Automated collection and analysis of phonological data

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

Presented at the International Conference on Linguistic Evidence 2008,
Tübingen, Germany

Revised May 26, 2008

1. Introduction

The past decade has seen renewed interest in the empirical basis of theoretical syntax, sparked in part by the publication of Schütze (1996) and Cowart (1997). However, a similar empirical revolution began in phonology about a decade earlier, with the publication of Ohala & Jaeger (1986) and Kingston & Beckman (1990). Today a significant proportion of the theoretical phonology papers published in the major linguistics journals employ experimental methods (e.g. S. Myers and Hansen 2007), quantitative corpus analysis (e.g. Uffmann 2006), or both (e.g. Zuraw 2007).

This dramatic change in the course of the phonological mainstream raises the key question of how the new phonological methods relate to the old traditional ones. Whatever the answer may be, it is not being made clearly to many phonology students, who in introductory classes continue to practice testing hypotheses against small data sets from dictionaries, and yet when they begin to read the contemporary literature they are confronted with the very different empirical vocabulary of psycholinguistic experiments, electronic corpora, and quantitative analysis.

Fortunately for those living through this age of methodological transition, differences between the old and new ways are more a matter of degree than of kind. In particular, native speaker intuitions of acceptability represent experimentally elicited psycholinguistic data (as noted even in the otherwise highly critical review by Labov 1975), and the analyses of small data sets that continue to dominate phonology textbooks represent a form of corpus analysis. Moreover, the arguments that phonologists make on the basis of such data are implicitly quantitative, as when rarity is used to identify exceptionality.

Clearly, however, the quantitative arguments used in the new methods are of a much higher degree of sophistication than the informal rules of thumb that have traditionally been used. It is precisely here where many theoretical linguists and their students face their greatest stumbling blocks in adapting to the new methodological world. Linguists are generally the sort of people who love words, not numbers, and it is my impression that those few who love both are naive to think that the rest will follow their lead by example alone, without bridges clearly linking the old and new worlds.

In this chapter I describe one way in which such bridges might be built. On the theoretical side, the key is to recognize that traditional phonologists are already doing quantitative corpus analyses and psycholinguistic experiments, even though they don't think of them in these terms. On the practical side, the key is automation: phonologists would benefit from special-purpose software that allows them to maintain many of their familiar habits, while, mostly hidden from view, powerful algorithms put their theoretical hypotheses through rigorous quantitative tests.

While the ideal versions of such software still lie a bit in the future, working prototypes already exist: MiniCorp (Myers forthcoming) for the analysis of phonological corpora, and

MiniJudge (Myers 2007a) for the design and analysis of linguistic judgment experiments. Both tools are free and open-source; currently the most frequently updated versions are written in JavaScript running in the user's Web browser, and have been tested most extensively in Firefox for Windows. The statistics are handled by R, the free, open-source statistics package (R Development Core Team 2008) that is fast becoming the de facto standard in quantitative linguistics (Baayen 2008, Johnson 2008). MiniCorp and MiniJudge may be found at http://www.ccunix.ccu.edu.tw/~lngproc/MiniGram.htm.

I start the discussion in section 2 with a brief overview of traditional methods in phonology, and why they remain worthy of respect. Section 3 then describes the principles behind MiniCorp, the corpus analysis tool, and section 4 does the same for MiniJudge, the judgment experiment tool. These tools are then demonstrated in section 5 in a study of Mandarin phonotactics. Section 6 concludes and points towards the future.

2. Traditional phonological methods

The most common method in phonology has traditionally been the study of dictionaries (discussed in 2.1), though phonologists have sometimes also made use of acceptability judgments (2.2). In this section I show why phonologists have tended to prefer the former, and how both methods, even in their traditional forms, are akin to the more sophisticated techniques becoming more common in the phonological literature.

2.1 Corpus analysis

Despite early generative arguments that certain aspects of phonological competence can only be discovered via acceptability judgments, such as the preference of English speakers for the unattested *blick* over the equally unattested *bnick* (Chomsky and Halle 1965), phonologists have generally remained unaffected by the revolution in data sources sparked by Chomsky (1957). Even Chomsky and Halle (1968) rely heavily on dictionary data (specifically, Kenyon and Knott 1944). As I show in this section, however, the traditional favoring of dictionary data in phonology turns out to be surprisingly well justified.

Of course, as has often been pointed out, systematicity in a lexicon does not suffice to show that the patterns are part of synchronic phonological competence, since they could be relics of sound change operating without the help of grammatical competence (Ohala 1986, Blevins 2004). This alternative view has recently been challenged by experiments using nonlexical items (e.g. Zuraw 2007, Moreton forthcoming), but it is less well recognized that the relevance of corpus data to competence theories can actually be defended on the basis of corpus data alone. For example, Kiparsky (2006) argued that sound change unguided by universal grammatical principles predicts lexical patterns that are apparently unattested (contra Blevins 2006). More generally, ascribing synchrony to diachrony undermines diachronic reasoning itself, since many reconstructions depend on assumptions about what makes a plausible synchronic grammar.

Another important reason for phonologists to focus on corpus data is that language learners do so as well. Understanding language acquisition, often cited as a central goal of linguistic theory (Chomsky 1965), means finding the one "true" corpus analysis used by actual children. This insight has recently fueled research on phonological acquisition models (e.g. Tesar and Smolensky 1998, Boersma and Hayes, 2001, Hayes and Wilson forthcoming), which take phonological corpora as input.

Finally, the traditional study of dictionaries is also worthy of respect because it relies on the same sort of quantitative logic used in computational corpus linguistics. In particular, phonologists often test claims using corpus type frequencies and probability theory, even though they rarely recognize that this is what they are doing. Thus a generalization with few or no exceptions is acknowledged to be more convincing than one with many. Likewise, what the traditional notion of the systematic gap represents is a type frequency much lower (perhaps zero) than expected by chance, given the combinatorial possibilities of the phonological units in the language. For example, *bnick* represents a systematic gap in English because the free combination of /b/ and /n/ (the null hypothesis) predicts many more /bn/ words than are actually found (zero). Phonologists intuitively understand that just as a grammatical claim can be supported by a robust generalization, the absence of evidence can be interpreted as evidence of absence if one has a model of chance probability.

2.2 Native speaker judgments

Phonologists also occasionally use native speaker judgments. Experimental data, including elicited judgments, have many familiar advantages over corpus data, in addition to those noted above: They make it much easier to test synchronic productivity (e.g. Frisch and Zawaydeh 2001) and phrasal phonology (e.g. Keller and Alexopoulou 2001), and new experiments can be devised whenever new questions arise (Ohala 1986). Here I address a less often discussed advantage of collecting phonological judgments in particular, as well as some important limitations.

