Similarity and bias in phonological learning

Colin Wilson
UCLA

Rumelhart Prize Symposium
CogSci 2005, July 22
Overview

- What role does **phonetic substance** play in phonological cognition?

- Results from language game experiments on velar palatalization (k → tʃ "ch", g → dʒ "j") show **asymmetric generalization patterns**

- Modeling with GCM and Maxent supports the claim that substance functions as a **bias** (or **prior**) on phonological learning
The role of substance
Limits on phonological cognition

- **Formal universals**
  Architecture (e.g., rules vs constraints)
  Formal language theory (e.g., finite-state phon)

- **Substantive universals**
  Does the phonetic content of phonological symbols also impose limits on phon cog?

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Chomsky &amp; Halle 1968, Hayes et al. 2004</td>
</tr>
<tr>
<td>Ltd</td>
<td>Halle 2005: distinctive features only</td>
</tr>
</tbody>
</table>
Typological evidence for substance

- Certain phonological patterns are found in many, genetically-diverse languages while others are rare or unattested. 
  
  *Ex.* Word-final devoicing (Catalan, German, Ingush, Turkish, Wolof, . . .) vs. word-final voicing (Lezgian)

- The widely-attested phonological patterns can often be understood in terms of ease of articulation and/or perceptual distinctiveness
  
  *Ex.* Reduction of perceptual cues for the t/d contrast word-finally facilitates neutralization (Steriade 1999)
Case study: Velar palatalization

- Velar pal refers here to the change from a velar stop (k g) to a palatoalveolar affricate (ʧ ʤ) in certain vowel environments.

- Components of the case study
  1. Phonetic properties of velar stops
  2. Typological generalization
  3. Steriade’s law of similarity
1. Phonetic properties: Articulation

Velar stop + Vowel Coarticulation

\[ k / \_i \text{ (most fronted)} \]
\[ k / \_e \]
\[ k / \_a \text{ (least fronted)} \]

(pics from http://www.umanitoba.ca/faculties/arts/linguistics/russell)

1. Phonetic properties: Acoustics

Articulatory fronting in the environment of front vowels gives rise to acoustic similarity of velar stops and palatoalveolar affricates

\[ \text{k} / \_\_\_\_ \alpha \text{ (from 'kagē)} \quad \text{k} / \_\_\_\_ \varepsilon \text{ (from 'kגē)} \quad \text{k} / \_\_\_\_ \iota \text{ (from 'kגē)} \]

Burst spectra for velar stops and the palatoalveolar affricate \( \text{tʃ} \) (from 'tʃיגē)
Details of acoustic measurements

- Guion 1996, 1998 measured the peak frequency (kHz) in the average spectrum of the burst/aspiration (512 FFT points, +6dB preemphasis)

<table>
<thead>
<tr>
<th></th>
<th>back</th>
<th>front</th>
<th>high front</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k / g$</td>
<td>2.25</td>
<td>3.5</td>
<td>4.0</td>
</tr>
<tr>
<td>$\tilde{t}f / \tilde{d}3$</td>
<td>4.25</td>
<td>4.25</td>
<td>4.5</td>
</tr>
</tbody>
</table>

- These results (and measurements on the stimuli for the present experiments) show that velar stops are more similar to palatoalveolar affricates before front vowels, especially high front vowels
1. Phonetic properties: Perception

Acoustic similarity of velar stops and palato-alveolar affricates is correlated with rate of velar → pal errors in identification (Guion 1996, 1998)

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>[ki]</th>
<th>[tʃi]</th>
<th>[gi]</th>
<th>[dʒi]</th>
<th>[ka]</th>
<th>[tʃa]</th>
<th>[ga]</th>
<th>[dʒa]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ki]</td>
<td>43</td>
<td>35</td>
<td>10</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[gi]</td>
<td>4</td>
<td>4</td>
<td>71</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[ka]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>84</td>
<td>13</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>[ga]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>0</td>
<td>87</td>
<td>9</td>
</tr>
</tbody>
</table>

Similarity and bias in phonological learning – p. 10/55
2. Typological generalization

If velar palatalization ($k \rightarrow \tilde{t}\tilde{ʃ}$, $g \rightarrow \tilde{d}\tilde{ʒ}$) applies before a vowel V, then it also applies before all vowels that are more front than V

*Ex. velar pal before e  \Rightarrow  velar pal before i*

*Ex. velar pal before æ  \Rightarrow  velar pal before e,i*

3. Steriade’s law of similarity

- The cognitive system privileges alternations involving perceptually-similar sounds

\[ \Delta_P(x, y) < \Delta_P(x, z) \]

\[ \Rightarrow x \rightarrow y \text{ is preferred over } x \rightarrow z \]

