Properties of a Phonotactic Theory

Presley Pizzo

December 16, 2014

Chapter 1

Introduction

The goal of this dissertation is to lay groundwork for the further study of phonotactics. In Chapter 1, I discuss software that will facilitate the implementation of phonotactic experiments. In Chapter 2, I review literature that suggests that any accurate model of phonotactics must allow for the accumulation of violations, so that the grammaticality of a word depends on all of the violations it contains. I present an experiment with the ability to refine this statement by investigating how additional violations affect the grammaticality of the word, weighing on the question of whether linear Harmonic Grammar or Maximum Entropy grammar is a better model for phonotactics. In Chapter 3, I ask whether variables not usually considered could affect phonotactics and need to be included in future modeling. First, I consider that token frequency might correlate negatively with the weight of a constraint in a grammar, rather than being unrelated. Second, I consider whether a constraint may have a higher weight in the phonotactic grammar if it is active in alternations than if it is not. Thus, the dissertation does not propose a particular model of phonotactics, but aims to narrow the space of models that need to be considered in the future via empirical investigation, and increase the speed and accuracy of future experimental work in the field.

Chapter 2

Speriment

SurveyMan is a desktop application that communicates with Amazon's Mechanical Turk survey-running website in order to manage surveys and experiments with minimal human intervention. A working prototype exists but it has many shortcomings, particularly for linguistics experiments which often have very precise requirements. I propose to develop an experiment-specific version of the software that will meet the needs of our phonotactic and psycholinguistic web-based studies. This would benefit my own research and that of any linguist interested in running word or sentence judgment studies on the internet. Other social scientists would also be likely to find it useful.

This software would enable researchers without a programming background to set a wide variety of options on the mechanics and presentation of their experiments.

There are several existing programs to help researchers run experiments on the web, and they differ from SurveyMan in various ways.

Survey programs like SurveyMonkey and LimeSurvey are relatively easy to use, but lack some crucial experiment-centric features. In SurveyMonkey, randomization requires a premium membership, and neither of them handle Latin square designs for the experimenter.

There are also experiment-running programs developed by scientists, such as Keller et al. (2009), Becker and Levine (2010), and McDonnell et al. (2012). These tend to require more programming but have more of the features needed for experiments. However, they still lack some of the features planned for SurveyMan. WebExp, for instance, lacks support for acoustic stimuli and constrained randomization of question order. Experigen does not record reaction time. PsiTurk is compatible with any of these features, but the experimenter has to be able to write JavaScript.

Drummond (2011) is a strong candidate for use in web experiments, as it was designed for linguists and thus addresses most of the features needed for our experiments. However, SurveyMan has plans for advanced features, such as training periods that continue until the participant performs well enough to advance.

Thus, SurveyMan will balance the ability to design experiments quickly and easily with minimal programming knowledge with a wide array of features tailored to the needs of phonologists and other social scientists.

2.0.1 Plan

In order to make an experimental version of SurveyMan that is fully functional for word and sentence judgment studies, the following features need to be added:

- Pseudorandomization as an alternative to pure randomization of question order.
- Counterbalancing of question types.
- Logging of reaction times and the times of events such as playing sounds or videos.
- Ability to place restrictions on events; for instance, allowing a sound to be played only once or allowing a choice to be made only after the sound has played.
- Training loops.
- Ability to randomly pair images, audio, and video with questions for each participant.

Currently, Emma Tosch is working on the backend of the survey-based version of SurveyMan and Molly McMahon is writing a Python front-end, which will enable an R interface later on. I will adapt SurveyMan to include the necessary experiment-based features.

I have implemented all of the features necessary to run my own experiments, and will focus on integrating those changes with the broader SurveyMan framework before adding additional features.