The view of linguistic theory as the search for the child’s preferred corpus analysis algorithm, alluded to in the previous section, predicts two types of mismatches between corpus patterns and judgments. On the one hand, speakers may be able to distinguish between forms that are equally unattested in the corpus, which is what the *blick* vs. *bnick* contrast is intended to show. This type of mismatch is related to the argument from the poverty of the stimulus (Chomsky 1980), and in this guise it has recently received renewed attention in the experimental phonology literature (e.g. Zuraw 2007). On the other hand, speakers may also ignore information in the input if their grammar learning algorithm is not designed to pick up on it. A version of this is seen in the child language literature, where children may reject negative evidence even when it is explicitly offered (e.g. Morgan, Bonamo, and Travis 1995). The result for the adult are judgment patterns that neglect certain statistically robust corpus patterns.

Although such theoretically important mismatches can only be detected with the help of judgments, they depend just as much on corpus data. Moreover, if we seriously view a mature grammar as the result of the single “true” corpus-analysis algorithm, the baseline condition in a judgment experiment should not be chance alone, but rather the predictions made by alternative corpus-analysis algorithms. This is why formal phonological judgment experiments typically control for phonotactic probability, relating to the probability of the internal components of target items relative to real words, and neighborhood density, relating to the overall similarity of target items to real words (e.g. Frisch and Zawaydeh 2001). Though both are known to affect phonological processing (e.g. Vitevitch and Luce 1999), they are assumed to reflect extra-grammatical factors. Thus the interpretation of a phonological judgment experiment typically depends on a corpus analysis quantifying these potential confounds.

The importance of corpus analysis to phonology means that the design of judgment experiments poses greater challenges to the phonologist than to the syntactician. As shown later in this chapter, the vagaries of actual lexicons and the rigidity of phonotactic constraints typically make it difficult to follow strict factorial designs when creating wordlike materials for a judgment experiment. Despite the many benefits of experimentation for testing phonological hypotheses, then, it is entirely understandable that phonologists continue to focus less on judgment data than syntacticians.

3. Automating phonological corpus analysis

The main purpose of the MiniCorp software tool is to extend and automate the quantitative logic underlying the traditional analysis of dictionary data (for related discussion, see Myers 2007b, 2008, forthcoming). The “mini” in the name indicates that it is designed for limited analyses of small corpora. In particular, MiniCorp tests whether a hypothesized phonological grammar is statistically supported by patterns in an electronic dictionary.

MiniCorp is not an automated grammar learner. Rather than exploring the corpus for patterns, it starts with a user-entered grammar and computes the probability that observed differences in the sizes of corpus categories (i.e. type frequencies), as defined by the proposed grammar, could have arisen by chance. To constrain the space of possible grammars, MiniCorp (at least in its current version) adopts the theoretical framework of Optimality Theory (OT; Prince and Smolensky 2004). As explained in section 3.1, not only is OT the contemporary lingua franca of theoretical phonology, but it has mathematical properties that make it convenient for statistical hypothesis testing; the relevant statistical techniques are explained in 3.2. MiniCorp also uses a standard search algorithm to annotate corpus items for analysis, as explained in 3.3. Finally, as discussed in 3.4, MiniCorp calculates certain (presumably extra-grammatical) lexical statistics so that they can be factored out in judgment experiments.

3.1 Modeling grammar

Despite lingering challenges (opacity being the most notorious), OT has proven itself a highly productive and flexible tool for describing patterns in dictionary data and beyond. Moreover, it has two very convenient mathematical properties: OT constraints describe surface forms (though ironically, this is why OT has trouble with opacity), and they are ranked so that lower-ranked constraints only have a say if higher-ranked constraints are noncommittal. These properties link OT to the older connectionist-inspired theory of Harmonic Grammar (HG; Legendre, Sorace, and Smolensky 2006), have led to recent advances in OT acquisition modeling (Hayes and Wilson forthcoming, Coetzee and Pater forthcoming), and allow MiniCorp to test the statistical significance of grammatical hypotheses.

OT constraint ranking is a special case of the constraint weighting of HG, where the overall evaluation score for a candidate output form is given by summing the products of each weight with the severity of its violation; the higher this score, the worse the candidate, with the lowest-scoring (most “harmonic”) candidate chosen as final output. For example, the grammar in the first row in the tableau in (1a), with the constraints Cons$_1$, Cons$_2$, Cons$_3$ and weights w_1, w_2, w_3, will choose Out$_1$ as output. This is because (1a) is equivalent to the equations in (1b).

(1) a.

	Cons$_1$ $w_1 = 3$	Cons$_2$ $w_2 = 1$	Cons$_3$ $w_2 = 1$
Out$_1$		*	*
Out$_2$	*		

b. Evaluation(Out$_i$) = $\Sigma\, w_j \times$ Violation(Out$_i$, Cons$_j$), i.e.:
Evaluation(Out$_1$) = $(3)(0) + (1)(1) + (1)(1) = 2$ (more harmonic)
Evaluation(Out$_2$) = $(3)(1) + (1)(0) + (1)(0) = 3$

Note that the winning candidate in (1a) would also win in an OT grammar where CONS$_1$ » {CONS$_2$, CONS$_3$}. This follows from the fact that $w_1 > w_2 + w_3$, so the two lower-ranked constraints cannot override the higher-ranked one (Prince 2007). A different choice of constraint weights may not have this property, so OT is a special case of HG.

Automated HG learners set the constraint weights by exposure to a corpus (Hayes and Wilson forthcoming, Coetzee and Pater forthcoming). Although MiniCorp is not a grammar learner, it also sets weights on the basis of corpus data, but here the weights are taken as measures of a pre-given grammar's statistical reliability rather than as components of the grammar itself. That is, the constraint weights set by MiniCorp reflect the type frequencies associated with the generalizations, exceptions, systematic gaps, and accidental gaps predicted by the user-defined grammar. Only if a weight is sufficiently different from zero is the associated constraint considered to be statistically reliable, and only if the weight of one constraint is significantly higher than that of another is a hypothesized constraint ranking considered to be supported by the data.

MiniCorp thus provides a quantitative formalization of core aspects of traditional phonological methodology, expressed in terms of the currently most familiar phonological framework. As explained in the next section, setting and evaluating constraint weights consistent with this logic can be accomplished using well-established statistical methods.

