Proceedings of Symposium on 90-92 NSC Projects, 26-45. Taipei, Taiwan.

The processing of phonological acceptability judgments[*]

James Myers and Jane Tsay, National Chung Cheng University

1. Introduction

In this paper we review a series of experiments on Mandarin syllable judgments that has had the aim of understanding the nature of phonological competence and performance in terms of explicit quantitative models. To provide some background for our research, we begin by addressing basic questions about the nature of evidence in generative phonology, arguing that native speaker judgments should play a more important role than they have so far. This then raises two central questions about the influence of extra-grammatical factors in judgments, namely what these factors may be and how they may be removed so that "true grammar," if it exists, can be made clear. We then describe our experiments, in which we manipulated a variety of processing parameters to study their effect on judgment patterns.

1.1 Phonological judgments

Judgments have not been the primary data source used in phonology (generative or otherwise), which has instead relied on attestation data, most usually dictionaries. This is so despite general acknowledgment of its inconclusive nature. As pointed out by Kenstowicz and Kisseberth (1979) and many others, attested phonological forms may simply be memorized as wholes, already patterned from diachronic sound change, thereby making synchronic phonological grammar superfluous (Blevins, 2004, argues that this may in fact be very close to correct). In sharp contrast to syntacticians, then, the only phonologists who seem interested in collecting grammaticality judgments are those who are already predisposed to laboratory experimentation (e.g., Ohala & Ohala, 1986; Coleman & Pierrehumbert, 1997; Wang, 1998; Hammond, 1999, 2004; Bailey & Hahn, 2001; Myers, 2002).

Even if all phonologists suddenly begin to rely on judgments like syntacticians, they will still have to be careful to recognize that judgments are not grammar, but (at best) mere reflections of grammar. This is in fact one motivation for the competence/performance distinction (Chomsky, 1965), since a judgment is a performance act, within which competence is (at best) just one component. Native speakers may feel that a form is acceptable, or unacceptable, for many reasons other than its grammatical status, such as ease of processing or similarity with established forms (i.e., analogy).

Since judgment making is an experimental task (as often emphasized by syntacticians; e.g., Phillips & Lasnik, 2003), it is important to recognize that an experiment is an attempt to test the relationship between hypothesized causal factors (the independent variables) and their hypothesized effects (the dependent variables). Here, judgments represent the dependent variable, and the independent variables include factors relating to both competence and performance. From a generative perspective, the performance factors are "nuisance variables." Their mere presence

[*] The continuing research reviewed in this paper is supported by National Science Council grants NSC-91-2411-H-194-022, NSC-92-2411-H-194-003, NSC-93-2411-H-194-003. Some of it has previously been presented in Myers (2004) and Myers and Tsay (2004). We would like to thank those who gave us comments on these presentations, Mike Hammond and José Benkí for inspiration and technical advice, and our many research assistants for their work both on the experiments and the development of the Southern Min spoken corpus.

does not make judgments useless as evidence about grammar, but they have to be understood or the researcher will never know if a judgment pattern is really due to competence and not performance. Below we discuss the most important nuisance variables in phonological judgments, and then address how they may be dealt with.

1.2 Extra-grammatical effects on phonological judgments

The most fundamental nuisance variable in phonology is lexical status (real vs. nonce). From a generative perspective, the lexical status of a form is irrelevant to its phonological grammaticality. Gaps in the lexicon may be systematic (ungrammatical), but they may also be accidental (grammatical). If a form is a lexical word, it may be highly likely to be grammatical (otherwise neither child nor linguist could infer the grammar in the first place), but this isn't guaranteed, since it may be an exception. Despite this logic, it is obvious that real words are much more likely to be judged as grammatical than nonwords.

Among real words, another factor expected to affect judgments is lexical frequency: more common words are likely to be judged as more acceptable. Again this must be recognized as a performance bias rather than a grammatical fact, since lexical exceptions aren't necessarily distributed in accordance with frequency. In fact, for certain types of phonological patterns, exceptions actually have higher frequency than most regular forms, since if they were less common speakers would tend to forget their exceptional status and regularize them (see, e.g., Bybee, 2001).

Judgments may also be expected to be affected by modality, i.e., whether they are made on spoken or written forms. Orthographic systems have rules that don't necessarily match those of the spoken language, and even if speakers are asked to judge how written forms sound rather than how they are spelled, spelling will likely have some sort of influence as well, despite its irrelevance to grammar.

1.3 Phonotactic probability and neighborhood density

Two further factors affecting phonological judgments are more subtle, but have received a lot of attention in the literature on phonological processing, as summarized below. Phonotactic probability represents the probability of a subsequence of phonological units across the lexicon; thus /sf/ is rare in the English lexicon (*sphere*, *sphinx*) while /st/ is common. Neighborhood density relates to the number of similar words; thus the form *lat* has many lexical neighbors in English (e.g., *cat*, *lap*), while *zev* has fewer. These two factors are highly confounded: forms with high values for one will also tend to have high values for the other. Moreover, both are derived from calculations made on the lexicon. Nevertheless, they have been claimed to be distinct in phonological processing: phonotactic probability acts as a filter prior to lexical access, providing an overall measure of the typicality of the target form being listened to, whereas neighborhood density only affects processing after forms in the lexicon themselves have been activated and begin to compete with the target form.

Evidence for processing distinctions between the two factors comes from a variety of sources. First, phonotactic probability has basically facilitatory effects, speeding up responses as listeners just begin to search for forms in memory, while the effects of neighborhood density may be inhibitory, since neighbors may compete with the target that must be responded to. Moreover, experimental tasks differ in the relative importance of the two factors. In a series of studies on English, Luce and colleagues (Vitevitch & Luce, 1998, 1999; Luce & Large, 2001) have pointed out that the auditory lexical decision task requires listeners to discriminate between words and nonwords, whereas tasks like auditory naming and auditory same-different matching tasks do not require the activation of neighbors. Thus lexical decisions should be inhibited by neighborhood density while the other two tasks should be facilitated by phonotactic probability. In general outline, these predictions have been supported. Another argument comes from time course studies: phonotactics is claimed to be prelexically while neighbors are activated postlexically (i.e., involving processes after access of words was completed, such as decision making). In support of these claims, Marantz and colleagues (Pylkkänen, et al., 2002; Pylkkänen & Marantz, 2003; Stockall, et al., 2004) report a magnetic brainwave component occurring about 350 milliseconds after stimulus presentation that seems to correlate with phonotactic probability but not neighborhood density.

