Forthcoming in J. Myers (Ed.) *In search of grammar: Empirical methods in linguistics*. Language and Linguistics Monograph Series 48. Taipei, Taiwan: Language and Linguistics.

Testing phonological grammars with lexical data*

James Myers
National Chung Cheng University
Lngmyers@ccu.edu.tw

July 11, 2011

1. Introduction

Most phonological research is actually corpus linguistics. This claim may sound odd, given that it is still relatively rare for theoretical phonologists to employ the sort of quantitative methods associated with corpus analysis (e.g. Frisch et al. 2004, Uffmann 2006). But think of the typical "problem set" in a phonology class: the data are not native-speaker judgments of novel forms, as in syntax, but fixed sets of lexical items, usually taken from linguist-compiled word lists. These data sets are corpora, in the sense that they are preexisting rather than generated through experimental manipulation (properly speaking, they are databases, summarizing data from corpora of fluent speech and quasi-experimental native-speaker elicitation). Phonologists have much to gain from experiments (see Hammond's contributions to this volume), whether they involve native-speaker judgments (e.g. Frisch and Zawaydeh 2001), phonetic measurements (e.g. Port and Leary 2005), or something else (e.g. Ohala 1986), and for phrasal phonology there has long been a tradition of analyzing productively generated forms, just as in syntax (e.g. Selkirk 1972). Yet the study of lexical phonology, which has played a key role in phonological research, especially since Chomsky and Halle (1968), relies heavily on the assumption that lexical patterns reveal something about the mental grammar that went into producing them.

Despite the importance of lexical data in phonological research, there is still some confusion over how corpus-based arguments should work in the study of grammar. In this chapter I hope to clarify the logic underlying phonological argumentation from word lists, and then show how the logic can be formalized as a set of simple computerized procedures. Unlike several recent attempts to automate phonological grammar analysis (e.g. Tesar and Smolensky 1998, Boersma and Hayes 2001, Hayes and Wilson 2008, Pater et al. 2007a, Coetzee and Pater 2008), my proposal is not a learning algorithm. Instead, it is meant as a tool for testing the statistical reliability of hypothesized grammars proposed as part of a traditional phonological analysis. Thus my proposal is intended to build on and clarify traditional methods, rather than representing a new theory of phonological competence or

* Work on this paper was supported by National Science Council (Taiwan) grants NSC 95-2411-H-194-005, NSC 95-2411-H-194-002, NSC 97-2410-H-194-067-MY3. My thanks to James S. Adelman, Tsung-Ying Chen, Bruce Hayes, Huichuan Hsu, Hui-chuan Huang, Joe Pater, and two anonymous reviewers for comments on various points discussed here. I apologize if evidence of their help may not always be obvious in the end result.

performance (further discussion of related issues is given in Myers 2007b, 2008, 2009a).

I lay the groundwork in section 2 by describing what role lexical data should play in phonological argumentation. I then describe, in section 3, how quantitative analysis can be naturally incorporated into traditional phonological methodology. In particular, I show how corpus data can be used to test grammatical hypotheses expressed in the framework of Optimality Theory (Prince and Smolensky 1993/2004), against the statistical null hypotheses that the proposed constraints or proposed ranking are not necessary. Explicit instructions are given for implementing the statistical methods in the free statistical software R, and these methods are applied to an analysis of a simple pattern in the Formosan language Pazih. Section 4 provides conclusions and considers future prospects of this way of looking at phonological methodology.

2. Corpus analysis in phonology

Because science is often naively thought of as inherently experimental, and since pronouncements against corpus analysis have been a staple of the generative literature for decades, I begin by reviewing why corpora are indeed valid data sources in the study of grammar. The purpose is not to argue that generative phonologists should take corpora analysis seriously, since in practice they already rely heavily on word lists, which are a kind of corpus (database). Rather, the purpose is to justify this standard practice, showing that it is entirely consistent with generativist philosophy, despite the superficial appearance of conflict with it.

In this section I first distinguish three distinct roles played by corpus analysis, only one of which is relevant here. I then show that despite well-known arguments in the generative literature, corpus frequency does indeed correlate with grammaticality. Finally I show that corpus analysis, despite its reputation as empiricist, actually requires a healthy dose of rationalism as well.

2.1 Corpora as performance evidence for competence

The goal of this chapter is to understand how phonologists can test grammatical hypotheses against lexical data. Though this is the most common way corpus data (i.e. databases) are used in phonology, it must be contrasted with two other goals that many researchers seem to think are more prototypical of corpus analysis. Namely, some researchers think of corpus analysis as concerned solely with the study of performance, not competence, while others focus on the use of corpora, not as data for testing hypotheses, but as input to grammar learners. In this section I clarify the distinctions among these three goals of corpus analysis, and explain how the goal of using corpora as evidence for grammar may be

justified.

One goal guiding corpus analysis is the description of performance for itself. This approach is founded on the correct observation that corpus data represent performance rather than competence. Chomsky (1957:15) pointed out that "the set of grammatical sentences cannot be identified with any particular corpus of utterances," and the same holds for a lexicon, or database of words. Whether or not a word ends up in a lexicon depends on many factors other than grammar, including historical accidents (borrowings and the invention and loss of concepts) and processing constraints (words cannot be too long nor too numerous). But these truisms do not mean that corpora can only be used to study performance itself. A corpus is the record of successful applications of grammar by native speakers, so it would be very surprising if corpora showed no trace of grammar at all, mixed in with those traces of historical accidents and processing constraints (this is in fact the central premise of a growing body of research in corpus linguistics, exemplified by the title of the new journal *Corpus Linguistics and Linguistic Theory*, founded in 2005).

Another goal that may be adopted in corpus research is to use corpora as input to automatic grammar-learning algorithms. At least since Chomsky (1965), generativists have claimed that this task is impossible without the learner having considerable built-in knowledge, so contemporary learning models are meant not as general-purpose discovery procedures, but empirical proposals about how actual children are innately built to acquire language (e.g. Tesar and Smolensky 1998, Boersma and Hayes 2001, Hayes and Wilson 2008, Pater et al. 2007a; Coetzee and Pater 2008).

This chapter, however, has a third goal, namely to use corpora (again, specifically word list databases) as evidence for grammatical hypotheses proposed by linguists, not child learners. The examination of preexisting word lists has historically been central in the arguments for virtually all innovations in theoretical phonology, from phonemes to features to rules to Optimality-Theoretic constraints. Even Chomsky and Halle (1968), both fluent English speakers (one native), and explicitly interested in productive mental grammar rather than lexical items for their own sake, cite Kenyon and Knott (1944), a pronunciation dictionary, as their gold standard for evidence.

Why should phonologists favor corpus analysis over the experimental data (elicited native-speaker judgments of acceptability) common in syntax? In the previous paragraph (and in the introduction to this book) I cited the burden of tradition: Phonological theory is founded on well over a century of corpus-based analysis, and it would be highly impractical to rederive all of the basic concepts using experimental data. Less trivially, even if the patterns in lexicons actually reflect diachronic processes more than synchronic grammar (Ohala 1986), these diachronic processes are also systematic and mental, and arguably are built of the same components (e.g. features, constraints) that compose synchronic mental grammar (see Blevins 2004, Hansson 2008, and Boersma and Hamann 2008 for arguments

against this traditional view, and Hayes and Steriade 2004, Kiparsky 2006, Martin 2007, and Moreton 2008 for empirical defenses of it).

Corpus data also make cross-linguistic comparisons much easier than would a reliance solely on judgment data. For this reason, phonologists have a much longer history of addressing typological issues than do syntacticians. Belying their titles, *The sound pattern of English* (Chomsky and Halle 1968) ranges far beyond English, whereas *Aspects of the theory of syntax* (Chomsky 1965) is almost exclusively restricted to aspects of English syntax. It is not surprising, then, that Optimality Theory, the currently dominant theory in phonology, is so typologically oriented.

Finally, given the dual status of a corpus as output of language use and input to language learning, shifting to judgment-based analyses would not avoid the necessity of corpus analysis. Speakers cannot judge words or nonce forms without checking them for familiarity or comparing them with other words (Bailey and Hahn 2001). In other words, speakers make phonological judgments at least partly by conducting a sort of naive corpus analysis. Researchers who do not anticipate them with formal corpus analyses of their own might misinterpret the effects of superficial analogy for true grammar.

The increasing use of experimental data in the testing of phonological hypotheses is certainly welcome. Nevertheless, given the well-justified role of corpus data in phonological theory, it remains important to understand how they may best be analyzed.

2.2 Grammaticality and corpus frequency

Despite the solid reasons for testing phonological hypotheses with corpora, implicitly recognized throughout the history of phonology, corpus data cannot be treated as intrinsically "direct" any more than any other sort of performance data. In particular, whether or not a given form is attested in a corpus, and how frequent it is if it is attested, is logically independent of its grammaticality. Nevertheless, the traditional anti-corpus rhetoric in the generative literature overstates the case. In this section I explain why, and show how attestation and frequency actually provide very useful information about grammaticality (similar points are made regarding corpus-based syntactic analyses by Stefanowitsch 2005).

The nonidentity of grammaticality with corpus attestation and frequency follows immediately from the recognition of the latter as performance data, affected by numerous factors other than grammar. In lexical phonology the most important of these factors is rote memory. Lexical exceptions survive in a corpus, not by virtue of grammar, but in spite of it. Hence they are, strictly speaking, ungrammatical. Grammatical theory per se has no need to explain them, since they are already sufficiently explained by the mind's ability to memorize (on display, for example, whenever someone learns the correct pronunciation of a foreign name). Therefore, efforts such as those of Inkelas et al. (1997) and Pater (2009) to develop a

grammar-internal theory of exceptions may be missing the point, by obliging grammar to account for every aspect of performance (unless such efforts are seen as models of competence/performance interactions).

