sounds, do you think this word could be a word of English?" followed by one of the stimuli.

Participants indicated their choice by pressing a key on their keyboard: 'f' for the choice on the left and 'j' for the choice on the right. The order in which the test and filler words were presented varied randomly across items and participants.

0.4.2 Results

As expected under any model, participants accepted the test item the most often when it had no violations, less often when it had one violation, and very rarely when it had two violations. Of the two types of words containing one violation, those with a coda violation and those with an onset violation, coda violation words were chosen less often, suggesting that the coda violations were on average more egregious than the onset violations. Thus, there is a total ordering of word types, from those with no violations (called GG for good onset, good coda), to those with an onset violation (BG for bad onset, good coda), to those with a coda violation (GB for good onset, bad coda), to those with two violations (BB for bad onset, bad coda).

The question this experiment seeks to answer is not about the ranking of the percentages for each condition, but the quantitative relationships among them. In order to analyze these relationships, we need a linking hypothesis to map from predictions

Table 1. Percent acceptance by condition.

Condition	Percent ‘Yes’
GG	83
BG	30.9
GB	17
BB	6.6

about psychological states to predictions about performance on the task. I hypothesize that participants use the output of whichever model they are using — harmony in Harmonic Grammar and probability in Maximum Entropy — as the input to a probabilistic process that governs whether they choose ‘yes’ or ‘no’ on the task. There are many forms this probabilistic process could take; I will adopt the assumption that the percent of times a participant accepts a word can be treated as a direct proxy for the output of the model for that word. That is, I assume that if participants calculate probabilities for words, they say ‘yes’ to a word with the same probability they assign to the word, and if they calculate harmonies, they say ‘yes’ with a probability that is proportional to the harmony the assign the word. I assume that the scaling necessary to convert harmonies to probabilities in the case of Harmonic Grammar is constant across words, so that the differences in the percents of ‘yes’ answers is the same, modulo the noise of the probabilistic process, as the differences in the harmonies. This is equivalent to normalizing the harmonies for all stimuli for a given subject and item set consistently. With our current limitations in understanding the transformations that apply to model outputs as they are used to direct behavior in an experimental task, the results of this experiment must be interpreted as dependent on this assumption. A more conclusive understanding of the phenomenon will depend on future studies that investigate it using different tasks to determine if the findings are dependent on a particular linking hypothesis, or if they are robust to different ways of framing the question and to the particulars of different experimental tasks.

The difference between the model outputs of GG words and another category of words can be viewed as the penalty for the violation(s) in the second category of words.

Harmonic Grammar predicts that the penalty for BB is equal to the sum of the penalties for BG and GB, while Maximum Entropy predicts that the penalty for BB is less than this sum. Table 1 shows that descriptively, the data appear to support the prediction of a Maximum Entropy model, as the penalties decrease at each step. Both report the percent of the time that participants responded ‘yes’ to a word, for each type of word.

The hypothesis that participants are employing Maximum Entropy rather than Harmonic Grammar predicts that there will be an interaction between the effects of onset violations and coda violations.

If a positive interaction is present, this would support a model that takes probability away from words in greater and greater amounts as the number of violations increases. I do not know of such a model, and this result is not predicted.

A negative interaction would support models like Maximum Entropy, in which each additional violation subtracts less probability from the word than the last violation did.

A linear relationship would support models like linear Harmonic Grammar, in which each additional violation subtracts the same amount of probability from a word as the last violation did.

Figure 4 shows the interaction of the effects of coda violations and onset violations on ‘yes’ responses. The difference between the slopes of the two lines shows that the addition of a coda violation decreases the acceptance rate less when an onset violation is already present than when it is not, as predicted by Maximum Entropy.

The interaction plot doesn’t show the shape of the distributions, though, so we can also break down the data by subject and item in order to view violin plots of the conditions. A violin plot is a box plot where the sides of the boxes are replaced with kernel density plots. Thus, the white dots represent the median subject or item mean, the black bars represent the interquartile range — the values in the medial two quartiles of the data. Figure 5 is a violin plot where each data point is the percent of ‘yes’ answers given to that word type by a particular participant. It shows us the distribution of participants for each condition. Figure 0.4.2 is a similar plot where the aggregation is done by item set. Recall that an item set is a set of four test words, one in each condition, where the words share substrings and constraint violations.

The violin plots reflect the same interaction as the interaction plot, indicating that it is not depending on viewing the data in a particular way. To test the significance of this interaction, a mixed effects model was fitted to the data. The dependent variable was the percent of ‘yes’ responses. OnsetViolation, CodaViolation, and their interaction served as fixed effects, in addition to TrialNumber and YesPosition, representing whether the ‘Yes’ option was presented on the left or right side of the screen on a particular trial. Random intercepts for subject and test item and random slopes for OnsetViolation, CodaViolation, and their interaction by subject and item were included in the model.

```
library(lme4)
cdata$OnsetViolation[which(cdata$OnsetViolation == 0)] = -1
cdata$CodaViolation[which(cdata$CodaViolation == 0)] = -1
cdata$TrialNumber = cdata$TrialNumber - mean(cdata$TrialNumber)
cdata$YesPosition = cdata$YesPosition - mean(cdata$YesPosition)
#didn’t converge but close, add iterations
full_model = glmer(Response ~ OnsetViolation * CodaViolation + TrialNumber +
YesPosition + (1|Subject) + (0+OnsetViolation|Subject) + (0 + CodaViolation |
Subject) + (0+ OnsetViolation:CodaViolation|Subject) + (1|Item) +
(0+OnsetViolation|Item) + (0 + CodaViolation | Item) + (0+
OnsetViolation:CodaViolation|Item), data = cdata, family =
binomial(link="logit"), glmerControl(optCtrl=list(maxfun=10000), optimizer =
"bobyqa" ) )
```

