Bridging the Gap:
MiniCorp Analyses of Mandarin Phonotactics

James Myers
National Chung Cheng University

March 10, 2008
[To appear, repaginated, in the WECOL 2007 proceedings]

1 Introduction

Despite skepticism about corpus data going back to the beginnings of generative linguistics (Chomsky 1957), most phonological research is actually a form of informal corpus linguistics. That is, unlike syntacticians, phonologists do not rely primarily on elicited native-speaker judgments of novel forms, but rather on collections of preexisting lexical items (e.g. dictionaries).

Corpus data are limited in what they can say about phonological knowledge, as has often been pointed out (e.g. Ohala 1986), and there has been growing interest among phonologists in testing hypotheses with native-speaker judgments (e.g. Coetzee to appear), phonetic measurements (e.g. Morén and Zsiga to appear), and other types of experimentally collected data (e.g. Moreton to appear). Nevertheless, the continued use of dictionary data in phonology is justifiable. The most important reason is that like acceptability judgments, a lexicon represents the output of processes that arguably include grammatical knowledge as a component (Blevins 2004 presents a contrary view, but see Kiparsky 2006, Zuraw 2007, Moreton to appear for responses). Moreover, dictionary analyses have provided key evidence for the most empirically robust concepts in phonological theory, from phonemes to constraints and beyond.

Though the corpus analyses in theoretical phonology typically do not use sophisticated quantitative methods (cf. Frisch et al. 2004, Uffmann 2006), they do rely on the implicitly quantitative assumption that type frequency is informative about grammatical status. Not only are exceptions dismissed if their type frequency is sufficiently low, but the distinction between systematic and accidental gaps depends on whether the gaps are rarer than would be expected by chance. Note that the logic here runs from the grammar to the corpus, not the other way around; type frequency does not directly indicate grammatical status any more than acceptability is identical to grammaticality in syntax. Instead, in phonological argumentation, type frequency is cited to support or challenge a grammar that has been motivated at least partly by a priori considerations.

This paper introduces a software tool, MiniCorp (Myers 2008a), that attempts to bridge the gap between this traditional logic (as applied in the Optimality-Theoretic framework) and truly quantitative corpus analysis. Virtually unique among OT software, MiniCorp is not an automatic grammar learner, but rather it follows the traditional logic in testing a proposed OT grammar against dictionary data. Specifically, MiniCorp tests whether the proposed constraints are obeyed more reliably than chance and whether the relative strengths of competing constraints are sufficiently different to support the proposed constraint ranking. Not only is MiniCorp the only program that tests OT grammars for statistical significance, but it also includes special tools to simplify the annotation of corpus items. It is also both freely available (www.ccunix.ccu.edu.tw/~lngproc/MiniCorp.htm) and open-source (the current version is written in JavaScript, with statistics handled by R, the free, open-source statistics program: R Development Core Team 2008).

The remainder of this paper describes the application of MiniCorp to the analysis of a phonotactic pattern in Mandarin. The grammatical proposal is introduced in section 2. Section 3 gives a step-by-step overview of how MiniCorp was used to test it, from corpus annotation to the output of the statistical analyses. Section 4 explains the algorithm behind MiniCorp's output report. Section 5 sums up and looks to the future.

2 Tone and voicing in Mandarin

The four lexical tones in Mandarin are often illustrated with the set of words shown in (1).

(1)	Tone 1 (high):	ma^{55}	"mother"
	Tone 2 (rising):	ma^{35}	"hemp"
	Tone 3 (low):	ma^{214}	"horse"
	Tone 4 (falling):	ma^{51}	"scold"

This set is misleading, however, since it is not typical for high tone to appear in syllables with voiced onsets like /m/. This is demonstrated in (2) below, which shows the number of morphemes with different combinations of onset and tone (Mandarin morphemes are almost always monosyllabic). Note that morpheme counts are relatively low when high Tone 1 appears with a voiced onset.