To tease apart the effects of these two correlated factors, Bailey and Hahn (2001) took a different approach in their analysis of subjective wordlikeness judgments in English. Rather than attempting to control all nuisance variables, they measured a variety of such variables and included them in the analysis from the beginning so that they could be factored out statistically at the end, using regression analysis (this involves the mathematical derivation of "best-fit" lines from raw data showing the separate contribution of each independent variable to the dependent variable). Their results showed that phonotactic probability and neighborhood density both had separate, positive effects of judgments. This approach is the one most relevant to our own research, since not only did they use judgments as their dependent measure, but we also followed them in relying on regression analyses.

A simple serial model with two distinct processes seems problematic, however, given that phonotactic probability and neighborhood density interact with each other, as Luce and Large (2001) and Stockall, et al. (2004) have shown for English. Moreover, the fact that both are derived from the lexicon is consistent with the fact that the "phonotactic" brainwave component identified by Pylkkänen, et al. (2002) is also sensitive to lexical frequency (Embick, et al., 2001). Thus there may well be no such thing as truly "prelexical" processing.

Not only are phonotactic probability and neighborhood density hard to distinguish from each other, but it is also not clear what their relationship is to grammar. Hammond (2004) argues that both can be formalized within Optimality Theory, but following Boersma and Hayes (2001), he rejects the competence/performance distinction as standardly understood, so that his grammar generates judgments directly (see Keller and Asudeh, 2002 for a critique of this approach). Perhaps closer to a mainstream approach would be to suppose that neighborhood effects are truly extra-grammatical, since they essentially reflect analogy, whereas phonotactic probability includes generalizations that have traditionally been considered part of grammar. A third possibility is to consider neither phonotactic probability nor neighborhood density parts of grammar. As pointed out by Inkelas, Orgun, and Zoll (1997), mere systematicity isn't sufficient to give lexical generalizations the status of "linguistically significant," since they can also arise through historical accident. However, their point actually applies to all lexical patterns, even ones that seem "natural" to a phonologist. The above research shows that the mind is capable of mirroring probabilistic phonotactic patterns, derived mechanically from a specific lexicon, in judgments. Thus it seems that judgments cannot give us any more information about grammar than the lexicon itself, unless we also factor out phonotactic probability. We return to this issue at the end of the paper.

1.4 The scope of our research

The goals of our project are to explore the processing of phonological judgments in a regression framework, with a central focus being the roles of phonotactic probability and neighborhood density. However, we also go beyond previous literature in various ways. Specifically, we have manipulated task, modality, the time given for making judgments, and memory load. We have also searched for evidence that a purely grammatical factor, independent of all of these other factors discussed above, is necessary to account for phonological judgment patterns.

The language we report on in this paper is Mandarin (we have also extended the tasks and models to Southern Min). Aside from convenience, there are least two other factors that make Mandarin interesting. First, its phonology is relatively simple: the number of phonemic units and their combinations is quite limited, its syllabary very small (around 1,500 characters, according to Ho, 1976), and there are virtually no cross-syllable interactions or morphologically sensitive patterns. Thus we can concentrate on general processes involved in judgment making rather than getting bogged down in language-specific complexities; the small syllabary also means that we can run larger subsets of it in experiments than is possible to do in English. Second, as a language very different from English, we can explore possibly universal and language-specific aspects of the judgment process.

2. General methods

The experiments described in this paper overlap greatly in their participant pool, materials, independent variables, procedures, and analyses.

2.1 Participants

Our experimental participants are all undergraduates at National Chung Cheng University with no training in linguistics. All are (near) native speakers of Mandarin, though like most Taiwanese, some also know Southern Min. Altogether, the experiments reported in this paper involved 200 participants.

2.2 Materials

Since our primary goal is to understand the judgment process, not Mandarin phonology, we chose materials solely for their simplicity (e.g., they did not involve medial glides, the subject of much phonological interest). Our experimental items are all syllables conforming to a CV(C) template, derived from all possible combinations of the eight onset phonemes /p, p^h, m, f, t, t^h, n, l/, four vowel phonemes /a, i, u, ə/, three endings /n, ŋ/ and ∅, and four tones. Of these 384 syllables, 235 appear in real words in Mandarin (i.e., lexical syllables) and 149 do not (i.e., nonlexical syllables).

The analyses here use syllable frequencies calculated from the character frequencies of Li, Li, and Tseng (1997) combined with character pronunciations from Tsai (2000). Since they are derived from written corpora, they are not ideal (by contrast, the frequency counts for our parallel research on Southern Min come from our own spoken corpus, currently containing 400,000 transcribed words).

In most of our experiments, the 384 syllables are divided into two sets of 192, balanced to contain approximately the same proportion of lexical vs. nonlexical syllables and systematic vs. accidental gaps (as judged by us phonologists). This is both to reduce participant fatigue and to permit counterbalanced designs in some experiments. We have found no consistent differences in the two sets, and so in the analyses below we collapse across them.

2.3 Defining phonotactic probability and neighborhood density

Adding to the complexity of their interpretation, both phonotactic probability and neighborhood density have been defined in more than one way, even by the same authors. We have primarily built on the definitions used by Bailey and Hahn (2001), with some additional modifications of our own. Here we briefly explain our choice.

Regarding the phonotactic probability for a target item, Bailey and Hahn (2001) used the conditional probability that some segment follows a given segment (e.g., the probability of /t/ being followed by /a/ rather than /i/). To assign a single value to an item, they computed the geometric mean (the *n*th root of the product of *n* values) of all conditional probabilities in the item. We followed this procedure, so that vowels were conditioned off of onsets and codas off of vowels. Tones were conditioned off of onsets, since tones interact phonologically with onsets more than other parts of the syllable in Mandarin (e.g., voiced onsets tend to disfavor the high-level tone). The probabilities used for these calculations are usually based on type frequencies (the number of lexical syllables containing a phoneme sequence), but as Bailey and Hahn (2001) note, if speakers learn probabilities from experience, then token frequency (how often a lexical item is encountered) should also play a role. They failed to find evidence for this in their study of English, but in our study we found a better fit with Mandarin judgments when phonotactics was defined using token frequencies; thus for us the conditional probability for /ta/ was the sum of the token frequencies of all words containing /ta/ divided by that for lexical syllables containing /t/.

Neighborhood density has also been defined in more than one way, though they are all built around the notion of "edit distance," which measures the number of deletions, insertions, or replacements that can change one form into a lexical item (e.g., using phonemic units, /pa/ has an edit distance of 1 from /a/, /pan/, and /ba/). This measure is then weighted by the the (log) token frequency of each neighbor, so that an item will have a higher density score if its neighbors are more common.

The definition used by Bailey and Hahn (2001) embeds these ideas in a much more complex model; we only adopted those aspects that seemed to make a difference with our data. Thus we used a simple weighting by frequency (rather than their quadratic weighting function, which allows for a non-monotonic effect that they found, but we didn't), a simple phoneme-based edit distance metric that ignored features (they didn't find a major role for features anyway), and built-in parameters (rather than deriving them from the data by nonlinear regression as they did, since our study involved too many different data sets). However, we did adopt their assumption that that all lexical items are neighbors to all targets, but just to different degrees, with neighbor influence rapidly dropping off with distance (according to an exponential function, widely applicable in perceptual psychology).