Chapter 3

Experiment 1: Cumulativity of Violations

3.1 Overview

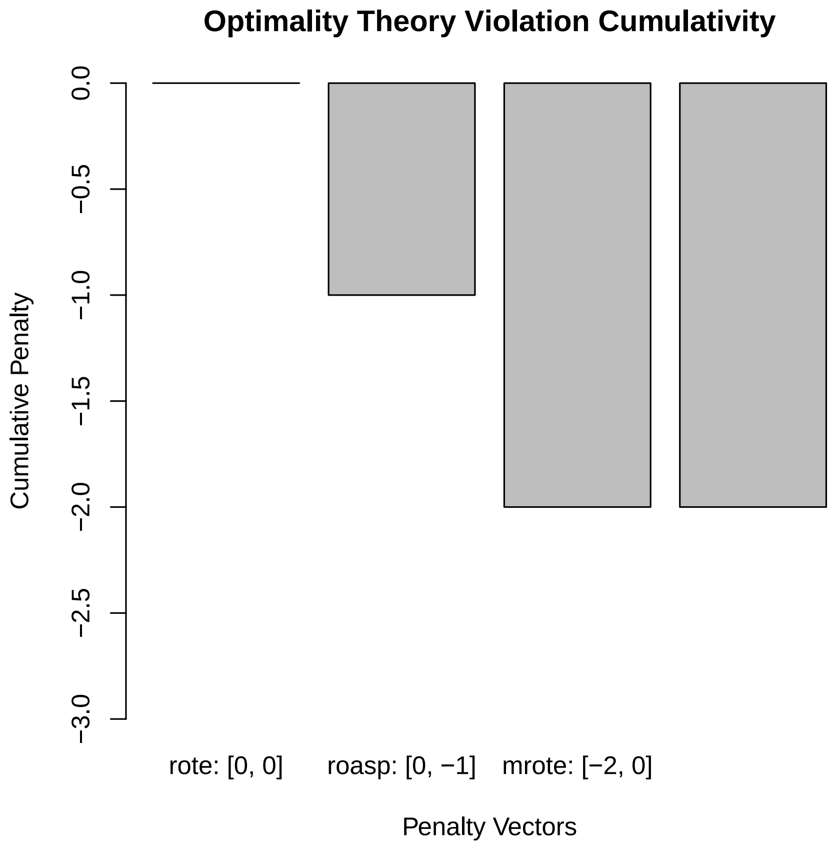
In modeling constraint-based phonotactics, there are three broad kinds of decisions to be made: which framework to use, which parameters (constraints) to use, and how to tune (rank or weight) the parameters. This experiment will weigh in on the question of framework choice, using the function they use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

Popular constraint-based frameworks include Optimality Theory (Prince and Smolensky, 2004), Harmonic Grammar (taken here to mean linear Harmonic Grammar, in which harmony scores are not subject to exponentiation) (Legendre et al., 1990; Smolensky and Legendre, 2006; Pater, 2009; Potts et al., 2010), and Maximum Entropy (Goldwater and Johnson, 2003).

As these frameworks are all constraint-based, they recognize units of violation. Every locus of a constraint violation is such a unit. Identical violations of identical constraints are recognized as "the same" across words, so that violations can be compared across words. Multiple violations can occur in the same word, and the ungrammaticality of the word is derived by combining the ungrammaticalities of each violation in the word according to some process. These frameworks differ in how they define that violation-combining process, and so 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.

3.2 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 **??** 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, Ohala and Ohala (1986) 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, Coleman and Pierrehumbert (1997) 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.

Albright (2008) 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 phonotactic violations in the onset as well as milder violatinos 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, and I turn to Harmonic Grammar and Maximum Entropy.

