Isolating grammar in phonological data*

James Myers
National Chung Cheng University
Lngmyers@ccu.edu.tw http://www.ccunix.ccu.edu.tw/~lngmyers/

Overview

- **Isolating competence in performance**
- **Modeling phonological data**
- **Modeling phonological grammar**

Isolating competence in performance

(1) The goals of generative linguistics (including phonology):

- Analyze data on individual languages to determine their grammars
- Analyze grammars cross-linguistically to figure out Universal Grammar (UG)

  • NOTE: These goals focus on *grammars*, not languages, lexicons, or phonetic/psychological motivation:

  "We are advocating that phonologists, qua phonologists, attempt to explain less, but in a deeper way." (Hale & Reiss, 2000, p. 167).

(2) Diagnostics of phonological competence:

Competence (grammar)	Performance (not grammar)
Stable knowledge	Context-dependent processing
Innate language-specific devices (e.g. rules, constraints, principles)	General-purpose devices (e.g. memory, analogy, physics)
Synchronically active	May reflect historical relics
Has certain formal properties…?	Doesn't have those properties…?

(3) Challenge: Linguistic data reflect both competence and performance, e.g.:

Lexical data = Phonological grammar + History + Unknown factors
Judgment data = Phonological grammar + Memory + Analogy + Unknown factors
Cross-linguistic data = UG + History + Phonetics + Unknown factors

(4) Solution: Put all components into detailed competence-performance linking models, using statistics to distinguish systematic effects from apparent randomness (unknown factors) and from each other (grammar vs. extra-grammatical confounds).

- Rather than forcing linguists to use the statistical conventions followed in other disciplines, the statistics should be designed specifically for linguistic hypotheses (about competence) and linguistic data (usually not reaction times, accuracy, etc).
  • Linguistic data are usually categorical: attested vs. not, acceptable vs. not.
- The process should be made as easy as possible by automating it in software and linking it to tools already familiar to linguists (e.g. Optimality Theory tableaux).

* Funded by NSC 95-2411-H-194-005. Thanks to codirectors Jane Tsay and Jhishen Tsay, assistants Tsung-Ying Chen, Chen-Tsung Yang, Yu-Kuang Ko, Ying-Hsing Li, Jih-Ho Cha, Yu-Ying Huang, and the participants.

Modeling phonological data

(5) There are only two types of data in science:

Experiments	**Corpora**
Test effect of controlled input on output	Analyze output for possible signs of input
E.g. physics, chemistry, psychology	E.g. astronomy, archeology, epidemiology
E.g. **syntax**: acceptability judgments (Schütze, 1996; Cowart, 1997)	E.g. **phonology**: attestations (Uffman, 2006; Duanmu, 2004; Shih, n.d.)
Statistical model fixed by the design	No inherently "best" statistical model

(6) Two fundamentals of the generative view of corpus data:

❶ A corpus merely reflects the use of grammar; grammarians aren't in the business of describing corpora (Chomsky, 1957; cf. Richness of the Base in Optimality Theory).

• Corollary: True lexical exceptions are *ungrammatical*: they are accidental *fillers* (parallel to accidental gaps) created in the noisy path from competence to corpus (hence grammatical theories of exceptions, like Pater, to appear, miss the point).

❷ There is one and only one inherently "best" corpus analysis: the one performed by babies applying their innate language acquisition device (Chomsky, 1965).

(7) Identifying accidental fillers (hence testing grammaticality) involves statistical modeling.

Pure chance model: Probability(Patterned form) = Random
Baseline model: Probability(Patterned form) = Baseline (Grammar?) + Random

(8) Pure chance model is like flipping a coin, with equal chance for a form to enter the lexicon as "Patterned" or "Exceptional". E.g. for 4 forms including 1 exception:

$\boxed{\text{PPPP}}$	$\boxed{\text{PPPE}}$	$\boxed{\text{PPEP}}$	PPEE	$\boxed{\text{PEPP}}$	PEPE	PEEP	PEEE	*p* = 0.3125 (5/16)
$\boxed{\text{EPPP}}$	EPPE	EPEP	EPEE	EEPP	EEPE	EEEP	EEEE	with 1 *or fewer* E

There can also be *more* exceptions than chance, so the *p* value is (0.3125)×2 = 0.625. By convention statistical significance is shown by $p < .05$ (= 1/20), so this isn't significant.