Yet given their origins in language use by people who know their grammar, lexicons must also reflect grammatical knowledge as well. In particular, we expect there to be a correlation between the grammaticality of a form type (i.e. a class of words defined by phonological structure) and its type frequency in the corpus. This assumption is no different in kind from the standard assumption of a correlation between grammaticality and native speaker judgments of acceptability.

The fact that this rather banal observation has sometimes been resisted by generative linguists can be explained, I think, by confusion over the three different goals of corpus analysis described in the previous section. For example, when Chomsky (1957:16) says that grammaticality cannot be equated "with the notion 'higher order of statistical approximation'," or when Chomsky (2002:102) says that "[i]f you took a videotape of things happening out the window, it would be of no interest to physical scientists," the argument assumes that corpus linguistics necessarily involves either a non-analytic description of the corpus itself or, at best, automatic grammar learning of a highly superficial sort. Yet corpus analysts who use observations to test prespecified hypotheses, rather than merely describing surface forms or building automatic grammar learners, entirely side-step these sorts of criticisms.

The logic runs like this. The only way a grammar can reveal its existence in a corpus is by frequency differences among competing form types, since a grammar that results in equal numbers of "grammatical" and "ungrammatical" forms cannot be detected. To take a schematic example, suppose a phonologist hypothesizes that structure [+X] is grammatical, and that [-X] is ungrammatical. A critic points out that examples of [-X] words actually exist in the lexicon. A reasonable defense of the original hypothesis is still possible, however, if the number of [-X] words is "sufficiently low" to dismiss them as lexical exceptions.

In one of the rare discussions in the generative literature of the logic of phonological corpus analysis, Duanmu (2004) describes the correlation between grammaticality and frequency in a more complex way. He acknowledges that the higher the frequency of a structure in a corpus, the more likely it is to be grammatical. Yet he suggests that for ungrammaticality, the inference runs in the opposite direction, from grammar to corpus; namely, the less grammatically well-formed a structure, the less likely it is to appear in the corpus. This apparently follows from the assumption that there is only one way for a form type to be common in a corpus, namely by being grammatical, but there are two ways for a form type to be rare, either by being ungrammatical (systematic gaps or exceptions) or by being grammatical but disfavored for some other reason (accidental gaps like historical accidents).

As a practical matter, however, there is no need to assume this asymmetry. First, there are also extra-grammatical ways for a form type to become common, such as superficial analogy and borrowing of lexical classes. Secondly, commonness and rarity are defined relative to each other, so the crucial factor is the difference in number between the two types, not the size of each category separately. Nonsystematic forces like borrowings and exceptions are expected to influence these proportions relatively randomly, making it possible, at least in principle, for systematic forces like grammar to stand out from the noise. These considerations suggest that quantification might indeed be useful in the analysis of phonological corpora, as explained more fully in section 3.

2.3 Universals in phonological corpus analysis

Very likely the main reason for the confusion over the different goals of corpus analysis is the fact that a corpus is intended to contain observations collected with a minimum of a priori assumptions. This fools many linguists, on both sides of the rationalist/empiricist divide, into thinking that corpus analysis itself must also be conducted with a minimum of a priori assumptions. However, this certainly is not how corpus analysis is usually employed in generative phonology, which adopts a decidedly rationalist, universalist approach. In fact, as I argue in this section, grammar-oriented corpus analysis cannot be done without making prior assumptions about what one is looking for.

To set the scene for the quantitative models in section 3, consider a simple statistical example. The chance probability of choosing any particular playing card from a deck is 1/52. This means that no matter which card one chooses, the result is so unlikely to have occurred by chance that it meets the usual criterion for statistical significance ($p = 1/52 = .019 < .05$). Yet clearly such a "finding" is meaningless without establishing ahead of time which card is expected. In a similar way, any set of n items in a corpus is just as unlikely to be selected for discussion as any other set of n items, and without prespecifying why one set represents a pattern and not the others (i.e. why the theorist considers them "grammatical"), there is no point trying to make an argument on p values alone.

In other words, to make the case that a corpus pattern is statistically significant, one first has to define what will count as a "success." There are essentially two ways to do this. One is more rationalist: even before looking at the corpus, declare what pattern is predicted based on universal principles. The other is more empiricist: look at the corpus, see what patterns seem to be present, and then test whether they are truly statistically significant (highly unlikely to have arisen by chance). In practice, both methods are often used in conjunction. For example, the more empiricist method is often used when studying a new phonological pattern in a previously unfamiliar language, but if the pattern can be expressed in terms of a universal principle, the more rationalist method is used to test the principle in other languages as well.

It might be objected that there actually is a single "correct" analysis of a phonological corpus, namely the grammar acquired by the child exposed to it, using her innate biases. These innate biases may, in principle, cause the learner to ignore statistical significance entirely. For example, some corpus could have equal numbers of [+X] and [-X] words, but the child may still innately know that [+X] must be grammatical and [-X] ungrammatical. Such considerations are important (see Myers 2009a for further discussion), but they certainly do not undermine the case for corpus analysis. If the grammar is productive, it should leave a record in the corpus through earlier adult productions, so situations like this should be quite rare. Even if they do arise, they would do nothing more than add yet another argument for supplementing corpus analysis, which can only tell us what language producers did in the past, with experimental tests of contemporary speakers, which can tell us what language producers and comprehenders are doing today.

In summary, then, the analysis of corpora as evidence for grammar deserves its key role in the methodology of generative phonology, since contrary to the anti-corpus rhetoric, frequency does correlate with grammaticality, and it is not a purely empiricist exercise because the researcher must start with prespecified hypotheses. In the remainder of this paper, we will see how these notions can be implemented quantitatively.

3. Quantitative evidence for grammar in phonological corpora

In this section I show how a hypothesized phonological grammar, represented in the familiar notation of Optimality Theory, can be tested against the statistical null hypothesis that the grammar does no work in accounting for the lexical data. The mathematical background is given in section 3.1, where I review the notion of constraint weights and the statistical models designed to find them. In section 3.2 I describe methods for testing the statistical significance both of individual constraints and of their ranking. The methods are illustrated in analyses of a pattern in the Formosan language Pazih (further applications to Mandarin phonotactics are described in Myers 2008, 2009a). Section 3.3 addresses some possible limitations of this approach.

3.1 Mathematical modeling of Optimality Theory

We want a model of phonological grammar that defines empirical success in terms of explicit prespecified components (as argued in section 2.3), and it would be even better if the model were mathematically simple and already familiar to practicing phonologists. Fortunately, such a model exists: Optimality Theory. OT, the lingua franca of contemporary theoretical phonology, claims that a grammar consists of the strict ranking of a set of universal constraints. This not only defines empirical success in an explicit way (a constraint

either plays a statistically significant role in accounting for the data, or not, and likewise for a ranking), but the notions of constraints and constraint ranking turn out to provide a convenient framework for statistical formalism.

I start in 3.1.1 with background on the implicit role of constraint weights in Optimality Theory, then review the automatic fitting of weights in 3.1.2.

3.1.1 Optimality Theory and harmony theory

The idea underlying the methods described below is the relationship between OT and Harmonic Grammar (HG) (Prince and Smolensky 1993/2004, 1997). Since this relationship may not be familiar to all readers, I first review it here.

Both OT and HG assume that knowledge (e.g. of a grammar) can be described in terms of violable constraints. The major difference is that unlike HG, competition between constraints in OT is resolved via strict ranking: if constraint Con_1 outranks constraint Con_2, then Con_2 can only choose the output if Con_1 happens not to discriminate between the candidate outputs. This logic is familiarly expressed in so-called tableaux like those in (1), where the stars represent violations by the given candidate outputs of the given constraints, ranked by the grammar (here $Con_1 >> Con_2$). In each tableau, the pointing finger indicates the candidate output that is in the set of candidates that least violate the highest-ranked constraint, and among those, is also in the set of candidates that least violate the next-highest-ranked constraint, and so on.

(1) a.

InA	Con_1	Con_2
☞ $OutA_1$		*
$OutA_2$	*	

b.

InB	Con_1	Con_2
$OutB_1$		*
☞ $OutB_2$		

c.

InC	Con_1	Con_2
$OutC_1$	*	*
☞ $OutC_2$	*	

There is a mnemonic for teaching the logic of OT that is also quite useful in explaining

its relation to HG. Namely, one imagines that the blank cells in the tableau are zeroes, and the stars are ones (or whatever the number of stars is). Then one reads the rows of digits, from left to right, as comprising a number in the ordinary decimal system (as long as no cell has more than nine stars). The optimal candidate will then be the one associated with the lowest valued number (the most "harmonic" candidate). In the case of the tableaux in (1), this procedure yields the values and winning candidates shown in (2).

(2) a. $OutA_1$: 01 = 1
$OutA_2$: 10 = 10 1 < 10, so $OutA_1$ wins

b. $OutB_1$: 01 = 1
$OutB_2$: 00 = 0 0 < 1, so $OutB_2$ wins

c. $OutC_1$: 11 = 11
$OutC_2$: 10 = 10 10 < 11, so $OutC_2$ wins

This procedure works because a value notated in the decimal system represents the sum of the digits weighted so that the further left a digit is, the heavier its weight. Thus "24" represents 2 times ten plus 4 times one; the leftmost digit is associated with a larger weight (ten) than the rightmost digit (one). Like the decimal system, OT ranking is a kind of weighting system that gives higher weights to the higher-ranked constraints. Also like the decimal system, OT ranking is designed so that no lower-ranked constraint, or gang of constraints, can "outvote" a higher-ranked constraint. The decimal system does this by using digits ("violations") that never go over nine, and the implicit weights in OT have the same effect (see Prince 2007 for a more sophisticated variation on this basic idea).

The logic behind (2) can be expressed more formally as in (3a), where $Star_i$ represents the number of violations of constraint Con_i and w_i represents the weight of Con_i. The systematic increase in constraint weight that forces constraints to be strictly ranked is ensured by the conditions in (3b).