- The general law forms part of an explanation of the typological implication on velar pal

<table>
<thead>
<tr>
<th>articulation</th>
<th>acoustics</th>
<th>perception</th>
<th>phonology</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V_1 &gt; V_2 )</td>
<td>( \hat{t} \hat{j} \geq kV_1 &gt; kV_2 )</td>
<td>( \Delta_P(k, \hat{t} \hat{j}</td>
<td>V_1) )</td>
</tr>
<tr>
<td>(frontness)</td>
<td>(spectral peak)</td>
<td>( &lt; \Delta_P(k, \hat{t} \hat{j}</td>
<td>V_2) )</td>
</tr>
</tbody>
</table>
Evolutionary alternative

• An alternative explanation for the typological generalization would invoke (non-cognitive) diachronic change or “evolution”

• Greater acoustic/perceptual similarity of palatoalveolars and velars before more front vowels could make $k \rightarrow \text{tʃ} \_\_i$ more likely as a sound change (“error pattern”) than $k \rightarrow \text{tʃ} \_\_e$

Proposal

- Knowledge of substance functions as a bias (not absolute restriction) on phon grammars

- Weaker version of phonetically based phonology that avoids the empirical problems noted by Blevins and others

- Expect bias to be revealed most strongly when the input to the learner is impoverished
Language game experiments
Language games

- Language games are naturally-occurring phenomena that systematically alter the pronunciation of words (see Bagemihl 1995)

- Experiments reported here use artificial language games with impoverished input

- The measure of interest is how participants generalize from the input to a new vowel context (“poverty of the stimulus method”)
Experiment 1: High vs Mid

- High exposure
  kime ... tjieme ×8
  pila ... pila

- Mid exposure
  kene ... tjenen ×8
  pila ... pila

- Testing (both conditions; also included fillers)
  kime ... ? ×8
  kise ... ? ×8
  kine ... ? ×8
  keze ... ? ×8

Similarity and bias in phonological learning – p. 17/55
Qualitative predictions

- If Steriade’s law forms part of a bias (or prior) on phonological learning, then:
  + participants in the Mid group \((k \rightarrow \hat{t}j / \_\_ e)\)
    should generalize to the novel context \((\_\_ i)\)
  - but do not expect the same degree of
    generalization in the High group \((k \rightarrow \hat{t}j / \_\_ i)\)

- The relevant statistic is the interaction between condition (High vs Mid) and vowel environment (exposure vs novel)
Results of Exp 1: High (N = 11)

i vs e,
\[ t(10) = 3.0, \]
\[ p < .05 \]
Results of Exp 1: Mid (N = 11)

\[ t(10) = 0 \]

- i vs e,
- Initial CV:
  - ki, gi
  - ke, ge
  - ka, ga

Proportion palatalized
Results of Experiment 1

- Repeated-measures ANOVA (High vs Mid × Voiceless vs Voiced × Exposure vs Novel)

- Greater generalization in the Mid condition than in the High condition
  Condition × Context: $F(1, 20) = 8.3, p < .01$

- Main effect of consonant voicing (k vs g):
  $F(1, 20) = 8.0, p < .05$
  Why? Practice items (n=2) contained voiced g
Experiment 1A: High vs Mid

Palatalization by Vowel and Consonant

Proportion palatalized:
0.0 0.2 0.4 0.6 0.8 1.0

CV: ki gi ke ge ka ga

Similarity and bias in phonological learning – p. 22/5

Similarity and bias in phonological learning – p. 22/5
Experiment 2: Voiceless vs Voiced

- Voiceless exposure
  kimə ... ṭjimə ×4  kẹnə ... ṭjẹnə ×4  kapə ... kapə ×3
  pilə ... pilə  pebə ... pebə  parə ... parə

- Voiced exposure
  gimə ... ᵋjimə ×4  gerə ... ᵋjẹrə ×4  gapə ... gapə ×3
  pilə ... pilə  pebə ... pebə  parə ... parə

- Testing (both conditions; also included fillers)
  kimə ... ? ×8  kẹnə ... ? ×8  kapə ... ? ×3
  gimə ... ? ×8  gerə ... ? ×8  gapə ... ? ×3
Qualitative predictions

- Guion 1996, 1998 found greater similarity for **voiceless** k and tʃ than for **voiced** g and dʒ

  + participants in the Voiced group (g → dʒᵢ/e) should generalize to the novel consonant (k)

  - but do not expect generalization to the same degree in the Voiceless group (k → tʃᵢ/e)