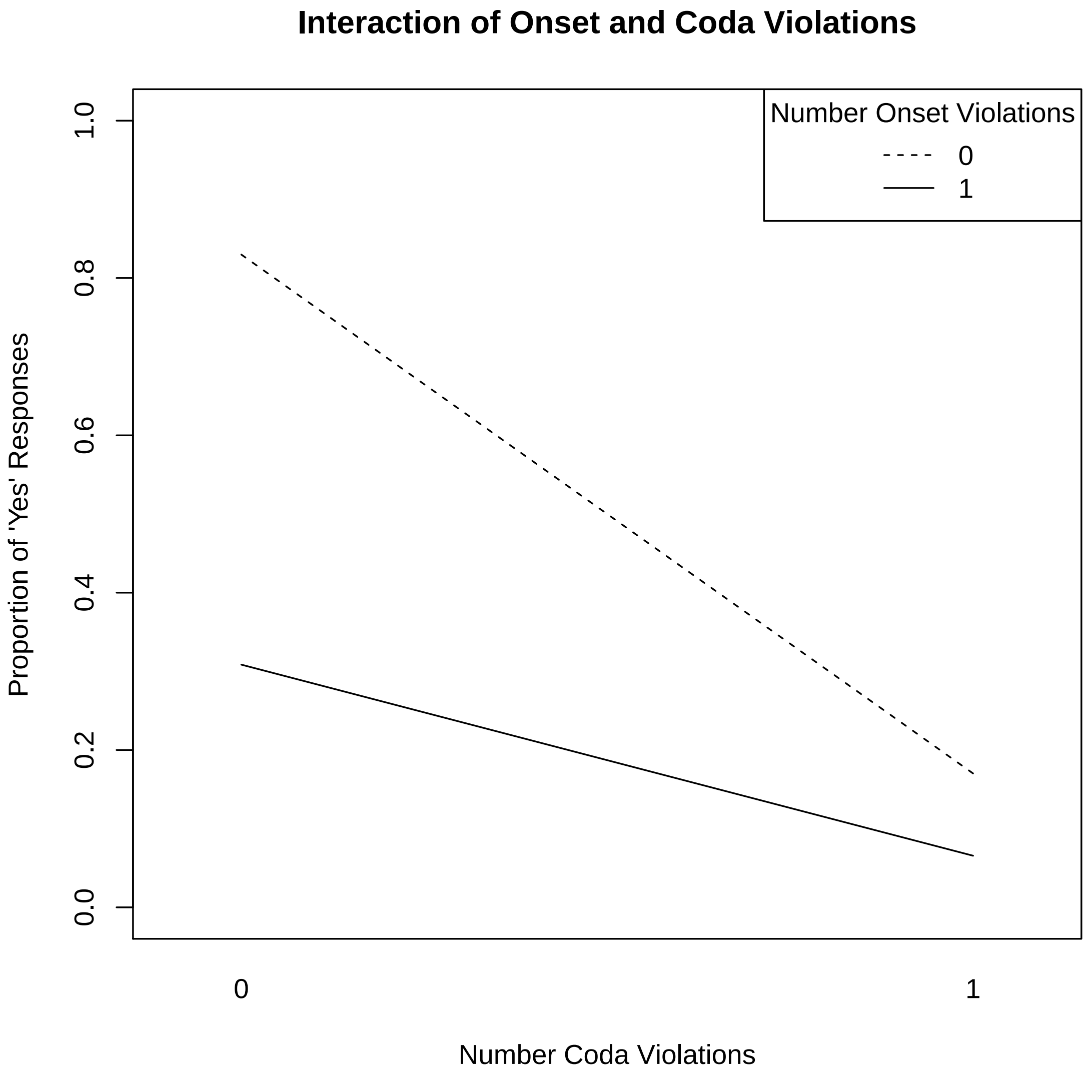

Figure 4. Interaction between onset and coda violations in predicting acceptance rate in Experiment 1.

