Investigating Properties of Phonotactic Knowledge Through Web-Based Experimentation

Presley Pizzo

July 31, 2015

Section 1

Introduction

Overview

Speriment A software package to make experiments easier to express, inspired by SurveyMan (Tosch and Berger, 2014) and built to work with psiTurk (McDonnell et al. 2012).

Cumulativity of Violations Experiments investigating how the presence of one violation affects the impact of another violation on acceptability.

Effect of Alternations on Phonotactics Experiments investigating whether learning an alternation affects phonotactic judgments.

Is Speriment For You?

Intended for linguists especially, social scientists generally

Requires simple programming

Feature requests welcome

How Speriment Works

1. Write your materials.
2. Describe your experiment in Python with Speriment's classes.
3. Run the Python script to generate a JavaScript file.
4. Start psiTurk to launch the JavaScript online and manage your participants.
5. Use Speriment to download and format your data for R.

For More Information

https://github.com/presleyp/Speriment

Section 2

Cumulativity of Violations

Subsection 1

Background

Cumulativity

"penalty": constraint weight * number of times it's violated

Constraint penalties $\rightarrow$ score assigned by grammar to entire word

OT Cumulativity

Optimality Theory doesn't define scores for words, but we want them for comparison.

Albright (2008) argues:

Grammaticality of word = maximum constraint penalty

Words eliminated by the same constraint have the same score, regardless of their milder violations

Mild violations can't gang up on one severe violation

Testing the Prediction

Ohala and Ohala (1986): 1 violation better than same violation + lesser violation

Testing the Prediction

Ohala and Ohala (1986): 1 violation better than same violation + lesser violation

Coleman and Pierrehumbert (1997): mrupation better than spleitisak

Testing the Prediction

Ohala and Ohala (1986): 1 violation better than same violation + lesser violation