Note that our definitions for these two factors rely on phonemic representations, which has worked well for English. However, for reasons that will become clearer below, we are also currently exploring an alternative approach (not reported here) relying on psychoacoustically detailed representations computed automatically using the algorithm of Kirchner (2004).

2.4 Procedures

In the judgment tasks, participants were presented with syllables auditorily and/or visually (written in the BPMF system), and asked to judge how similar to Mandarin the item is on a scale from 1 to 6, where 6 represents "most like Mandarin" (*zui xiang Guoyu*). In most experiments, there was no time pressure, since we were interested in modeling the processes used by linguists when pondering the grammaticality of forms. Syllables were displayed and judgments collected by computers running E-Prime (Schneider, Eschman, & Zuccolotto, 2002). An alternative way of collecting judgments that we are currently exploring is magnitude estimation (Bard, Robertson, & Sorace, 1996), which is more sensitive to subtle effects and also generates continuous rather than ordinal values, simplifying statistical analysis and interpretation.

2.5 Analyses

Although judgments were made on an ordinal scale, we did not use tests designed for such data (see Agresti, 1989, for a review) due to their computational complexity and relatively low power. Instead, we followed Bailey and Hahn (2001) in using standard parametric statistics after transforming the raw data with an arcsine function to distribute the variability more evenly across the entire scale. The resulting distributions were not normal, but this is much less of a problem for parametric statistics than once thought (see, e.g. Rasch & Guiard, 2004). Also following Bailey and Hahn (2001), wherever possible we used the repeated-measures regression algorithms of Lorch and Myers (1990), which involve computing regression equations separately for each participant and testing the significance of the coefficients with one-sample t tests across participants. Overall model fits were measured with R^2, which reflects the proportion of variance in the dependent measure captured by the model; like Bailey and Hahn (2001), we computed it on cross-participant averages.

3. Experiments

The experiments described below examined the effects on judgment patterns of phonotactic probability, neighborhood density, and their interaction with lexical status; the relationships of judgment patterns with patterns in other phonological tasks; the effects on judgments of modality, judgment speed, and working memory load; and the role (if any) of "pure grammar" in the judgment process.

3.1 Lexical status, frequency, phonotactic probability, and neighborhood density

We began our investigation with an experiment in which 40 participants made judgments of syllables, presented simultaneously in auditory and written (BPMF) form, prior to conducting other phonological processing tasks on the same syllables. This experiment revealed effects of lexical status, (log) frequency, phonotactic probability and neighborhood density that were replicated, with variations depending on other factors, in all of the other experiments. In Tables 1 and 2 we report two of the regression analyses we conducted, one containing only these four factors (nonlexical syllables were given a log frequency of 0), and a second that also contained terms for the interactions of phonotactic probability and neighborhood density with lexical status (coded as 1 for lexical items and -1 for nonlexical items). Because only the signs of the coefficients are meaningful, not their magnitudes (and coded variables can't be standardized), we report only the associated by-participant *t* values.

Table 1. Judgments: simple model

R^2 = .72***	Lexical status	Frequency	Phonotactics	Neighbors
t	9.42***	14.42***	15.09***	5.69***

***p < .00001

Table 2. Judgments: interaction model

$R^2 = .75$***	Lexical status	Frequency	Phono-tactics	Neighbors	Status × Phonotactics	Status × Neighbors
t	2.37*	15.32***	15.34***	-0.37	-8.64***	9.84***

*p < .05, ***p < .001

Both models capture a respectable amount of the variance across judgments (over 70%). Overall, judgments are higher for syllables that are lexical, more common, phonotactically more typical, and (at least in the simpler model) in denser neighborhoods. This basic pattern replicates that found in English by Bailey and Hahn (2001). In particular, the independent effects of phonotactic probability and neighborhood density in the simpler model implies that they may well be measuring distinct psychological processes, as is claimed in the literature.

Additional support for their different natures emerges in Table 2. Bailey and Hahn (2001) only analyzed results for nonlexical items, but we included lexical items as well. Thus we can examine the interaction between lexical status and the two key factors to determine whether they play different roles in the two types of target items, as Luce and colleagues have claimed in their research (starting with Luce & Pisoni, 1998). Indeed, both interactions are not only significant, but go in opposite directions: the interaction with phonotactic probability is negative, meaning that judgments of nonlexical items show greater influence of this factor than do lexical items (recall that nonlexical items were coded as -1), while the interaction with neighborhood density is positive. In fact, the interaction between lexical status and neighborhood density was so great that it cancelled out any overall effect of neighbors across all items.

This pattern can be seen even more clearly when we look at regressions run on lexical and nonlexical syllables separately, as summarized in Tables 3 and 4.

Table 3. Judgments: lexical syllables only

$R^2 = .48$***	Frequency	Phonotactics	Neighbors
t	15.32***	3.73***	9.80***

***p < .001

Table 4. Judgments: nonlexical syllables only

$R^2 = .21$***	Phonotactics	Neighbors
t	14.33***	-6.23***

***p < .001

Both factors are significant for both types of syllables, but the signs of the coefficients show that the effect of phonotactic probability is positive for both, while the effect of neighborhood density is positive for lexical syllables but negative for nonlexical syllables. This is consistent with the standard model, where phonotactics is assumed to have a prelexical effect, where lexical status should be irrelevant. Similarly, the opposite directions of the neighborhood effect for lexical and nonlexical syllables is consistent with its being available only after the lexicon is contacted. The reason for the particular directions is not immediately obvious, but one might speculate that similarity to other real syllables may boost a lexical syllable's "real" status, hence its judgment score, whereas similarity to a real syllable may cause confusion, and hence lower scores, if the target syllable is in fact not lexical.

Note that the interaction between lexical status and phonotactics in Table 2 shows that its effect is significantly stronger in nonlexical syllables, which should not be possible if this effect arises before the processor has determined the lexical status of the syllable. This suggests that there may be a tradeoff in the two effects for nonlexical syllables, and a further regression demonstrates it: judgments for nonlexical syllables show a significant interaction between phonotactic probability and neighborhood density ($t(39) = -7.32$, $p < .001$), just as Luce and Large (2001) had found for English with non-judgment tasks, and also like their study, there was no such interaction for lexical syllables. Interestingly, however, Bailey and Hahn (2001) found no such tradeoff in their English judgment experiments on nonlexical syllables, where both neighborhood and phonotactics showed positive effects.