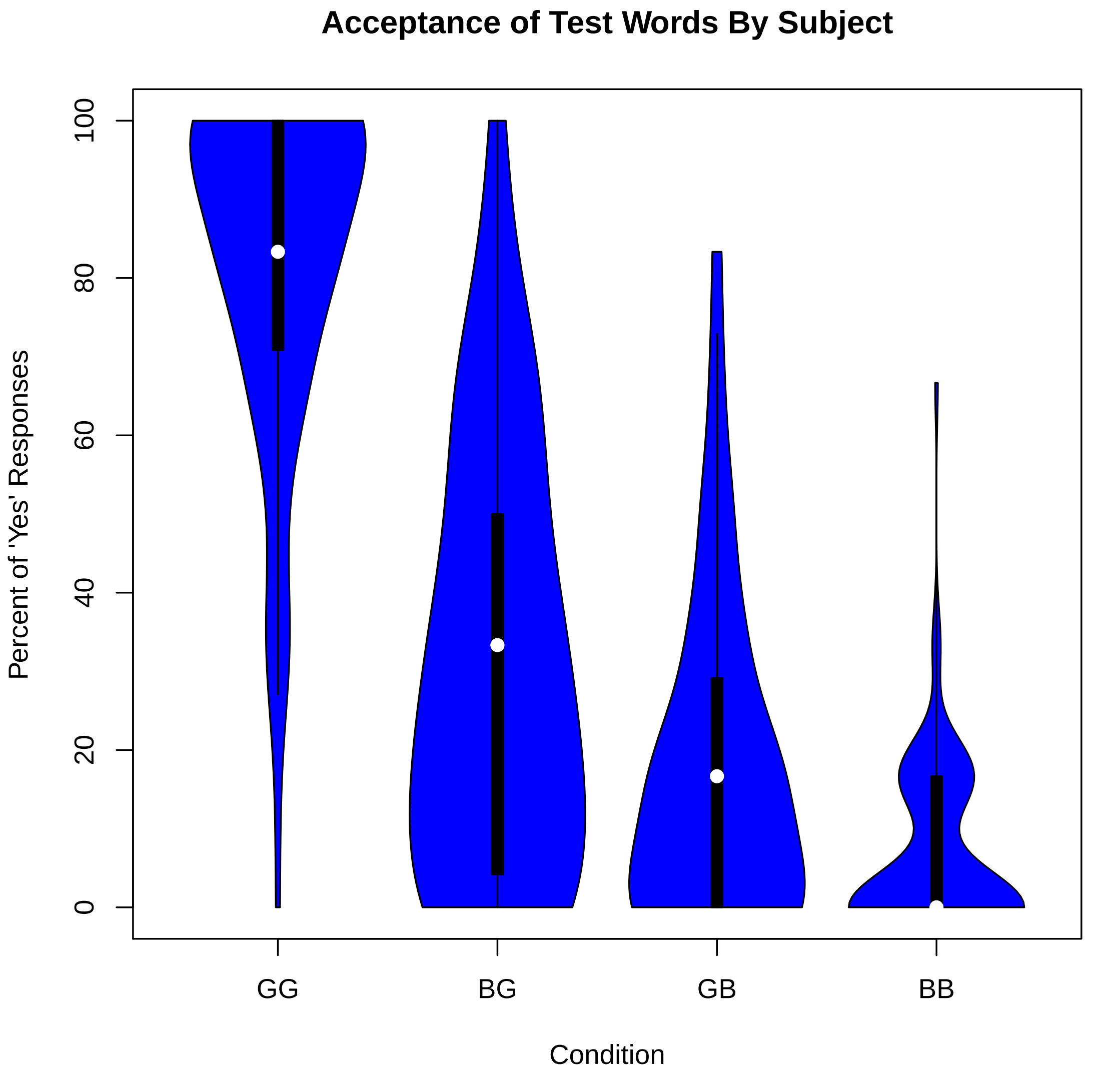

Figure 5. Distribution of mean percent acceptance by subject for each condition in Experiment 1.

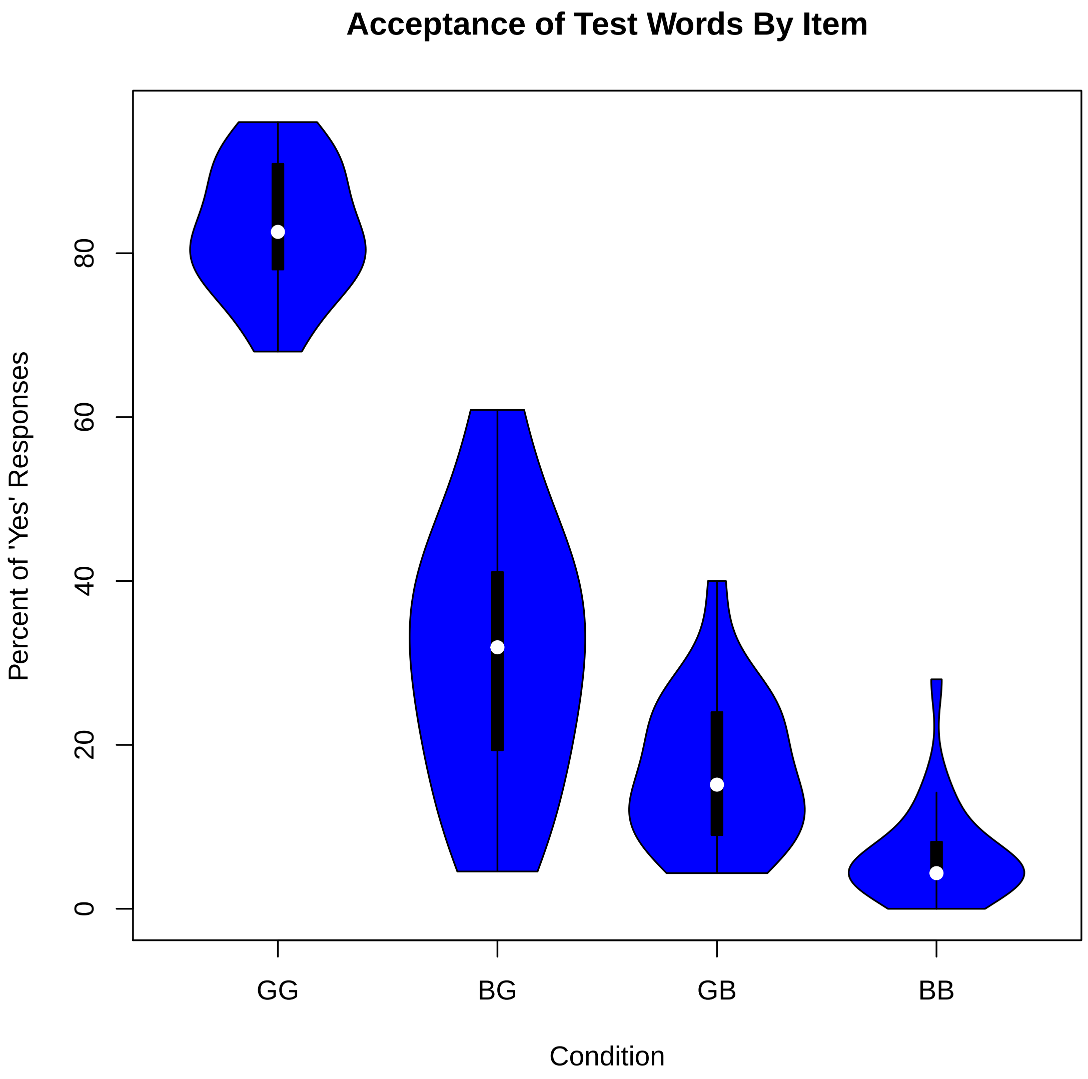

Figure 6. Distribution of mean percent acceptance by item for each condition in Experiment 1.

Table 2. Coefficients of the mixed effects model in Experiment 1.