Coleman and Pierrehumbert (1997): mrupation better than spleitisak

Albright (2008): models that take all violations into account fit data better

Cumulativity Effects

Cumulativity effects are observed experimentally

Let's look at models that define how they should work

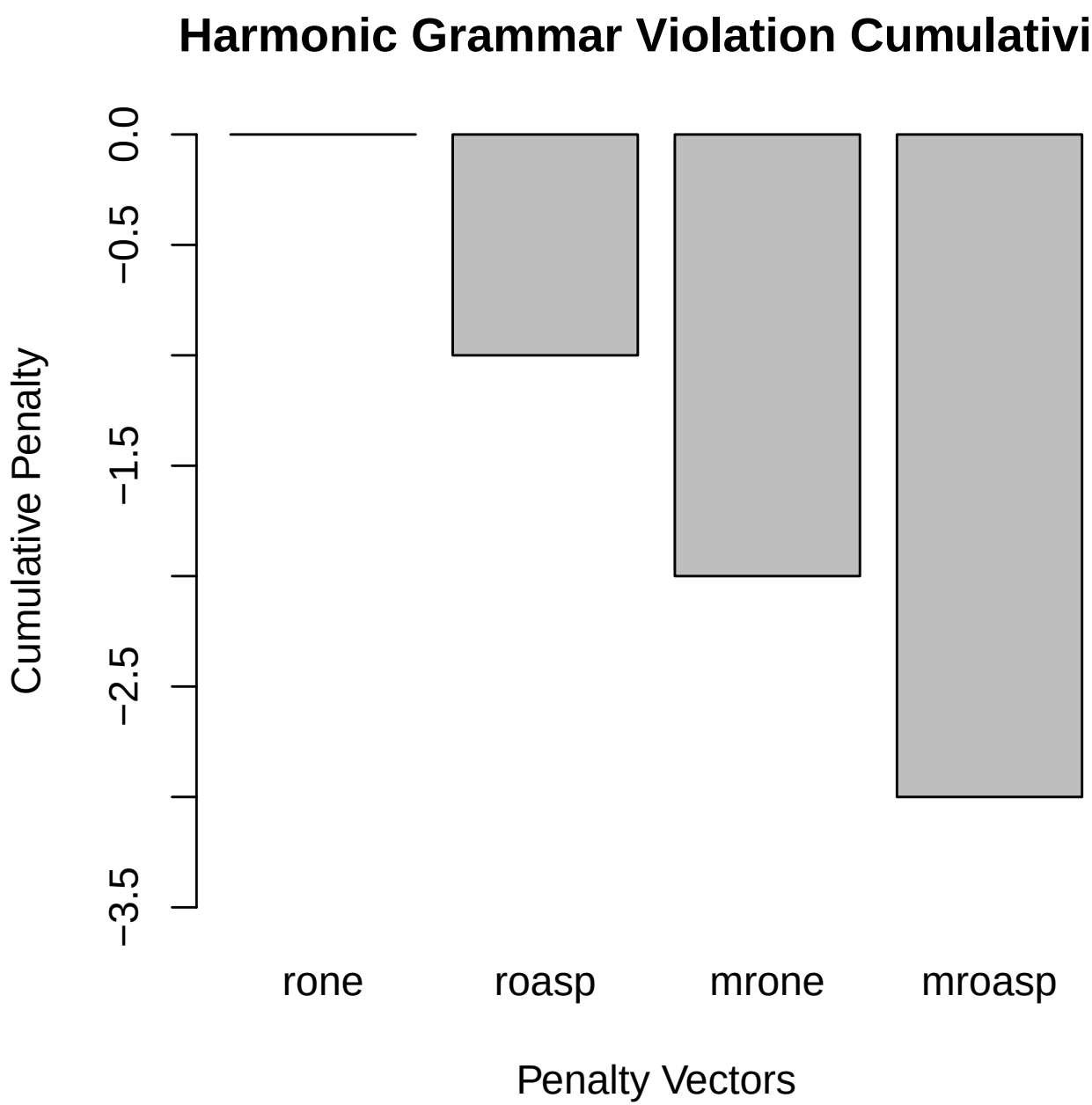
Example

Imagine four words: rone, roasp, mrone, mroasp

And two constraints: *[mr], weight -2; *[osp], weight -1

Linear HG Cumulativity

Harmony of word = sum of all constraint penalties


Harmonic Grammar Violation Cumulativity
Cumulative Penalty
0.0
−0.5
−1.5
−2.5
−3.5
rone
roasp
mrone
mroasp
Penalty Vectors

MaxEnt Cumulativity

Probability of word = normalized exponentiated sum of constraint penalties

Differences:

MaxEnt Cumulativity

Probability of word = normalized exponentiated sum of constraint penalties

Differences:

Scores are positive

MaxEnt Cumulativity

Probability of word = normalized exponentiated sum of constraint penalties

Differences:

Scores are positive

Scores are between 0 and 1

MaxEnt Cumulativity

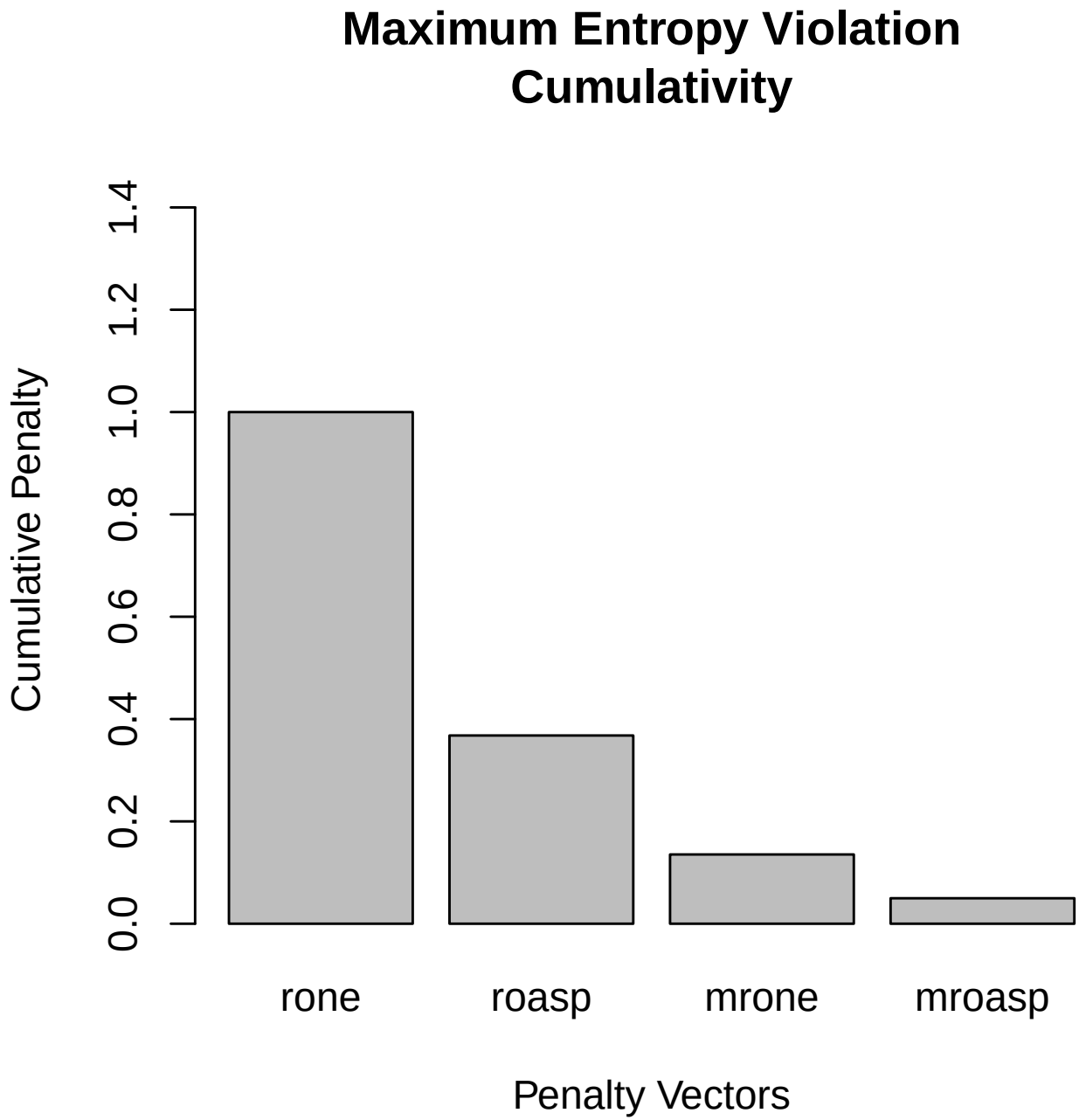
Probability of word = normalized exponentiated sum of constraint penalties

Differences:

Scores are positive

Scores are between 0 and 1

Decreasing marginal "returns" on penalties


Maximum Entropy Violation Cumulativity
Cumulative Penalty
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
rone
roasp
mrone
mroasp
Penalty Vectors

Subsection 2

Experiment 1

Question

Harmonic Grammar or Maximum Entropy?