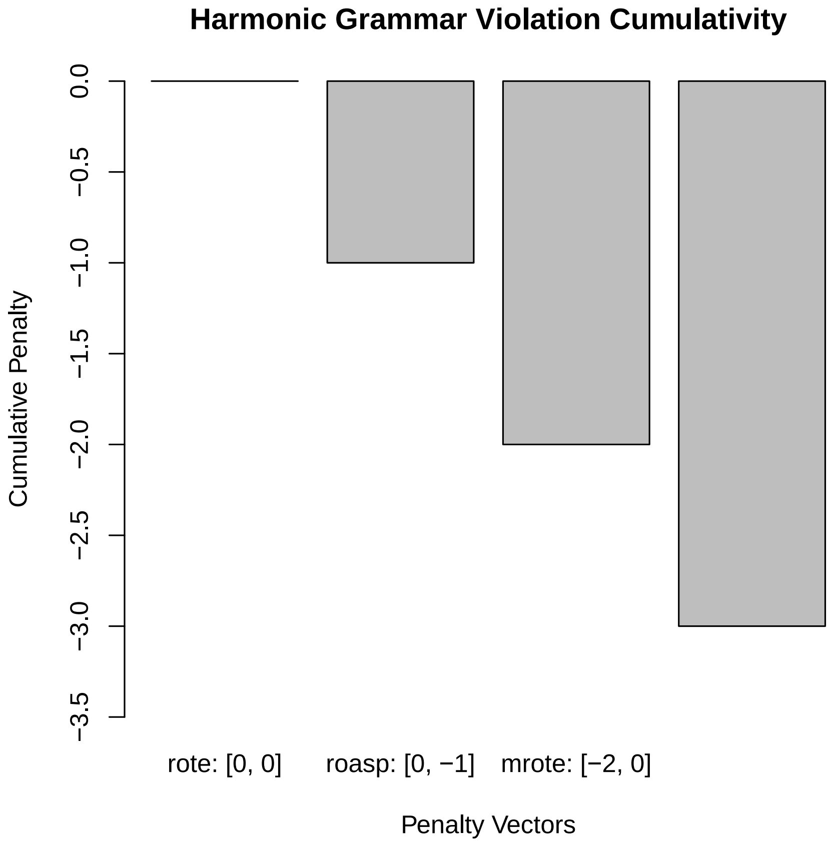
3.3 Linear vs. Exponential Combination

Albright (2008) 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 Pater (2008) points out, Harmonic Grammar predicts a well-restricted set of cumulativity effects, unlike Optimality Theory with Local Constraint Conjunction (Smolensky, 2006). 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.

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

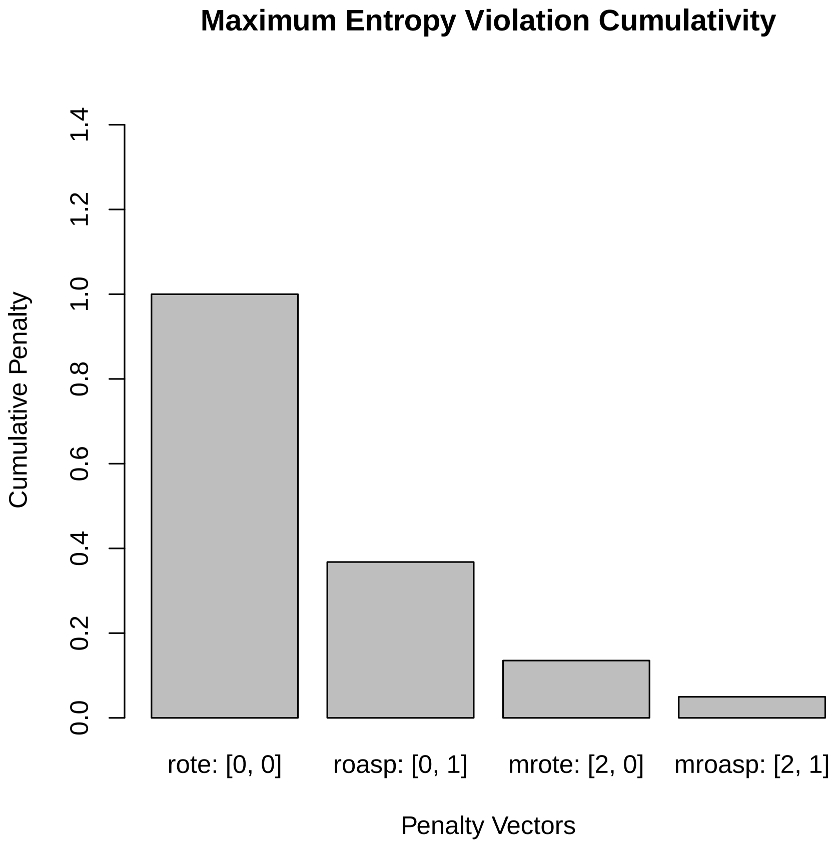
(2) Maximum Entropy: The probability p of a word x is the exponentiated negative harmony of the word, normalized relative to the candidate set X.