(9) The *p* value for this (exact) binomial test can easily be computed for any data set:

Excel: = MIN(1, 2 * BINOMDIST(MIN(*#Patterned*, *#Exceptional*), *Total*, 0.5, TRUE))
R (www.r-project.org): min(1, 2 * pbinom(min(*#Patterned*, *#Exceptional*), *Total*, 0.5))

(10) For example, Li & Tsuchida (2001) cite 45 reduplicated morphemes in the Formosan language Pazih. Most show vowel harmony on the epenthesized vowel, but 12 don't. Do the exceptions prove that vowel harmony is just random? (See Myers 2006a.)

a. [33]	bak-a-bak	"native cloth"	b. [12]	bar-e-bar	"flag"
	hur-u-hur	"steam, vapor"		hur-a-hur	"bald"

Answer: No. The chance probability is *p* = 0.0024589, which is far below .05.

(11) Li and Tsuchida (2001, p. 21) note that among the 12 exceptions "/a/ appears to be the most common" epenthetic vowel. But this makes only 7 "patterned": $p = 0.774414 > .05$.

(12) But is this the right statistical model? Pazih has four vowels (/i/, /u/, /e/, /a/), so the chance probability of *avoiding* /a/ is 1/4 (=0.75). And in fact:

p = min(1, 2 * pbinom(min(7, 5), 12, **0.75**)) = 0.02850556 < .05. Significant!

(13) The Baseline model also ignores the *harmony*: the epenthetic vowel is not merely non-random, but predicted (partly) by the stem vowel:

Prob(EpentheticA) = $w_0 + w_1 \times$**StemA** + Random

- w = "weight", w_0 = Baseline (boring). The statistics then test if $w_0 \neq 0$, $w_1 \neq 0$.
- The baseline and StemA effects should be tested *independently of each other*.

(14) The independence requirement means we need more sophisticated statistics. The most commonly used is <u>logistic regression</u> (羅吉斯迴歸), which is at the heart of the Labovian variable rule analysis program VARBRUL (Mendoza-Denton et al., 2003).

- <u>R command</u>: summary(glm(EpentheticA~StemA, family = binomial, data = pazih))
  where "pazih" has the two variables EpentheticA and StemA (1 if /a/, 0 otherwise)
- <u>Results</u>: Baseline has a negative weight: /a/ is less than half of data (boring).
  StemA has a positive weight: Stem /a/ predicts epenthetic /a/ (interesting).

(15) Another example: Does Mandarin grammar have a cooccurrence constraint *H/VOICED restricting the appearance of Tone 1 (HH) on syllables with voiced-initial onsets?

- Syllable type frequencies (data from Li et al., 1997 and Tsai, 2000):

			Tone 1 (HH)	Tone 2 (LH)	Tone 3 (LL)	Tone 4 (HL)
[–voice]	[–cor]	p^h...	105	182	45	89
		k...	223	44	144	129
		k^h...	122	26	82	136
	[+cor]	t...	154	114	93	267
		t^h...	103	290	87	117
		ʂ...	180	57	72	206
		p...	167	73	105	243
[+voice]	[–cor]	**m...**	**13**	210	113	171
	[+cor]	**n...**	**7**	101	78	99
		l...	**23**	440	161	329
		ʐ...	**1**	103	47	60

(16) As always with corpus data, there are many analyses to choose among, including:

- Prob(**Tone 1 vs. all others**) = $w_0 + w_1 \times$**Voiced** + $w_2 \times$**Cor** + Random

  <u>Voiced</u>: Negative weight: Voicing makes Tone 1 rarer.
  <u>Cor</u>: Negative weight: Coronals also disfavor Tone 1 (probably a boring confound).