3.2 Loglinear modeling

Type frequencies are discrete, countable values, and thus represent categorical data; such data are often handled statistically with loglinear modeling (Agresti 2002). Loglinear modeling is a generalization of linear regression, so-called because it attempts to relate the independent (predictor) and dependent (predicted) variables in terms of a straight line. Regression can be applied to categorical data with the proper transformation of the dependent variable and the proper random distribution. Loglinear models use the logarithmic transformation, and when the dependent variable represents type frequencies, the appropriate distribution is the Poisson distribution, which unlike the normal distribution is discrete and tends to be positively skewed (because counts cannot go below zero).

The relevance here of these well-established techniques is that the right side of a (log)linear regression equation is highly reminiscent of the right side of the HG equations in (1b). Namely, they contain the sum of the products of regression coefficients (here, constraint weights) and independent variables (here, constraint violations). The weights are set so that the right side of the equation fits the observed (transformed) type frequencies as closely as possible, and the contribution of each constraint is evaluated in the context of all of the others.

Loglinear modeling (though not Poisson regression) is also used to set constraint weights in the HG learner proposed by Hayes and Wilson (forthcoming), but since the goal of MiniCorp is hypothesis testing rather than grammar learning, there are two important differences. First, as with regression models generally, Poisson regression allows MiniCorp to test the statistical significance of each constraint (relevant to hypothesis testing but not necessarily to learning). Second, Poisson regression also makes it possible to test the statistical significance of a proposed OT constraint ranking (irrelevant to HG and thus to HG-based learners).

MiniCorp tests a ranking hypothesis by comparing a regression equation in which the constraints are free to take any weight, as in (2a), with an equation in which the weights must be identical, as in (2b) (the notation Y ~ X means "Y varies as a function of X"). Only if the model in (2a) does a significantly better job at fitting the data (as evaluated by a likelihood ratio test, another standard statistical technique) can we reject the null hypothesis that $w_1 = w_2$ and conclude that the constraints may indeed be ranked. (A bit of algebra shows that (2a) is the same as (2b) with the addition of a term, the significance of which is what the likelihood ratio test is actually testing, and with further algebra we can generalize the logic to grammars with multiple and multiply violated constraints; see Myers 2008.)

(2) a. Counts ~ $w_1\text{Cons}_1 + w_2\text{Cons}_2$
b. Counts ~ $w_1\text{Cons}_1 + w_2\text{Cons}_2$, $w_1 = w_2$

In its current version, MiniCorp (like MiniJudge, described below) runs the analyses in R, the free statistical programming language (R Development Core Team 2008).

3.3 Corpus annotation

As in most corpus analyses, the analyses performed by MiniCorp depend on annotations marking linguistically relevant abstract features, in this case, which OT constraints are violated by which lexical items. Corpus annotation can be the most labor-intensive aspect of corpus preparation, but fortunately mathematical properties again make it possible for MiniCorp to automate the task through a well-established algorithm.

It is an empirical fact about human language that phonological patterns can be described with regular expressions (Bird and Ellison 1994). Regular expressions also happen to be commonly used for pattern matching in text searches. Regular expression notation systems include symbols representing wildcards (which match to any string), repetition, disjunction, the start and end of strings, and so on. Since violations of OT structure constraints represent classes of substrings, Karttunen (1998) noted that they can also be encoded as regular expressions.

MiniCorp exploits these observations in a tool (using the regular expression engine built into JavaScript) that automatically searches for, and then annotates, corpus items for OT constraint violations. This method works best for output structure constraints. It cannot reliably annotate faithfulness constraints, which reflect relationships with representations not available in the corpus itself, and its success depends partially on manual annotations like syllable boundaries (relevant to constraints like ONSET and NOCODA). Nevertheless, as will be demonstrated in 5.2, the regular expression tool does greatly simplify the annotation of constraint violations.

3.4 Quantifying lexical confounds

Though the central purpose of MiniCorp is to test a pre-specified grammar, it also helps compute extra-grammatical lexical statistics to be factored out in phonological judgment experiments. Here I discuss only one of these lexical statistics, neighborhood density.

The reason for focusing on this particular variable is to limit the risk of throwing out the baby with the bathwater. Just because certain patterns can be detected in a corpus by a presumably extra-grammatical algorithm does not mean that they are not also detected by the child’s grammar-learning algorithm. Factoring out an extra-grammatical lexical variable that mimics the results of grammar-learning too closely may cause us to miss genuine evidence for grammar in judgments.

With neighborhood density the risk of this happening is low, since this lexical statistic seems to reflect exemplar-driven analogy, not grammar. First, neighborhood density evaluates forms holistically; phonotactic probability, by contrast, is similar to OT constraints in analyzing forms into substrings. Second, psycholinguistic experimentation suggests that neighborhood density only affects phonological processing after the lexicon has been contacted, whereas phonotactic probability plays a prelexical role (Vitevitch and Luce 1999), and thus is, like grammar, partially independent of the lexicon. Finally, there has recently been some interest in incorporating probabilistic phonotactics inside grammar itself (e.g. Coetzee and Pater forthcoming).

The current version of MiniCorp applies the simplest possible definition of neighborhood density, namely the number of lexical items differing from a target item by deletion, insertion, or replacement of one phonological unit (Luce 1986). This definition is not only simple, but it is akin to the MAX, DEP, and IDENT correspondence constraints familiar to OT phonologists.

4. Automating phonological judgment experiments

As with MiniCorp, the purpose of MiniJudge is to build on traditional linguistic methodology, in this case the collection of native speaker judgments of acceptability (Myers 2007a). Its scope is also minimalist, helping the linguist to design, run, and statistically analyze experiments with relatively few speakers judging relatively few items on a binary good/bad scale. The tool was originally developed for syntax, where judgments have historically played a more important role, but in this section I highlight the special characteristics of MiniJudge when used for phonological judgments, in terms of material design (4.1) and the collection and statistical analysis of data (4.2).

4.1 Material design

MiniJudge guides the researcher to choose the experimental factors and materials instantiating them, and includes tools to deal with the special challenges posed by phonological judgment experiments.

Because grammatical hypotheses often involve the relationship between two elements, the typical MiniJudge experiment involves two factors, each representing one of the elements; the theoretical hypothesis then relates to the interaction between them. For example, an experiment on the constraint against **bnick* in English could involve two factors, one representing onset /b/ (in contrast to /s/, say) and the other onset /n/ (in contrast to /l/, say). The hypothesized constraint would predict lower acceptability for /bn/ relative to /sn/, /bl/, /sl/; the results could not then be explained away as constraints against /b/ and/or /n/ themselves.

Theoretical linguists are already familiar with the basic logic of factorial experimental design, as instantiated by the minimal pairs and minimal sets of examples cited in research papers. Starting a MiniJudge experiment thus involves entering such a basic set of matched materials. To help generate the additional sets needed for generalizability, MiniJudge detects the structural contrasts implicit in the initial material set, so that the user only has to enter in new matching components rather than create new sets from scratch (risking typos).

