The components of phonological data

James Myers
National Chung Cheng University

International Workshop on Grammar & Evidence
April 14, 2007
Lngmyers@ccu.edu.tw

1

Acknowledgments

- NSC 94-2411-H-194-018
- Co-directors Jane Tsay & Woody Tsay
- MiniJudge is co-copyrighted by National Chung Cheng University
- So will MiniCorp (if it ever gets finished)

2

Goals

- Link phonological methodology to standards followed in the rest of cognitive science…
  - ② · Quantification and statistical analysis
  - ④ · Factoring out nuisance variables
- … without losing phonological insights
  - ① · Dictionary data are useful
  - ③ · Universal constraints are necessary
- Describe software tools to help with all this

3

Phonological corpora

- Phonologists rely primarily on corpora (dictionary attestations)…
  - · Convenience: dictionaries are already available
  - · Theoretical interest in lexical knowledge (contrastiveness, morphological interface)
- … despite serious limitations (Ohala 1986)
  - · Historical residue vs. synchronic grammar
- Should phonologists dump corpora?

4

The case for phonological corpora

- Banning corpora means throwing out centuries of phonological theorizing
- Corpora ease cross-linguistic typology
- Lexical judgments strongly mirror the lexicon anyway (e.g. Bailey & Hahn 2001)
- The lexicon represents an important target in phonological development
- Historical change itself is probably shaped by grammar (e.g. Kiparsky 2006)

5

An extended example: Pazih

- Li & Tsuchida (2001) provide a corpus of 45 morphemes of the form CVC-V-CVC
  - · CVC root is reduplicated
  - · All examples show epenthetic -V-
- In most items, -V- is same as base vowel
  - · e.g. hur-u-hur (steam, vapor)
- Yet there are many exceptions: 12/45 (27%)
  - · e.g. hur-a-hur (bald)

6

Handling the exceptions

- By definition, all items epenthesize, so no items need to have a vowel slot in the input
- But vowel quality in the exceptions must still be specified: **floating vowels**…?

“steam”	“bald”
[hur-u-hur]	[hur-**a**-hur]
u u	u **a** u
/hVr-hVr/	/hVr-hVr/

7

An Optimality Theory analysis

- Epenthesis is driven by syllable structure (Pazih disfavors word-internal codas)

  NoCoda » DepV
- Features on epenthetic vowels are filled in by vowel harmony, unless blocked by features of floating input vowel

  IdentV » AgreeV

8

Unpacking the logic

- How justified is this analysis?
- This tiny corpus has the only available data
  - · Too few native speakers to run experiments
  - · History is unknown
- Crucial assumptions:
  - · Exceptions don’t undermine harmony claim
  - · Epenthesis and harmony are independent
  - · Exceptional vowels are unpredictable

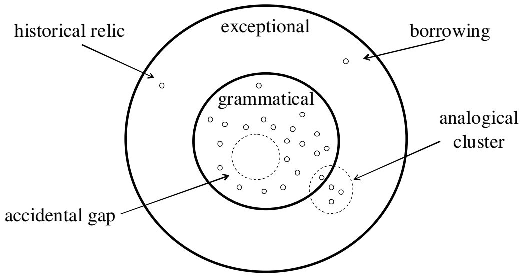
9

Seeking grammar in corpus data

- Raw data are **attestations**
  - · Is the predicted type in the corpus or not?
- The path to attestation is partly random
  - · **Accidental gaps:** Coinage isn’t obligatory
  - · **Exceptions:** Memory side-steps grammar
- It also reflects systematic non-grammar
  - · **Analogy:** Superficial bottom-up generalization

10

A target metaphor



11

Quantifying the metaphor

- Key correlations (cf. Duanmu 2004)
  - · Higher type frequency in the corpus implies higher probability of grammaticality
  - · Ungrammaticality implies lower type frequency in the corpus (all else being equal)
- Probability of predicted type in corpus ~ Baseline bias + Other systematic biases (including analogy) + **Grammar** + Chance

12

Testing Pazih harmony

- The chance probability of harmonizing
  - · The **null hypothesis** is 50% harmonizing, 50% not
  - · So the chance probability of 33/45 harmonizing items is like getting 33 heads in 45 coin flips
  - · If chance probability is low enough ($p < 0.05$), we can reject the null hypothesis
- **Binomial test** in R (www.r-project.org)
  - · `min(1, 2*pbinom(min(33, 12), 45, 0.5)) # (2-tailed)`
  - · $p = 0.0024 < 0.05$: Significant!