- Prob(**Voiced**) = $w_0 + w_1 \times$**Cor** + $w_2 \times$**T1H** + $w_3 \times$**T2H** + $w_4 \times$**T1H×T2H** + Random
  [The <u>interaction</u> T1H×T2H tests if Tone 1 is special, e.g. HH vs. LH]

  <u>Cor</u>: Positive weight: Coronals prefer to be voiced (in this sample: boring).
  <u>T1H</u>: Negative weight: H in first position disfavors voicing.
  <u>T2H</u>: Positive weight: H in second position *favors* voicing (historically, HH > LH).
  <u>T1H×T2H</u>: Negative weight: HH disfavors voiced onsets over LH.

(17) Yet another example: Does Mandarin grammar have constraints OCP[rd] and OCP[bk] blocking two high round (or front) vowels from appearing in the same triphthong?

- Syllable type frequencies (data from Li et al., 1997 and Tsai, 2000):

			...i	...u
i...	Onset	...e/o...	**0**	191
		...a...	**0**	411
	No onset	...e/o...	**0**	78
		...a...	**4**	83
u...	Onset	...e/o...	378	**0**
		...a...	58	**0**
	No onset	...e/o...	137	**0**
		...a...	2	**0**

(18) A weakness of logistic regression: The algorithm crashes if the correlation is too strong, as it is here! (*Exact* logistic regression, an extension of the exact binomial test, doesn't have this problem, but R can't do it yet. SAS can, but it's not free.)

(19) Due to historical confounds, corpus data are limited in what they say about synchronic grammar (e.g., Hsu, 2005). Hence phonologists also need to run experiments, including judgment experiments on fake words (e.g., Ohala, 1986; Frisch & Zawaydeh, 2001).

- With only one pair of items, testing binary judgments ("good"/"bad") from multiple judges, the statistics work the same as in the binomial test: the exact sign test:

  For each judge, count [+Gram] > [–Gram] as "Patterned" and [+Gram] < [–Gram] as "Exceptional". Ignore [+Gram] = [–Gram]. Apply formulas in (9).

(20) Of course, since any single pair of items has its own idiosyncratic properties, more than one pair of items should be tested. This requires a more complex type of statistics.

- Binary judgments on multiple items and speakers can be analyzed with generalized linear mixed effect modeling (GLMM), an extension of logistic regression.
- MiniJudge (Myers, 2006b; http://www.ccunix.ccu.edu.tw/~lngproc/MiniJudge.htm) simplifies the design, running, and analysis (with R) of judgment experiments.

(21) A MiniJudge experiment on *H/Voiced in Mandarin.

- Several sets of (mostly) nonlexical syllables (thanks to my native-speaking assistants):

	[+Voiced]		[–Voiced]	
	[+HH]	[–HH]	[+HH]	[–HH]
Set 1	ㄇㄧㄡ [哂]	ㄇㄧㄡˊ	ㄅㄧㄡ	ㄅㄧㄡˊ
Set 2	ㄋㄣ	ㄋㄣˇ	ㄊㄣ	ㄊㄣˇ
Set 3	ㄋㄟ	ㄋㄟˊ	ㄉㄟ	ㄉㄟˊ
Set 4	ㄌㄣ	ㄌㄣˊ	ㄎㄣ	ㄎㄣˊ
Set 5	ㄌㄨㄞ	ㄌㄨㄞˊ	ㄎㄨㄞ	ㄎㄨㄞˊ
Set 6	ㄖㄡ	ㄖㄡˇ [糅]	ㄅㄡ	ㄅㄡˇ

- 20 naive students gave "yes"/"no" judgments to randomly ordered items in BPMF.
- We predict an interaction between [Voiced] and [HH]: [+V+H] should be disfavored.
- Results: A total failure. Nothing was significant at all!

(22) A MiniJudge experiment on OCP[bk] and OCP[rd] in Mandarin.

- Again, we predict an interaction: [+u1+u3] and [–u1–u3] should both be disfavored.

	[+u1]		[–u1]	
	[+u3]	[–u3]	[+u3]	[–u3]
Set 1	ㄉㄨㄡˊ	ㄉㄨㄟˊ	ㄉㄧㄡˊ	ㄉㄧㄟˊ
Set 2	ㄉㄨㄠˊ	ㄉㄨㄞˊ	ㄉㄧㄠˊ	ㄉㄧㄞˊ
Set 3	ㄋㄨㄡˊ	ㄋㄨㄟˊ	ㄅㄧㄡˊ	ㄅㄧㄟˊ
Set 4	ㄋㄨㄠˊ	ㄋㄨㄞˊ	ㄅㄧㄠˊ	ㄅㄧㄞˊ

- Same 20 naive students (counterbalanced order with other experiment).
- Results: [u1]×[u2] interaction: Significant negative weight, as predicted.
  [u1]: Negative weight: Judges disfavored medial /u/ compared with /i/ (?)