HG: violations accumulate linearly

MaxEnt: violations accumulate sublinearly

Hypothesis

Two factors of interest, OnsetViolation and CodaViolation.

Hypothesis

Two factors of interest, OnsetViolation and CodaViolation.

Binary, crossed

Hypothesis

Two factors of interest, OnsetViolation and CodaViolation.

Binary, crossed

Linking hypothesis: linear transformation from grammaticality to acceptability

Hypothesis

Two factors of interest, OnsetViolation and CodaViolation.

Binary, crossed

Linking hypothesis: linear transformation from grammaticality to acceptability

HG: no interaction

Hypothesis

Two factors of interest, OnsetViolation and CodaViolation.

Binary, crossed

Linking hypothesis: linear transformation from grammaticality to acceptability

HG: no interaction

Maximum Entropy: subadditive interaction

Subsection 3

Method

Participants

100 participants paid $0.75 each for five minute-ish experiment.

Excluded if:

- not a native speaker of English
- choose whichever answer was on one side
- answered too quickly
- didn't prefer good fillers to bad fillers.

94 participants used in the analysis.

Materials

Nonce words in four conditions:

	Good Onset	Bad Onset
Good Coda	GG	BG
Bad Coda	GB	BB

Materials

24 item sets

For each item set, there is one good onset, bad onset, good coda, bad coda, and vowel.

These were combined into the four types of words.

Orthographic presentation to avoid perceptual repairs.

Example

Good onset, good coda (GG): plag

Bad onset, good coda (BG): tlag

Good onset, bad coda (GB): plavb

Bad onset, bad coda (BB): tlavb

Procedure

Run with Speriment on Mechanical Turk.

Latin square over the conditions.

Yes/no questions.

Example Item

Based on how it sounds, do you think this could be a word of English?