(2) Morpheme type counts in Mandarin (data from Li et al. 1997 and Tsai 2000)

	Onset	High	Rising	Low	Falling	Toneless
[-voice]	p	167	73	105	243	3
	pʰ	105	182	45	89	0
	f	108	146	62	102	0
	t	154	114	93	267	2
	tʰ	103	290	87	117	1
	k	223	44	144	129	2
	kʰ	122	26	82	136	0
	x	116	283	56	255	1
	ts	122	43	72	79	1
	tsʰ	68	69	33	92	0
	s	126	4	49	132	0
	tɕ	339	240	176	328	0
	tɕʰ	197	272	63	123	0
	ɕ	329	193	87	259	0
	tʃ	268	129	124	258	0
	tʃʰ	142	241	89	130	0
	ʂ	180	57	72	206	0
[+voice]	m	13	210	113	171	4
	n	7	101	78	99	0
	l	23	440	161	329	1
	ʐ	1	103	47	60	0
	Onsetles	384	561	388	644	0

Tone-voicing cooccurrence restrictions are not typologically unusual, as shown by tone split (Hombert et al. 1979) and depressor consonants in Bantu languages (Laughren 1984). The apparent depressors in Mandarin are somewhat unusual in being sonorants (including the voiced retroflex-like fricative; see Wang 1993:113), but depressor sonorants are also found in other language families (Bradshaw 1999). Note also that the high level tone (i.e. H) is affected (becoming LH) while the high-initial falling tone (HL) is not, consistent with the well-known tendency of Asian contour tones to act as if unitary (Yip 1995).

We thus have reason to hypothesize that the Mandarin pattern is consistent with the universal markedness constraint in (3). According to the view expressed earlier, whereby a lexicon is only partially predicted by a grammar, we may interpret violations of this constraint in Mandarin as simply ungrammatical, since speakers can memorize them using extra-grammatical components of the speech processing system.

(3) *Voice/H

Setting the exceptions aside enables us to propose a simple grammatical analysis. *Voice/H potentially competes with faithfulness constraints protecting either voicing or tone. Since Mandarin does not use voicing phonologically (in lexical contrasts or in alternations), but does contrast and manipulate tone (in tone sandhi), we have some justification for the ranking shown in (4). Note that this ranking further permits us to assume that in voiced-initial morphemes that surface with the rising Tone 2, the underlying tone may be high level Tone 1.

(4) Ident(Voice) >> *Voice/H >> Ident(H)

One may criticize some of the a priori assumptions motivating this grammar, in particular the synchronic derivation of rising tones from high tones, given that Lexicon Optimization (Prince and Smolensky 2004) should have made them underlyingly rising. However, such criticisms rely on a priori assumptions themselves. A more objective test of the proposed grammar would be to see how well it describes the data set.

Here quantification becomes crucial. *Voice/H is proposed to account for the rarity of voiced-initial high-toned morphemes, but there are still 44 of them. Is that really rare enough to ignore? Ident(H) is claimed to be ranked low, but the more lowly ranked a constraint, the fewer items will obey it in a corpus. Doesn’t this weaken any language-internal evidence for it? Finally, (4) claims not only that Ident(Voice) is undominated, but that it can override the effects of both of the lower-ranked constraints put together. Does Ident(Voice) really provide such an overwhelmingly robust a description of the corpus?

In short, the question here is whether the proposed grammar describes the corpus better than chance. This is consistent with the implicit logic of traditional phonological argumentation, where the degree of empirical coverage (rarity of exceptions and accidental gaps) is a crucial factor in convincing skeptics of the validity of an analysis. MiniCorp automates the steps needed to make this kind of argument statistically sound, even in corpora too large to analyze by hand.

3 Using MiniCorp

A MiniCorp session starts with an electronic dictionary and ends with a statistical analysis testing the reliability of each proposed OT constraint and their proposed ranking. Currently the only version of MiniCorp is MiniCorpJS, written in JavaScript and run in the user’s web browser (it’s been tested in Firefox, Internet Explorer, Safari, and Opera).