This last point hints at a possible cross-language difference that may have its origins in the different syllabary sizes in Mandarin vs. English. English allows so many phoneme combinations that the boundaries of the syllabary are quite fuzzy; the very next borrowing or coinage may introduce a new one. This is not the case in Mandarin, which strongly resists change in its lexical syllabary; even borrowings are forced into the already-existing set. Thus if the boundaries of the lexicon are perceived by Mandarin speakers as quite sharp, nonlexical syllables with global similarity to lexical syllables (i.e., high neighborhood density) may be considered particularly anomalous rather than particularly acceptable. That is, in contrast to the way English judges perceive them, Mandarin judges may take the apparent lexicality of such syllables as a misleading illusion that must be suppressed. Obviously this is rather speculative at the moment, but it seems worthy of further study.

For our purposes the major implications of these findings are as follows. First, phonological judgments are indeed influenced by phonological factors operating independently of lexical status and frequency, consistent with fundamental generative assumptions. Second, if we take neighborhood density effects as analogical effects, we can conclude both that analogy does indeed affect phonological judgments, but also that phonotactic probability, which is not equivalent to analogy, affects judgments as well. Third, the behavior of phonotactics indicates that it is nevertheless also influenced by lexical processing, given that it interacts with neighborhood density and lexical status. Unless "pure grammar" can be found elsewhere, then, all elements contributing to phonological judgments involve lexical processing.

3.2 Judgments vs. other phonological processing tasks

As we saw in the review earlier, phonological processing is task-specific, with different factors playing different roles in different experimental tasks. This is just what a generative phonologist would expect: competence may be constant, but performance is variable. Yet judgment making is also a specific experimental task. Can we really assume that it is a more reliable reflection of competence than other tasks? Critics claim that it may actually be much less reliable, since as a metalinguistic task, it is more sensitive to context and other strategic factors (Ohala, 1986; Edelman and Christiansen, 2003).

To address this issue, we included three other phonological tasks in the experiment described above, each associated with a different mode of processing: a perception task, a production task, and a recall task. Each of the 40 participants thus performed four different tasks on the same syllable set. The judgment task was always conducted first, in order to avoid contamination from the other tasks, and the recall task was always last, for reasons explained shortly, while the order of the perception and production tasks was counterbalanced across participants. The perception task (run using a Matlab program written by José Benkí as part of a collaborative project with our lab) required participants to listen to their assigned set of 192 syllables presented in noise (3 dB signal-to-noise ratio) and write down what they heard in BPMF. The dependent measure was a binary correct/incorrect score. The production task (run using DMDX[1]) required participants to read aloud each of the 192 syllables in their assigned set, written in BPMF on a computer screen. The main dependent measure was naming latency (response time or RT) for correctly named syllables. In the recall task (run using E-Prime), participants were simultaneously presented with auditory and visual (BPMF) forms of all 384 syllables, and they were asked to indicated, through button presses, which had been presented in the previous three experiments. The dependent measure was response time (RT) for correctly recognized syllables. Since the perception and production tasks involved reading, we also included visual forms in the judgment and recall tasks to reduce confounding between task and modality.

To simplify cross-task comparisons, the analyses below are based on scores averaged across participants. This gave, for each item, an average (arcsine transformed) judgment score, an average production RT, an average recall RT, and a proportion correct (PC) score for the perception task. To help normalize the PC scores, they were also transformed using the arcsine function. The within-participant design meant that differences across tasks must really be due to the tasks and not cross-participant differences (though possible task×participant interactions remain hidden).

We were interested to know the relationship between the tasks, independent of the effects of lexical status, frequency, phonotactic probability, and neighborhood density. Hence we conducted partial correlations, separately for lexical and nonlexical syllables, on results from the four tasks plus phonotactic probability, neighborhood density, and frequency (for lexical syllables). Like regression, partial correlations describe the relationship between variables separately from the other variables in the analysis. In fact, the *p* values they provide are identical to those obtained from a series of regressions using the four task measurements as dependent variables and all remaining factors as independent variables. These regressions also provide R^2 values as an indication of how much of the variance in each task is explained by the other variables. Due to lack of space, we will focus on patterns relating to the tasks themselves, rather than their different sensitivities to phonotactic probability vs. neighborhood density. The results are summarized in Tables 5 and 6.

Table 5. Partial correlation *r* and regression R^2 for lexical syllables

			Recall R^2 = .25***
		Prod R^2 = .43***	r = -.33***
	Perc R^2 = .15***	r = -.08	r = -.11
Judge R^2 = .62***	r = -.03	r = -.45***	r = -.40***

***$p < .001$

[1] This is free software available at http://www.u.arizona.edu/~kforster/dmdx/dmdx.htm.

Table 6. Partial correlation *r* and regression R^2 for nonlexical syllables

			Recall R^2 = .09*
		Prod R^2 = .28***	r = -.19*
	Perc R^2 = .07	r = .06	r = .03
Judge R^2 = .44***	r = .21*	r = -.49***	r = -.23**

*$p < .05$, **$p < .01$, ***$p < .001$

With the exception of the perception measure for nonlexical syllables, all models were significant. The fact that judgments were consistently better explained by the other variables than were the other three tasks is intriguing, but may have a mundane explanation, since the judgment measure inherently had the least variance to be explained (derived as it was from an ordinal scale); in any case, comparing regression models with different independent and dependent measures is not recommended.

More revealing are the partial correlations. These show that judgments are consistently related with scores from the other three tasks in reasonable ways (again with the exception of perception, which showed no correlation with judgments for lexical syllables). Thus when judgments are higher, production and recall are both faster (i.e., RT scores are lower, resulting in negative correlations), and perception PC is higher as well (though only for nonlexical syllables). Since RT and PC both indicate ease of processing, this pattern is consistent with the generative assumption that competence underlies all language processing, making grammatical forms easier to process. However, given that correlation is a symmetrical relationship that does not indicate causality, the results just as well support the conclusion that participants were relying on ease of processing, rather than grammaticality per se, to make their judgments. In either case, the fact that the results make sense shows that judgments do not provide a misleading picture of phonology.

In fact, they seem to be more stable than the other three task measurements. Perception is particularly erratic, showing no correlations with the production or recall tasks, as if they tap into totally distinct processes. Similarly, while production and recall are consistently correlated, the pattern is somewhat mysterious; even though both represent ease of processing, their correlation is negative, as if they trade off from one another. This confirms that if one’s main interest is in phonological competence rather than specific processing modalities, judgments do indeed seem to be the right measure to use.

3.3 Modality

Our first judgment experiment involved items that were presented simultaneously auditorily and visually (BPMF). One reason was to help equate materials across tasks, but another was that presenting visual forms may help put participants into the appropriately metalinguistic frame of mind, treating syllables as objects of study rather than communicative elements. Still, it is natural to ask is whether modality affects judgment patterns. Curiously, Bailey and Hahn (2001) did not find strong evidence for this in their study of English judgments, a finding relied on by Hammond (2004), who used only written stimuli. Marantz and colleagues (starting with Pylkkänen et al., 2002) have also relied on written stimuli, better suited to their brain-imaging methods than spoken stimuli.