(3) a. Candidate harmony value = $w_1 \times Star_1 + w_2 \times Star_2 + ... + w_n \times Star_n$
b. $w_1 = base^{n-1}$, $w_2 = base^{n-2}$, ... $w_n = base^0$, and base > max(Star)

Harmonic Grammar is a generalization of Optimality Theory (or, historically speaking, OT is a special case of HG), where there are no built-in restrictions on the weighting system. That is, it assumes (3a) but not (3b). Thus it is possible for a lower-weighted constraint to "outvote" a higher-weighted one, or for a set of lower-weighted constraints to "gang up" on higher-weighted ones.

Although HG predates OT, there has recently been a resurgence of interest in weight-based models, for two major reasons. First, some argue that the "fuzzy" ranking of constraints does a better job act capturing actual language data (e.g. Keller 2000; Pater et al. 2007b; Coetzee and Pater 2008). Second, some are attracted by the mathematical connection to well-established techniques in computational linguistics (e.g. Goldwater and Johnson 2003; Pater et al. 2007a; Hayes and Wilson 2008). The latter reason is more relevant to the goals of this chapter, since the automatic setting of constraint weights from corpus data underlies the statistical tests of OT grammar that I describe later.

3.1.2 Loglinear models

With corpus data, the most useful type of weight-fitting model is the family of so-called loglinear models. In this section I unpack the term "loglinear," starting with "linear."

Loglinear models are a kind of linear regression model. The goal of linear regression is to analyze a scatterplot of data, where each dot in the scatterplot represents a pair of observed input (x-axis) and output (y-axis) values, in order to find the best linear fit (straight line) through the cloud of dots. The closer the cloud of dots to the line, the better the line describes the actual data. The associated *p* value then indicates that how probable it is for a cloud this linear to arise by chance.

Like all lines, this best-fitting line can be described using a linear equation of the form shown in (4), where the slope is expressed as a weight on the input variable and the intercept represents the value of the output when the input value is zero (i.e. where the line crosses the y-axis).

(4) Output = Intercept + Weight × Input

The weight and intercept are not chosen by hand, but are calculated automatically from the data by a method called maximum likelihood estimation. This method finds the parameter values for the line maximizing the likelihood that this line underlies the cloud of data.

A similar procedure applies if we are trying to predict the output from two or more input values simultaneously. With two input variables, for example, we would have a boxlike three-dimensional scatterplot, with the base plane defined by axes representing the two input values, and elevation off the base representing the output values. The line that best fits the three-dimensional cloud of data points can be expressed with a regression equation like that in (5). Among other things, this equation expresses the fact that the two input axes are orthogonal (at right angles) to each other by adding their values together, so that the height of a data point off the base plane is the sum of the independent effects of each input.

(5) $\text{Output} = \text{Intercept} + \text{Weight}_1 \times \text{Input}_1 + \text{Weight}_2 \times \text{Input}_2$

In the case of testing an OT grammar, the input variables are the numbers of violations of our proposed constraints, and the weights are the constraint weights. To explain what the output represents, we must now unpack the "log" part of "loglinear."

It is not really appropriate to treat type frequencies in a corpus the same way as a continuous-valued measure like naming time. Type frequencies can be thought of as category sizes or probabilities (e.g. the probability of a random corpus item falling into one category rather than another). These measures relate to discrete observations (i.e. the individual corpus items). If we simply add up the factors affecting these category sizes or probabilities, as in ordinary linear regression, we will not describe the output correctly. For one thing, neither category sizes nor probabilities can go below zero, yet the right side of the equation describes an infinitely long line that goes below zero. Another problem is that independent probabilities should not be added, but multiplied (e.g. in a dice game the probability of rolling a six is 1/6, and the probability of rolling two sixes is (1/6) × (1/6)). Putting these points together, we really should be looking at an equation with a product on the right side, not a sum. Unfortunately, products are mathematically harder to deal with than sums.

The solution is to exploit the mathematical trick shown in (6). Namely, a product containing two variables x and y can be converted into the sum of x and y if we start out with x and y as powers of some base value (here, the famous constant $e = 2.71828+$) and then take the logarithm of the same base. Hence if we define our regression equation carefully, we can take the logarithm of both sides and convert the right side to a conveniently linear sum (and the left side to a value that can, in principle, range infinitely below and above zero). This is a loglinear model.

(6) $\log_e[(e^x)(e^y)] = \log_e[e^{x+y}] = x + y$

Loglinear models are all closely related to each other mathematically, but different variants have different strengths. For most linguists, the most familiar type of loglinear model is logistic regression, the workhorse at the heart of the VARBRUL program widely used in sociolinguistics (Mendoza-Denton et al. 2003). Logistic regression is used when observations are binary, such the choice among two competing allomorphs in a sociolinguistic corpus. This type of regression does not suit our purposes, however, since we are interested in the sizes of the categories attested in a corpus, not in competing variants.

Another type of loglinear model is more familiar to computational linguists, who express maximum likelihood in terms of the related information-theoretic notion of maximum entropy. Recently there have been a growing number of studies applying maximum entropy models to Harmony Grammar, including Goldwater and Johnson (2003)

and Hayes and Wilson (2008), with Pater et al. (2007a) and Coetzee and Pater (2008) using different algorithms for a similar purpose. Like non-loglinear OT learning algorithms (e.g. Tesar and Smolensky 1998, Boersma and Hayes 2001), a maximum entropy model is given a set of harmonic constraints, a set of input forms, a set of candidate outputs for each input, and an indication about which of these candidates is the winning output. The models then learn the best weights consistent with the trained input-output pairs, if there are any.

These types of loglinear models are automatic grammar learners, and unlike the logistic regression models used by sociolinguists, are not intended to test for statistical significance (though there is no principled reason why they could not be modified to do so; Wilson and Obdeyn 2009 has been cited as providing statistical significance tests for lexical patterns, though the paper is not yet publicly available). The general lack of interest in significance testing may be partly because the traditional focus of the maximum entropy literature has been in engineering rather than science per se, but perhaps also because the modelers recognize the possibility, noted at the end of section 2.3, that actual child language learners may not care about statistical significance. These models also do not test for the statistical reliability of constraint ranking, which makes sense given that the algorithms are designed to learn harmonic grammars, not strictly-ranked OT grammars.

What kind of loglinear regression model would be most appropriate for testing linguist-proposed OT hypotheses about corpus data? Like logistic regression, the left side of the equation should represent corpus observations and the right side our grammatical hypotheses about them. Unlike logistic regression, however, the left side of the equation should relate to the sizes of categories, namely the type frequencies for structures predicted to be grammatical vs. structures predicted to be ungrammatical.

Fortunately, there is a well-established type of loglinear model ideal for our purposes. It is called Poisson regression, named after the French mathematician Siméon-Denis Poisson (Agresti 2002). Chance probability in this type of regression is defined in terms of the Poisson distributions associated with count data. Poisson distributions are asymmetrical, since counts cannot go below zero and small values are more common than large ones (counting up to a higher number entails counting up to lower ones first, but the reverse is not the case).

In the following section, then, I show how Poisson regression can be applied to test the statistical significance of OT constraints and their ranking.

3.2 Testing OT grammars

In this section I show how OT-like grammatical hypotheses can be tested statistically, giving all of the commands necessary for running the analyses using R, the free statistics software package (R Development Core Team 2011; see also Baayen 2008, Johnson 2008,

Hammond this volume). The procedures are illustrated with data from the Formosan language Pazih (or Pazeh; Li and Tsuchida 2001). Other aspects of the analytical problems posed by the Pazih data are discussed in Myers (2007), and applications of the basic logic to other data sets are given in Myers (2008, 2009a). Pazih represents a good test case of corpus analysis methodology because like a growing number of languages studied by phonologists, Pazih is nearly extinct, and very soon all the world will have left is a fixed record of the language as it was once spoken.

The logic works roughly as follows. Since the mathematics of linear modeling (loglinear or otherwise) treats components on the right side of the equation as orthogonal to each other, we should be able to test the independent contributions of each constraint; we may find, for example, that one constraint is statistically significant but another is not. To test a ranking hypothesis, we can compare the constraint weights; the constraint ranking $Con_1 >> Con_2$ would be supported if we found that the weight w_1 was significantly larger than the weight w_2. Testing individual constraints is a more basic task, so I describe this first, in section 3.2.1. The test for ranking is discussed in section 3.2.2.

3.2.1 Testing OT constraints

Pazih has a set of 45 morphemes consisting of reduplicated CVC syllables with an epenthetic medial vowel (listed in Li and Tsuchida 2001:20-21). In 33 of these morphemes, all of which are listed in (7), the epenthetic vowel is identical to that of the reduplicated syllables. However, in the twelve morphemes listed in (8), the epenthetic vowels do not show vowel harmony (note the minimal pair in (7dd) vs. (8j)). It is this small corpus of 45 items that is the focus of the next few sections.