In the present case study, the corpus was a file listing the 13,607 Mandarin monosyllabic morphemes described in (2), transcribed in IPA, except that the four lexical tones were transcribed 1-4 as in (1). The choice of transcription system is up to the researcher, as is the choice of corpus. Such choices may affect the analysis, as well they should, since a transcription represents a hypothesis about phonological representation, and different corpora represent different levels of the grammatical system (e.g. a morpheme inventory, as is analyzed here, as opposed to a syllable inventory, conflating all homophones).

The next step is to tag (i.e. annotate) the corpus for the theoretically relevant aspects. All we care about these Mandarin morphemes is which of the proposed OT constraints are violated by them. As first pointed out by Golston (1996), tagging a word for its constraint violations serves as a sort of representational system. For example, marking the morpheme [ma1] "mother" as violating *Voice/H is equivalent to saying that it is a voiced-initial high-toned syllable.

This link between constraint violations and representations is convenient because it offers a way to tag corpus items automatically (tagging 13,607 items by hand would not only be time-consuming, but error-prone as well). In order to tag all violations of some constraint, we merely have to encode this constraint in terms of the class of character strings that violate it. Fortunately, as Karttunen (1998) realized, there is already a well-established mathematical tool for transcribing classes of character strings, called regular expressions. The most familiar element of regular expression notation is the "wildcard" symbol offered by many search systems, but it goes far beyond this, with symbols marking the starts and ends of strings, repetition, and set union, among other things.

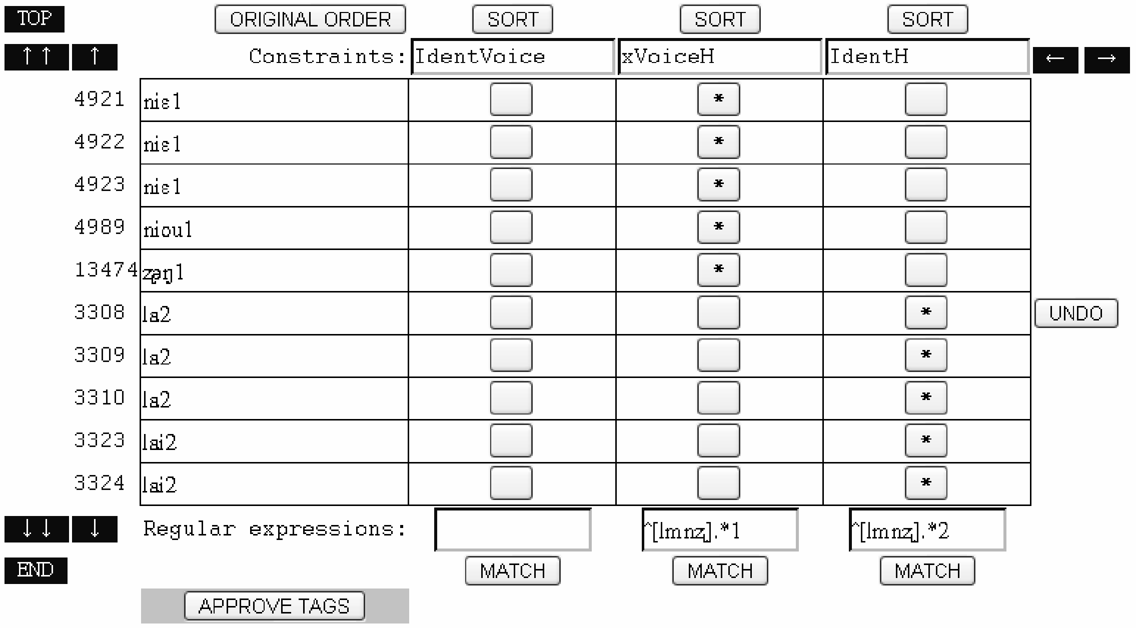
For example, (5) gives regular expressions that encode violations of the three constraints in (4). The faithfulness constraint Ident(Voice) is, by hypothesis, never violated, so it requires no encoding. The markedness constraint *Voice/H is violated by syllables with any of the four voiced onsets and the high tone, transcribed as 1 in the corpus (for Tone 1). The set union of the different onsets is indicated by placing them inside square brackets, and restriction to onset position is indicated by the caret (/n/ also appears in coda position, where it doesn't interact with tone). The dot is a wildcard symbol, and the star after it indicates repetition. Thus the expression in (5b) picks out items containing both voiced onsets and Tone 1. Finally, the faithfulness constraint Ident(H) is encoded with a regular expression indicating all voiced-onset syllables with rising tone (Tone 2), based on the simplifying assumption that all such syllables underlyingly have high Tone 1. This assumption helps because faithfulness violations involve representations that are not available in the corpus (e.g. inputs) (Golston 1996 actually rejects the very notion of faithfulness).