Given the apparently greater mismatch between English phonology and spelling than between Mandarin phonology and BPMF, one might expect Mandarin judgments to show no modality effect either. If an effect is found, however, its patterning could provide further insights into the judgment process and Mandarin phonology. Note first that the BPMF system has a few systematic mismatches from the actual production of Taiwan Mandarin, most notably in the representation of labial-initial syllables; for example, "wind" is spelled /f-əŋ-1/, but most Taiwanese (including our participants) actually pronounce it [foŋ1]. The transcriptions used in our definitions of lexical status, phonotactic probability, and neighborhood density reflected these actual pronunciations, so "wind" was transcribed /foŋ1/, making /fəŋ1/ a nonword. Hence if such mismatches are relevant, we expect that judgments of spoken forms will be predicted better by our models than judgments of written forms.

An opposite prediction emerges from another possible difference between BPMF and actual Mandarin phonology. Like the English alphabet, the BPMF system assumes that syllables can be split up into discrete units, and this assumption was adopted in the transcriptions used in our model (though they didn't follow BPMF in treating rimes as indivisible units). Now, it is conceivable that the greater complexity of the English syllabary makes such an analysis more plausible as a phonological processing model for English speakers than for Mandarin speakers. English speakers have many syllables to memorize and the relatively free distribution of segments in them makes it easier to derive a segmental analysis from the data. By contrast, Mandarin speakers have fewer syllables to memorize and the distribution of their phonemes is more restricted. Thus if Mandarin listeners rely more on whole-syllable processing when presented spoken forms, we expect our models to do worse with judgments on them than with judgments on written forms, which share the segmental assumption of our models.

To test these competing predictions, we had 20 participants judge spoken forms and 20 judge written forms (BPMF), and then conducted a series of regressions on their (transformed) judgments, separately by modality and by lexical status. The results are summarized in Tables 7 and 8.

Table 7. Effect of modality: lexical syllables only

		Frequency	Phonotactics	Neighbors
Auditory R^2 = .28***	t	9.46***	2.74*	5.35***
Visual R^2 = .36***	t	11.95***	4.38***	10.95***

*p < .05, ***p < .001

Table 8. Effect of modality: nonlexical syllables only

		Phonotactics	Neighbors
Auditory R^2 = .04*	t	4.82***	-1.22
Visual R^2 = .17***	t	9.68***	-3.00**

*p < .05, **p < .01, ***p < .001

The above results replicate the basic pattern we saw from the first judgment experiment: both phonotactic probability and neighborhood affect judgments for both lexical and nonlexical syllables, but while phonotactics always has a positive effect on judgments, neighbors have a positive effect for lexical syllables but a negative effect for nonlexical syllables. What's new is the role of modality. The most obvious effect is that the influences of both factors seem more reliable when participants are judging visually presented syllables. Thus for lexical syllables, the phonotactic effect has a higher *p*-value (a valid measure of effect size given that the numbers of participants were matched across conditions), and for nonlexical syllables, the neighborhood effect is robust with visual stimuli but absent with auditory stimuli. This pattern is more consistent with the second of the two predictions given above: our models better reflect BPMF judgments than auditory judgments because the segmental assumption is not appropriate for the representations used in purely phonological processing.

Given the different patterns across lexical and nonlexical syllables, we next decided to conduct an ANOVA on the (transformed) judgments averaged across items by participant, with modality (auditory vs. visual) as between-group factor and lexical status (lexical vs. nonlexical) as within-group factor. The results can be seen in Table 9 below. The analysis showed the usual main effect of lexical status ($F(1, 38) = 527.54$, $p < .001$), with lexical syllables receiving higher scores. It also found a main effect of modality ($F(1, 38) = 9.27$, $p < .01$), with visually presented syllables receiving higher scores, and a significant interaction ($F(1, 38) = 11.75$, $p < .01$), since the modality effect was restricted to lexical syllables.

Table 9. Effects of modality and lexical status on judgments

	Lexical	Nonlexical	Average
Auditory	0.701	0.330	0.516
Visual	0.839	0.338	0.589
Average	0.770	0.334	

Since this analysis makes no reference to our models of phonotactic probability and neighborhood density, the modality effect here cannot be due to our transcriptions, but reflects a more general issue: Mandarin speakers are biased to give higher judgments to items presented orthographically than visually, but only if they are lexical items. The simplest explanation of this pattern begins with the assumption that judgments are made primarily on lexical status, a reasonable assumption given our instructions. Modality affects detection of lexical status because written forms are unambiguous while spoken forms are more liable to misperception (e.g., a /p/ may happen to sound like a /t/). Judgments of spoken lexical syllables drop because they have an increased chance of being misheard as nonlexical syllables. To make this work, we must also assume that spoken nonlexical syllables do not have an increased chance of being misheard as lexical syllables, which may be a corollary of the “sharp syllabary boundary” proposal made above. Of course, for now this has to be considered an unproven, though intriguing, speculation.

Regardless of their ultimate interpretation, these results lead to two interesting conclusions. First, phonological representations in Mandarin may be less segmentalized than those used by English speakers. Second, researchers conducting phonological judgment experiments cannot assume that auditory stimuli will necessarily provide stronger, more reliable results.

3.4 Speed

As a metalinguistic process, judgment making is considered to be nonautomatic. Hence we expect it to operate more slowly than other sorts of phonological processing, which is why we gave no time pressure to the participants in the above experiments.

To explore how important the speed of decision making really is, we decided to manipulate it experimentally. Aside from mere curiosity, another motivation was that speed seemed to offer a way of distinguishing the effects of phonotactic probability and neighborhood density. If the former really operates prelexically (or at least automatically), it should affect judgments regardless of how fast they are made. By contrast, neighborhood density has been claimed to play a role only at a later stage of processing, affecting decision making itself but not earlier automatic stages (see, e.g., Vitevitch & Luce, 1999; Pylkkänen et al., 2002).

In our experiment, we had 20 participants make judgments on the syllables quickly and 20 others make judgments on them slowly (all stimuli were auditory, to control modality and make them more natural). The participants in the "quick" condition were given only two seconds after the onset of a stimulus to give their response. Participants in the "slow" condition were forced to wait five seconds from the onset of the stimulus before their responses would be accepted; after this time, they again had two seconds in which to respond. To train participants to obey these time constraints, responses received outside of the permitted windows (after two seconds in the "quick" condition and before five seconds in the "slow" condition) would cause the program to freeze for ten seconds and display a bright red (and hopefully highly irritating) warning message.