(7)	a.	bakabak	'native cloth'	b.	dapadap	'to rub'
	c.	hayahay	'stalk of miscanthus'	d.	ma-hakahak	'itchy in taste'
	e.	ŋaraŋar	'to bite'	f.	bələbəl	'banana'
	g.	bəŋəbəŋ	'loom'	h.	dəkədək	'to step on ground'
	i.	dəŋədəŋ	'to boil in water'	j.	i-dəkədək	'to sink'
	k.	ləŋələŋ	'to aim'	l.	ma-bətəbət	'suffer from gas pain'
	m.	maa-bəzəbət	'help each other'	n.	rəxərəx	'to wrestle'
	o.	ma-hərəhər	'suffer from asthma'	p.	pa-gənəgən	'to hum'
	q.	sərəsər	'to repeat'	r.	zəkəzək	'to erect'
	s.	girigir	'to saw'	t.	hirihir	'to grind'
	u.	ŋiriŋir	'to nibble'	v.	mu-tinitin	'to weigh'
	w.	pa-kihikih	'dry cough'	x.	bukubuk	'bamboo pipe'
	y.	busubus	'smoke'	z.	dukuduk	'ginger'

aa.	duŋuduŋ	'drum'	bb.	dusudus	'to grate'
cc.	gunugun	'bucket, to measure'	dd.	huruhur	'steam, vapor'
ee.	kulukul	'type of bird'	ff.	ŋuruŋur	'to scold'
gg.	pa-surusur	'to masturbate'			

(8)	a.	barəbar	'flag'	b.	dakədak	'to scrape off with feet'
	c.	parəpar	'paper'	d.	ma-lədalət	'to tremble'
	e.	birabir	'tough as of meat'	f.	ma-bidabit	'to wobble'
	g.	ma-ŋiraŋir	'easy'	h.	buhabuh	'to powder one's face'
	i.	burabur	'dust'	j.	hurahur	'bald'
	k.	ma-buxibux	'very tired'	l.	muximux	'to gargle'

Setting aside a number of interesting phonological issues (see Myers 2007), one way to analyze this pattern would be to describe regular (non-exceptional) words like those in (7) as having no underlying value for the epenthetic vowel, which receives its surface form by obeying a universal vowel harmony constraint AgreeV. Exceptions like those in (8), where the value of the epenthetic vowel appears to be unpredictable, specify this value underlyingly, and preserve it via a faithfulness constraint IdentV. To prevent AgreeV from nullifying IdentV in (8), we can hypothesize the ranking IdentV >> AgreeV. This analysis can be summarized in the tableaux (9) and (10) for regular words and exceptions, respectively.[1]

(9) 'steam'

hur-V-hur	IdentV	AgreeV
☞ hur-u-hur		
hur-a-hur		*

(10) 'bald'

hur-a-hur	IdentV	AgreeV
hur-u-hur	*	
☞ hur-a-hur		*

Note that implicit quantification plays a crucial role in this argument, allowing us to classify the words in (7) as regular and those in (8) as exceptions. If it had been the case that

[1] An anonymous reviewer suggests an alternative analysis using ParseV (requiring realization of the epenthesized vowel) and DepV (banning insertion of new material). Together they favor vowel harmony in regular words (fulfilling ParseV without violating DepV) and the realization of the underlying nonharmonic vowel in exceptions (again obeying both constraints). Since both constraints are obeyed in all words, no ranking is necessary, making this analysis irrelevant for my goal of showing how OT ranking hypotheses can be tested statistically. However, the existence of this alternative analysis does highlight that the purpose of my statistical model is to test linguist-invented hypotheses, not to learn grammars directly from data.

the great majority of words had epenthetic vowels different from the base syllables, we would never have considered any role for AgreeV in this language, with apparently "harmonizing" examples dismissed as statistical flukes.

Now we want to put this implicitly quantitative argument on firmer statistical footing. Namely, are 33 vowel-harmonizing morphemes out of 45 truly enough to justify AgreeV? After all, 12/45 exceptions is 27%, quite a substantial proportion, especially given that Pazih only has four phonemic vowels to work with. Justifying IdentV seems even more difficult, given the argument in section 2.2 that lexical exceptions should be considered ungrammatical. For example, if there were 44 harmonizing morphemes and only one exception, the proportion of data points in support of IdentV would drop to 1/45, which seems to be too low to justify its existence in the grammar proper, given the mind's extra-grammatical ability to memorize arbitrary exceptions.

Poisson regression can help clarify the situation. The first step is to classify words in the corpus in terms of the grammar. As in the schematic example in 2.2., we can think of each proposed constraint as dividing the corpus into two categories, grammatical vs. ungrammatical relative to this constraint. In the case of the Pazih data, the hypothesized grammar puts lexical items into four categories, defined by the two evaluations associated with AgreeV (violate vs. obey) times the two evaluations associated with IdentV (violate vs. obey). By crossing the two categorization schemes, we define orthogonal axes that decompose their contributions independently, allowing us to test both constraints simultaneously in the same statistical model.

To make the constraint evaluations consistent with the procedure described in section 3.1.1, we code violations with 1 (star) and use 0 (blank cell) for instances where the constraint is obeyed. We then count the number of items of each type in the corpus, and tabulate the information as in (11).

(11) Raw count table

Count	AgreeV	IdentV
33	0	0
0	0	1
12	1	0
0	1	1

The table in (11) should not be confused with an OT tableau. First, it summarizes information about many items, not just a single item. Second, it does not represent input-output pairs. Rather, it encodes item representations in terms of the constraints they violate. In doing this, it is similar to the OT-based representational scheme proposed in

Golston (1996), except that it also permits faithfulness constraints like IdentV to be included as part of the description. Thus the descriptions encode both output forms (via markedness constraints) and input forms (via faithfulness constraints). From the very start, then, the data are encoded in a theory-governed way; formally, each row in (11) represents a so-called violation profile (Samek-Lodovici and Prince 2005) implied by the constraints of the proposed grammar. As argued in section 2.3, a theory-governed description is always necessary for hypothesis testing in corpus analysis, in order to avoid circularity.

Now we want to use Poisson regression to fit an equation like that in (12). Since the values of AgreeV and IdentV represent violations, we expect that the weights for both will be negative, since counts should be higher for categories where constraints are obeyed (evaluated as 0) and lower for categories where constraints are violated (evaluated as 1). In addition, we hope that both weights are significantly different from zero. That is, the chance probability of getting weights with magnitudes as great (i.e. as far from zero) as ours should be "sufficiently low" (conventionally, $p < .05$ is considered to be statistically significant).

(12) $\log_e(\text{mean}_{\text{Poisson}}) = \text{Intercept} + \text{Weight}_1 \times \text{AgreeV} + \text{Weight}_2 \times \text{IdentV}$

Unfortunately, before continuing this example we must make an adjustment to accommodate an intrinsic limitation of the maximum likelihood estimation method. Namely, this method assumes that none of the correlations are perfect, so if any is, the algorithm will crash and the results will be nonsense. In our case, Count = 0 always occurs whenever IdentV = 1. The effect is that IdentV = 0 holds for all attested tokens, which makes phonological sense (vowel harmony presumably does not change underlying feature values), but it wreaks mathematical havoc. There are alternative methods that can avoid this limitation; one promising one is exact loglinear regression, which has the added advantage of giving equally reliable results for any sample size (e.g. Agresti 2002). Unfortunately it is computationally intensive, since *p* values are computed by counting all logically possible outcomes rather than relying on distribution-based estimates, and R currently does not implement the necessary algorithms.

One simple way out is to tweak the data to make the correlations less than perfect. This can be done by replacing all zero counts with one, which makes all categories non-empty while minimizing the effect on the total count. In the current example, this gives us the modified data table in (13). Though this adjustment slightly reduces pattern strength (e.g. IdentV is actually exceptionless, but (13) implies that it is violated twice), the drop in strength generally is not so great as to make a real pattern non-significant, even in rather small data sets like this one (e.g. in the adjusted data set IdentV is still obeyed far more often than violated).

(13) Adjusted count table

Count	AgreeV	IdentV
33	0	0
1	0	1
12	1	0
1	1	1

Now we are ready to run the Poisson regression. To do this in R, we must save the table in (13) as a tab-delimited text file. The easiest way to do this is by creating the table in a spreadsheet program like Excel, including the headings, and then copying the cells and pasting them into a text editing program (like Window's Notepad). Let us call this file Pazih.txt.

After R has been downloaded, installed, and started up, we change its file directory to the location of the data file so that R can find it. R is a command-line program, so we must now type in commands in the main R window to load the data and name it, as in (14) (file names must be given inside quotes, since they are treated as text strings). For our file Pazih.txt, the command would be as follows. In these examples, italicized elements represent names that should be changed depending on the particular study, while the rest should be entered exactly as shown. Note that in R syntax, $Y = x$ is equivalent to $Y <- x$ (where <- stands for a leftward pointing arrow), which means that the value x is assigned to the variable Y. Text following # is ignored by R, so we can add explanatory comments. The *F(X)* syntax represents a function F operating on an argument X.

(14) *Tableau* = read.table("*Pazih.txt*", T) # Load data (with "header" set to TRUE)
attach(*Tableau*) # Make column names (headers) available to further commands

We can even remove zero counts (changing (11) to (13)) in R automatically by using the command in (15), which searches through Count for values equal to zero ($X == Y$, with the double equals sign, is a logical statement meaning "X is equal to Y"), and when it finds one, changes this value to one (assigning it via the single equals sign).

(15) *Count*[*Count* == 0] = 1

Once this is done, running the Poisson regression simply involves typing the command in (16), where the variables Count, AgreeV, and IdentV match the column names in the data file. The $Y \sim X1 + X2$ syntax represents the regression equation, the "glm" command runs a generalized linear model (i.e. a regression model where the linear part is created by

transformation, in this case by taking the logarithm), the "family" tag indicates that we want to use Poisson distributions for count data, the "summary" command generates a compact overview of the results, the "$coefficients" tag pulls out from this summary just the table showing the constraint weights and their associated *p* values, and the round() function rounds the values to five decimal places so that the results are displayed in neat columns.

(16) round(summary(glm(*Count* ~ *AgreeV* + *IdentV*, family = poisson))$coefficients, 5)

In the case of the adjusted data in Pazih.txt, the command in (16) gives the table in (17).

(17)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.48288	0.17423	19.98968	0.00000
AgreeV	-0.96141	0.32609	-2.94830	0.00320
IdentV	-3.11352	0.72264	-4.30851	0.00002

The estimates are the constraint weights and *Pr(>|z|)* represents the *p* values (the standard errors and *z* values are used to compute the *p* values). Note first that the weights for AgreeV and IdentV are both negative, as desired: violations of them (coded 1) are associated with lower type frequencies in the corpus. Moreover, the *p* values for both constraints represent statistically significant results, since both are below the traditional cut-off value of .05. This is so despite the need to add spurious data points that weakened evidence for the two constraints.