tlavb

Yes

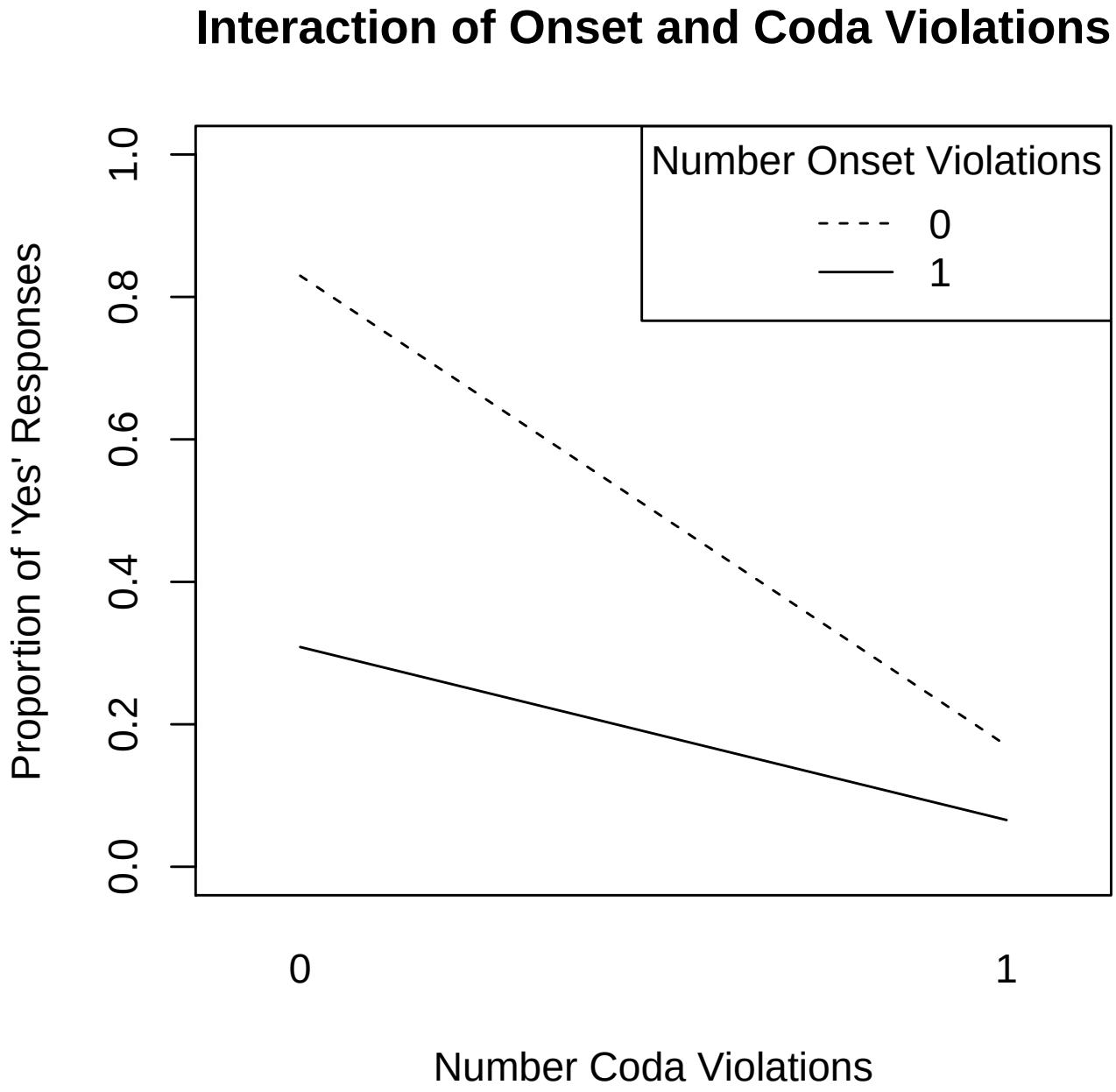
No

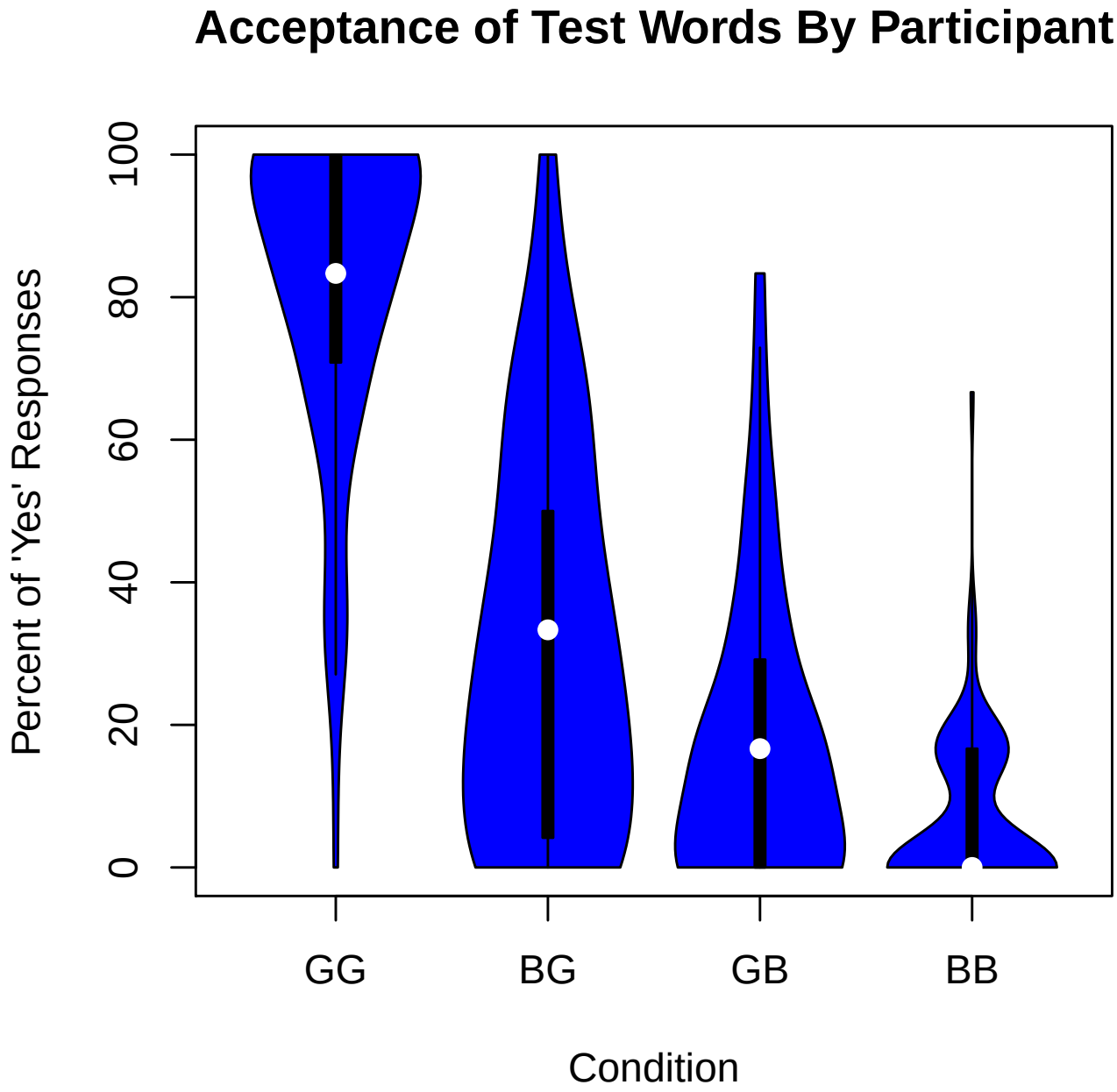
Subsection 4

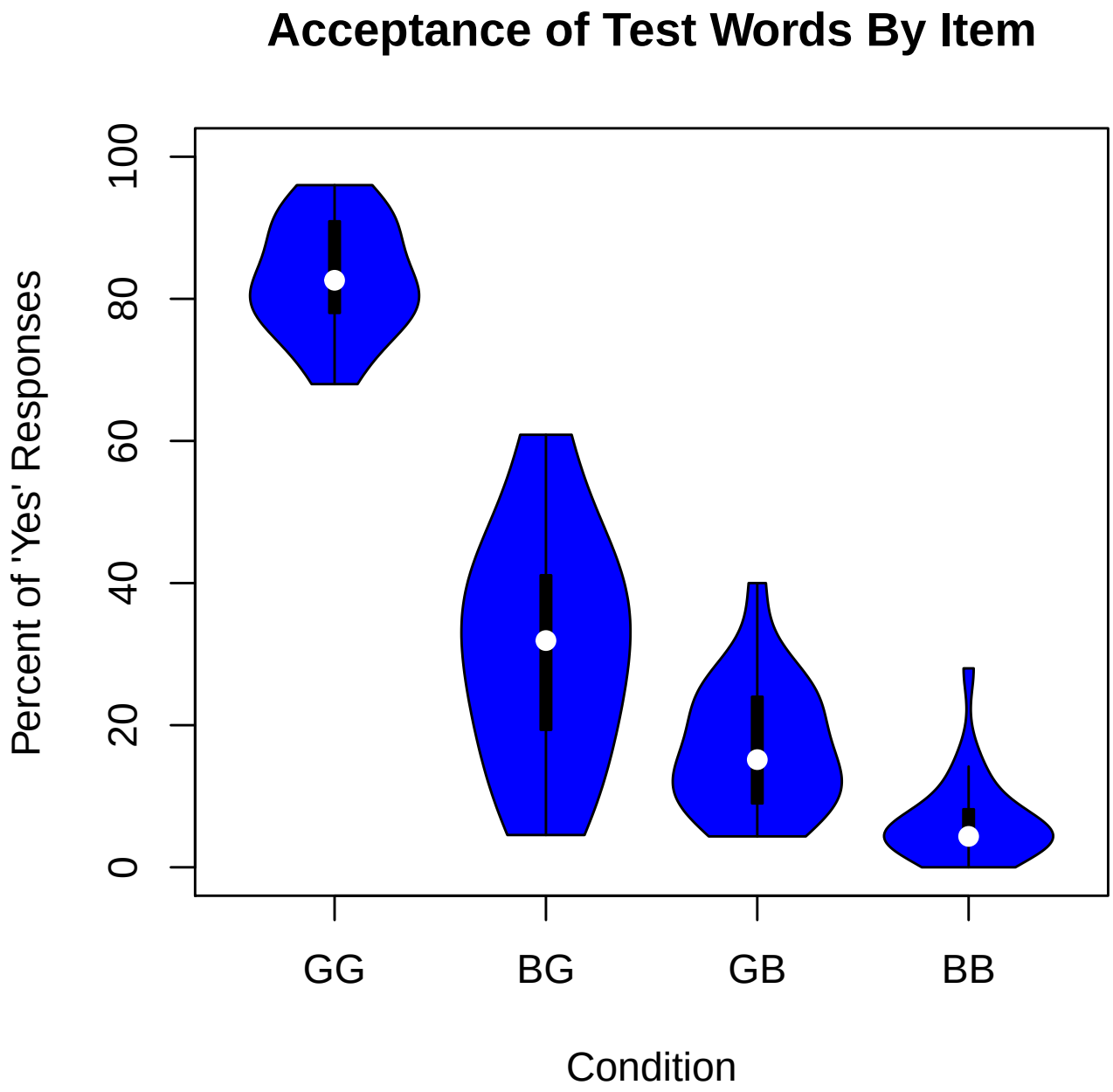
Results

Acceptance Rates

Condition	Mean Percent 'Yes'	SD
GG	83	37.6
BG	30.9	46.2
GB	17	37.6
BB	6.6	24.8


Interaction of Onset and Coda Violations
Number Onset Violations
0
1
Proportion of 'Yes' Responses
0.0
0.2
0.4
0.6
0.8
1.0
0
1
Number Coda Violations


Acceptance of Test Words By Participant
Percent of 'Yes' Responses
0
20
40
60
80
100
GG
BG
GB
BB
Condition


Acceptance of Test Words By Item
Percent of 'Yes' Responses
0
20
40
60
80
100
GG
BG
GB
BB
Condition

Analysis

Logistic mixed effects model:

Response $\sim$ OnsetViolation * CodaViolation

random slopes and intercepts for main effects and interaction by participant and item

Analysis

Factor	Estimate	p-value
Intercept	-0.87	< 0.001
OnsetViolation	-0.84	< 0.001
CodaViolation	-1.34	< 0.001
OnsetViolation:CodaViolation	0.82	< 0.001

Discussion

Hypothesis supported:

OnsetViolation adds penalty (significant main effect)

CodaViolation adds even bigger penalty (significant main effect)

Both at the same time add a penalty lower than the sum of the main effects (significant subadditive interaction)

Problem

Floor effect:

What if the penalty for both "should" be lower, but we couldn't measure that because they can't go below zero?

Subsection 5

Experiment 2

Same idea with tweaks to avoid the floor effect problem.

Subsection 6

Method

Participants

101 participants were run

96 were included

Filler words were not an exclusion criterion

Materials

New nonce words

Half of bad onsets and bad codas removed

The other half used twice each

Fillers

Good fillers: no known violations plus real English suffixes

bressic

Bad fillers: bisyllabic with 3 violation sites

lbafthrizk

Procedure

Extra instruction page:

Before we start, here are examples of the kind of words you'll see.

blickity is the kind of word you might want to say "yes" to. It's not an English word, but it sounds like it could be.