Since the basic design of the experiment was the same as for the modality experiment, our analyses followed the same structure. Thus we first conducted a series of regressions on the (transformed) judgments, separately by modality and by lexical status. To our disappointment, however, the results revealed no effect of speed on the roles of phonotactic probability and neighborhood density. Regardless of speed, lexical syllables showed effects of both factors, while nonlexical syllables only showed effects of phonotactics, which we already knew was typical for auditory stimuli.

More interesting results emerged from the ANOVA, conducted on the (transformed) judgments averaged across items by participant, with speed (quick vs. slow) as between-group factor and lexical status (lexical vs. nonlexical) as within-group factor. As summarized in Table 10, judgments showed the usual main effect of lexical status ($F(1, 38) = 342.00$, $p < .001$) but no overall effect of speed ($F < 1.7$, $p > .2$), though there was a trend for scores to be higher for slower judgments. There was, however, a significant interaction ($F(1, 38) = 8.53$, $p < .01$). Namely, the speed effect seemed to be localized to lexical syllables, which received higher acceptability scores the more time was available to process them. No real difference was found with nonlexical syllables.

Table 10. Effects of speed and lexical status on judgments

	Lexical	Nonlexical	Average
Quick	0.624	0.336	0.480
Slow	0.716	0.321	0.519
Average	0.670	0.329	

What could be responsible for this pattern? Building on the speculations sketched in the previous section, we can again assume that judgments are primarily made on the basis of lexical status. Some words take longer to recognize than others, so if insufficient time is given to make a judgment, a higher proportion will be miscategorized as nonlexical and thus given lower scores.

The lack of overall speed effects seems to provide both theoretical and practical lessons. The theoretical lesson is that the properties picked up by judgments are processed quite fast, which may give comfort to linguists who assume that judgments tap into deep aspects of language, rather than superficial, consciously controlled ones. The practical lesson is that researchers with access to experimental control programs can run their studies faster with no apparent reduction in data quality, while researchers who rely on slower paper-and-pencil methods don’t have to worry that their data is less reliable than those from their more high-tech colleagues

3.5 Working memory load

Because the speed manipulation did not succeed in teasing apart the processing roles of phonotactic probability and neighborhood density, we tried another tack. Our idea was that if the former is (mostly) prelexical and automatic, while the latter involves (semi-conscious) activation of lexical neighbors, they should have different effects on working memory load. Processing neighborhood density should involve placing non-target lexical items (i.e., the neighbors) in working memory as the target itself is processed, but processing phonotactic probability should not. Thus if we load participants’ working memories with distracter syllables, irrelevant to the task at hand, we should find no change in the phonotactic effect but a reduction in the neighborhood effect, since there would be less room available in working memory for activating neighbors.

Since we were primarily interested in any possible interactions between memory load on the one hand and phonotactic probability and neighborhood density on the other, we manipulated memory load within participant groups to increase statistical power. Thus all participants were asked to judge syllables from the entire collection of 384, divided (by the two matched sets) into two blocks: one with a light memory load and one with a heavy memory load. The sets and the order of the blocks were both counterbalanced across participants. We knew that modality interacts with phonotactic probability and neighborhood density, so 40 participants made judgments on spoken syllables and 40 on written syllables (BPMF).

Memory load was manipulated as follows. Within each block, syllables were divided into subblocks of 17, with proportions of lexical and nonlexical syllables reflecting the collection as a whole. Prior to presentation of each subblock of syllables, participants were presented (in the same modality as the experiment as a whole) either with one (light load) or three (heavy load) distracter syllables. These were taken from the original collection of 384 syllables, carefully chosen so that they were never neighbors of any syllable in the associated subblock (at most matching only vowel, tone, or coda). Participants were told to memorize these syllables, and after each subblock, their recall was tested by asking them to name them aloud (if the stimuli were auditory) or writing them (if the stimuli were visual). Thus participants were forced to hold the distracter syllables in memory while making judgments on the remaining syllables.

Analysis involved separate regressions by lexical status and by modality, with memory load included as a coded variable (1 = high load, -1 = low load). The regressions also included the interaction factors Load × Phonotactics (L×P) and Load × Neighborhood (L×N). The regression results are summarized in Tables 11 and 12.

Table 11. Effect of memory load: lexical syllables.

		Freq	Phon	Neigh	Load	L×P	L×N
Aud R^2 =.27***	t	16.22***	6.89***	-4.29***	0.91	-1.58	1.45
Vis R^2=.47***	t	18.82***	1.32	11.64***	0.75	-3.24**	1.37

**p < .01, ***p < .001

Table 12. Effect of memory load: nonlexical syllables.

		Phon	Neigh	Load	L×P	L×N
Aud R^2 =.10*	t	-2.88**	8.59***	0.87	-0.28	0.08
Vis R^2 =.25***	t	16.20***	-5.01***	0.86	-0.03	0.06

*p < .05, **p < .01, ***p < .001

The first thing to note is that the additional task, along with the factoring out of interactions with memory load, affected the patterns for phonotactic probability and neighborhood density. For lexical syllables, the neighborhood effect is now negative for spoken items and the phonotactic effect for written items has disappeared; for nonlexical syllables, both effects are significant for both modalities, but go in opposite directions. We won’t attempt to interpret these patterns here.

Given the goals of our experiment, the more relevant findings relate to the interactions with memory load. Only one significant interaction was found, with written lexical syllables, and as expected, the interaction was negative. Contrary to expectations, however, the interaction involved phonotactic probability, not neighborhood density. This was matched by consistently negative trends for the three other tests of the memory load × phonotactic probability interaction. This pattern implies that loading up working memory with distracter syllables inhibited processing of phonotactic probability, resulting in a weakening of its ability to affect judgments, but had no detectable effect on the influence of neighborhood density on judgments.

What can be made of this? One interpretation is that our memory load manipulation actually measures relative processing difficulty across the two factors, rather than activation of competing lexical items. This could be the case if phonotactic probability must be computed online while neighborhood density is an emergent property of the lexicon, latent in permanent memory. Independent support for this interpretation is the success of the Bailey and Hahn (2001) measure of neighborhood density, which assumes that all lexical items are neighbors to a target to varying degrees; thus neighborhood effects cannot literally involve copying neighbors into working memory. If this interpretation is on the right track, our findings support a view in which “grammar” is more closely associated with phonotactic probability than neighborhood density, since only the former involves online computation, as a grammar might be expected to do. Perhaps the most solid implication of this experiment, however, is that it provides further confirmation that phonotactic probability and neighborhood density are measuring different things, since they show different properties with respect to memory load.

3.6 The search for grammar

We started this paper by discussing how to detect a role for “true grammar” in the judgment making processing, but the “nuisance variables” have proven so complex that we haven't said anything yet about grammar at all (aside from the speculation at the end of the previous section). Here we report an attempt to address the grammar issue via regression analyses of judgments from the first experiment.