Factor	Estimate	p-value
Intercept	-0.87	4.5767767×10^{-6}
OnsetViolation	-0.84	7.4229744×10^{-8}
CodaViolation	-1.34	$4.2921481 \times 10^{-17}$
OnsetViolation:CodaViolation	0.82	4.9355623×10^{-7}
TrialNumber	0.01	0.2541263
YesPosition	-0.02	0.8428708

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,
:  Model is nearly unidentifiable:  large eigenvalue ratio
##  - Rescale variables?
```

(4) Mixed effects model formula
$, Response, OnsetViolation * CodaViolation + TrialNumber + YesPosition +$
$(1|Subject) + (0 + OnsetViolation|Subject) + (0 + CodaViolation|Subject) +$
$(0 + OnsetViolation : CodaViolation|Subject) + (1|Item) + (0 + OnsetViolation|Item) +$
$(0 + CodaViolation|Item) + (0 + OnsetViolation : CodaViolation|Item)$

Table 2 gives the coefficients of the mixed effects model. It shows significant effects of both OnsetViolation and CodaViolation, meaning that the presence of an onset violation significantly decreases acceptance rate, as does the presence of a coda violation. Additionally, it shows a significant subadditive interaction between the two, showing that the combination of both violations is less potent than would be expected if they acted completely independently. This is in line with the Maximum Entropy prediction.

```
    cc = summary(full_model)$coefficients
    #
```

In Figure 5, we see that the distribution for BB appears cut off at the bottom. This suggests a floor effect: the BB words are so unacceptable that we cannot get an accurate sense of how unacceptable they are, because we already hit zero percent ‘yes’ responses, and it’s not possible to give a negative number of ‘yes’ responses. This result is problematic for the interpretation of this experiment, which hinges on an accurate estimation of the acceptability of BB words. If participants are using an HG-like grammar, could it map to a negative number of ‘yes’ responses, which is then forced up to zero, making it appear that they employed a MaxEnt-like grammar instead?

One problem with this alternative interpretation is that it’s unclear how harmonies would map to a negative number of intended ‘yes’ responses. We would need a more sophisticated linking hypothesis than the one offered above, which assumed

Table 3. Central tendencies of log-transformed reaction times for rejecting GB and BB words in Experiment 1.

Condition	Mean	Median
Rejected GB	7.8901758	7.786126
Rejected BB	7.8504456	7.7454356

that harmonies are transformed into probabilities in a way that preserves their proportions. Implicit in that assumption was the idea that they were normalized to fit into the probability space and translated into positive space, keeping the proportions among them constant. Perhaps they are scaled but not fully pushed into the positive range, but it remains unclear what that would mean for our hypothesized probabilistic process to have a target of a negative number. The fact that harmonies must be shoehorned into a positive space may be the very motivation for a Maximum Entropy-like model which transforms harmonies and results in the attentuation of differences at the bottom of the scale.

Furthermore, we might expect a more dramatic floor effect, and one consistent whether we look at the data aggregated by subject or by item, if indeed the subjects intended to assign BB words a harmony score as low as HG would predict for this data.

One clue to participants' true assessments of the BB words may lie in their reaction times. If BB words were far worse, not just slightly worse, than BG and GB words, and participants merely ran out of ways to express this, we might expect them to have shorter reaction times in responding to BB words than to BG and GB words, because it was so obvious that they should be rejected. In particular, we would expect shorter reaction times for rejecting BB words than for rejecting the second worst category, GB words. In fact, however, the violin plots for log-transformed reaction times for rejected GB and BB words look fairly similar, as shown in Figure 7.

```
rtt = t.test(log(rejectgb), log(rejectbb), alternative = 'greater')

#rt_model = lmer(log(ReactionTime) ~ Condition + TrialNumber + YesPosition +

#rt_model2 = lmer(log(ReactionTime) ~ Condition + TrialNumber + YesPosition
```

The mean and median reaction times for BB words are lower than those for GB words, as given in Table 3, but a one-tailed t-test finds that the difference is not significant, with a t-score of 0.9745602 and a p-value of 0.16501. Thus, reaction times do not offer support for the Harmonic Grammar hypothesis. However, we cannot accept the null hypothesis that BB and GB reaction times are the same. The evidence of a floor effect is inconclusive, motivating a second experiment to both replicate the interaction and differentiate it from a task effect.