For example, the factors in a **bnick* experiment, schematized in (3a), could be instantiated with the material set in (3b), where the items are identical except for the properties defined by the factors. The repeated elements are those listed in (3c), each of which can be replaced by a functionally equivalent element as in (3d) (assuming that **bnick* is a special case of *[-cont][+nas]). By substituting these new elements for the old ones, MiniJudge derives the new set in (3e).

(3)	a.	[+b+n]	[+b-n]	[-b+n]	[-b-n]
	b.	bnick	blick	snick	slick
	c.	b s	n l	ick	
	d.	k s	m l	oss	
	e.	kmoss	smoss	kloss	sloss

It should be obvious from this example that the design of wordlike materials for judgment experiments face serious challenges (such problems do not arise for syntax or phrasal phonology). First, there is nothing to prevent an item generated by the above algorithm from being an actual word (e.g. *slick*). Testing phonological judgments on real words is notoriously problematic because real words have many memorized properties that are difficult to control for, including lexical frequency and semantics (Bailey and Hahn 2001). If real words and nonwords are mixed together, lexical status becomes a confounding factor as well (though perhaps it may be explicitly recognized and factored out, as in Myers and Tsay 2005).

A related challenge, ironically, is phonotactics itself, which makes it difficult to avoid real words and maintain the experimental design at the same time. For example, if *bnick* and *blick* are included in a two-factor design, we really have no choice but to make the other two items *snick* and *slick*, despite the fact that both are real words. This is because the only consonant that appears before a nasal in English is /s/, and the only sonorant available as a control (to be minimally different from the nasal), and which appears after /s/ as required by the factorial design, is /l/ (aside from non-nativized borrowings like *Sri Lanka*).

MiniJudge's sister program MiniCorp provides some assistance with such challenges, since as a collateral benefit of calculating the neighborhood density for each experimental item, it also detects whether this item is listed in the lexicon. The researcher is then alerted to any sets containing real words (including less familiar ones like *snick*), and may choose to replace them or counterbalance the distribution of real words across sets to minimize confounding with the experimental factors.

Neighborhood density itself is easier to deal with. After computing the values with MiniCorp, the MiniJudge user may choose either to match materials on this lexical statistic, or to keep the original materials and allow MiniJudge to factor out neighborhood density as a covariate in the statistics (as explained in 4.3).

4.2 Data collection and analysis

After MiniJudge has helped to create a judgment experiment, it then helps to run and analyze it. Here I briefly review these steps, highlighting the special characteristics of phonological judgment experiments where relevant.

MiniJudge generates surveys presenting items in different random orders for each experimental participant to reduce the confounds of fatigue, practice, and cross-item priming. While this is standard psycholinguistic practice, the current version of MiniJudge has three built-in limitations that are somewhat nonstandard. First, experiments can have at most two factors, and factors must be binary. Secondly, there is currently no option for filler items or counterbalanced lists, methods recommended in the experimental syntax literature to prevent participants from detecting, and perhaps responding atypically to, the patterns of theoretical interest (e.g. Cowart 1997). Finally, judgments must be made on a binary good/bad scale, rather than on the ordinal or continuous-valued scales often advocated in the experimental syntax literature (though see Weskott and Fanselow 2008 for arguments that a binary scale can suffice).

All three limitations exist solely to keep MiniJudge experiments as simple as possible, both conceptually and practically, for both the experimenter and the participants. In particular, just as the algorithm for generating materials described in 4.1 allows users to start with the sort of minimal set already familiar from traditional linguistic practice, the choice of a binary judgment scale is intended to link MiniJudge with the most commonly used acceptability diacritics in the linguistics literature (* vs. blank). Future versions of MiniJudge will convert these limits into mere defaults, while providing more flexible options for the more experienced experimenter.

MiniJudge surveys themselves can currently be distributed on paper or by email, as long as participants understand that they must judge items in the order in which they are presented. While these modes are presumably sufficient for collecting syntactic judgments from literate participants, it is reasonable to wonder whether phonological judgments would be better elicited using auditory stimuli (or video, in the case of sign languages). Implementing this suggestion would merely require a bit more software; the deeper question is how to interpret modality effects if any are found. Bailey and Hahn (2001) found very little effect of modality (auditory vs. written) in English nonword judgments, whereas Myers and Tsay (2005) found a stronger modality effect in judgments on Mandarin syllables (auditory vs. written in a quasi-phonemic orthography, described below). Auditory stimuli presumably engage the phonological processor more directly than written stimuli, but written stimuli have the advantage of eliminating acoustic ambiguity, and they perhaps also encourage judgments to be made at a more abstract, amodal level, rather than solely at a perceptual level.

After the raw results have been collected, MiniJudge reformats them so that they can be analyzed using mixed-effects logistic regression, another member of the loglinear family (Agresti 2002, Baayen 2008). This is a generalization of logistic regression, the statistical technique at the core of the sociolinguistic software tool VARBRUL (Mendoza-Denton, Hay, and Jannedy 2003), but as a mixed-effects model it takes random variation across participants (and items) into account along with the fixed experimental factors. It therefore permits by-participants and by-items analyses to be run simultaneously, and because these random variables are included inside the model, their contributions can be tested by likelihood ratio tests. Thus it may sometimes happen that item sets don’t differ much in their effect on judgments, and a by-participants analysis is sufficient. Mixed-effects models have the further advantage over separate by-participants and by-items analyses in that they are more sensitive in small experiments, since statistical reliability depends on the total number of observations, which is the product of the numbers of participants and items.

As a species of regression, mixed-effects logistic regression also permits non-categorical independent variables. By default MiniJudge includes the order of item presentation as a covariate, to help reduce the influence of shifting judgment scales over the course of the experiment, and interactions between presentation order and the experimental factors can be analyzed as well, if desired (see Myers 2007a,c for why this may be useful).

Phonologists are also given the option to factor out lexical covariates, in particular neighborhood density. A hypothesized constraint that continues to have a significant effect on judgments even when neighborhood density is factored out is more likely to represent an actual grammatical component, rather than merely the effects of exemplar-driven analogy.

5. A demonstration

Now that the philosophical underpinnings and technical details of MiniCorp and MiniJudge have been clarified, we can turn to an application with real data, involving a phonotactic pattern in Mandarin. The decision to look at phonotactics is dictated solely by the choice of language; Mandarin has relatively few alternations or prosodic phenomena (Duanmu 2007). MiniCorp and MiniJudge are not limited to examining phonotactics, however; any hypothesis that can be expressed in a standard OT grammar can be studied with MiniCorp, and any hypothesis predicting judgment contrasts can be studied with MiniJudge.