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

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.

I predict that, in accordance with the Maximum Entropy model, additional violations will have smaller effects on the grammaticality of the word, as the grammaticality approaches a floor.


Maximum Entropy Violation Cumulativity
Cumulative Penalty
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
rote: [0, 0]
roasp: [0, 1]
mrote: [2, 0]
mroasp: [2, 1]
Penalty Vectors

3.4 Method

3.4.1 Participants

One hundred participants were recruited from Mechanical Turk and paid for their participation. They were all located in the United States and claimed to be native speakers of English. Catch trials asked participants to choose the more English-like of two words where one word is free of violations and the other has a severe violation; any participants who preferred the severe violation were removed from the analysis. Additionally, participants whose reaction times were too short to be human were removed.

3.4.2 Materials

Each of 24 test items will appear in four conditions. The four conditions are created by crossing two factors: presence of an onset violation and presence of a rime violation. Within an item, the onset violation will be the same whenever it is present, and likewise for the rime violation. This way, comparing an item with only an onset violation to the same item with both violations shows the effect of adding the rime violation. The vowel is held constant across all conditions of an item.

Aside from test items, there will also be filler items. These will be nonce words of medium acceptability; they will have mild violations that are different from the kinds of violations found in the test words. Instead of violating cluster phonotactics, they will violate long-distance OCP constraints or constraints on the cooccurrence of certain vowels and codas. The goal is to keep participants from comparing filler and test items piecewise and encourage them to use the filler only to get a sense of a baseline of grammaticality against which to compare the test words.

3.4.3 Design

A Latin square design is applied to the test items so that each participant only sees one word from each item set. Each test word is then randomly paired with a filler word. This random assignment is done independently for each participant.

3.4.4 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 a test word and a filler word, for 24 trials. In each trial, the participant was asked to choose the more English-like of the two words.

3.5 Results

A mixed effects model was fitted to the results. The dependent variable was the proportion of times the test item was chosen. OnsetViolation, RimeViolation, and their interaction served as fixed effects. Random effects for subject and item were included in the model.

The purpose of the experiment is to test whether an interaction exists between the two fixed effects.

3.6 Discussion

Chapter 4

Experiment 2: Effect of Alternations on Phonotactics

Infants show evidence of knowing phonotactic patterns before they begin producing words (Werker and Tees, 1984; Kuhl et al., 2006), and thus it may be the case that the phonotactic grammar is learned independently of alternations, which depend on a lexicon. Adriaans and Kager (2010), for instance, developed a phonotactic learner based on this assumption. However, it is also possible that infants have learned about the lexicon of their language before they begin producing words; furthermore, they may first learn phonotactics in a vacuum but later incorporate morphological knowledge into their phonotactic grammar. Thus, it is possible that participation in an alternation improves the salience of a constraint and causes learners to more heavily weight that constraint, even for the purposes of phonotactic judgments.

On the other hand, it has not been proven that alternations do increase the learnability of a phonotactic pattern, and if we can determine that the two seem to be independent, we could safely model them separately. Such a finding would simplify future work on phonotactics and would also have implications for arguments that the facts of both systems should be derived from the same machinery. Pater and Tessier (2003) gave evidence that alternation learning cannot be fully modeled without reference to phonotactics, but it is unknown whether phonotactics can be captured without reference to alternations.

Thus, it is of considerable interest whether the presence of alternations to or from a sound sequence affects judgments of the grammaticality of that sequence, holding the sequence's type frequency in the lexicon constant.

It is difficult to find a case in natural language with these properties in order to do a well-controlled test of the idea, so I will use an artificial language learning experiment to test the effect of alterations elsewhere in a language on identical words.

The results of this study will inform future modeling as well as theories about the interaction between phonotactics and alternations.

4.1 Method

I will address this question with an experiment in which participants are trained in an artificial language and then asked for judgments about the probability that novel words could belong to the artificial language.