To operationalize the notion of "grammar," we coded each of our 384 syllables as 1 if we, as phonologists, judged it to be grammatical, and -1 if not. Thus we adopted the standard generative assumption that grammaticality is a categorical property, even if judgments can vary gradiently. All lexical syllables were assumed to be grammatical (each of our experimental syllables represents more than one morpheme, so none can be a lexical exception). We also assumed that the grammar allows tones to combine freely with any onset. Despite the simplicity of our syllable set, some phonological constraints remained relevant; all related to labial onsets. Specifically, we assumed that it is ungrammatical for labial onsets to combine with /ə/ in an open syllable (e.g., */pə/) or (according to the grammar of Taiwan Mandarin) when followed by /ŋ/ (e.g., */fəŋ/), or for labial onsets to be followed by /u/ in closed syllables (e.g., */mun/). In addition, we assumed that the sequence */fi/ is ungrammatical. This made 56 of our syllables ungrammatical.

There are two things to note about our definition of "grammar." First, though the codings are distinct from the scores given by phonotactic probability and neighborhood density, they also overlap with them somewhat. For example, since the phoneme sequence /fi/ never occurs in the lexicon, items containing this sequence have a phonotactic probability score of 0 and a relatively low neighborhood density score (depending on how well the item matches lexical forms in other ways). Regression analyses nevertheless make it possible to separate out the effects of "grammar" itself. Second, our definition of grammar is based solely on patterns observed in the Mandarin lexicon, with no consideration given to "naturalness." Therefore, while the constraints disallowing */pə/, */fəŋ/, and */mun/ would probably be considered natural (the first two due to autosegmental feature spreading, the third resulting from the OCP), the constraint against */fi/ is harder to justify in terms of naturalness. If grammars express only "natural" patterns, we might want to consider */fi/ to be an accidental rather than systematic gap.

We began our analyses with the same simple model summarized earlier in Table 1, which showed that 72% of the variance in judgments could be explained in terms of lexicality, frequency, phonotactic probability, and neighborhood density, all of which had significant positive effects. Then we added the grammar as defined above, producing a new model that captured 73% of the variance, a significantly greater amount ($F(1, 378) = 12.88$, $p < .001$). All of the four original factors still had significant positive effects, but so did the grammar ($t(39) = 6.96$, $p < .001$). This means that grammar, as we defined it, contributed something above and beyond what was encoded in phonotactic probability and neighborhood density. Before concluding that we have proven that grammar really plays a role in judgments, however, we must remember that our definitions for phonotactic probability and neighborhood density were very coarse-grained, referring only to tone and segment-sized units, with no reference to the natural classes involved in our posited grammatical constraints.

One way to see if our "grammar" really deserves the name is to test if it operates entirely independently of the other "performance" factors. Because a fundamental assumption behind the competence/performance distinction is that competence is modular, if we find that our "grammar" interacts with other factors, it doesn't really behave like grammar. Indeed, when we created a third regression model containing the four original factors, plus both “grammar” and a grammar × phonotactic probability interaction factor, we found that all six factors were significant. All had positive effects except for the interaction, indicating that grammar and phonotactic probability complemented each other. However, this model still only accounted for 73% of the judgment variance, no more than the first model containing grammar. By contrast, a model containing the four original factors, plus “grammar” and a grammar × neighborhood density interaction factor, accounted for 75% of the judgment variance, significantly more than the first grammar model, with all factors remaining significant. Now, however, not only was the interaction factor negative, but the effects of grammar and neighborhood density themselves became negative. A final model, containing all seven factors (including both interaction factors) accounted for 76% of the variance, significantly more than the two previous models; all factors within it were significant, and again grammar and neighborhood density had negative effects, but now the interaction with phonotactic probability was negative while the other interaction was positive.

Putting these results together implies that grammar does indeed interact with both phonotactic probability and neighborhood density, but in different ways. Note that the interaction with phonotactic probability is consistently negative (competitive) and that adding this interaction to the original grammar model neither changed the direction of the individual effects of the grammar and phonotactics, nor explained any more of the judgment variance. These observations imply that grammar and phonotactics capture different aspects of the same thing. The finding that grammar interacts with neighborhood density, but only in a complex way, also makes sense if grammar is partly lexicalized like phonotactic probability, which also interacts with neighborhood density.

In short, these analyses suggest that what phonologists consider “true grammar,” at least with regard to phonotactic constraints, is really just an aspect of phonotactic probability, which is partly lexicalized. On the one hand, this fits reasonably well with the assumption that grammar is external to the lexicon, assuming that phonotactic probability is “more prelexical” than neighborhood density. On the other hand, phonotactic probability is a quantitative measure that can be extracted mechanically from the lexicon without the use of grammar as traditionally understood. Its definition, as we used it, does rely on phonological units (tones and phonemes in our case, as well as features in the case of Bailey & Hahn, 2001), but not generative rules or constraints, and we are currently testing the hypothesis that even these units can be dispensed with, by defining it in terms of psychoacoustic representations, as mentioned in the introduction.

We are not arguing that we have proven that grammar doesn’t really exist, but our results do suggest that researchers should consider looking for it in new places. The question shouldn’t be: “What grammar best describes this lexicon, and/or this set of native-speaker judgments?” Rather, it should be: “What processes do actual human brains use to extract patterns from lexicons?” It seems unlikely that brains actually carry out the arcane algorithms we used to compute phonotactic probability, but instead do something else that happens to have a similar effect. If we understand what that something is, we’ll understand the true nature of phonological competence.

4. Conclusions

In this paper we have reviewed some of our recent findings relating to syllable judgments in Mandarin. First, they are strongly affected by a variety of factors that would never be considered parts of grammar (e.g., lexical status, frequency, and modality) and others that may be, but which also have lexical components (phonotactic probability and neighborhood density). Second, judgments nevertheless seem to be a reliable source of information about phonology, correlating consistently with data from other phonological tasks, whereas data from these tasks don’t always correlate coherently with each other. Third, lexical status not only interacts with the other factors but is perhaps the essential element in the judgment-making process, at least for Mandarin-speaking judges, as indicated by the consistent positive effects of lexicality across a variety of tasks and the finding that judgment scores for lexical syllables increase the more time judges are given to think about them. Fourth, Mandarin speakers differ from English speakers in being more sensitive to modality, indicating a possible difference in the role of discrete, subsyllabic phonological units across the two languages. Finally, phonotactic probability behaves a great deal like phonological grammar: it is processed early (according to previous literature), is less sensitive to lexical factors (though not fully “prelexical”), may involve active use of working memory space (as would be expected for online computation), and overlaps in data coverage with our operationally defined “grammar.”