13

Epenthesis vs. harmony

- What about the second assumption, that epenthesis and harmony are independent?
  - · *A priori* assumption of autosegmental theory…?
  - · Or is it empirically testable…?
- Prob(Epenthesis given Harmony) = Prob(Epenthesis given Non-Harmony) = Prob(Epenthesis) = 100%
- This fits formal definition of independence

14

Testing non-harmonic vowels

- The third crucial assumption is that the non-harmonizing vowels are unpredictable
- Yet Li & Tsuchida (2001:21) observe that "/a/ appears to be the most common" (7/12)
- We can compare /a/ vs. non-/a/ exceptions
  - · `min(1, 2*pbinom(min(7, 5), 12, 0.5))`
  - · $p = 0.774 > 0.05$
- No need to reject null hypothesis

15

But which null hypothesis?

- Pazih has 4 different vowels (/a/, /i/, /u/, /e/)
- Chance probability of 7/12 /a/ items in exceptions isn't really like flipping a coin…
- … but more like rolling a 4-sided die
  - · `min(1, 2*pbinom(min(7, 5), 12, 0.75))`
  - · Now $\boldsymbol{p}$ **= 0.029** < 0.05: Significant!
- So is the /a/ pattern significant or not…?

16

The corpus paradox

- Experimental hypotheses follow by design
  - · Design: e.g. Factor [+F] vs. [-F]
  - · Hypothesis: e.g. [+F] > [-F]
- But in corpus analysis there's no design
  - · Corpus data suggest hypotheses…
  - · … then the same data are used to test them!
- It's like playing cards with no rules
  - · Any hand of cards is equally "significant"

17

Universals resolve the paradox

- Begin with *a priori* **framework** ("the rules")
  - · Induced from other languages (empiricism)
  - · Inherently necessary (rationalism, functionalism)
- Li & Tsuchida's "cophonology" framework
  - · Exceptions reflect an earlier historical stratum
  - · In this reference frame, 7/12 /a/ implies $p = 0.028$
- Simpler framework rejecting cophonology
  - · 12/45 /a/ is 27%, close to chance of 25% (1/4)

18

Universals in Optimality Theory

- OT is the universalist framework assumed by most phonologists today
  - · Constraints describe regularities…
  - · … but also representations (Golston 1996)
    e.g. [hur-a-hur] = ✓IdentV, ×AgreeV, …
- Can we test statistical models of the form
  - · Data ~ Constraint_1 + Constraint_2 + …?

19

Regression modeling of corpus data

- **Loglinear regression**
  - · Y ~ $w_0 + w_1X_1 + w_2X_2 + \ldots$
  - · Y is probability or count data, Xs are predictors
  - · *w*s are weights (w_0 = baseline); $w > 0$: positive correlation; $w < 0$: negative; $w = 0$: no correlation
- Most familiar type: **logistic regression**
  - · Y = probability (log odds) of some property
  - · E.g. VARBRUL (Mendoza-Denton et al. 2003)
- Another type: **poisson regression**
  - · Y = count data (size of a category)

20

AgreeV in logistic regression

- Model: Harmonizing ~ Baseline
  - · `Harmonizing = c(rep(1, 33), rep(0, 12))`
  - · `summary(glm(Harmonizing~1, family=binomial))`
  - · p = 0.0027 (very close to binomial test)
- Limitations
  - · Y = property of attested items (typically binary), but we lose the patterning of non-attested items
  - · Also, the Xs don't look like OT constraints