	[+u1]	[-u1]	Total
[+u3]	***16***	36	52
[-u3]	30	***27***	57
Total	**46**	63	109

(23) However, despite consisting (almost) entirely of nonlexical syllables, both experiments have a serious lexical confound: analogy.

- Grammar: Induced by babies from a corpus by an innate learning mechanism.
- Analogy: Directly generalized from a corpus by adults, unrestrictedly.
- Analogy is known to affect judgments independently of (apparent) grammar (e.g., Bailey & Hahn, 2001; Frisch & Zawaydeh, 2001; Myers & Tsay, 2005).
- Analogy can even affect the corpus itself: Note that all 4 /iai/ exceptions in (17) are homophones (onsetless with Tone 2), hence analogically similar with each other.

(24) The most basic measure of analogical similarity is edit distance: the number of deletions, insertions, or replacements of units needed to change one item into another:

ED(/tan/, /kan/) = 1 [1 replacement]
ED(/tan/, /kuan/) = 2 [1 replacement, 1 insertion/deletion]

- A lexical item is a neighbor of a target item if their edit distance is 1; summing the type frequencies of all neighbors gives the target item's neighborhood density.

(25) To test how neighborhood density affected judgments in the *H/Voiced and OCP experiments, I extended the default MiniJudge analyses. [Note: A*B = A + B + A×B]

Prob("yes") = Baseline + [Voiced] * [HH] * Neighbors + Random
Prob("yes") = Baseline + [u1] * [u3] * Neighbors + Random

- OCP results: Neighbors: Positive weight: More neighbors meant more acceptable.
  [u1] effect: Gone, and replaced by a negative [u3] effect. (?)
  OCP ([u1]×[u3] interaction): Gone! So pattern was just analogy!

- *H/Voiced: [Voiced]×[HH]×Neighbors interaction: ??
  [Voiced]×[HH]: Significant *positive* effect on judgments, perhaps because Neighbors and *H/Voiced constraint actually overlap (??)

<u>Modeling phonological grammar</u>

(26) Optimality Theory (OT) makes it possible to extend the linking model notion to test grammar-internal components (e.g., constraints and ranking). Three reasons why:

❶ Analogy can be represented in OT as output-output faithfulness (Myers, 2002).
❷ OT researchers already use corpus-analysis tools, namely, learning algorithms like the <u>Gradual Learning Algorithm</u> (GLA; Boersma & Hayes, 2001; Boersma & Weenink, 2006). Such models attempt to describe the inherently "best" (innate) corpus analysis.
❸ OT is mathematically related to statistical modeling (e.g., Keller, to appear).

(27) Regarding ❶, note the parallel between edit distance and the three basic correspondence constraints: MAX (don't delete), DEP (don't insert), IDENT (don't replace).

- Paradigmatic analogy is already well-established in OT, as output-output (OO) correspondence between morphologically related words (Benua, 1997).
- Myers (2002) extended OO constraints to handle cross-paradigmatic analogy:

tan	FAITH-OO(kan)	FAITH-OO(tuan)	FAITH-OO(kuan)
tan	*	*	**
Edit distance:	1	1	2

- This suggests analogical constraints of the form NEIGHBOR(Trigger): Target incurs one star if ED(Target, Trigger) > 1. Now we can test if some hypothesized grammatical constraint GRAM is truly superior to mere analogy by testing if GRAM » NEIGH:

tiai2	OCP[bk]	NEIGH(iai2)	NEIGH(tai2)	NEIGH(tiau4)	…
tiai2	*			*	…
tiau2		*	*		…

(28) There are too many NEIGH constraints to do this by hand. Can GLA help?