(5) a. Ident(Voice): {not applicable: no violations}
b. *Voice/H: ˆ[lmnz].*1
c. Ident(H): ˆ[lmnz].*2

After entering the corpus and automatically tagging the items, the MiniCorp user is able to scroll around and sort the tagged corpus, making it easier to find any mistagged items. This is done via the tabular display in (6). Constraint names are modified to serve as legal variable names for the statistical analysis.

The user then defines the grammar in terms of a ranking of the constraints, and MiniCorp generates analysis code to run in R, the free statistics program rapidly becoming a standard tool in quantitative linguistics (Baayen 2008, Johnson 2008). The R code runs two types of tests, one for the contribution of each constraint to the description of the corpus data, and one for the ranking. The two sets of results for the present analysis are shown in (7) and (8) below.

(6) MiniCorp tagging table



TOP   ORIGINAL ORDER   SORT   SORT   SORT

↑↑ ↑

	Constraints:	IdentVoice	xVoiceH	IdentH	← →
4921	niɛ1	□	*	□	
4922	niɛ1	□	*	□	
4923	niɛ1	□	*	□	
4989	niou1	□	*	□	
13474	zəŋ1	□	*	□	
3308	la2	□	□	*	UNDO
3309	la2	□	□	*	
3310	la2	□	□	*	
3323	lai2	□	□	*	
3324	lai2	□	□	*	
↓↓ ↓	Regular expressions:		^[lmnz].*1	^[lmnz].*2	
END		MATCH	MATCH	MATCH	

APPROVE TAGS

(7) Constraint test:

Constraints	Weights	p	
IdentVoice	-8.1321	0	*
xVoiceH	-5.6651	0	*
IdentH	-2.7002	0	*

(* significant constraint)

(8) Ranking test:

Constraints	p	
IdentVoice	0.6654	
xVoiceH	0	*

(* significant ranking)

The significant results (*p* < .05) in (7) show that each of the constraints does better than chance, independently of the others, at describing the data. The constraint weights are also all negative, as they should be if the constraints are obeyed more often than violated (as described in the next section, the statistical analysis is attempting to predict type frequencies from constraint violations).

The ranking tests examine the partial rankings in (9), implied by the grammar in (4), where the ranking in (9a) indicates that the topmost constraint strictly outranks all of the others. In general, MiniCorp encodes the ranking hierarchy of a grammar with *n* constraints in terms of the *n*-1 non-terminal constraints. According to the report in (8), then, the ranking in (9b) describes the corpus data better than chance, but the ranking in (9a) fails to reach statistical significance.

(9) a. Ident(Voice) >> {*Voice/H, Ident(H)}
b. *Voice/H >> Ident(H)

Putting these results together, the constraints do seem to describe genuine patterns in the observed data, and part of the constraint ranking is supported as well, namely *Voice/H >> Ident(H). However, Ident(Voice) does not provide such a robust description of the data to rank it confidently above both of the other constraints. This calls into question the assumption that potential *Voice/H violations are avoided at the expense of tone rather than voicing.

As this example shows, MiniCorp formalizes and automates aspects of traditional phonological argumentation, even with large data sets.

4 How MiniCorp works

The validity of the above conclusions depends on the validity of the algorithm used to generate them. As it happens, this algorithm not only builds on well-established statistical techniques, but is also reasonably easy to understand, even without much statistical background.