21

AgreeV in poisson regression

- Model: Count ~ (Baseline) + AgreeV

```
·Count = c(33, 12)
·AgreeV = c(0, 1) # (1 = violation)
·Pazih = data.frame(Count, AgreeV)
·summary(glm(Count~AgreeV, family =
poisson, data = Pazih))
```

  - · AgreeV: p = 0.0027 (same as logistic regression)
- Now we can factor out multiple constraints
  - · E.g. Count ~ Cons_1 + Cons_2 tests both constraints' independent contributions

22

AgreeV and IdentV

- Model: Count ~ AgreeV + IdentV

```
·Count = c(33, 0, 12, 0)
·AgreeV = c(0, 0, 1, 1)
·IdentV = c(0, 1, 0, 1)
·Pazih = data.frame(Count,
AgreeV, IdentV)
·summary(glm(Count ~ AgreeV + IdentV,
family = poisson, data = Pazih))
```

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

· AgreeV: p = 0.0027 (same as before)
· IdentV: p = 0.9996…
…because regression can't handle perfect correlations

23

AgreeV and IdentV in Pazih2

- Model: Count ~ AgreeV + IdentV

```
·Count = c(33, 1, 12, 0)
·AgreeV = c(0, 0, 1, 1)
·IdentV = c(0, 1, 0, 1)
·Pazih2 = data.frame(Count,
AgreeV, IdentV)
·summary(glm(Count ~ AgreeV + IdentV,
family = poisson, data = Pazih2))
```

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

· AgreeV: p = 0.0019 (basically the same as before)
· IdentV: p = 0.0002… significant, as it should be

24

Weights and constraint ranking

- We can also learn something from the **weights** associated with each constraint
  - · AgreeV: *weight* = -1.04
  - · IdentV: *weight* = -3.81
- A curious parallel
  - · $|weight_{\mathrm{IdentV}}| > |weight_{\mathrm{AgreeV}}|$
  - · IdentV » AgreeV
- Coincidence? Not quite….

25

OT as regression modeling

- If you treat stars like digits, the "lowest" candidate wins (Prince & Smolensky 2002)

In	$Cons_1$	$Cons_2$	Value	
Out_A		*	= 01 = **1**	Lowest (winner)
Out_B	*		= **10**	

· Value = $weight_1 Star_1 + \ldots + weight_n Star_n$, where $weight_1 = b^{n-1}, \ldots, weight_n = b^0$ for some $b > \max(\mathrm{Star})$

26

Exploiting the OT/regression link

- Used in learning models (e.g. Keller 2000, Goldwater & Johnson 2003, Lin 2005, Pater et al. 2006)
  - · Goldwater & Johnson (2003) use a type of loglinear regression (based on conditional probability of output candidate given an input)
- But none test statistical significance
  - · Most model child language acquisition
  - · Keller (2000) focuses on judgments, not corpora

27

Poisson regression as grammar modeler

- Proposal:
  - · Accept $Cons_i$ » $Cons_j$ only if $|weight_i| >> |weight_j|$
- Justification:
  - · If $Cons_i$ » $Cons_j$, then there should be "significantly more" items where $Cons_i$ is obeyed but $Cons_j$ isn't than the other way around
  - · Hence if $|weight_i| \approx |weight_j|$, then the claim of $Cons_i$ » $Cons_j$ is doubtful

28

Acquisition as corpus analysis

- But is this really on the right track?
- The need for *a priori* framework in corpus analysis fits with a key nativist claim
  - · The only "true" corpus analysis is the grammar acquired by the child (Chomsky 1965)
- So the one "true" grammar learner…
  - · … may have nothing to do with regression
  - · …e.g. it could be the Gradual Learning Algorithm (GLA) of Boersma & Hayes (2001)

29

The corpus paradox: Version II

- Statistical significance may be entirely irrelevant in grammar learning
  - · E.g. A = 60 vs. B = 40: not significant (p = 0.057)
  - · But if children are highly sensitive to *any* A vs. B asymmetry, this may be grammaticalized anyway
- Yet surely we don't need children (or history)
  - · Astronomers rely on corpora alone; why can't we?
  - · A phonological corpus is more than the **input** to acquisition; it's also the **output** of grammar use