We are well aware of the limitations of our study. The number of parameters that can be varied (e.g., language, material set, frequency counts, phonological transcription systems, formal definitions of phonotactic probability and neighborhood density, factor choice for regressions and other statistical choices) is so vast that we face a factorial explosion of possibilities, only a tiny proportion of which we have yet examined. There are also other factors that we hope to begin exploring soon, including markedness, perhaps the central concept in current phonological theorizing.

In addition to our future theoretical goals, we also have a more practical item on our “to-do” list: develop tools that make it easier for generative phonologists to carry out experimental research of this kind without having to set up a lab, study computer programming and statistics, and pay hundreds of participants. After all, as Blevins (2004, p. 258) observes about experimentally collected phonological judgments, “it is likely that it is only through continued work of this kind that the true nature of phonological knowledge will be understood.”

References

Agresti, A. (1989). Tutorial on modeling ordered categorical response data. *Psychology Bulletin, 105* (2), 290-301.

Baayen, R. H. (2004). Statistics in psycholinguistics: A critique of some current gold standards. In G. Libben & K. Nault (eds.) *Mental Lexicon Working Papers, Vol. 1*, pp. 1-45. www.mpi.nl/world/persons/private/baayen/submitted/statistics.pdf

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory & Language, 44*, 569-591.

Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language, 72* (1), 32-68.

Blevins, J. (2004). *Evolutionary phonology: The emergence of sound patterns*. Cambridge, UK: Cambridge University Press.

Boersma, P., & Hayes, B. (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry, 32*, 45-86.

Bybee, J. (2001). *Phonology and language use*. Cambridge University Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Coleman, J., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. In *Computational Phonology: Third Meeting of the ACL Special*

Interest Group in Computational Phonology, pp. 49-56. Association for Computational Linguistics, Somerset.

Edelman, S., & Christiansen, M. H. (2003). How seriously should we take Minimalist syntax? *Trends in Cognitive Science, 7* (2), 60-61.

Embick, D., Hackl, M., Schaeffer, J., Kelepir, M., & Marantz, A. (2001). A magnetoencephalographic component whose latency reflects lexical frequency. *Cognitive Brain Research, 10* (3), 345-348.

Halle, M. (1978). Knowledge unlearned and untaught: What speakers know about the sounds of their language." In M. Halle, J. Bresnan, and G. A. Miller (Eds.), *Linguistic theory and psychological reality*, pp. 194-303. MIT Press.

Hammond, M. (1999). Lexical frequency and rhythm. In M. Darnell, E. Moravcsik, F. Newmeyer, M. Noonan, & K. Wheatley (eds.), *Functionalism and formalism in linguistics, vol. 1: General papers*, pp. 329-358. Amsterdam: John Benjamins.

Hammond, M. (2004). Gradience, phonotactics, and the lexicon in English phonology. *International Journal of English Studies, 4*, 1-24.

Ho, K.-C. (1976). A study of the relative frequency distribution of syllabic components in Mandarin Chinese. *Journal of the Institute of Chinese Studies, 8*, 275-352.

Inkelas, S., Orgun, C. O., & Zoll, Cheryl. (1997). The implications of lexical exceptions for the nature of grammar. In I. Roca (Ed.) *Derivations and constraints in phonology*, pp. 393-418. Oxford: Clarendon Press.

Keller, F., & Asudeh, A. (2002). Probabilistic learning algorithms and Optimality Theory. *Linguistic Inquiry, 33*, 225-244.

Kenstowicz, M. & Kisseberth, C. (1979) *Generative Phonology: Description and Theory.* Academic Press.

Kirchner, R. (2004, June). Exemplar-based phonology and the time problem: A new representational technique. Talk presented at Laboratory Phonology 9, University of Illinois at Urbana-Champaign, USA. http://roa.rutgers.edu/view.php3?id=930

Li, H., Li T.-K., & Tseng J.-F. (1997). *Guoyu cidian jianbianben bianji ziliao zicipin tongji baogao*. [Statistical report on Mandarin dictionary-based character and word frequency] Ministry of Education, Republic of China. http://140.111.1.22/clc/dict/htm/pin/start.htm

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16* (1), 149-157.

Luce, P. A., & Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language & Cognitive Processes, 16* (5/6), 565-581.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing, 19* (1), 1-36.

Myers, J. (2002). An analogical approach to the Mandarin syllabary. *Journal of Chinese Phonology, 11*, 163-190.

Myers, J. (2004, May). Modeling phonological acceptability judgments in Mandarin. Invited colloquium talk, National Tsinghua University.

Myers, J., & Tsay, J. (2004, June). *Exploring performance-based predictors of phonological judgments in Mandarin.* Poster presented at Laboratory Phonology 9, University of Illinois at Urbana-Champaign, USA.

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

Ohala, J. J., & Ohala, M. (1986). Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. In J. J. Ohala and J. J. Jaeger

(Eds.) *Experimental phonology*, pp. 239-252. Academic Press: New York.

Phillips, C., & Lasnik, H. (2003). Linguistics and empirical evidence: Reply to Edelman and Christiansen. *Trends in Cognitive Science, 7* (2), 61-62.

Pulvermüller, F., Kujala, T., Shtyrov, Y., Simola, J., Tiitinen, H., Alku, P., Alho, K., Martinkauppi, S., Ilmoniemi, R. J., & Näätänen, R. (2001). Memory traces for words as revealed by the mismatch negativity. *NeuroImage, 14*, 607-616.

Pylkkänen, L., & Marantz, A. (2003). Tracking the time course of word recognition with MEG. *Trends in Cognitive Science, 7* (5), 187-189.

Pylkkänen, L., Stringfellow, A., & Marantz, A. (2002). Neuromagnetic evidence for the timing of lexical activation: An MEG component sensitive to phonotactic probability but not to neighborhood density. *Brain & Language, 81*, 666-678.

Rasch, D., & Guiard, V. (2004). The robustness of parametric statistical methods. *Psychology Science, 46* (2), 175-208.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime reference guide*. Pittsburgh: Psychology Software Tools Inc.

Stockall, L., Stringfellow, A., & Marantz, A. (2004). The precise time course of lexical activation: MEG measurements of the effects of frequency, probability, and density in lexical decision. *Brain and Language, 90*, 88-94.

Tsai, C.-H. (2000). Mandarin syllable frequency counts for Chinese characters. Available online at http://technology.chtsai.org/syllable/

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in spoken word recognition. *Psychological Science, 9*, 325-329.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory & Language, 40*, 374-408.

Wang, S. H. (1998). An experimental study on the phonotactic constraints of Mandarin Chinese. In B. K. T’sou (Ed.), *Studia Linguistica Serica* (pp. 259-268). Language Information Sciences Research Center, City University of Hong Kong.