The first insight exploited by the algorithm is that an OT grammar is a species of harmonic grammar (HG), in which constraints are weighted rather than strictly ranked (Prince and Smolensky 2004). Strict ranking emerges when weights are chosen such that the weight of each constraint is greater than the sum of the weights of all lower-ranked constraints (Prince 2007).

The second insight is that constraint weights of the HG type can be set automatically from corpus data through a technique called loglinear modeling (Goldwater and Johnson 2003, Hayes and Wilson to appear; Pater et al. 2007 use a related approach). In the case of HG, the loglinear model is an equation relating constraint violations to type frequencies, where the weights are equation coefficients. All else being equal, the larger the magnitude of a weight, the better the associated constraint is at predicting the type frequencies.

As the above-cited works show, the relationship between HG and loglinear modeling makes it possible to create an automatic HG learner. However, the purpose of MiniCorp is not to learn an HG grammar, but to test an OT grammar. Thus MiniCorp uses loglinear modeling to compute the chance probabilities (*p*) that constraint weights differ from zero (i.e. help describe the data) and that constraint weights differ from each other (i.e. are ranked).

More precisely, MiniCorp converts the information in a tagging table like (6) into a type frequency table like (10), where each category is defined by a different combination of constraint violations. It then runs a standard sort of loglinear model called Poisson regression (Agresti 2002) to model the fit between the type frequencies and the constraints attempting to predict them.[1] This is how the weights and *p* values in (7) were computed.

(10) Type frequency table

Count	Ident(Voice)	*Voice/H	Ident(H)
12709	0	0	0
854	0	0	1
44	0	1	0
0	0	1	1
0	1	0	0
0	1	0	1
0	1	1	0
0	1	1	1

To see how the *p* values in (8) were computed, note first that the weights in (7) are partially consistent with the claimed ranking, with the weight magnitude of *Voice/H (-5.67) greater than that of Ident(H) (-2.70). However, the weight magnitude of Ident(Voice) (-8.13) is not greater than the sum of the other two (-8.37), conflicting with the claim that it outranks both of them put together. To test hypothesized rankings for statistical significance, MiniCorp uses likelihood ratio tests to compare the data fit of the model in (11a), where the constraint weights for *Voice/H and Ident(H) are identical, with the model in (11b), where they need not be. Similarly, the ranking of Ident(Voice) over both of the other constraints is tested by comparing the models in (12).[2]

(11) a. Counts ~ w_1xVoiceH + w_1IdentH
b. Counts ~ w_1xVoiceH + w_2IdentH

(12) a. Counts ~ (w_1 + w_2)IdentH + w_1xVoiceH + w_2IdentH
b. Counts ~ w_1IdentH + w_2xVoiceH + w_3IdentH

All of these statistical techniques are well-established. The unique contribution of MiniCorp is to automate them in a user-friendly package designed for OT grammars assuming strict constraint ranking.

5 Conclusions

Phonological argumentation traditionally relies on comparing type frequencies in dictionary corpora. MiniCorp expands on and automates this idea so that it can be applied to large and complex data sets. In the case of the hypothesized grammar tested in this paper, MiniCorp was able to confirm some aspects (e.g. the role of *Voice/H) while calling other aspects into question (e.g. the undominated ranking of Ident(Voice).

MiniCorp isn’t restricted to phonotactics; if a grammatical proposal can be expressed as a fixed ranking of constraints, MiniCorp can test it. It does have some limitations, however. One that should be overcome soon is that it assumes that each item violates each constraint at most once; extending the algorithm to allow any number of constraint violations merely requires a bit more algebra. Other planned extensions include techniques for testing variable grammars (e.g. Boersma and Hayes 2001) and grammars incorporating derivational ordering (e.g. Kiparsky 2000). One extension that has already been implemented is a tool for computing neighborhood density (i.e. number of similar lexical items of a target), known to influence acceptability judgments (Bailey and Hahn 2001). These values can then be used in the analysis of judgments collected with the help of MiniCorp’s sister program MiniJudge, a tool for designing, running, and analyzing acceptability judgment experiments (Myers 2007, 2008b).