```
rejectgb = rtdata[rtdata$Condition == 'GB' & rtdata$Response == 0,]$ReactionTi
rejectbb = rtdata[rtdata$Condition == 'BB' & rtdata$Response == 0,]$ReactionTi
vioplot(log(rejectgb), log(rejectbb), col='blue', names = c('GB', 'BB'))
title(main = 'Log Reaction Times for Rejected Words',
          xlab = 'Condition',
          ylab = "Distribution of Log Reaction Times (ms)")
```

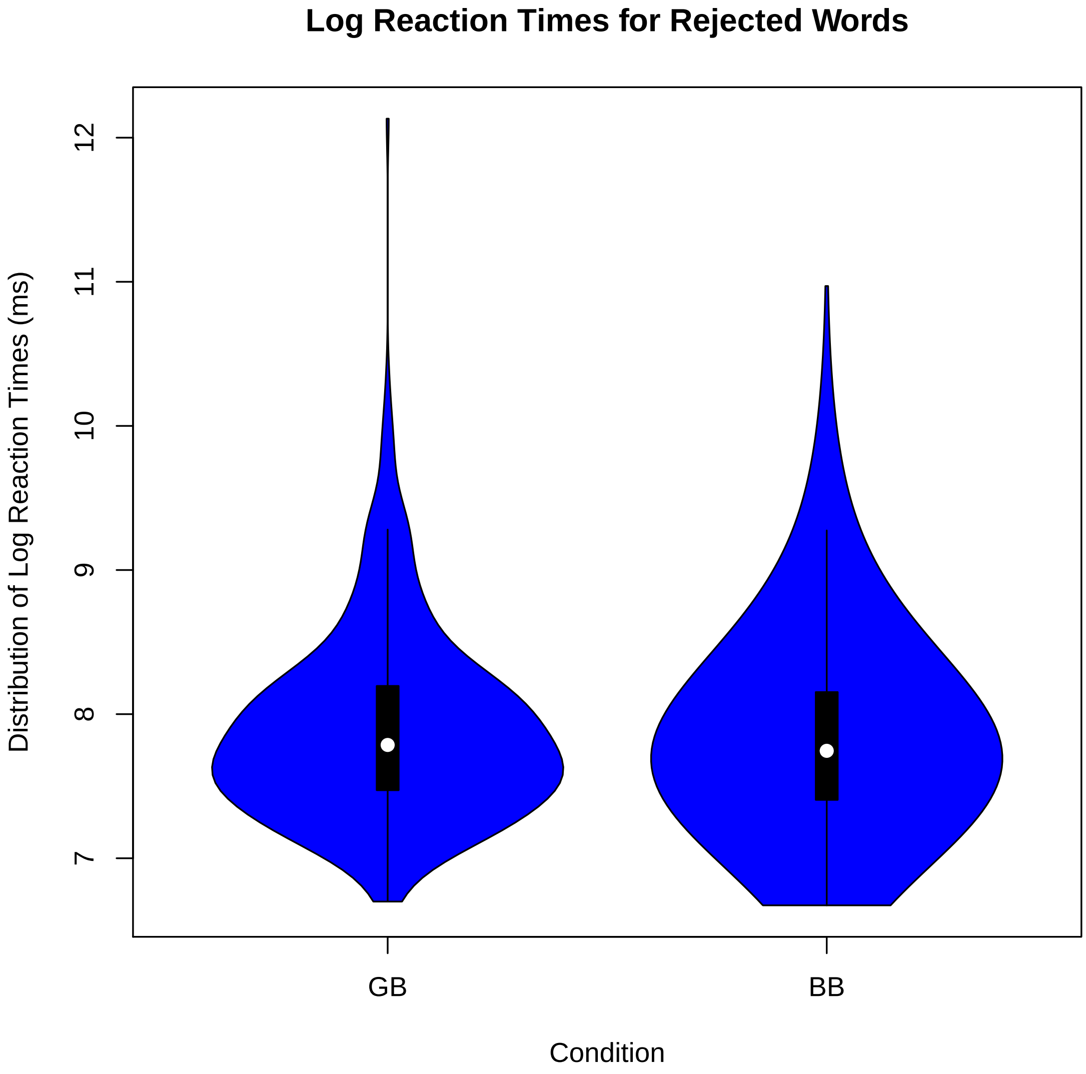

Figure 7. Distribution of log reaction times for rejecting GB and BB words in Experiment 1.

0.4.3 Discussion

The results support the hypothesis that Maximum Entropy predicts participants' preferences better than Harmonic Grammar. The interaction between the effect of an onset violation and the effect of a coda violation was both statistically significant and of a considerable effect size. However, it is possible that this is a task effect, where the inability to accept BB words a negative number of times artificially inflated its measured acceptability, creating the subadditive interaction.

0.5 Experiment 2

Experiment 1 showed evidence of a floor effect, which interfered with the ability to interpret the results as conforming to the prediction of Harmonic Grammar or Maximum Entropy. Experiment 2 seeks further evidence to distinguish between these theories by using different materials to make a floor effect less likely and increase our ability to distinguish between a floor effect and an accurate measurement of acceptability. Experiment 2 is largely the same as Experiment 1, but with the materials altered to increase the distance in acceptability space between the violation-containing test words and the violation-containing fillers. This adjustment may encourage participants to distribute their responses differently, avoiding the low extreme of the acceptability space when assessing BB words, and it will also increase the likelihood that we can differentiate statistically between the BB words and an even less acceptable category of words, the bad fillers. If there is a category below the BB words, then they must not be at the "floor" of the task's measurable acceptability values.

0.5.1 Method

0.5.1.1 Participants

One hundred participants were run on Mechanical Turk, each paid $0.75, as the experiment was expected to take the same amount of time as Experiment 1. Two participants were excluded for answering too quickly, implying they may be automated workers or not paying attention. Three participants were excluded for not being native or regular speakers of English. A total of 96 participants were included.

Participants were not excluded from this experiment on the basis of their responses to filler words, because the fillers serve a different purpose in this experiment and doing so could bias the results.

0.5.1.2 Materials

The materials for Experiment 2 have the same structure as those in Experiment 1 but differ in their exact makeup.

In order to construct these materials, the results of Experiment 1 were analyzed. The mean acceptance rates of each BG word and each GB word were calculated and sorted. The least preferred half of the bad onsets and the least preferred half of the bad codas were removed. In order to construct the same number of test words, the

Table 4. Percent acceptance in Experiment 2 by condition.

Condition	Percent 'Yes'
GG	88.7
BG	46.9
GB	29
BB	14.6

remaining bad onsets and bad codas were each used in two items. The good and bad onsets and codas were shuffled and recombined into new nonce words, which were filtered for real words or words with noticeable OCP violations.

The good fillers were the same as in Experiment 1: nonce words with no known violations and real English suffixes. The bad fillers were made worse. They were extended to two syllables long, with violations in initial, medial, and final consonant clusters.

0.5.1.3 Procedure

The procedure in Experiment 2 was the same as in Experiment 1, except that an additional instructional page was used. This page was intended to anchor participants' expectations for acceptable and unacceptable words by giving them an example of a word they would choose "Yes" for (a word constructed in the same manner as the good fillers) and an example of a word they would choose "No" for (a word constructed in the same manner as the bad fillers).

0.5.2 Results

Figure 0.5.2 shows the familiar pattern of evidence for the interaction between OnsetViolation and CodaViolation. It also supports the idea that the BB condition could have gotten lower acceptability scores on this task than it did, because the bad fillers have a noticeably lower distribution.

As in Experiment 1, a mixed effects model was fitted to the data. The model did not converge when fixed effects for trial number and the position of the "Yes" option on the page were included. The summary of the nonconvergent model indicated that their effects were non-significant. They were removed. The resulting model, including the fixed effects of theoretical interest and a full random effects structure for those fixed effects, did converge, and showed significant effects of OnsetViolation, CodaViolation, and their interaction. As predicted by Maximum Entropy, and in agreement with Experiment 1, the interaction is subadditive. The coefficients of the model are given in Table 5.