In contrast with this procedure, OT and harmonic grammar models that automatically learn grammars from corpus data only assign weights, but do not test if they are statistically significant (again, aside from the model under development by Wilson and Obdeyn 2009). For example, with the (original unaltered) Pazih data, the Harmonic Grammar with Linear Programming (HaLP) model of Pater et al. (2007a) assigns a weight of 2.0 to IdentV and a weight of 1.0 to AgreeV, but it does not give any associated *p* values. Moreover, these same weights will be derived even if only one of the 45 items is non-harmonizing (giving almost no evidence for IdentV), or if all but one of the 45 items is non-harmonizing (giving almost no evidence for AgreeV). Other learning algorithms, such as the Gradual Learning Algorithm of Boersma and Hayes (2001), as implemented in Praat (Boersma and Weenink 2011), share this behavior. Whether or not actual children show one-trial learning of grammars, such models are certainly not helpful for researchers interested in what a corpus reveals about the grammar underlying it. Since a corpus is performance data, it is inherently noisy. Thus a complete analysis must include a measure of this noisiness, which, very roughly speaking, is what the *p* values represent.

Because of the way regression factors out the contributions of individual input variables, the procedure described here can be generalized to any number of constraints. For example, with three constraints, the Count column would represent eight (2^3) type frequencies. The procedure is a bit more complex if the analysis includes constraints that evaluate some items with more than one star, such as inherently gradient constraints or categorical constraints that happen to be violated more than once by the same item (e.g. NoCoda in a word with several codas). The difficulty is not gradience itself, since for any given constraint the regression approach can test whether the number of stars and the associated category sizes are inversely correlated. However, multiply violated constraints will increase the number of categories that must be tested. For example, an analysis with two constraints, each of which can give evaluations of up to two stars, defines nine (3^2) categories, in contrast to the four categories of the two-constraint analysis of Pazih. If constraint evaluations can be continuous, as suggested in Kirchner (1997), the procedure breaks down completely.

Note that so far, the procedure is actually consistent with Harmonic Grammar as well as OT, since both assume constraint weighting. It should also be noted that even if a constraint proves to be unsupported statistically (i.e. empirically), there may be a priori (i.e. rationalist) reasons for maintaining it in the analysis anyway. In OT, for example, markedness constraints that never apply in some language because they are dominated by faithfulness constraints are assumed, nevertheless, to exist, since it keeps the overall (universal) theory simpler by allowing grammars to be described as varying only in constraint ranking, not in constraint inventory. Another way to put it is that absence of evidence is not the same as evidence of absence.

3.2.2 Testing OT constraint ranking

The observant reader will have noticed that not only are the weights in (17) for AgreeV and IdentV negative and statistically significant, as desired, but the magnitude of the weight for IdentV (-3.11) is greater than that for AgreeV (-0.96). This is also consistent with the OT analysis, since we expect that AgreeV, though active in the grammar, is nevertheless ranked lower than IdentV. With only two constraints Con_1 and Con_2, each of which is violated no more than once, finding that the difference in the associated weights is in the appropriate direction ($w_1 > w_2$) is sufficient to demonstrate the strict OT ranking of $Con_1 >> Con_2$. This is demonstrated in (18) and (19) using the original OT tableaux for Pazih with the regression-determined weights.

(18) 'steam'

a.

hur-V-hur	IdentV w = -3.11	AgreeV w = -0.96
☞ hur-u-hur		
hur-a-hur		*

b. hur-u-hur: (-3.11)(0) + (-0.96)(0) = 0
hur-a-hur: (-3.11)(0) + (-0.96)(1) = -0.96 |0| < |-0.96|, so hur-u-hur wins

(19) 'bald'

a.

hur-a-hur	IdentV w = -3.11	AgreeV w = -0.96
hur-u-hur	*	
☞ hur-a-hur		*

b. hur-u-hur: (-3.11)(1) + (-0.96)(0) = -3.11
hur-a-hur: (-3.11)(0) + (-0.96)(1) = -0.96 |-0.96| < |-3.11|, so hur-a-hur wins

Yet how can we be sure that this difference in constraint weights is not a statistical fluke? Fortunately, there turns out to be a very simple way to find out (as pointed out to me by James S. Adelman, personal communication, September 7, 2007). The technique builds on the fact that linear regression equations involve sums of components on the right side. Given a pair of equations that are identical except for an extra component in one of them, we can conclude that the extra component makes a significant contribution if the more complex equation fits the observed data better. The general tool for comparing simple and complex equations is called a likelihood ratio test, and in the case of loglinear regression, the version we want relies on the analysis of deviance (analogous to the more familiar analysis of variance or ANOVA, used when comparing ordinary linear regression equations).

To see how this leads to a test for comparing weights, let us start with the two equations in (20), where O stands for the output, w_0 for the intercept, w_i for the weight associated with input variable I_i, and w_iI_i for their product. Equation (20a) represents the null hypothesis (symbolized H_0) that the two constraints have identical weights (no ranking), and equation (20b) represents the alternative hypothesis that they have different weights (symbolized H_r, the *r* standing for "ranked"). Though output O and input I_i values are identical across the two equations because they come from the same data set, this is not the case for the weights, which are fit separately via maximum likelihood estimation (this point will become crucial

later).

(20) a. H_0: $O = w_0 + w_1I_1 + w_2I_2$, where $w_1 = w_2$
b. H_r: $O = w_0 + w_1I_1 + w_2I_2$

Now we rewrite (20a) as (21a), since this equation assumes that the weights for both inputs are identical. We transform (20b) into (21b) in a more complex way, rewriting the weights w_1 and w_2 in terms of new values as $w'_1 + w'_2$ and $w'_1 - w'_2$, respectively (that is, $w'_1 = (w_1 + w_2)/2$ and $w'_2 = (w_1 - w_2)/2$). The reason for doing this will become clear shortly.

(21) a. H_0: $O = w_0 + w_1I_1 + w_1I_2$
b. H_r: $O = w_0 + (w'_1 + w'_2)I_1 + (w'_1 - w'_2)I_2$

Then we use algebra to reorganize the equation components around the weights rather than around the input variables, as in (22).

(22) a. H_0: $O = w_0 + w_1(I_1 + I_2)$

b. H_r: $O = w_0 + (w'_1 + w'_2)I_1 + (w'_1 - w'_2)I_2$
$= w_0 + (w'_1I_1 + w'_2I_1) + (w'_1I_2 - w'_2I_2)$
$= w_0 + (w'_1I_1 + w'_1I_2) + (w'_2I_1 - w'_2I_2)$
$= w_0 + w'_1(I_1 + I_2) + w'_2(I_1 - I_2)$

Finally, we relabel $I_1 + I_2$ as X_1 and $I_1 - I_2$ as X_2. Because the weights are derived separately for each equation, we can also rewrite w'_1 as w_1 and w'_2 as w_2. These substitutions produce the equations in (23).

(23) a. H_0: $O = w_0 + w_1X_1$
b. H_r: $O = w_0 + w_1X_1 + w_2X_2$

The end result is that we now know that the equation in (20b), which rejects the no-ranking hypothesis that the constraint weights are the same, is actually just an additive extension of the equal-weight equation in (20a). Hence it is legitimate to use a regression equation comparison method like the likelihood ratio test to compare the two models.

In the case of the Pazih data, the two models can be expressed in R as shown in (24). The model in (24a) assumes identical weights, and the model in (24b) does not. By default, R interprets arithmetic expressions in regression models as abbreviations for equations like that in (23b), so in order to express ordinary addition of the input values as in (22a), we must put

the addition operation inside the "as is" function *I*() in the command in (24a).

(24) a. *NoRanking* = glm(Count ~ I(*AgreeV* + *IdentV*), family = poisson)
b. *Ranking* = glm(Count ~ *AgreeV* + *IdentV*, family = poisson)

Running the likelihood ratio test (analysis of deviance) on the two models is done using the command in (25), where "Chisq" represents the chi-square test (another well-known count-based test).

(25) anova(*NoRanking*, *Ranking*, test = "Chisq")

The output of this command, shown in (26), indicates that the more complex model (not assuming equal constraint weights) has a much smaller residual deviance, meaning that it fits the data better. This improvement is statistically significant, as shown by the chi-squared $p = .0012 < .05$.

(26) Model 1: Count ~ I(AgreeV + IdentV)
Model 2: Count ~ AgreeV + IdentV

	Resid. Df	Resid. Dev	Df	Deviance	P(>|Chi|)
1	2	11.0185			
2	1	0.4677	1	10.5508	0.0012

Hence we can conclude that not only are both IdentV and AgreeV individually supported by the corpus data, but their hypothesized ranking is also supported. More precisely, this test shows that the two constraint weights are significantly different. Since we already know that the absolute value of IdentV's weight is greater than that of AgreeV's weight, this significant difference is consistent with the claim that IdentV >> AgreeV.

A bit more algebra is necessary if the hypothesized grammar involves constraints that can give evaluations of more than one star. In this case, the requirement that Weight_1 merely be larger than Weight_2 is not sufficient, since we do not want a strong violation of a lower-ranked Con_2 (many stars) to override the opinion of a higher-ranked Con_1, even if this opinion is expressed less strongly (e.g. with one star). Thus to demonstrate $\text{Con}_1 >> \text{Con}_2$, we actually have to demonstrate the relationship shown in (27a), where the smallest violation of Con_1 (i.e. one star) must outvote Con_2 even if maximally violated. In other words, we have to falsify the null hypothesis in (27b).