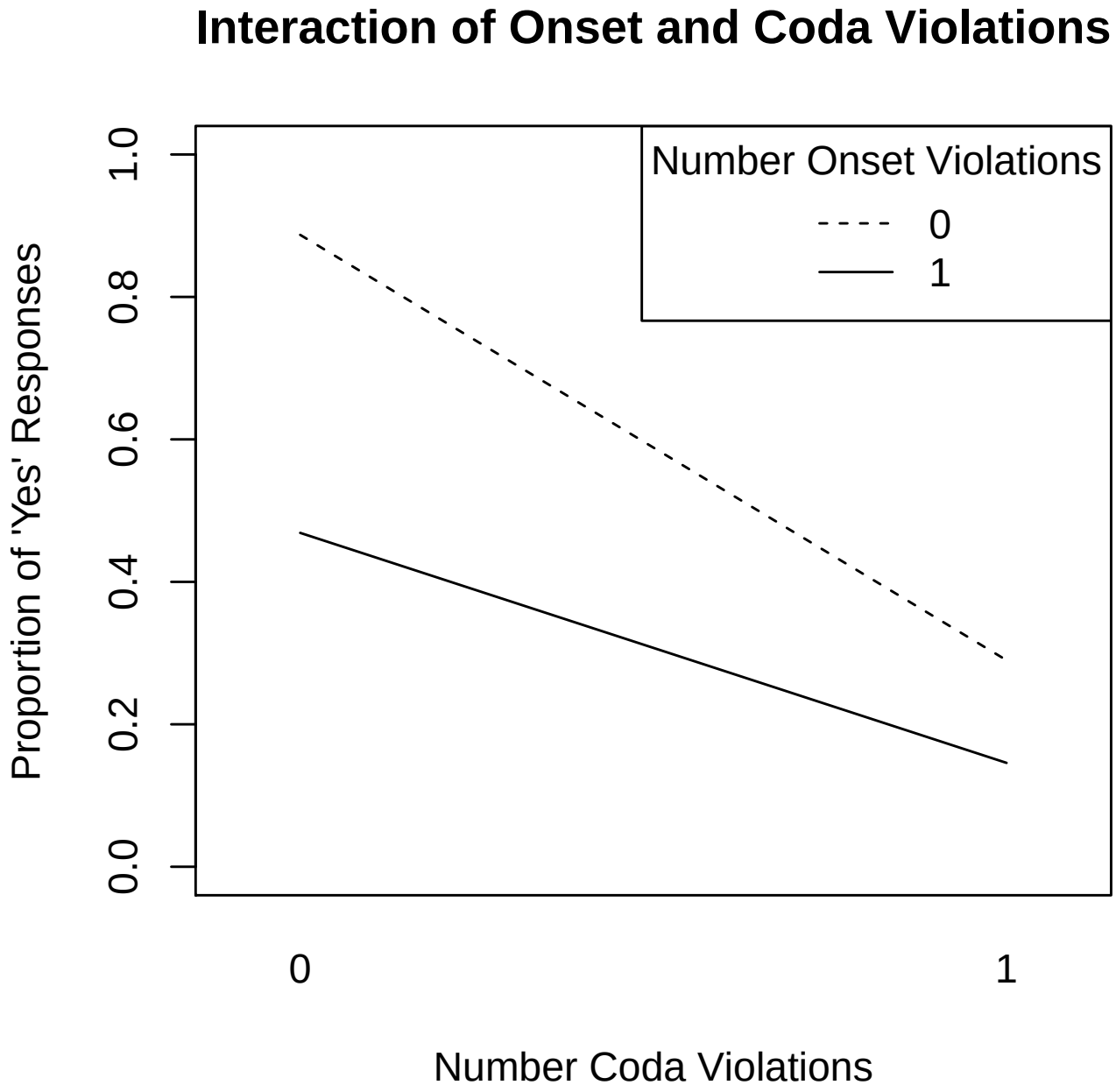
rzbesgathv is the kind of word you might want to say "no" to. It's not an English word, and it doesn't sound like it could ever be one.