- Grammatical constraints: OCP[rd], OCP[bk]
- Analogical constraints: NEIGH(iai2), NEIGH(tai2), ...
  - 486 distinct syllables are neighbors with at least one item in the training set.
- Training set: All 1342 real syllables in the table in (17).
  - Inputs: Surface forms (e.g., /iau2/ for [iau2] and /iai2/ for [iai2]).
  - Candidates: Violate vs. obey OCP (e.g., [iai2] vs. [iau2] for both /iau2/ and /iai2/).
- <u>Results</u>: OCP[rd] and OCP[bk] are ranked high, but the latter doesn't outrank all analogical constraints, presumably because of those [iai2] exceptions:

  OCP[rd] » {NEIGH(iou3), OCP[bk], NEIGH(iou1)} » ...

(29) A weakness of GLA as a corpus analyzer is that it doesn't test statistical significance.

- Training GLA on an abstract "Pazih" corpus of 33 patterned (P) and 12 exceptional (E) forms, using FAITH and the "markedness" constraint *E, makes FAITH dominate over *E, thus "killing" it, even though we know that the pattern is statistically significant.

(a) [33]

P	FAITH	*E
☞ P		
E	*	*

(b) [12]

E	FAITH	*E
P	*	
☞ E		*

(30) To see how to test significance, first note that constraint ranking implies an equation.

- Count the stars and treat them like digits. The winner is the "lowest" candidate:

InA	Cons_1	Cons_2
☞ OutA_1		*
OutA_2	*	

InB	Cons_1	Cons_2
☞ OutB_1		
OutB_2		*

InC	Cons_1	Cons_2
☞ OutC_1	*	
OutC_2	*	*

☞ OutA_1	0 1 = ***1***
OutA_2	1 0 = 10

☞ OutB_1	0 0 = ***0***
OutB_2	0 1 = 1

☞ OutC_1	1 0 = ***10***
OutC_2	1 1 = 11

- More generally (Prince & Smolensky, 2002, p. 219):

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

(31) Keller (to appear) builds on this by noting that if we measure harmony for each candidate (from judgments), we can model the paired differences between candidate values with an OT equation (cf. Goldwater & Johnson, 2003; Pater et al., 2006):

$$\text{Value}_{\text{CandA}} - \text{Value}_{\text{CandB}} = w_1\times(\text{Star1}_{\text{CandA}} - \text{Star1}_{\text{CandB}}) + w_2\times(\text{Star2}_{\text{CandA}} - \text{Star2}_{\text{CandB}}) + \ldots$$

- This means that data confirm Cons_i » Cons_j only if:

$$\text{Data} = w_i\times\mathbf{Star_i} + w_j\times\mathbf{Star_j}, \qquad \text{where } w_i \neq 0,\ w_j \neq 0,\ |w_i| > |w_j|.$$

(32) Keller's approach can be extended to test significance in corpus data (I think).

- Candidate values would be binary (1 = attested, 0 = absent), as would candidate differences (if $\text{Value}_{\text{CandA}} < \text{Value}_{\text{CandB}}$, reverse the signs of the weights).
- Represent each item by an OT tableau with all logically possible candidates.
- Model candidate differences, grouped by item, using GLMM.

- For example, the "Pazih" data would start with the following tableaux:

(a) [33]

P		FAITH	*E
i	1	0	0
ii	0	1	0
iii	0	0	1
iv	0	1	1

(b) [12]

E		FAITH	*E
i	0	0	0
ii	0	1	0
iii	1	0	1
iv	0	1	1

- These would be encoded as paired candidate differences like so:

(a) [33]

P		FAITH	*E
i-ii	1	-1	0
i-iii	1	0	-1
i-iv	1	-1	-1
ii-iii	0	1	-1
ii-iv	0	0	-1
iii-iv	0	-1	0

(b) [12]

E		FAITH	*E
i-ii	0	-1	0
i-iii	1	0	1
i-iv	0	-1	-1
ii-iii	1	-1	1
ii-iv	0	0	-1
iii-iv	1	-1	0

(33) Does it work? Well, aside from the logistic regression "crashing" problem, I think so:

- For the "Pazih" data, $|w_{\text{FAITH}}| = 1.288 > |w_{*\text{E}}| = 0.606$, so the ranking is motivated.
- Crucially, however, *both* weights are statistically significant ($p < .05$).