After proposing an OT analysis of the phonotactic pattern in Mandarin (5.1), I then describe how it was tested in a MiniCorp analysis (5.2) and in a MiniJudge experiment (5.3).

5.1 A pattern in Mandarin phonotactics

Mandarin syllable structure may be schematized as in (4), where C represents a consonant, V a vowel, and X either; the only obligatory element is a nuclear vowel.

(4) (C)(V)V(X)

Virtually all Mandarin morphemes are monosyllabic, so phonotactic patterns are syllable-internal. The most relevant to the one tested here are the following. As in many languages, vowels outside the sonority peak must be high, namely /i/ (front unrounded), /u/ (back rounded), or /y/ (front rounded); two high vowels cannot be adjacent in a syllable. In diphthongs and triphthongs, the nuclear vowel must be low (/a/) or mid, in the latter case agreeing in voicing and backness with the following vowel. The vowel /y/ can appear prevocalically, but not postvocalically. Thus the only two possible syllable-final diphthongs are /ou/ and /ei/.

With these exceptionless patterns as background, the phonotactic pattern tested here concerns the combinations of high vowels permitted in triphthongs. The pattern is illustrated in Table 1 (superscripts indicate the conventional numbering for the four lexical tones: 1 = high level, 2 = rising, 3 = low dipping, 4 = falling). The cells marked * represent unattested triphthongs. The pattern here is also seen with consonantal-initial syllables.

Table 1. Cooccurrence restrictions on Mandarin triphthongs

		First vowel		
		i	u	y
Last vowel	i	*iei *iai (some speakers)	uei^{4} ‘for’ uai^{4} ‘outside’	*yei *yai
	u	iou^{4} ‘again’ iau^{4} ‘want’	*uou *uau	*you *yau

The generalization is clear: triphthongs cannot start and end with vowels identical in backness or rounding (Duanmu 2007). However, some speakers have an apparent exception in the morpheme for ‘cliff’, pronouncing it /iai^{2}/. There are three further low-frequency exceptions cited in Mandarin dictionaries, all homophonous with this one. Other speakers pronounce the morpheme for ‘cliff’ as /ia^{2}/, consistent with the generalization.

This generalization is an instantiation of the Obligatory Contour Principle (OCP; S. Myers 1997). Within the OT framework, the fact that the OCP can be violated for some Mandarin speakers implies a grammar in which it is blocked by a higher-ranked faithfulness constraint. This blocking constraint must be indexed to apply only in an arbitrary lexical class (see Pater forthcoming for analyses like this). Thus we end up with an OT grammar with the structure in (5) (the faithfulness constraint is kept vague since theory-internal details are not relevant).

(5) FAITH$_{\textit{Exceptions}}$ » OCP

While this grammar is trivially simple, and the arguments for it familiar and rather banal, it raises two sets of difficult methodological questions. The first concerns the reliability of the grammar in (5) as a description of the Mandarin lexicon, the very data source that suggested it in the first place. Doesn't the mere existence of lexical exceptions cast doubt on the OCP being a genuine component of Mandarin grammar? Yet at the same time, aren't the number of exceptions too few (a mere four morphemes) to provide convincing evidence for an exception-specific FAITH$_{Ex}$ constraint? Finally, even if both constraints prove to give statistically reliable descriptions of the Mandarin lexicon, is their claimed ranking supported as well? After all, the very fact that the OCP is rarely violated implies that it provides better coverage of the data than the hypothesized exception constraint. Can this state of affairs truly be handled by ranking the latter over the former?

The second set of questions concerns the synchronic relevance of the grammar. Even if one or both of the constraints accurately describes the Mandarin lexicon, are they still reflected in contemporary native speaker judgments? If so, does the evidence for grammar in judgments remain even if analogy, as measured by neighborhood density, is taken into account? Finally, if the corpus and judgments prove to differ in what they say about the hypothesized grammar, how should this mismatch be interpreted?

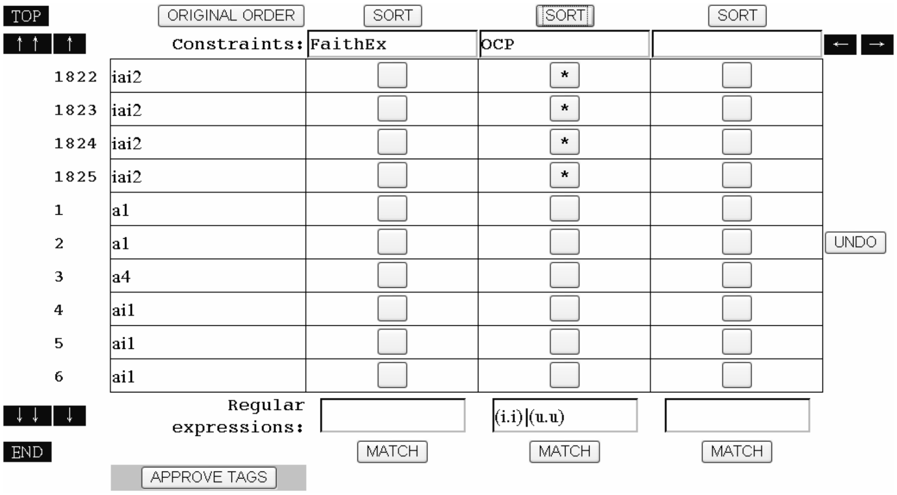
5.2 A MiniCorp analysis

The MiniCorp analysis began by entering a list of 13,607 Mandarin monosyllabic morphemes (Tsai 2000), transcribed using IPA for the segments and using the conventional single-digit notation for the four tones.

The next step was to annotate the corpus in terms of constraint violations. Given the grammar proposed in (5), there should be no violations of the undominated constraint FAITH$_{Ex}$. The OCP should be violated by morphemes both ending and beginning in /i/ or /u/ (/y/ can be ignored, since it cannot appear in triphthongs at all). Violations can thus be found with help of the regular expression in (6), where "." is the wildcard symbol (here representing the nucleus vowel) and "|" represents disjunction.

(6) (i.i)|(u.u)

This regular expression adds a violation mark to four items, namely the morpheme for 'cliff' and its homophones. The researcher can sort the corpus items according to violations to make sure that the annotations are correct; changes can be made by toggling violations on and off in the interface shown in Figure 1 (currently, this interface assumes that each form violates a constraint at most once, an obvious limitation to be fixed in the next version).