(27) a. $\text{Weight}_1 > \text{Weight}_2 \times \max(\text{Star}_2)$
b. $\text{Weight}_1 = \text{Weight}_2 \times \max(\text{Star}_2)$

Since it is defined by the hypothesized grammar and data set as a whole, not by individual corpus items, max($Star_2$) is a constant; call it MaxStar. A bit of algebra then results in the no-ranking hypothesis being expressed in R as in (28a), where the "variable" MaxStar × Con1 is assumed to have the same weight as Con2. The ranking model in (28b) does not need to include MaxStar since in this model the absolute values of the weights are not as important as its ability to capture the data as compared with the no-ranking model. We can then perform the comparison using (28c), as before.

(28) a. *NoRanking* = glm(Count ~ I(*MaxStar* * *Con1* + *Con2*), family = poisson)
b. *Ranking* = glm(Count ~ *Con1* + *Con2*, family = poisson)
c. anova(*NoRanking*, *Ranking*, test = "Chisq")

Testing ranking hierarchies with more than two constraints requires still more work. First, we need more algebra. Take the hypothesized constraint ranking Con_1 >> Con_2 >> Con_3. Among other things, this implies that the lower-ranked constraints Con_2 and Con_3 cannot "gang up" and collectively override Con_1. In terms of constraint weights, the equation in (29a) must hold. That is, we want to falsify the null hypothesis in (29b) (returning, for simplicity, to the case where no constraint can assign more than one star).

(29) a. $Weight_1 > Weight_2 + Weight_3$
b. $Weight_1 = Weight_2 + Weight_3$

After some algebraic manipulation (generalizing the argument that led to (23)), the no-ranking and ranking models would be expressed in R as in (30a) and (30b), respectively.

(30) a. *NoRanking* = glm(Count ~ I(*Con1* + *Con2*) + I(*Con1* + *Con3*), family = poisson)
b. *Ranking* = glm(Count ~ *Con1* + *Con2* + *Con3*, family = poisson)

A second complexity is that the ranking Con_1 >> Con_2 >> Con_3 implies more than one ranking claim, namely both Con_1 >> {Con_2, Con_3} and Con_2 >> Con_3. Thus testing a strict ranking of *n* constraints actually requires testing *n*-1 sub-claims. These sub-claims are independent, in the sense that the truth or falsity of one does not depend on the truth or falsity of any other (e.g. Con_1 >> {Con_2, Con_3} may be true while Con_2 >> Con_3 is false, or vice versa). Hence we can perform separate statistical tests on each sub-claim and get separate *p* values.

It is reasonable to ask whether the procedure proposed here for comparing constraint weights is really testing constraint ranking in standard OT. The simple answer is no, not exactly. This is because there is no single set of "true weights" in an OT analysis. To go back

to the mnemonic introduced in section 3.1.1, instead of using the decimal system we could have used binary or a system with different bases for each constraint. Indeed, the many competing OT and Harmonic Grammar algorithms all produce different constraint weights from the same data. This calls into question the assumption that a significant difference in Poisson regression weights necessarily supports ranking in the traditional OT sense.

Nevertheless, the proposed procedure has an intuitive plausibility. A Poisson regression test of individual constraints is, in a sense, a formalization of the informal analysis performed by phonologists when they check to see how well patterns generalize across a corpus. Moreover, ranking and data coverage are related, even in the traditional methodology. For example, an undominated constraint not only has a ranking status (at the top) but it also provides a true description of all of the corpus data. Similarly, a constraint with a slightly lower rank should be expected to be violated more often in the corpus than one with a much lower rank. More generally, the proposed procedure provides perhaps the statistically most straightforward way of measuring the strength of a phonological pattern in a lexicon, an aspect of phonological evidence that is likely to be crucial to theoretical analysis regardless of theoretical framework.

3.3 Limitations

The above discussion is merely meant to suggest that analyses based on word lists can be formalized, not that this traditional data source is inherently superior to all others. It is undeniable that phonological analyses based on lexical data are limited simply because lexicons are finite. This property may make it difficult to distinguish between systematic and accidental gaps (Duanmu 2008, 2009), though if one's sample is representative enough, the procedure above should suffice. After all, the difference between an accidental and a systematic gap is a matter of statistics: only the former goes significantly beyond chance probability. Of course, as in corpus linguistics generally, one can never be sure that one's sample is truly representative (here, that a speaker's lexicon is a nonbiased sample of all possible words of the speaker's language). An even more serious limitation, noted earlier, is the difficulty of distinguishing between grammatical and extra-grammatical sources of systematicity in lexicon data (Blevins 2004). Since both of these sources are systematic, statistical significance by itself is not a useful diagnostic, though if the researcher includes both as predictors in a single statistical model, it may be possible to tease apart their effects, if these effects are truly independent of each other.

In addition to these general limitations of lexicon-based analyses, the regression-based method proposed above has other limitations as well, though some of these may only be apparent, not actual, problems. One is the implication that the more items in a lexicon that obey a universal constraint, the stronger the evidence for this constraint, even if the number

of violations is held constant. This is illustrated in (31) below, where the constraint in the dataset in (31a) is predicted to be weaker than that in (31b).

(31) a.

Count	Cons
5	
1	*

Weight = -1.61
$p = .14$

b.

Count	Cons
10	
1	*

Weight = -2.30
$p = .03$

Since an item can obey a constraint vacuously (e.g. *cat* obeys the English constraint against the cluster **bn*), this has the counterintuitive implication that we can learn about a universal constraint by studying data irrelevant to it. This paradox is not created by the regression-based approach per se (see Reiss 2008 for a related critique of constraint-based theories more generally), but quantification may help make the problem more explicit.

Interestingly, this problem parallels the famous raven paradox of Hempel (1945). Since confidence in the statement "All ravens are black" is increased by finding a black raven, and since this statement is logically equivalent to "All non-black things are non-ravens", then, paradoxically, we can support "All ravens are black" with a green apple. Hempel's own solution was to say that green apples do indeed support the black raven statement, but it is only obvious if the universe is small. That is exactly the case with the above data sets. With data sets of more realistic sizes, there is no real difference, as shown in (32a) vs. (32b).

(32) a.

Count	C
5000	
1	*

Weight = -8.52
$p < .0001$

b.

Count	C
10000	
1	*

Weight = -9.21
$p < .0001$

Another potential problem with the regression-based approach follows from its attempt to derive ranking from constraint weighting alone, when actually the logic does not run in

that direction. That is, ranked constraints must differ in weight by at least a certain amount, but it does not follow that constraints whose weights differ by that amount can be considered "ranked" as understood in standard OT. In particular, standard OT would not rank two constraints if they never interact within any item (e.g. one bans segment X and the other bans a distinct segment Y), even if the lexical type frequencies show that one is violated much more often than the other.

However, there is a positive way of looking at this, namely as an empirical prediction. Standard OT says nothing about the ranking of noninteracting constraints; hence the present proposal fills a gap in the theory, rather than contradicting it. Suppose that OT is right in saying that grammar learning involves ranking constraints, and then we add the hypothesis that children take relative type frequencies into account when ranking. Testable predictions then follow from transitivity. For example, if a type frequency analysis of the noninteracting constraints A and B shows significantly stronger weighting for A, then if we discover a third constraint C that interacts with A and B, the ranking A >> B must remain respected (e.g. C >> A and B >> C would be an impossible combination).

Yet another problem with the proposed procedure is that it seems to neglect what has become a standard convention the quantitative analysis of phonological corpus data (see e.g. Frisch et al. 2004, Coetzee and Pater 2008). Rather than defining the statistical null hypothesis as equal probability for a constraint to be violated vs. obeyed, as is done here, the standard convention defines it in terms of the free combination of the relevant phonological units (e.g. segments in a phonotactic study). The researcher then divides the number of observed violations of a constraint (O) by the number expected by free combination (E). The further the O/E ratio is from 1, the stronger the constraint is (i.e. a value close to zero implies a bias against a combination, while a value far above one implies a bias in favor of a combination).

This convention differs from the proposed Poisson regression method in at least two important ways. First, the calculation of E is restricted to just the units mentioned in the constraint (e.g. testing the English **bn* constraint would involve calculating the probability only for free combinations of /b/ and /n/), so the total sample size (e.g. including words with neither /b/ nor /n/) is irrelevant. This has the advantage of avoiding raven-like paradoxes entirely, though as noted above, this may not be a practical advantage in real data sets.

Second, according to the O/E convention, the calculation of chance probability is done without explicit reference to the researcher's prior theoretical assumptions (i.e. the hypothesized constraint set is ignored). This may seem to make the method more objective, but the key word here is "explicit": the researcher is still free to build in theoretical assumptions implicitly. For example, in a phontactic study, the researcher may choose to transcribe items in terms of phonemes (e.g. /k/, /æ/, /t/) or as subsyllabic units (e.g. onset /k/ and rime /æt/), or to class segments in various ways (e.g. treating all obstruents as a group).

This freedom grows still further when we go beyond the phonotactic studies to which the O/E convention has thus far been restricted, and consider more complex phonological patterns like stress distributions or allomorphic alternations, or indeed the epenthesis and vowel harmony patterns of Pazih. The fundamental problem here is that, as argued above in 2.3, no corpus analysis is truly objective. Since the researcher must start with some sort of a priori assumptions, it is better to state them in an explicit and principled way. This is just what the proposed regression-based approach does, by requiring the researcher to spell out the theoretical background in terms of constraints.

Another advantage of an approach based on multiple regression over simply calculating O/E values is that it breaks a pattern into its components while treating them all as a part of a system. As an example of this, consider that in their discussion of the unpredictable epenthetic vowels in the twelve exceptional Pazih forms, Li and Tsuchida (2001:21) observe that "/a/ appears to be the most common." This can be seen from the counts listed in (33) for each of Pazih's four vowels appearing in the exceptional items in (8).