```
cfdata$OnsetViolation[which(cfdata$OnsetViolation == 0)] = -1
cfdata$CodaViolation[which(cfdata$CodaViolation == 0)] = -1
cfdata$TrialNumber = cfdata$TrialNumber - mean(cfdata$TrialNumber)
```

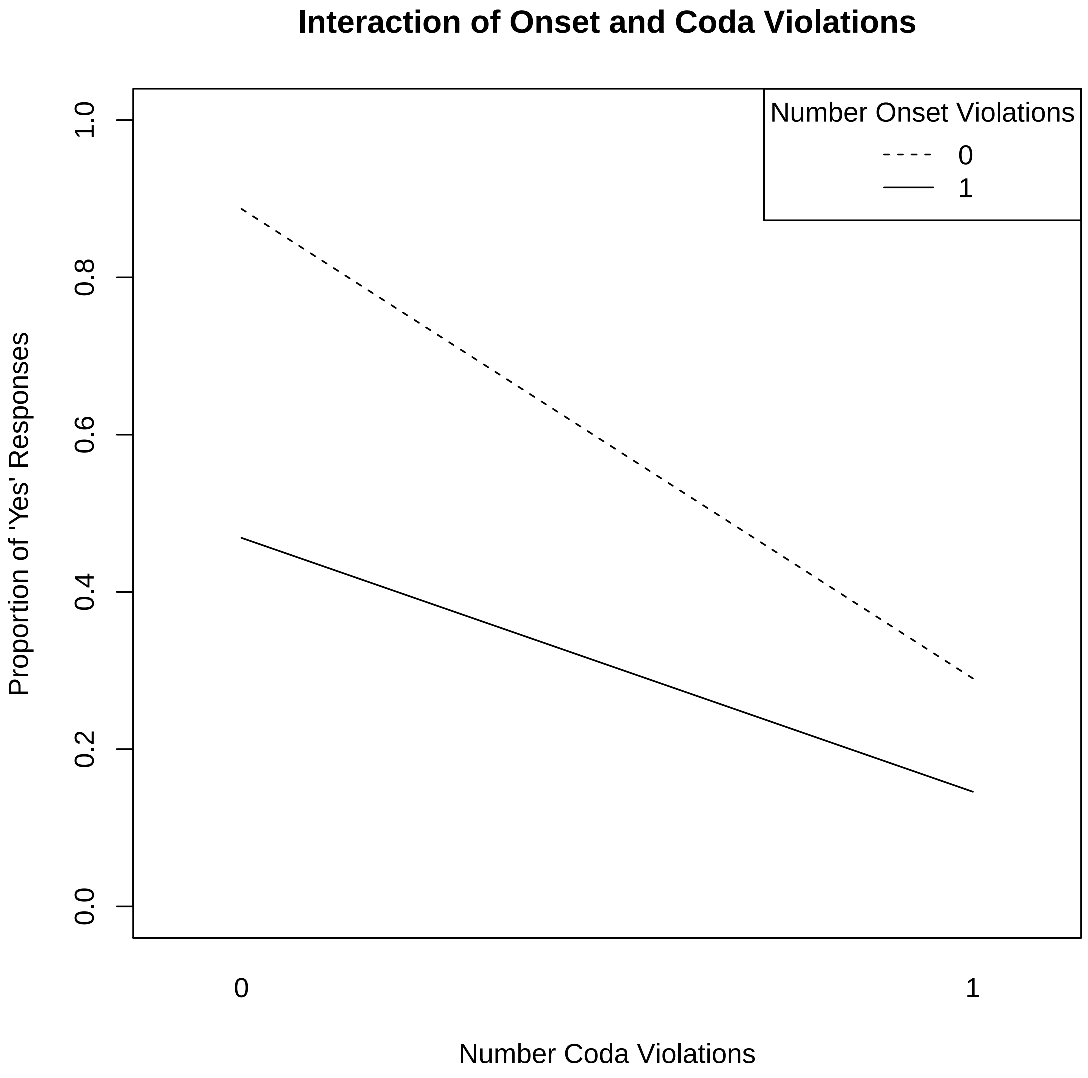

Figure 8. Interaction between onset violations and coda violations in Experiment 2.

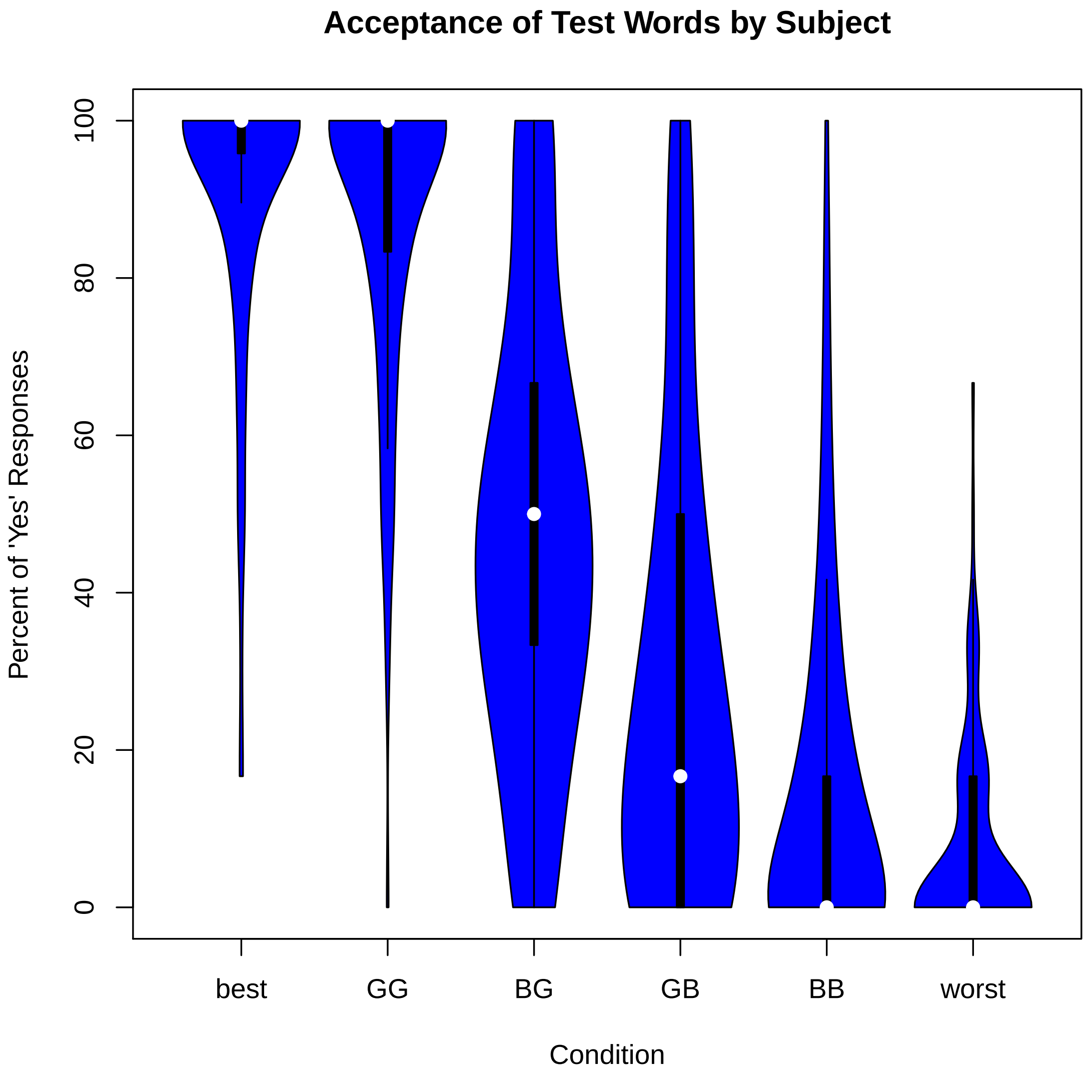

Figure 9. Distribution of mean percent acceptance by subject for each condition in Experiment 2.