MiniCorp is intended to help bridge the gap between traditional phonological argumentation and truly quantitative corpus analysis. While already useful and reasonably user-friendly, MiniCorp is always in need of further improvement. Collaborators and competitors are both most welcome!

Notes

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. MiniCorp is co-copyrighted by National Chung Cheng University.

[1] Note that the top-ranked constraint is almost perfectly correlated with the output; that is, it is (by definition) never violated, so input 1 values are always associated with output counts of

zero. Since (near) perfect correlations cause the weight estimation algorithm used in Poisson to crash (Agresti 2002), MiniCorp replaces each 0 count with 1 before running the analyses.

[2] Model equations can only be compared like this if one is contained within the other (e.g. $y \sim x$ vs. $y \sim x + z$), which algebraic manipulation shows to be true of the equations in (11ab) and (12ab).

References

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

Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics*. 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.

Blevins, Juliette. 2004. *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge, UK: Cambridge University Press.

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

Bradshaw, Mary. 1999. *A Crosslinguistic Study of Consonant-Tone Interaction*. Doctoral dissertation, Ohio State University, Columbus, Ohio.

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

Coetzee, Andries W. To appear. "Grammaticality and ungrammaticality in phonology", *Language*, **84**.

Frisch, Stefan. A., Janet B. Pierrehumbert and Michael B. Broe. 2004. "Similarity avoidance and the OCP", *Natural Language and Linguistic Theory*, **22**.1:179-228.

Goldwater, Sharon, and Mark Johnson. 2003. "Learning OT constraint rankings using a maximum entropy model", *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, ed. by J. Spenader et al. (pp. 111-120). Stockholm Univ.

Golston, Chris. 1996. "Direct Optimality Theory: Representation as pure markedness", *Language*, **72**.4:713-748.

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

Hombert, Jean-Marie, John J. Ohala, and William G. Ewan. 1979. "Phonetic explanations for the development of tones", *Language*, **55**.1:37-58.

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

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

Kiparsky, Paul. 2000. "Opacity and cyclicity", *The Linguistic Review*, **17**:351-67.

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

Laughren, Mary. 1984. "Tone in Zulu nouns", *Autosegmental Studies in Bantu Tone*, ed. by George N. Clements & John Goldsmith (pp. 183-230), Dordrecht: Foris.

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://www.edu.tw/EDU_WEB/EDU_MGT/MANDR/EDU6300001/allbook/pin/f11.html?open

Morén, Bruce and Elizabeth Zsiga. To appear. "The lexical and post-lexical phonology of Thai tones", *Natural Language and Linguistic Theory*.

Moreton, Elliott. To appear. "Analytic bias and phonological typology", *Phonology*.

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

12.2:175-194.

Myers, James. 2008a. MiniCorpJS (Version 0.5) [Computer software]. Accessible via http://www.ccunix.ccu.edu.tw/~lngproc/MiniCorp.htm

Myers, James. 2008b. MiniJudgeJS (Version 1.1) [Computer software]. Accessible via http://www.ccunix.ccu.edu.tw/~lngproc/MiniJudge.htm

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

Pater, Joe, Christopher Potts, and Rajesh Bhatt. 2007. "Harmonic grammar with linear programming", University of Massachusetts at Amherst ms. ROA 872-1006.

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

Prince, Alan. 2007. "Let the decimal system do it for you: a very simple utility function for OT", Rutgers University ms.

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.

Tsai, Chih-Hao. 2000. Mandarin syllable frequency counts for Chinese characters. http://technology.chtsai.org/syllable/

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

Wang, Jenny Zhijie. 1993. *The geometry of segmental features in Beijing Mandarin*. University of Delaware PhD thesis.

Yip, Moira. 1995. "Tone in East Asian languages", *The Handbook of Phonological Theory*, ed. by John A. Goldsmith (pp. 476-494). Oxford, UK: Blackwell.

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

James Myers
Graduate Institute of Linguistics
National Chung Cheng University
Minhsiung, Chiayi Taiwan 62102
Lngmyers@ccu.edu.tw