(33) /a/ 7 /ə/ 3 /i/ 2 /u/ 0

Is there really a preference for epenthetic /a/ in these twelve items? Following the O/E convention, O is 7 and E is (12) × (1/4) = 3 (i.e. by chance we expect three each of the four vowels, including /a/, distributed across these twelve items), so O/E = 7/3 = 2.3, quite a bit higher than the baseline of 1. This seems to suggest that the pattern may indeed be real.

However, the regression-based approach can place this simple observation into a much richer theoretical context. In OT terms, a preference for /a/ means that /a/ is less marked than the other vowels. This implies a hierarchy of vowel-specific constraints as in (34), where the proposed ranking can be empirically justified by conforming with the number of violations in (33). Interestingly, this ranking also seems to be justifiable a priori, if we hypothesize that epenthesis prefers vowels of higher sonority.

(34) *u >> *i >> *ə >> *a

We tested this OT mini-grammar using the procedure described above. We first generated all $2^4 = 16$ combinations of violations of these four constraints and counted type frequencies for each (i.e. the values in (33)). Because the sample size (just twelve exceptions) is so small, this time I decided not to convert all zero counts to one, choosing to tolerate misestimated weights and *p* values (due to the limitations of maximum likelihood estimation mentioned above in 3.2.1) in order to increase sensitivity to small effects. The Poisson regression analyses gave the results shown in (35).

(35)

Constraints	Weights	*p*
*u	-20.03	.996
*i	-1.61	.038
*ə	-1.10	.099
*a	0.34	.566
Ranking		***p***
*u >> {*i, *ə, *a}		.233
*i >> {*ə, *a}		.468
*ə >> *a		.094

While the weight and *p* value associated with *u are overestimated due to this constraint's being unviolated in the twelve exceptions, the weights and *p* values for the other three constraints are consistent with the minigrammar in (34). The weights for *i and *ə are both negative, the former significant at the traditional .05 level, while the latter comes close ($p < .1$). By contrast, the weight for *a is positive, since it is actually violated more often than obeyed in this data set, thereby falsifying any claim to be active here, even aside from its high *p* value ($p > .5$). The ranking tests are similarly clear: while the relative rankings of *u, *i, and *ə are indistinguishable, the ranking of *a with *ə, its closest neighbor in the hierarchy, is significant at the .1 level.

Nevertheless, the preceding analysis highlights one final limitation of the Poisson regression approach that I will discuss here, namely the need for samples of sufficient size. Of course, this holds for any sample-based research, and having to make this limitation explicit with the proposed approach may actually be an advantage. When phonologists ignore the sample size problem, they can be led into dead ends. For example, Chomsky and Halle (1968) and Borowsky (1990) propose a k-copying rule in English lexical phonology, based on the appearance of /k/ in the coda of the prefixes in several of the words in (36). However, note that this sample is complete: there are exactly seven such k-final-prefix words in English, along with two exceptions. Thus even if we accounted for this pattern with a single "super-constraint" obeyed by words like *access* but violated by words like *succinct*, the proposed method would find this constraint (with weight -1.25) to be nonsignificant ($p = .12$). The same null result is found if other standard statistical tests are used (e.g. the chi-square test or the sign test). A skeptic may thus be justified in setting aside this data set as irrelevant to synchronic phonological theory.

(36) /ad/ [ks] access, accident, accept, accelerate, accede, accent
/sub/ [ks] succeed vs. [s] susceptible, succinct

In order to get some sense of the effect of sample size on the phonologist's ability to detect patterns, I ran the Poisson regression procedure on a series of artificial data sets varying along four parameters. Figure 1 graphs the results, showing the minimum sample size needed to detect a pattern at the .05 significance level as a function of each of these parameters. While the precise values along the x and y axes depend partly on arbitrary assumptions in the operationalization of these parameters, it is hoped that the overall trends will generalize to realistic data sets.

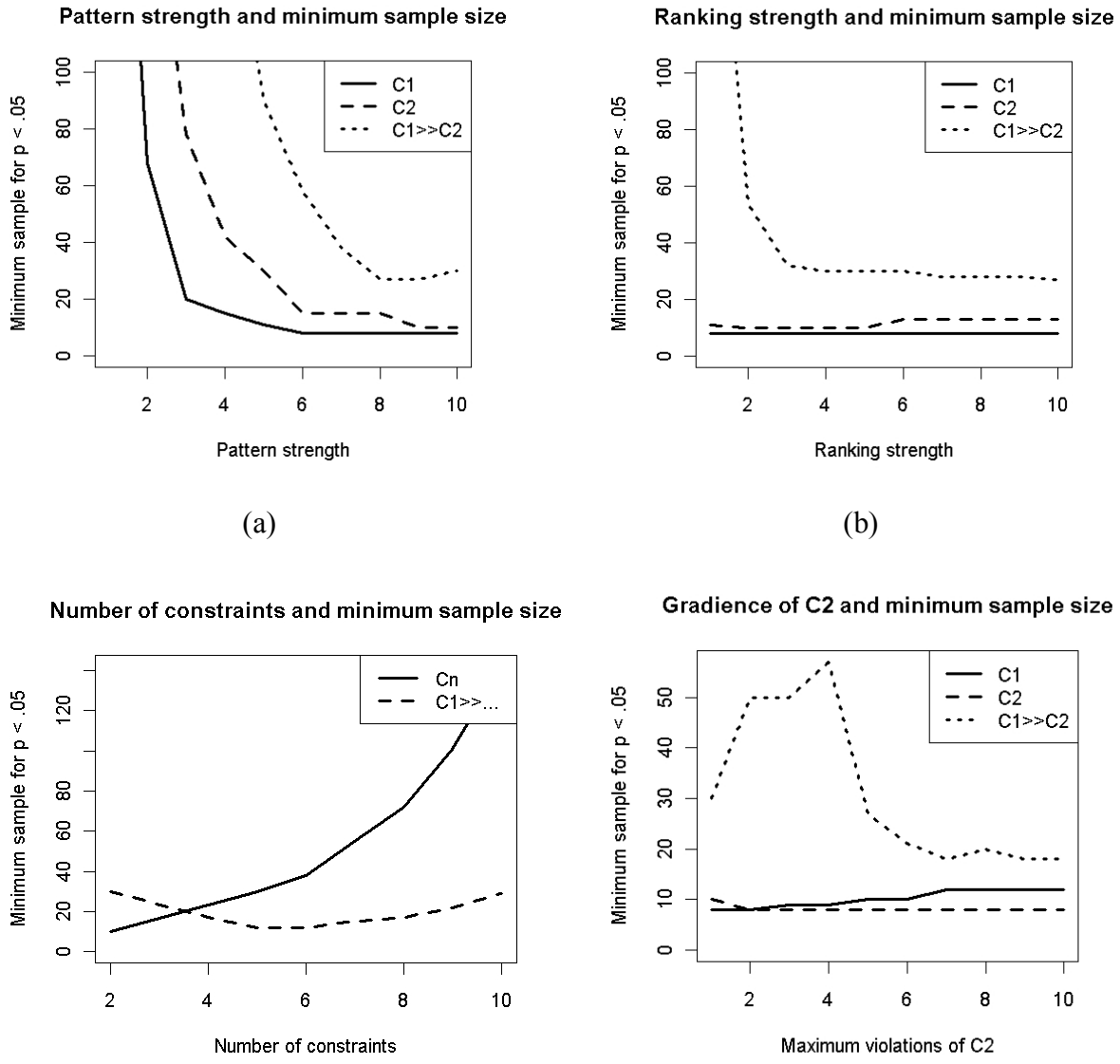

Figure 1. The minimum sample sizes needed to detect a pattern at the .05 significance level as a function of pattern strength (a), ranking strength (b), number of constraints (c), and gradience of the lower-ranked constraint (d). See text for details.

The first parameter is pattern strength, defined as the degree to which the four type frequencies in a two-binary-constraint analysis differ from each other. In Figure 1a, the values along the x-axis were used to generate data sets with increasingly large differences across type frequencies consistent with the constraint ranking C1 >> C2. The fact that the line for C1 is always lower than that for C2 implies that higher-ranked constraints are easier to detect than lower-ranked constraints (i.e. a smaller sample suffices to detect them). However, evidence for the ranking itself requires still larger samples, as shown by the relatively high position of the line for C1>>C2. Stronger patterns are of course easier to detect than weaker ones.

The second parameter is ranking strength, defined as the relative difference in weight between the ranked constraints C1 and C2, graphed in Figure 1b. Again we see that detecting ranking takes more data than detecting individual constraints, a finding that, perhaps surprisingly, holds even when the ranking difference gets quite large.

The third parameter is the number of constraints in the analysis. As shown in Figure 1c, while the detectability (i.e. required sample size) of the ranking of the topmost constraint over all others remains relatively constant regardless of the number of constraints, the detectability of the lowest-ranked constraint gets progressively more difficult, until quite large samples (by the standards of theoretical phonology) may be required.

The final parameter considered is the gradience of the lower-ranked constraint, defined as the maximum number of violations it accrues in the data set. The graph in Figure 1d shows that while the detectability of each constraint remains constant, the detectability of the ranking itself varies in a surprising way. As the gradience of the lower-ranked constraint C2 increases, this initially hinders detection of ranking (i.e. larger sample sizes are required), apparently because the weight of C1 must be large enough to outweigh the larger violations of C2 (see (27) above). After C2 is allowed to be violated to a certain degree, however, detecting ranking becomes easy again, apparently because the larger the degree of violability, the greater the number of cells in the counts table (i.e. we count the number of items obeying C1 but violating C2 once, then the number obeying C1 but violating C2 twice, and so on). Thus if the ranking C1 >> C2 is true, as the simulations assume, then with great enough gradience in C2, eventually the entire data set ends up in top half of the counts table (where C1 is obeyed and C2 is violated to varying degrees), making the ranking easy to detect.

In short, though there are some limitations to the proposed regression-based quantification of traditional lexicon-based phonological analysis, most if not all of these limitations are not as serious as they may seem at first, and even those that remain challenging tend to be inherited from the traditional approach. The proposed procedure at least has the advantage of bringing these latent limitations out into the open.