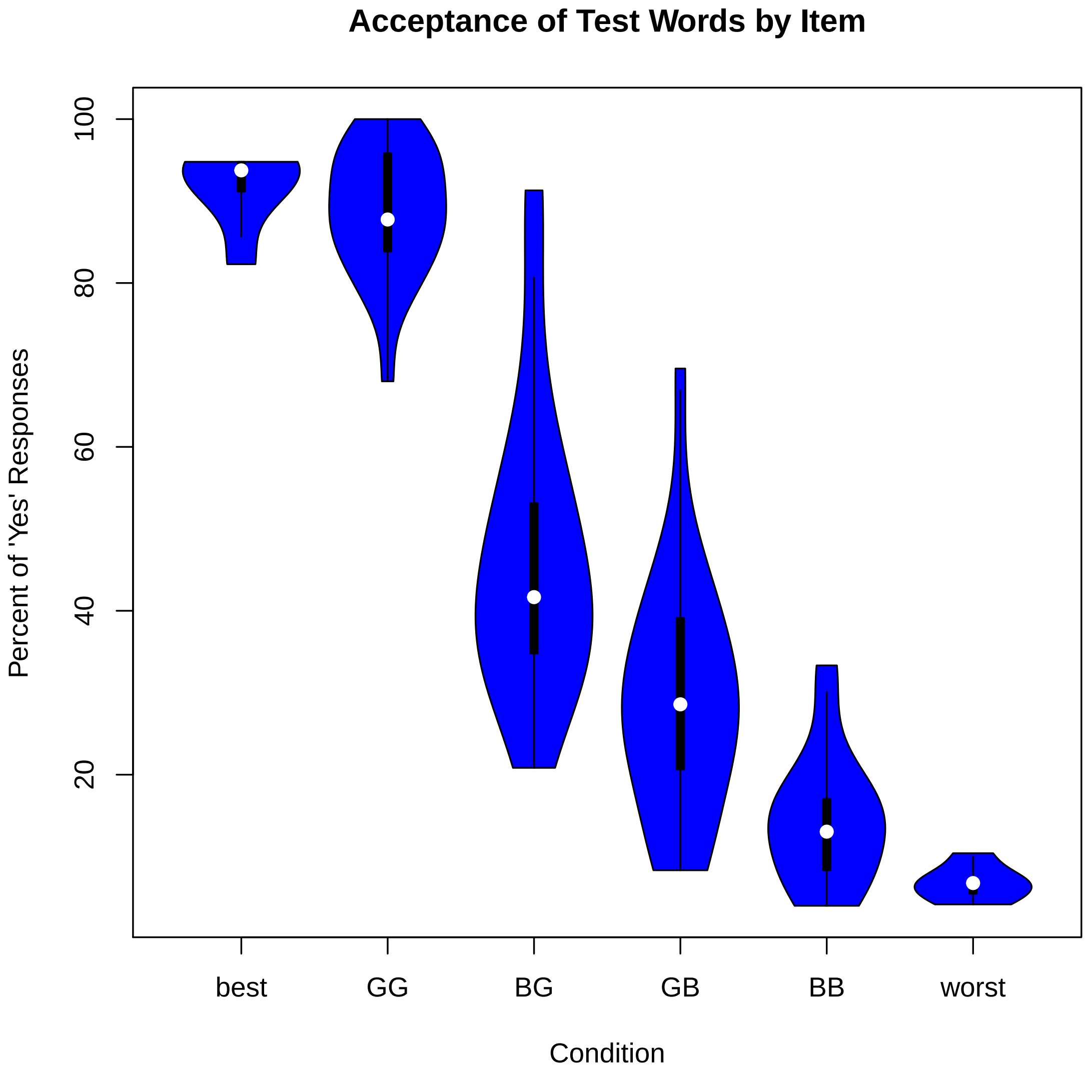

Figure 10. Distribution of mean percent acceptance by item for each condition in Experiment 2.

Table 5. Coefficients of the mixed effects model.

Factor	Estimate	p-value
Intercept	-0.29	0.0944687
OnsetViolation	-1.07	1.230616×10^{-22}
CodaViolation	-1.68	$2.0908335 \times 10^{-35}$
OnsetViolation:CodaViolation	0.46	4.4394417×10^{-5}