Conclusions

❶ Phonology is primarily corpus linguistics, with all of the challenges this implies.
❷ Collecting judgments doesn't avoid all lexical confounds, due to analogy.
❸ MiniJudge is an attempt to make the collection and analysis of judgments easy.
❹ GLA can be used as a phonological corpus analysis tool, though it has limitations.
❺ Hopefully MiniCorp will eventually exist to help simplify phonological corpus analysis, especially for researchers working within an OT framework.

References

Bailey, T. M., Hahn, U. (2001). Determinants of wordlikeness: Phonotactic or lexical neighbourhoods? *Journal of Memory and Language, 44*, 568-591.

Benua, L. (1997). *Transderivational identity: Phonological relations between words*. U. Mass PhD dissertation.

Boersma, P., & Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic Inquiry, 32*, 45-86.

Boersma, P., & Weenink, D. (2006). Praat. http://www.fon.hum.uva.nl/praat.

Chomsky, N. (1957). *Syntactic structures*. Mouton.

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

Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgments*. Sage.

Duanmu, S. (2004). A corpus study of Chinese regulated verse: Phrasal stress and the analysis of variability. *Phonology, 21*, 43-89.

Frisch, S. A., & Zawaydeh, B. A. (2001). The psychological reality of OCP-Place in Arabic. *Language, 77* (1), 91-106.

Goldwater, S., & Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In J. Spenader, et al. (Eds.), Proceedings of the Stockholm Workshop on Variation within Optimality Theory (p. 111-120). Stockholm Univ.

Hale, M., & Reiss, C. (2000). Substance abuse and dysfunctionalism: Current trends in phonology. *Linguistic Inquiry, 31*, 157-169.

Hsu, H.-C. (2005). A diachronic look into aberrant synchronic sound patterns. *Tsing Hua Journal of Chinese Studies, 35* (2), 435-456.

Keller, F. (to appear). Linear Optimality Theory as a model of gradience in grammar. In G. Fanselow et al. (eds.) *Gradience in grammar: Generative perspectives*. Oxford. http://roa.rutgers.edu.

Li, P. J.-K., & Tsuchida, S. (2001). *Pazih dictionary*. Taipei: Institute of Linguistics.

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] ROC MOE. http://140.111.1.22/clc/dict/htm/pin/start.htm

Mendoza-Denton, N., Hay, J., & Jannedy, S. (2003). Probabilistic sociolinguistics: Beyond variable rules. In R. Bod, J. Hay, & S. Jannedy (Eds.) *Probabilistic linguistics* (pp. 97-138). MIT Press.

Myers, J. (2002). Exemplar-driven analogy in Optimality Theory. In R. Skousen etal. (Eds.) *Analogical Modeling: An exemplar-based approach to language* (pp. 265-300). John Benjamins.

Myers, J. (2006a). *Linguistics as cognitive psychology. IACL 14 & IsCCL 10 Proceedings* (pp. 150-174). Taipei, Taiwan.

Myers, J. (2006b). MiniJudge: Software for minimalist experimental syntax. ROCLING 18 (pp. 271-285). Hsinchu, Taiwan.

Myers, J., & Tsay, J. (2005). The processing of phonological acceptability judgments. *Proceedings of Symposium on 90-92 NSC Projects* (pp. 26-45). Taipei, Taiwan.

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

Pater, J. (to appear). The locus of exceptionality: Morpheme-specific phonology as constraint indexation. In S. Parker (Ed.) *Phonological argumentation*. Equinox.

Pater, J., Potts, C., & Bhatt, R. (2006). Harmonic grammar with linear programming. UMass Amherst ms. ROA 872-1006. HaLP at http://web.linguist.umass.edu/~halp/

Prince, A., & Smolensky, P. (2002). *Optimality Theory: Constraint interaction in generative grammar*. Rutgers & John Hopkins ms. ROA 537-0802.

Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. U. Chicago Press.

Shih, C. (n.d.) Corpus methods for Mandarin phonology. U. Illinois, Urbana-Champaign ms. https://netfiles.uiuc.edu/cls/www/course/notes.pdf

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

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