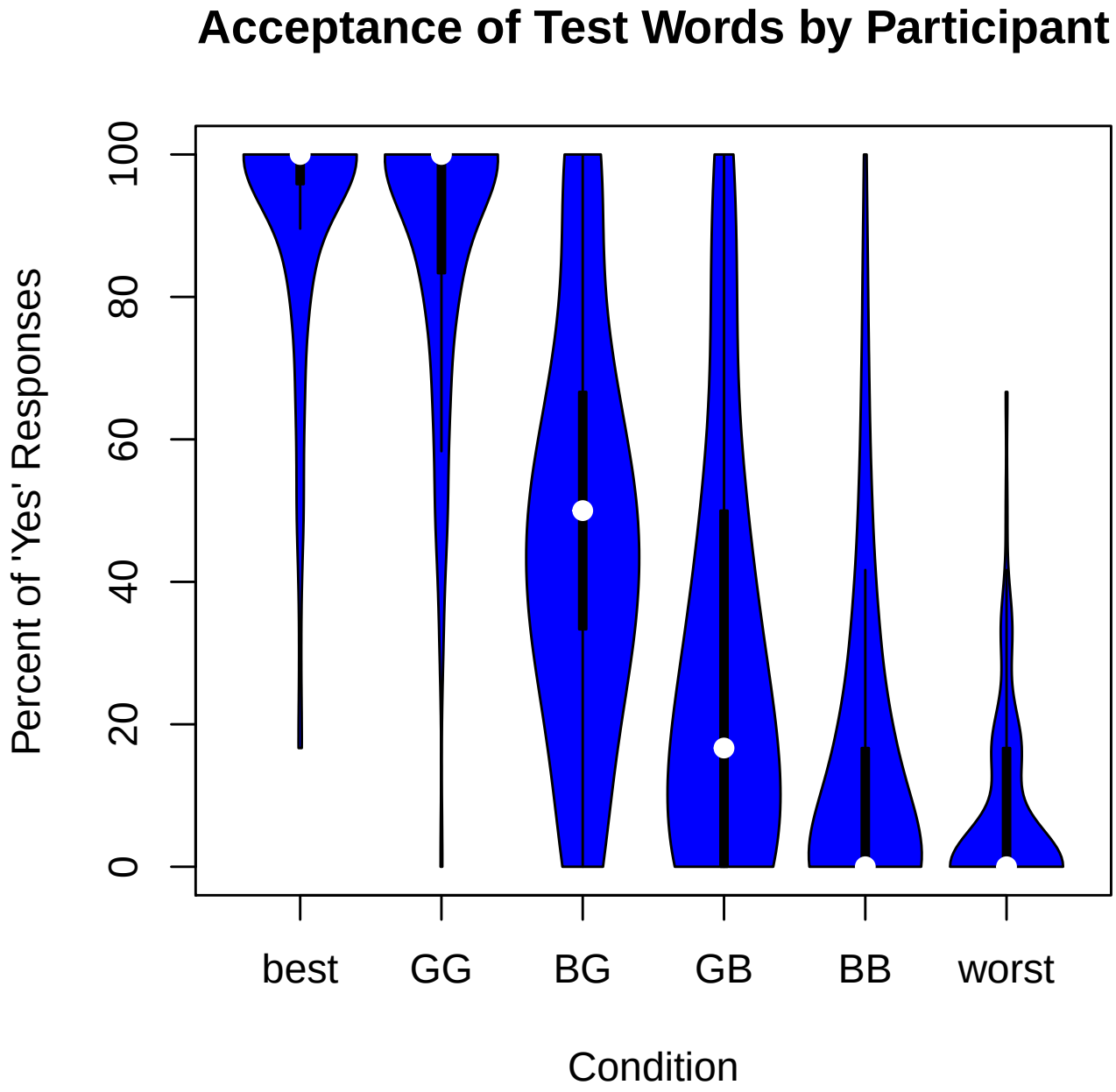
Subsection 7

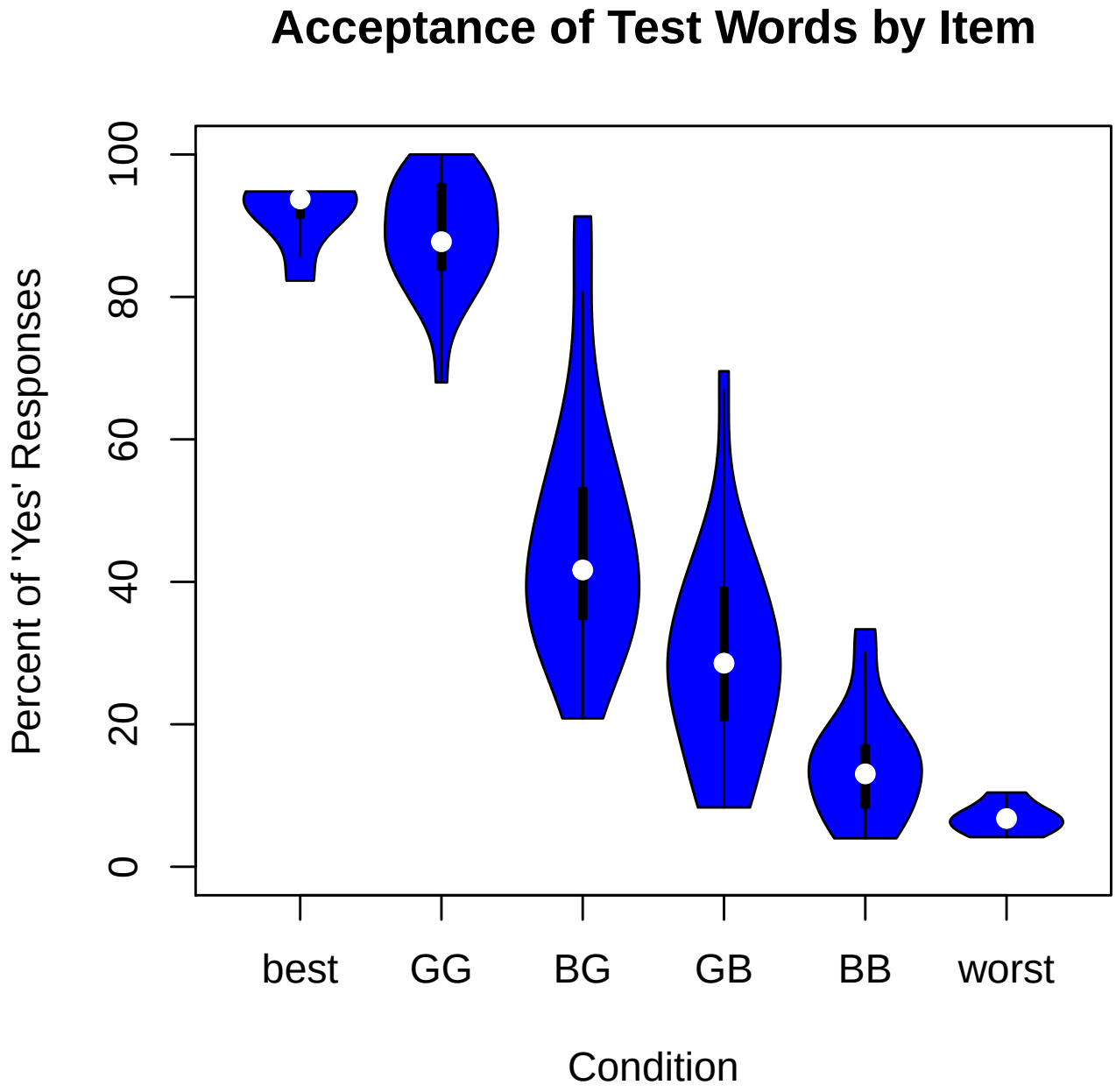
Results

Acceptance Rate

Condition	Mean Percent 'Yes'	SD
Good Filler	91.7	27.7
GG	88.7	31.7
BG	46.9	49.9
GB	29	45.4
BB	14.6	35.3
Bad Filler	6.8	25.1