```
cfdata$YesPosition = cfdata$YesPosition - mean(cfdata$YesPosition)
#didn't converge but close, add iterations
cf_model = glmer(Response ~ OnsetViolation * CodaViolation + (1|Subject) +
(0+OnsetViolation|Subject) + (0 + CodaViolation | Subject) + (0+
OnsetViolation:CodaViolation|Subject) + (1|Item) + (0+OnsetViolation|Item) + (
+ CodaViolation | Item) + (0+ OnsetViolation:CodaViolation|Item), data =
cfdata[cfdata$PageType == 'Test',], family = binomial(link="logit"),
glmerControl(optCtrl=list(maxfun=10000), optimizer = "Nelder_Mead" ) )
```

```
cf = summary(cf_model)
ccf = cf$coefficients
```

(5) Mixed effects model formula
Response $\sim$ OnsetViolation $*$ CodaViolation + (1 | Subject) + (0 + OnsetViolation—Subject) + (0 + CodaViolation — Subject) + (0+ OnsetViolation:CodaViolation—Subject) + (1—Item) + (0+OnsetViolation—Item) + (0 + CodaViolation — Item) + (0+ OnsetViolation:CodaViolation—Item)

```
ftt = t.test(cfdata[cfdata$Condition == 'BB',]$Response, cfdata[cfdata$Conditi
```

In order to assess whether the BB condition was subject to a floor effect, meaning that the interaction is evidence of a task effect rather than evidence of the way the phonotactic grammar computes acceptability, a t-test was performed on the BB words and the bad fillers. The bad fillers contain multiple onset and coda violations each, rendering them worse than the BB words. Therefore, if the task is capable of showing a statistically significant difference between the bad fillers and the BB words, then we can conclude that the performance on the BB words is not due to a floor effect. Indeed, a one-tailed paired t-test gives a t-score of 4.4630896 and a p-value of 4.8618589×10^{-6}. In order to control for Type I error, we should apply a Bonferroni correction to the results of the mixed effects model and the t-test, adjusting the α level to $\frac{0.05}{2} = 0.025$. Both the p-value of the interaction and that of the t-test are below this new threshold, so we can conclude that both are statistically significant.

0.5.3 Discussion

This experiment finds support for the hypothesis that a violation lowers acceptability less in the presence of other violations than it does in isolation. Not only was this pattern observed and found to be significant, but it was also found to be distinguishable from a task effect. The concern that the condition containing multiple violations was assigned artificially high acceptability due to the nature of the task is undermined by the existence of a further condition assigned lower acceptability under the same task. The findings are consistent with Maximum Entropy models but not with Harmonic Grammar models.

0.6 General Discussion

Experiments 1 and 2 reach the same conclusion: there is evidence that a violation in the presence of another violation contributes a smaller penalty to word acceptability than that violation does in isolation. This conclusion is consistent with a Maximum Entropy model and challenges a Harmonic Grammar model.

The results of this study cannot, of course, single out Maximum Entropy as the correct grammar. There may be other models that would also fit the data gathered here. Furthermore, Maximum Entropy makes more specific predictions than that of subadditive cumulativity of violations, and further work is needed to test whether those predictions are also supported by the data.

However, these findings do pose a challenge to Harmonic Grammar as a model of cumulative phonotactics and support the idea that violations make more of a difference at the high end of the acceptability scale than at the low end. This idea can bear on the question of whether grammaticality is categorical or gradient, which **?** point out has been a challenging question for decades. If differences matter more in one region of the scale than another, this can make elicited intuitions appear to show a threshold between grammatical and ungrammatical data, offering a sort of reconciliation between categorical and gradient views of phonotactic well-formedness.

Another insight this experiment can bring has to do with the very high region of the acceptability scale, which is often overlooked. The especially acceptable fillers in this experiment, made from violation-free words with English suffixes, were rated slightly higher than the GG (violation-free) words. This raises challenges for a theory of grammar that uses only violations and no rewards, such as **?**. It is suggestive of analogical approaches to grammar, which are straightforwardly capable of rewarding forms that resemble existing words. However, it is possible that a grammar abstracted from the lexicon, such as constraint-based grammars, could also account for this phenomenon by incorporating constraints that reward the use of existing affixes in addition to penalizing the use of marked sequences.

The fact that the finding was replicated with a different set of materials can increase our confidence in the conclusion that violations accumulate nonlinearly. However, further variations on this experiment would be helpful. Auditory stimuli would help ensure that we are measuring phonology rather than orthographical effects, although it would be important to find stimuli that can be accurately perceived. Testing speakers of other languages on violations of their phonological constraints would further test the robustness of the effect, and show whether it reflects something about human grammar rather than a fact particular to English or the kinds of constraints available for testing on English speakers. Investigations of the cumulativity of constraint violations that rely on different tasks and designs would be helpful in reducing our reliance on the linking hypothesis adopted for this experiment.

BIBLIOGRAPHY