4.1.1 Participants

4.1.2 Materials

Two constraints will be constructed. They will be chosen to be of similar levels of attestion in English, segmental length, featural specificity, and phonetic naturalness, but perfect control is not necessary due to the experimental design. It is, however, necessary that neither be inviolable in English, or else the experiment will get ceiling effects. For purposes of exposition I will call these constraints *AB and *CD. At least initially, I plan for these constraints to actually be *bf and *nf, where the rules are b → p / _f and n → m / _f. The rules will apply across syllable boundaries, where these rules are not obligatory in English. However, even heterosyllabic instances of these bigrams are uncommon or unattested in English, so I will analyze the results for ceiling effects. If needed, I will switch to completely non-English constraints, such as vowel harmony and consonant harmony.

Two artificial languages will be constructed. The lexicons of the two languages will be as similar as possible, and neither language will allow sequences of AB or CD. However, Language 1 will use B as a plural ending, causing alternations from AB to XB, while Language 2 will use D as a plural ending, causing analogous alternations from CD to YD. Thus, in Language 1 AB is the alternating sequence, and in Language 2 CD is the alternating sequence.

Participants will be shown words and definitions from their assigned language until they have learned the meanings. They will then be asked to give judgments on novel words containing AB and CD.

(1) Languages

- a. Both Languages:
  No bf
  No nf
  plural: -fa
- b. Language I: b → p / _f
- c. Language II: n → m / _f

(2) Example training words

- a. Language I: dela, delafa, vem, vemfa, sirab, sirapfa, ...
- b. Language II: dela, delafa, vem, vemfa, siran, siramfa, ...

(3) Example testing words for both languages
malo, kamfin, labfu, . . .

4.1.3 Procedure

The most straightforward judgment task for this experiment would be asking participants to simply answer the yes/no question "Is this word possible in the language you learned?" Thus, in the first run of the experiment, I will ask participants this question for words containing AB, CD, XB, and YD, and filler words. However, there is some concern that when asked yes/no questions, participants try to balance the number of yeses and nos they give. I will instruct them not to do this, but I will also analyze the results of the experiment to look for such behavior. If I suspect that such a strategy has been used, I will rerun the experiment with a two-alternative forced choice task in which I compare words with AB to words with XB and words with CD to words with YD.

4.2 Results

The results will be analyzed with a mixed effects model predicting proportion of times the banned bigrams are chosen based on the interaction of language learned and constraint tested, with random effects for subject and item. If there is a significant effect of whether a sequence was an alternating sequence on the proportion of times the sequence was chosen in the forced choice task, I will conclude that evidence from alternations are used in the phonotactic grammar. I will look separately at whether the sequence that the rule eliminates is especially dispreferred and at whether the sequence that the rule creates is especially preferred, as these may yield different answers.

4.3 Discussion

Chapter 5

Simulation: Effect of Token Frequency on Phonotactics

5.0.1 Token Frequency

Motivation

Albright (2006, 2009) has shown that giving models the ability to use token frequency does not improve their performance, and can even hinder it. However, the way he incorporated token frequency into various models made the effect of a sound sequence on the grammar or analogical system correlate positively with the token frequency of the words those sequences appeared in. Thus it is possible that token frequency is uncorrelated with the productivity of a pattern, but it is also possible that token frequency is inversely correlated with it. On one hand, token frequency may simply not be used to determine grammaticality, as in a model where token frequency is represented in the lexicon and grammaticality is calculated from a grammar which only pulls patterns out of the lexicon, abstracting away from word-specific information like token frequency. On the other hand, high token frequency may serve to increase the degree to which the patterns in the word are associated with that particular word rather than with the grammar in general. Thus, the patterns found in very frequently used words may be memorized as exceptional, so that patterns associated with high token frequencies are less likely to be generalized than others.

I predict that constraints found mainly in very high token frequency words are not highly weighted in the phonotactic grammar. This would have implications for how we model phonotactics, as token frequency would need to be reintroduced as a factor, but as one whose effect is negative or found by the grammar rather than assumed to be positive. It would also have implications for theories of how the lexicon interacts with the grammar; token frequency is closely associated with access to the lexicon, so this finding would support a view in which the lexicon is involved in phonotactic judgment even if there is a separate grammar.

Plan

The hypothesis is that constraints that hold mostly of high token frequency words, which we could think of as constraints that encode exceptions, are not highly weighted in the general phonotactic grammar, the one applied to arbitrary novel words (rather than to exceptional words and conceivably, to novel words that have specific properties that cause analogy to exceptional words).