	Constraints:	FaithEx	OCP	
1822	iai2	□	*	□
1823	iai2	□	*	□
1824	iai2	□	*	□
1825	iai2	□	*	□
1	a1	□	□	□
2	a1	□	□	□
3	a4	□	□	□
4	ai1	□	□	□
5	ai1	□	□	□
6	ai1	□	□	□
	Regular expressions:		(i.i)|(u.u)	

Figure 1. MiniCorp interface for annotating lexical items for constraint violations

After the annotated corpus has been saved, MiniCorp generates an R script to classify lexical items by all possible combinations of constraint violations and count the associated type frequencies. The result here is shown in Table 2, where 1 in the constraint columns indicates violation and 0 indicates non-violation. Since loglinear models like Poisson regression cannot provide reliable coefficients if there are perfect correlations (Agresti 2002), the script converts all zero counts into one, as in the last two cells of the counts column (this weakens the statistical power only slightly).

Table 2. Adjusted type frequencies associated with constraint violations

Counts	Faith$_{Ex}$	OCP
13603	0	0
4	0	1
1	1	0
1	1	1

The R script then analyzes the frequency table using Poisson regression, as explained earlier in section 3.2, and outputs the results summary shown in (7); a more detailed statistical report is saved in an offline file. Ranking is tested in terms of the position of each constraint relative to all constraints hypothesized to be ranked lower. Thus for the grammar proposed in (5), we only need to test the ranking of FAITH$_{Ex}$ (relative to the OCP).

(7) a. Constraint test:

```
  Constraints  Weights  p
      FaithEx  -8.8252  0    *
          OCP  -7.9087  0    *
(* significant constraint)
```

b. Ranking test:

```
  Constraints  p
      FaithEx  0.2491
(No significant rankings)
```

The results in (7a) show that both constraints provide significantly reliable descriptions of the data (*p* values below .05), even the FAITH$_{Ex}$ constraint, which is only relevant in four morphemes. The weights for both constraints are negative, indicating that, as desired, they are obeyed more often than violated (that is, counts are lower when the independent variable is coded as 1 rather than 0).

Note also that the magnitude of the FAITH$_{Ex}$ constraint is slightly larger than that of the OCP, consistent with the ranking hypothesis. Unfortunately, as shown in (7b), this difference in constraint magnitude is not great enough to be significantly different by the likelihood ratio test. Thus we are not justified in positing the ranking in (5).

Without this ranking, however, the FAITH$_{Ex}$ hypothesis itself loses support, since the only reason this constraint was posited in the first place was to block the OCP in an arbitrary lexical class. The failure here does not mean that the concept of exception-specific faithfulness constraints is inherently flawed, though. Simulated data varying lexicon size and number of exceptions shows that a mere seven exceptions can suffice to provide statistically significant evidence (with the usual $p < .05$ criterion) for the undominated ranking of an exception-specific constraint.

5.3 A MiniJudge experiment

Despite problems with other aspects of the proposed OT grammar, the MiniCorp analysis showed that the OCP is a statistically reliable pattern in the Mandarin lexicon. To determine whether it remains active synchronically, native speaker acceptability judgments were collected and analyzed using MiniJudge.

Like many grammatical generalizations, the OCP involves the relationship between two elements, here the first and last vowel in a triphthong. These can be represented by the two binary factors [±FirstU] (whether or not the first vowel is /u/ rather than /i/), and [±LastU] (likewise for the last vowel). The OCP predicts an interaction between these two factors, such that forms with same-sign factor values should be judged worse than forms with different-sign factor values. These predictions are illustrated in (8) with a set of syllables unattested in the Mandarin lexicon (transcriptions again use IPA, other than the tone marks, so /t/, like /p/ in the examples to follow, represents an unaspirated plosive).

(8)	[+FirstU, +LastU]	tuau^2	[unacceptable?]
	[+FirstU, -LastU]	tuai^2	[acceptable?]
	[-FirstU, +LastU]	tiau^2	[acceptable?]
	[-FirstU, -LastU]	tiai^2	[unacceptable?]

In order to test whether the OCP applies beyond this single quartet, three further item sets were created. However, the need to avoid lexical items while respecting other phonotactic constraints meant that the perfect matching seen in (8) was not possible for these other sets. The variant sets generated with the help of MiniJudge thus had to be adjusted manually, resulting in the material list in Table 3 ([F] and [L] stand for [FirstU] and [LastU], respectively). The nucleus varies across the items in Sets 2 and 3 in order to obey the constraint, noted earlier, that a mid vowel in a triphthong must agree in rounding and backness with the final vowel. Similarly, the onset /n/ substitutes for /p/ in the first two columns in Sets 3 and 4 because of an independent phonotactic constraint against labial-round-vowel sequences (Duanmu 2007). The variation in mid vowels is thus confounded with the [LastU] factor, while the variation in onsets is confounded with the [FirstU] factor. This situation is not ideal, but at least neither is confounded with the crucial

[FirstU]×[LastU] interaction predicted by the OCP.

Table 3. Materials in the MiniJudge experiment

Factors:	[+F+L]	[+F-L]	[-F+L]	[-F-L]
Set 1:	tuau^2	tuai^2	tiau^2	tiai^2
Set 2:	tuou^2	tuei^2	tiou^2	tiei^2
Set 3:	nuou^2	nuei^2	piou^2	piei^2
Set 4:	nuau^2	nuai^2	piau^2	piai^2

These sixteen items were written in *zhuyin fuhao*, the quasi-phonemic Mandarin orthography used in Taiwan (functionally much like the *Hanyu pinyin* system used across the Taiwan Strait, but written in non-Roman symbols representing onsets, rimes, and tones rather than segments). MiniJudge was then used to generate printed surveys with the items in different random orders. Twenty native speakers of Mandarin in Taiwan, without any linguistic training, were asked to judge each item, in order, as being "like Mandarin" (*xiang Guoyu*) or not.

After the judgments were collected, MiniJudge created a data file and wrote an R script to run mixed-effects logistic regression on it, including a likelihood ratio test to determine whether cross-item variation needed to be taken into account. This script generated the summary report in (9), along with a more detailed statistical report saved offline. Crucially, the results revealed a significant interaction between the two factors ($p < .05$). There was also a main effect of [FirstU], but there is no theoretical significance of this; it could relate somehow to the failure to match onsets in two of the four sets, though the last line of the results summary indicates that including cross-item variation in the statistical model did not affect its fit with the data.

(9) Results summary for the initial analysis, generated by MiniJudge's R script

> The factor FirstU had a significant negative effect.
> The interaction between FirstU and LastU had a significant negative effect.
> There were no other significant effects.
>
> The above results do not take cross-item variability into account because no confound between items and factors was detected (p > .2).