Interaction of Onset and Coda Violations
Number Onset Violations
0
1
Proportion of 'Yes' Responses
0.0
0.2
0.4
0.6
0.8
1.0
0
1
Number Coda Violations


Acceptance of Test Words by Participant
Percent of 'Yes' Responses
0
20
40
60
80
100
best
GG
BG
GB
BB
worst
Condition


Acceptance of Test Words by Item
Percent of 'Yes' Responses
0
20
40
60
80
100
best
GG
BG
GB
BB
worst
Condition

Analysis

Factor	Estimate	p-value
Intercept	-0.29	0.09
OnsetViolation	-1.07	< 0.001
CodaViolation	-1.68	< 0.001
OnsetViolation:CodaViolation	0.46	< 0.001

Floor effect?

One-tailed, paired t-test of BB and Bad Fillers

$t = 4.463$

$p < 0.001$

Significant with Bonferroni correction

Discussion

Subadditive interaction found as predicted

BB words have not hit a floor because Bad Fillers have a measurably lower acceptance rate

We can interpret the interaction as support for the hypothesis

Discussion

Results consistent with Maximum Entropy

HG is not supported

Any other frameworks with sublinear (but nonzero) cumulativity are also consistent

Future Research

Replication with auditory stimuli

Cross-linguistic work

Investigation into linking hypotheses:

What effect do other words in the experiment have on a yes/no judgment?

How do grammaticalities transform into acceptance/choice rates? Ratings? Reaction times?

Jumping Off Points

Are Good Fillers significantly better than GG words, or could even better Good Fillers be?

Could suggest positive constraints or analogical processes

Maybe judgments seem categorical to some because the distinctions among less acceptable words are smaller

Section 3

Conclusion

Summary

- More help writing experiments! Good for efficiency, transparency, replicability.
- Violations seem to accumulate sublinearly
- Stay tuned, but alternations seem to affect phonotactics

Bigger Picture

- Maximum Entropy seems like a better fit than Harmonic Grammar
- Phonology may have modules for phonotactics and alternations, but if so, they can probably communicate
- Experiments can help us answer architectural questions about phonological knowledge

Albright, A. (2008). From clusters to words: Grammatical models of nonce word acceptability. Handout of talk presented at 82nd LSA, Chicago.

Coleman, J. and Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. In Computational Phonology: Third meeting of the ACL special interest group in computational phonology, pages 49–56.

McDonnell, J., Martin, J., Markant, D., Coenen, A., Rich, A., and Gureckis, T. (2012). psiTurk (Version 1.02) [Software]. New York, NY. Available at https://github.com/NYUCCL/psiTurk.

Ohala, J. J. and Ohala, M. (1986). Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. In Ohala, J. J. and Jaeger, J. J., editors, Experimental Phonology, pages 239–252. Academic Press, Orlando.

Tosch, E. and Berger, E. D. (2014). Surveyman: Programming and automatically debugging surveys. In Proceedings of the 2014 ACM International Conference on Object Oriented Programming Systems Languages & Applications, pages 197–211. ACM.

Thank you!

Thanks especially to my committee and the Sound Seminar attendees for help and feedback.

Thank you!

Thanks especially to my committee and the Sound Seminar attendees for help and feedback.

So long, and thanks for the fish.

Exp. 1 Sans Outliers

	Estimate	Std. Error	z value	Pr($>|z|$)
(Intercept)	2.0511	0.1919	10.687	$<$ 2e-16 ***
OnsetViolation	-3.1863	0.2669	-11.940	$<$ 2e-16 ***
CodaViolation	-4.0934	0.2471	-16.568	$<$ 2e-16 ***
Interaction	2.6886	0.5710	4.708	2.5e-06 ***

Exp. 1 Reaction Times

	Estimate	Std. Error	t value
(Intercept)	7.80629	0.04738	164.76
OnsetViolation	0.09621	0.03767	2.55
CodaViolation	0.11089	0.04131	2.68
Interaction	-0.16058	0.04991	-3.22

Exp. 1 Model Comparison

	AIC	BIC	logLik	deviance	Chisq	Pr
null	5036.4	5110	-2506.2	5012.4		
alt	5028.3	5108	-2501.2	5002.3	10.121	0.001466

Exp. 2 Reaction Times

	Estimate Std.	Error	t value
(Intercept)	7.74154	0.03412	226.90
OnsetViolation	0.19294	0.03467	5.57
CodaViolation	0.18763	0.03419	5.49
Interaction	-0.30936	0.04631	-6.68

Exp. 2 Model Comparison

	AIC	BIC	logL	dev	Chisq	Pr
null	5438.6	5512.4	-2707.3	5414.6		
alt	5403.7	5483.6	-2688.8	5377.7	36.891	1.249e-09

In both cases, the main effects' t-values become much smaller without the interaction.