This hypothesis predicts that a grammar learned from a lexicon containing highly frequent, exceptional words will generalize to nonce words less well, that is, less similarly to English speakers, than a grammar learned from a lexicon with the exceptional words removed.

I will gather a corpus of English words to train a Maximum Entropy grammar on. I will remove a certain number of words from this corpus several times, resulting in several different corpora of the same size, and then repeat this process with multiple sizes since there is no *a priori* reason to believe in a certain threshold on the degree of token frequency that would cause the effect in question. For each size, one corpus will be formed by removing the most frequent words, and at least one will be formed by removing a randomly selected set of words.

I will find or elicit judgment data on nonce words that do not seem to cause analogy to specific English words more than usual. Then, I'll test each learned grammar on these words and compare their results to the English speaker results. I'll run a regression to see if the grammars that were trained on lexica with high token frequency words removed systematically performed better than the ones based on lexica with arbitrary sets of words removed.

Chapter 6

Conclusion

I will conclude with recommendations for phonotactic modeling, regarding the function that we should fit the data to and two of the potential factors, token frequency and participation in alternations, that we could use in those models.

I will also discuss SurveyMan and recommendations for running phonotactic judgment experiments in a convenient and effective way.

Bibliography

Adriaans, F. and Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language*, 62(3):311–331.

Albright, A. (2006). Gradient phonotactic effects: Lexical? grammatical? both? neither? In *Talk handout from the 80th Annual LSA Meeting, Albuquerque, NM.*

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

Albright, A. (2009). Modeling analogy as probabilistic grammar. *Analogy in grammar*, pages 185–213.

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

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.

Drummond, A. (2011). IbexFarm.

Goldwater, S. and Johnson, M. (2003). Learning OT constraint rankings using a Maximum Entropy model. In Spenader, J., Eriksson, A., and Östen Dahl, editors, *Proceedings of the Workshop on Variation within Optimality theory*, pages 111–120. Stockholm University.

Keller, F., Gunasekharan, S., Mayo, N., and Corley, M. (2009). Timing accuracy of web experiments: A case study using the WebExp software package. *Behavior Research Methods*, 41(1):112.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., and Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental science*, 9(2):F13–F21.

Legendre, G., Miyata, Y., and Smolensky, P. (1990). Harmonic Grammar – A formal multi-level connectionist theory of linguistic well-formedness: An

Application. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages 884–891. Lawrence Erlbaum Associates, Mahwah, NJ.

McDonnell, J., Martin, J., Markant, D., Coenen, A., Rich, A., and Gureckis, T. (2012). *PsiTurk (Version 1.02).* New York University, New York, NY.

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.

Pater, J. (2008). Cumulative ill-formedness in typological and experimental data. In *Conference on Experimental Approaches to Optimality Theory, University of Michigan. http://people. umass. edu/pater/pater-michigan. pdf. Accessed*, volume 20.

Pater, J. (2009). Weighted constraints in generative linguistics. *Cognitive Science*, 33(6):999–1035.

Pater, J. and Tessier, A.-M. (2003). Phonotactic knowledge and the acquisition of alternations. In Solé, M. J., Recasens, D., and Romero, J., editors, *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 1177–1180. Universitat Autònoma de Barcelona, Barcelona.

Potts, C., Pater, J., Jesney, K., Bhatt, R., and Becker, M. (2010). Harmonic Grammar with Linear Programming: From linear systems to linguistic typology. *Phonology*, 27(1):77–117.

Prince, A. and Smolensky, P. (1993/2004). *Optimality Theory: Constraint Interaction in Generative Grammar.* Malden, MA, and Oxford, UK: Blackwell.

Smolensky, P. (2006). Optimality in phonology II: Harmonic completeness, local constraint conjunction, and feature-domain markedness. In Smolensky, P. and Legendre, G., editors, *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*, pages 585–720. MIT Press/Bradford Books, Cambridge, MA.

Smolensky, P. and Legendre, G. (2006). *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar.* MIT Press, Cambridge, MA.

Werker, J. F. and Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, 7(1):49–63.