The detailed report file gives the coefficient associated with the interaction in the best-fitting model as -0.469 ($p = .001$). The negative sign is consistent with the OCP because it means that same-sign items were judged worse than different-sign items. This interaction is much easier to appreciate from the graph in Figure 2, which is also automatically generated by the R script. As predicted by the OCP, triphthongs with identical first and last vowels tended to be judged worse than triphthongs beginning and ending in different vowels.

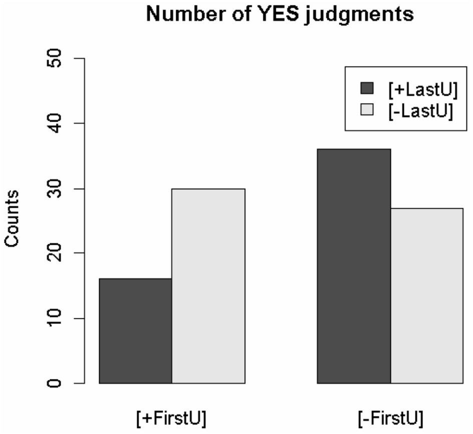


Figure 2. Graph generated by MiniJudge's R script

We have thus found evidence supporting the synchronic activity of the OCP in a very small and quick experiment, involving only twenty speakers judging sixteen items on a binary good/bad scale. However, as discussed earlier, a stricter test of claims about grammatical knowledge, as opposed to mere knowledge of superficial lexical statistics, would be to factor out analogical influences on judgments, as measured by neighborhood density. We have already seen an apparent example of the power of analogy in the Mandarin lexicon itself; recall that the exceptions to the OCP are all homophonous with each other.

MiniJudge took the neighborhood densities computed by MiniCorp for each of the sixteen experimental items, incorporated them into the data file, and generated a new R script taking them into account. The results summary was dramatically different, as seen in (10) (again the best-fitting model was by-participants only).

(10) Results summary for the analysis including neighborhood densities

Neighborhood density had a significant positive effect.
There were no other significant effects.

The detailed report file shows that neighborhood density was positively correlated with the probability of acceptance (coefficient = 0.013, *p* = .04); judgments were indeed affected by analogy with real lexical items. Meanwhile, the interaction predicted by the OCP, while still negative, was no longer significant (coefficient = -0.112, *p* = .62). The disappearance of the OCP effect when neighborhood density was taken into account raises the possibility that this effect was due primarily to analogical processes, not a grammatical constraint. Of course, as a null result in a small experiment, this conclusion cannot be conclusive, but it does seem suggestive given that adding neighborhood density caused such a large drop in significance for the OCP (from *p* = .001 to *p* = .62).

Putting the results from the MiniCorp and MiniJudge analyses together, then, it seems that although the OCP is consistent with type frequencies in the Mandarin lexicon and it correlates with native speaker judgments, these judgments may be sufficiently explained by analogy. If replicated in larger studies, the latter result may suggest that speakers do not need to actively process phonotactics in languages with syllable inventories small enough to memorize in toto (in contrast to languages with larger syllable inventories, where OCP-like constraints continue to affect judgments even when neighborhood density is controlled; see Frisch and Zawaydeh 2001 for Arabic, and Coetzee forthcoming for English). A more extreme possibility would be that the particular variety of the OCP seen in Mandarin triphthongs represents the kind of corpus pattern that cannot be learned by the child's grammar-learning algorithm, as alluded to earlier in 2.2 (in this scenario, the lexical pattern would be the result of extra-grammatical diachronic processes of the sort posited in Blevins 2004). Perhaps the most plausible possibility, however, is that Mandarin's very simple syllable structure, and consequently very small syllable inventory, means that neighborhood density is particularly tightly confounded with grammatical constraints, so factoring it out will tend to throw out evidence for genuine grammatical patterns even if they do exist. Indeed, the mean neighborhood density of the experimental items obeying the OCP (67.75) is much higher than that of the items violating it (12.25). Such observations raise the interesting methodological question of how grammatical patterns could be reliably detected in languages like Mandarin.

These conclusions, tentative though they are, are presumably of some relevance to theoretical phonology. Being based on inherently quantitative results, however, they could not have been reached without corpus analysis and formal experimentation. The contribution made by MiniCorp and MiniJudge is that they link such techniques directly with concepts and methods already familiar in theoretical phonology, including OT grammars, analogy, the analysis of dictionary data, minimally contrasting example sets, and binary acceptability judgments.

6. Conclusions and beyond

I began this chapter by asking whether the new methods currently sweeping the field of phonology are compatible with the traditional ones. I hope to have shown that they are, and that in fact the traditional methods can readily be "scaled up" to the same level of quantitative sophistication. To show how this process can be made simpler for theoretical phonologists without much quantitative experience, I described software tools designed to automate the most time-consuming and technically difficult steps, from corpus annotation to experimental material preparation to statistical analysis. The tools were then demonstrated in the testing of phonological hypotheses that could not have been tested with traditional methods alone.

The tools themselves, MiniCorp and MiniJudge, have already been used in a variety of linguistic studies, including, in the case of MiniJudge, studies on morphology and syntax (Myers 2007a,b,c, Ko 2007). However, they both continue to undergo refinement, and since both are open-source (under a GNU General Public License, permitting reuse of the code if it remains open-source), researchers impatient for upgrades are encouraged to borrow code or ideas for their own software tools.

Planned improvements include options for free corpus exploration (as in Uffmann 2006), ordinal and continuous-valued judgment scales, corpus-based testing of non-OT grammars (e.g. rule ordering tests following Sankoff and Rousseau 1989), and tools using electronic corpora to generate matched sets of nonword items for phonological judgments. Moreover, to make the programs easier to use for inexperienced users, future versions will not require R at all, though an R interface will remain available for those wanting to extend analyses. Linguists desiring quick results would also benefit from statistical tests optimized for extremely small samples (e.g. Myers, Huang, and Tsay 2007). Finally, the interface needs to be improved; eventually MiniCorp and MiniJudge will be integrated into a single package written in Java (MiniJudge already has an alternative Java implementation).

There are many practical advantages for linking traditional methods with the quantitative techniques standard in the rest of the cognitive sciences. Linguists may find it easier to collaborate with their colleagues across disciplines, theoretical linguistics students may be less intimidated by quantitative data, and certain controversies over empirical claims may be resolved more quickly and easily. Just as important, however, is a philosophical implication:

The old and new methods are truly part of a single, unified science of linguistics.