30

The corpus paradox: Version III

- The corpus's dual role
  - · ... as input to the child's innate corpus analyzer
  - · ... as output for people to analyze freely
- This suggests that "free" analyses can work their way into the corpus itself
- That is, corpus data may be "corrupted" by the diachronic operation of **analogy**

Grammar vs. analogy

- Analogy: Bottom-up and superficial
  - · **Bottom-up:** Patterns arise through similarity between items, not top-down by general rules
  - · **Superficial:** Similarity is defined in a concrete way (not via abstractions defined by grammar)
- It's an open question whether grammar and analogy are really so distinct (see e.g. Albright & Hayes 2003; Myers 2002; Wang 1995)

Analogy in Mandarin triphthongs

- Mandarin triphthongs ($V_1V_2V_3$) generally can't start and end with same vowel, but …

Syllable type frequencies (Li et al. 1997, Tsai 2000)

			V_3	
		V_2	... i	... u
V_1	i ...	... e/o ...	0	269
		... a ...	4	494
	u ...	... e/o ...	515	0
		... a ...	60	0

Homophones of /iai^{35}/

Analogy as null hypothesis

- A grammatical hypothesis is best supported
  - · if it's not only superior to chance...
  - · ... but also to analogy
- Is epenthetic vowel quality in Pazih predicted by overall **typicality**?
  - · If so, maybe harmony spread by analogy
  - · If not, we strengthen the claim that harmony is handled by a top-down, abstract grammar

Quantifying analogy

- This requires defining a similarity metric
- The most basic metric: **Edit Distance**
  - · The minimum number of **deletions**, **insertions**, or **substitutions** to change one string into another

    EditDist(abcd, abc_) = 1

    EditDist(abcd, abc**x**) = 1

    EditDist(abcd, ab**x**_) = 2
- Note parallels with Max, Dep, and Ident….

Does AgreeV go beyond analogy?

- Model: Count ~ Neighbor + AgreeV
  - · $Item_i$ & $Item_j$ are "neighbors" if EditDist($Item_i$, $Item_j$) = 1
  - · "Neigh": At least one neighbor

Count	Neigh	AgreeV
29	0	0
10	0	1
4	1	0
2	1	1

```
·Count = c(29, 10, 4, 2)
·Neigh = c(0, 0, 1, 1)
·AgreeV = c(0, 1, 0, 1)
·PazihAn = data.frame(Count, Neigh, AgreeV)
·summary(glm(Count ~ Neigh + AgreeV, family
= poisson, data = PazihAn))
```

· Typicality: *weight* = -1.87, $p < 0.0001$
· AgreeV: *weight* = -1.01, $p = 0.0027$

Other ways to define analogy

- Is counting only nearest neighbors enough?
  - · Bailey & Hahn (2001): Gradient neighborhoods
- Edit distance ignores some similarity
  - · EditDist(abcd, dcba) = 4: Too high?
- Similarity at what level? (Features?)
- Note parallel issues with faith constraints
  - · Linearity, Max-F, etc….
  - · An OT-based formalism for analogy…?

37

Analogy and ranking in OT

- Myers (2002): Analogy via faith constraints
  - · Analogy and "grammar proper" can be described within the same formalism
- Thus an OT learner can be used to test the relative ranking of analogy and grammar
  - · If Analogy » Grammar, is grammar justified?
- Even a non-regression-based learner like GLA could be used to run such tests

38

Automating all this (someday)

- **MiniCorp** is software for the analysis of small (phonological) corpora

  … cf. **MiniJudge**, for running, designing, and analyzing small (syntactic) judgment experiments

  [http://www.ccunix.ccu.edu.tw/~lngproc/MiniJudge.htm]
- Three steps
  - · Tagging corpus items for relevant properties
  - · Testing constraint significance (and ranking?)
  - · Testing for analogy