4. Conclusions and future prospects

If I have convinced the reader of nothing else, I hope I have shown that the anti-corpus rhetoric of generative linguistics, going back to its very foundations by Chomsky, has always had a powerful counterargument standing in plain sight. This counterargument is generative phonology, which, at least in its study of lexical patterns, actually depends on corpus (database) analysis almost exclusively. I then explained why generative phonologists are justified in treating corpus frequencies as information about grammar, as they implicitly do in their actual practice (e.g. when distinguishing regular patterns from exceptions), and gave this insight quantitative teeth by showing how prespecified OT grammars, composed of constraints and their ranking, can be tested statistically on lexical data using well-established statistical methods, namely loglinear (poisson) regression and likelihood ratio tests.

Much more could be said about all of this, but I will restrict myself to just two final points. First, the quantitative procedures assume that lexicons result solely from OT constraints and purely random noise. Since type frequencies are also affected by systematic extra-grammatical forces, the proposed methods may find statistically significant support for a hypothesized constraint when the data pattern actually results from borrowing or analogy. In principle, the solution to this challenge is simple: represent these extra-grammatical forces as additional input variables in the statistical model. If the grammatical and extra-grammatical factors are not fully confounded, the mathematics of regression should be able to distinguish their separate effects.

Second, though I have tried to make the background and instructions for the statistical procedures as simple as possible, it is likely that they are still too technically intimidating for the average working phonologist. It would therefore be beneficial if they could be carried out by a software tool running behind a fully intuitive, non-threatening interface. I am currently working on a program with this goal, called MiniCorp (complementing the judgment-analysis program MiniJudge; Myers 2007a, 2009a,b). MiniCorp is described more fully in Myers (2008, 2009a), and both MiniCorp and MiniJudge are available online at www.ccunix.ccu.edu.tw/~lngproc/MiniGram.htm.

References

Agresti, Alan. 2002. *Categorical Data Analysis*, 2nd ed. Hoboken, NJ: Wiley-Interscience.

Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge, UK: Cambridge University Press.

Bailey, Todd M., and Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory & Language* 44:569-591.

Blevins, Juliette. 2004. *Evolutionary Phonology: The Emergence of Sound Patterns*.

Cambridge, UK: Cambridge University Press.

Boersma, Paul, and Silke Hamann. 2008. The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology* 25:217-270.

Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32:45-86.

Boersma, Paul, and David Weenink. 2011. Praat: doing phonetics by computer (Version 5.2.28) [Computer program]. http://www.praat.org/

Borowsky, Toni. 1990. *Topics in the Lexical Phonology of English*. New York: Garland Publishing.

Chomsky, Noam. 1957. *Syntactic Structures*. The Hague: Mouton.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam. 2002. *On Nature and Language*. Cambridge, UK: Cambridge University Press.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Patterns of English.* Cambridge, MA: MIT Press.

Coetzee, Andries W., and Joe Pater. 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language and Linguistic Theory* 26.2:289-337.

Duanmu, San. 2004. A corpus study of Chinese regulated verse: phrasal stress and the analysis of variability. *Phonology* 21:43-89.

Duanmu, San. 2008. The spotty-data problem and boundaries of grammar. In *Interfaces in Chinese Phonology: Festschrift in Honor of Matthew Y. Chen on his 70th Birthday*, ed. Yuchau E. Hsiao, Hui-Chuan Hsu, Lian-Hee Wee, and Dah-an Ho, 261-278. Taipei: Institute of Linguistics, Academia Sinica.

Duanmu, San. 2009. *Syllable Structure: The Limits of Variation*. Oxford: Oxford University Press.

Frisch, Stefan A., Janet B. Pierrehumbert, and Michael B. Broe. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22:179-228.

Frisch, Stefan A., and Bushra Adnan Zawaydeh. 2001. The psychological reality of OCP-Place in Arabic. *Language* 77.1:91-106.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, ed. by Jennifer Spenader, Anders Eriksson & Östen Dahl, 111-120. Stockholm, Sweden: Stockholm University.

Golston, Chris. 1996. Direct Optimality Theory: Representation as pure markedness. *Language* 72.4:713-748.

Hammond, Michael (this volume). Empirical methods in phonological research.

Hansson, Gunnar Ólafur. 2008. Diachronic explanations of sound patterns. *Language and*

Linguistics Compass 2/5:859-893.

Hayes, Bruce and Donca Steriade. 2004. Introduction: The phonetic bases of phonological markedness. In *Phonetically Based Phonology*, ed. by Bruce Hayes, Robert Kirchner, and Donca Steriade, 1-33. Cambridge, UK: Cambridge University Press.

Hayes, Bruce, and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39:379-440.

Hempel, Carl G. 1945. Studies in the logic of confirmation (I). *Mind* 54.213:1-26.

Inkelas, Sharon, C. Orhan Orgun, and Cheryl Zoll. 1997. The implications of lexical exceptions for the nature of grammar. In *Derivations and Constraints in Phonology*, ed. by Iggy Roca, 393-418. Oxford: Clarendon Press.

Johnson, Keith. 2008. *Quantitative Methods in Linguistics*. Oxford: Blackwell Publishing.

Keller, Frank. 2000. *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. PhD dissertation, University of Edinburgh.

Kenyon, John Samuel, and Thomas A. Knott. 1944. *A Pronouncing Dictionary of American English*. Springfield, MA: Merriam.

Kiparsky, Paul. 2006. Amphichronic linguistics vs. Evolutionary Phonology. *Theoretical Linguistics* 32:217-236.

Kirchner, Robert. 1997. Contrastiveness and faithfulness. *Phonology* 14:83-111.

Li, Paul Jen-kuei, and Shigeru Tsuchida. 2001. *Pazih Dictionary*. Taipei: Institute of Linguistics.

Martin, Andrew Thomas. 2007. *The Evolving Lexicon*. PhD dissertation, UCLA.

Mendoza-Denton, Norma, Jennifer Hay, and Stefanie Jannedy. 2003. Probabilistic sociolinguistics: Beyond variable rules. In *Probabilistic Linguistics*, ed. by Rens Bod, Jennifer Hay and Stefanie Jannedy, 97-138. Cambridge, MA: MIT Press.

Moreton, Elliott. 2008. Analytic bias and phonological typology. *Phonology* 25.1:83-127.

Myers, James. 2007a. MiniJudge: Software for small-scale experimental syntax. *International Journal of Computational Linguistics and Chinese Language Processing* 12.2:175-194.

Myers, James. 2007b. Linking data to grammar in phonology: Two case studies. *Concentric: Studies in Linguistics* 33.2:1-22.

Myers, James. 2008. Bridging the gap: MiniCorp analyses of Mandarin phonotactics. In *Proceedings of the Thirty-seventh Western Conference on Linguistics*, ed. by R. Colavin, K. Cooke, K. Davidson, S. Fukuda, and A. Del Guidice, 137-147. California, US: University of California, San Diego.

Myers, James. 2009a. Automated collection and analysis of phonological data. In *The Fruits of Empirical Linguistics: Volume 1: Process*, ed. by S. Featherston and S. Winkler, 151-176. Berlin: Mouton de Gruyter.

Myers, James. 2009b. The design and analysis of small-scale syntactic judgment experiments.

Lingua 119:425-444.

Ohala, John J. 1986. Consumer's guide to evidence in phonology. *Phonology Yearbook* 3: 3-26.

Pater, Joe. 2009. Morpheme-specific phonology: Constraint indexation and inconsistency resolution. In *Phonological Argumentation: Essays on Evidence and Motivation*, ed. by Steve Parker, 123-154. London: Equinox.

Pater, Joe, Rajesh Bhatt and Christopher Potts. 2007b. Linguistic Optimization. Ms, University of Massachusetts, Amherst.

Pater, Joe, Christopher Potts, and Rajesh Bhatt. 2007a. Harmonic grammar with linear programming. Ms, University of Massachusetts, Amherst. ROA 872-1006. Software package: http://web.linguist.umass.edu/~halp/.

Port, Robert and Adam Leary. 2005. Against formal phonology. *Language* 85:927-964.

Prince, Alan, and Paul Smolensky. 1993. *Optimality Theory: Constraint interaction in generative grammar*. Published 2004, Oxford: Blackwell Publishing.

Prince, Alan, and Paul Smolensky. 1997. Optimality: from neural networks to universal grammar. *Science* 275:1604-1610.

Prince, Alan. 2007. Let the decimal system do it for you: A very simple utility function for OT. Rutgers University ms. ROA 943. http://roa.rutgers.edu/view.php3?id=1356

R Development Core Team. 2011. R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Reiss, Charles. 2008. Constraining the learning path without constraints, or the OCP and NoBanana. In *Rules, Constraints and Phonological Phenomena*, ed. by A. Nevins and B. Vaux, 252-302. Oxford: Oxford University Press.

Samek-Lodovici, Vieri, and Alan Prince. 2005. Fundamental properties of harmonic bounding. University College, London, and Rutgers University, New Brunswick ms. Available at Rutgers Optimality Archive (roa.rutgers.edu) 785.

Selkirk, Elisabeth O. 1972. *The Phrase Phonology of English and French*. PhD dissertation, MIT.

Stefanowitsch, Anatol. 2005. New York, Dayton (Ohio), and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory* 1.2:295-301.

Tesar, Bruce, and Smolensky, Paul. 1998. The learnability of Optimality Theory. *Linguistic Inquiry* 29:229-268.

Uffmann, Christian. 2006. Epenthetic vowel quality in loanwords: Empirical and formal issues. *Lingua* 116:1079-1111.

Wilson, Colin, and Marieke Obdeyn. 2009. Simplifying subsidiary theory: Statistical evidence from Arabic, Muna, Shona, and Wargamay. Ms., Johns Hopkins University.