Acknowledgments

This research was supported in part by National Science Council (Taiwan) grants 94-2411-H-194-018, 95-2411-H-194-005, 96-2411-H-194-002. All versions of MiniCorp and MiniJudge are co-copyrighted by National Chung Cheng University. My thanks to Tsung-Ying Chen for helping to run the MiniJudge experiment; he and Chen-Tsung Yang also programmed the Java version of MiniJudge.

References

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

Baayen, R. H.
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 and Language* 44: 569-591.

Bird, Steven, and T. Mark Ellison
1994 One level phonology: Autosegmental representations and rules as finite automata. *Computational Linguistics* 20: 55-90.

Blevins, Juliette
2004 *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge, UK: Cambridge University Press.
2006 A theoretical synopsis of Evolutionary Phonology. *Theoretical Linguistics* 32 (2): 117-166.

Boersma, Paul and Bruce Hayes
2001 Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32 (1): 45-86.

Chomsky, Noam
1957 *Syntactic Structures*. The Hague: Mouton.
1965 *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
1980 Rules and representations. *Behavioral and Brain Sciences* 3: 1-61.

Chomsky, Noam, and Morris Halle
1965 Some controversial questions in phonological theory. *Journal of Linguistics* 1 (2): 97-138.
1968 *The Sound Pattern of English*. New York: Harper and Row.

Coetzee, Andries W.
Forthcoming Grammaticality and ungrammaticality in phonology. *Language*.

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

Cowart, Wayne
1997 *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. London: Sage Publications.

Duanmu, San
2007 *The Phonology of Standard Chinese*, 2nd ed. Oxford, UK: Oxford University Press.

Frisch, Stefan A., and Bushra Adnan Zawaydeh
2001 The psychological reality of OCP-Place in Arabic. *Language* 77 (1): 91-106.

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

Hayes, Bruce, and Colin Wilson
Forthcoming A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*.

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

Karttunen, Lauri
1998 The proper treatment of Optimality Theory in computational phonology. *Finite-state Methods in Natural Language Processing*, pp. 1-12. Ankara.

Keller, Frank, and Theodora Alexopoulou
2001 Phonology competes with syntax: Experimental evidence for the interaction of word order and accent placement in the realization of Information Structure. *Cognition* 79:301-372.

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

Kingston, John, and Mary E. Beckman (eds.)
1990 *Papers in Laboratory Phonology*. Cambridge, UK: Cambridge University Press.

Kiparsky, Paul
2006 Amphichronic linguistics vs. Evolutionary Phonology. *Theoretical Linguistics* 32 (2): 217-236.

Ko, Yu-Guang
2007 Grammaticality and parsibility in Mandarin syntactic judgment experiments. National Chung Cheng University MA thesis.

Labov, William
1975 Empirical foundations of linguistic theory. In: Robert Austerlitz (ed.) *The Scope of American Linguistics*, pp. 77-133. Lisse: Peter de Ridder.

Legendre, Géraldine, Antonella Sorace, and Paul Smolensky
2006 The Optimality Theory - Harmonic Grammar connection. In: Paul Smolensky and Géraldine Legendre (eds.) *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*, Vol. 2, 339-402. Cambridge, MA: MIT Press.

Luce, Paul A.
1986 Neighborhoods of words in the mental lexicon. Doctoral dissertation, Indiana University, Bloomington, IN.

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

Moreton, Elliott
Forthcoming Analytic bias and phonological typology. *Phonology*.

Morgan, James L., Katherine M. Bonamo, and Lisa L. Travis
1995 Negative evidence on negative evidence. *Developmental Psychology* 31 (2): 180-197.

Myers, James
2007a MiniJudge: Software for small-scale experimental syntax. *International Journal of Computational Linguistics and Chinese Language Processing* 12 (2): 175-194.
2007b Linking data to grammar in phonology: Two case studies. *Concentric* 33 (2):

1-22.

2007c Generative morphology as psycholinguistics. In: Gonia Jarema and Gary Libben (eds.), *The Mental Lexicon: Core Perspectives*, pp. 105-128. Amsterdam: Elsevier.

2008 Testing phonological grammars with dictionary data. National Chung Cheng University ms.

Forthcoming Bridging the gap: MiniCorp analyses of Mandarin phonotactics. *Proceedings of the 35th Western Conference on Linguistics*.

Myers, James, Shih-Feng Huang, and Jhishen Tsay

2007 Exact conditional inference for two-way randomized Bernoulli experiments. *Journal of Statistical Software* 21, Code Snippet 1, 2007-09-02.

Myers, James, and Jane Tsay

2005 The processing of phonological acceptability judgments. *Proceedings of Symposium on 90-92 NSC Projects*, pp. 26-45. Taipei, Taiwan, May.

Myers, Scott

1997 OCP effects in Optimality Theory. *Natural Language and Linguistic Theory* 15: 847-892.

Myers, Scott, and Benjamin B. Hansen

2007 The origin of vowel length neutralization in final position: Evidence from Finnish speakers. *Natural Language and Linguistic Theory* 25 (1): 157-193.

Ohala, John J.

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

Ohala, John J., and Jeri J. Jaeger (ed.)

1986 *Experimental Phonology*. Orlando, FL: Academic Press.

Pater, Joe

Forthcoming The locus of exceptionality: Morpheme-specific phonology as constraint indexation. In: Steve Parker (ed.), *Phonological Argumentation: Essays on Evidence and Motivation*. London: Equinox.

Prince, Alan

2007 Let the decimal system do it for you: A very simple utility function for OT. Rutgers University ms.

Prince, Alan, and Paul Smolensky

2004 *Optimality Theory: Constraint Interaction in Generative Grammar*. Oxford, UK: Blackwell Publishing.

R Development Core Team

2008 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sankoff, David, and Pascale Rousseau

1989 Statistical evidence for rule ordering. *Language Variation and Change* 1 (1): 1-18.

Schütze, Carson T.

1996 *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.

Tesar, Bruce, and Paul Smolensky

1998 The learnability of Optimality Theory. *Linguistic Inquiry* 29 (2): 229-268.

Tsai, Chih-Hao

2000 Mandarin syllable frequency counts for Chinese characters. Kaohsiung Medical University, Taiwan, ms. http://technology.chtsai.org/syllable/

Uffmann, Christian

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

Vitevitch, Michael S., and Paul A. Luce

1999 Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* 40: 374-408.

Weskott, Thomas, and Gisbert Fanselow

2008 Different measures of linguistic acceptability: Not so different after all? Talk presented at the International Conference on Linguistic Evidence 2008, Tübingen, Germany, January 31 - February 2.

Zuraw, Kie

2007 The role of phonetic knowledge in phonological patterning: Corpus and survey evidence from Tagalog infixation. *Language* 83 (2): 277-316.