39

Tagging the corpus

- Electronic dictionary loaded in as text file
  - · Phonetic font can be applied
- User tags some items for relevant properties
  - · E.g. constraint violations (as in Golston 1996)
- MiniCorp uses analogy (edit distance) to "guess" tags for other items
- User checks and approves all tags

40

Testing constraints

- MiniCorp runs loglinear analysis on counts
  - · Poisson? Or Goldwater & Johnson (2003)?
- Weights are compared, to test for ranking
  - · But what's the best way to compare weights…?
- Option to predict one binary property from the others (logistic regression)
  - · Might reveal further unsuspected patterns

41

Testing for analogy

- Loglinear analysis is run with analogical score as one of the predictors
  - · But which definition of analogy?
  - · Count data require categorical predictors, so how to use continuous-valued neighborhood measures like that of Bailey & Hahn (2001)?
- What if Analogy and Grammar are both significant, but $|weight_{An}| >> |weight_{Gr}|$?

42

Summary

- Phonological corpora are informative; we don’t have to rely solely on experiments
- Yet corpus analysis faces serious paradoxes
  - · Resolving some requires universalist assumptions
  - · Others require testing grammar vs. analogy
- All of this requires quantitative methods
  - · Improved software could help make such methods part of the phonologist’s standard toolkit

43

References (1/3)

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition, 90*, 119–161.

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

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

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

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

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.

Golston, C. (1996). Direct Optimality Theory: Representation as pure markedness. *Language, 72*, 713-748.

44

References (2/3)

Keller, F. (2000). *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. PhD diss, University of Edinburgh.

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

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 Ministry of Education. http://140.111.1.22/clc/dict/htm/pin/start.htm

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

Lin, Y. (2005). Learning stochastic OT grammars: A Bayesian approach using data augmentation and Gibbs sampling. Paper presented at the 43rd Association for Computational Linguistics. Ann Arbor, MI. http://dingo.sbs.arizona.edu/%7Eyinglin/SOT_acl05.pdf

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.

45

References (3/3)

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.

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

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.

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

Wang, H. S. (1995). *Experimental studies in Taiwanese phonology*. Taipei: Crane Publishing.

46

Appendix: Pazih corpus

bak-a-bak (native cloth)
dap-a-dap (to rub)
hay-a-hay (stalk of miscanthus)
ma-hak-a-hak (itchy in taste)
ngar-a-ngar (to bite)
bel-e-bel (banana)
beng-e-beng (loom)
dek-e-dek (to step on ground)
deng-e-deng (to boil in water)
i-dek-e-dek (to sink)
leng-e-leng (to aim)
ma-bet-e-bet (suffering from gas pain)
maa-bez-e-bet (to help each other)
rex-e-rex (to wrestle)
ma-her-e-her (to suffer from asthma)
pa-gen-e-gen (to hum)
ser-e-ser (to repeat)
zek-e-zek (to erect)
gir-i-gir (to saw)
hir-i-hir (to grind)
ngir-i-ngir (to nibble)
mu-tin-i-tin (to weigh)

pa-kih-i-kih (dry, hacking cough)
buk-u-buk (bamboo pipe)
bus-u-bus (smoke)
duk-u-duk (ginger)
dung-u-dung (drum)
dus-u-dus (to grate)
gun-u-gun (bucket, to measure)
hur-u-hur (steam, vapor)
kul-u-kul (type of bird)
ngur-u-ngur (to scold)
pa-sur-u-sur (to masturbate)

harmonizing

bar-e-bar (flag)
dak-e-dak (to scrape off with one's feet)
par-e-par (paper)
ma-led-a-let (to tremble)
bir-a-bir (tough as of meat)
ma-bid-a-bit (to wobble)
ma-ngir-a-ngir (easy)
buh-a-buh (to powder one's face)
bur-a-bur (dust)
hur-a-hur (bald), mu-hur-a-hur (to pluck feather of a fowl)
ma-bux-i-bux (very tired)
mux-i-mux (to gargle)

exceptional

47