- As before, the relevant statistic tests for an interaction: condition (Voiceless vs Voiced) and focus consonant (exposure vs novel)
Results of Exp 2: Voiceless (N = 11)

Voiceless

Proportion palatalized

Initial CV

ki  gi  ke  ge  ka  ga

0.0  0.2  0.4  0.6  0.8  1.0
Results of Exp 2: Voiced (N = 11)

![Graph showing proportion palatalized for different initial CVs]

- **Proportion palatalized**
  - ki
  - gi
  - ke
  - ge
  - ka
  - ga

**Initial CV**
Results of Experiment 2

- Repeated-measures ANOVA (Voiceless vs Voiced × High vs Mid × Exposure vs Novel)

- Non-significant difference between the rates of generalization in the conditions
  Condition × Focus: $F(1, 20) < 1$

- All other main effects and interactions (except main effect of focus) n.s.
Summary of experiments

- Exp 1. Asymmetric generalization on the context is consistent with Steriade’s law:
  - $\Delta_P(k, t\hat{ʃ} / _i) < \Delta_P(k, t\hat{ʃ} / _e)$
  - $k \rightarrow t\hat{ʃ}/_e$ generalized to _i (and _α!)
  - $k \rightarrow t\hat{ʃ}/_i$ not generalized to _e

- These results are not explained by the “error” mechanisms postulated by evolutionary phon

- Exp 2. Lack of generalization on the focus suggests an architectural limit on the law
Modeling substantive bias
Overview

1. Quantifying perceptual similarity with the GCM (Nosofsky 1986, et seq)

2. Integrating substantive bias into the maximum entropy (maxent) formalism

3. Comparing predictions of the biased and unbiased with the experimental results
1. Generalized context model (GCM)

- Stimulus properties (Nosofsky 1986)
  \[ d_{ij} = c \left[ \sum_{m=1}^{M} w_m |x_{im} - x_{jm}|^r \right]^{1/r} \]
  \[ \sum w_m = 1 \]
  \[ c > 0 \]

- Perceptual similarity (Shepard 1957, 1987)
  \[ \eta_{ij} = \exp(-d_{ij}) \quad \text{we solve for this} \]

- Luce choice rule (Luce 1962)
  \[ P(\text{resp} = x_j \mid \text{stim} = x_i) = \frac{b_j \eta_{ij}}{\sum_{k=1}^{n} b_k \eta_{ik}} \]
  \[ \sum b_k = 1 \]
1. Applying the GCM

- How perceptually similar are velar stops and palatoalveolar affricates in vowel contexts?
- Stimulus dimensions
  - peak spectral frequencies (Guion / ours)
  - dummy-coded vowel quality and voicing
- Confusion matrix (Guion)
  - [contains data for /i and /a, but not /e]
- Maximum-likelihood fit of scale ($c$), attention weights ($\{w_m\}$), and response biases ($\{b_k\}$)
1. Response bias * similarity values

Inverse of the perceptual ‘cost’ of changing a velar stop to a palatoalveolar affricate

- Response bias($\hat{t}\hat{V}$) * similarity($kV,\hat{t}\hat{V}$)
  
<table>
<thead>
<tr>
<th>Rule</th>
<th>Response bias</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ki \rightarrow \hat{t}\hat{i}$</td>
<td>$9.23^{-1}$</td>
<td></td>
</tr>
<tr>
<td>$ke \rightarrow \hat{t}\hat{e}$</td>
<td>$12.68^{-1}$</td>
<td></td>
</tr>
<tr>
<td>$ka \rightarrow \hat{t}\hat{a}$</td>
<td>$88.72^{-1}$</td>
<td></td>
</tr>
</tbody>
</table>

- Response bias($\hat{d}\hat{z}\hat{V}$) * similarity($gV,\hat{d}\hat{z}\hat{V}$)

<table>
<thead>
<tr>
<th>Rule</th>
<th>Response bias</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$gi \rightarrow \hat{d}\hat{z}\hat{i}$</td>
<td>$21.13^{-1}$</td>
<td></td>
</tr>
<tr>
<td>$ge \rightarrow \hat{d}\hat{z}\hat{e}$</td>
<td>$40.60^{-1}$</td>
<td></td>
</tr>
<tr>
<td>$ga \rightarrow \hat{d}\hat{z}\hat{a}$</td>
<td>$126.93^{-1}$</td>
<td></td>
</tr>
</tbody>
</table>

- Cost for e context estimated with $b(Xe) = b(Xi)$
2. Maxent model: General

- Type of log-linear model in which entropy is maximized subject to \((\text{expected constraint violation}) = (\text{observed constraint violation})\)

- Closely related to random fields (Della Pietra et al. 1986, Lafferty et al. 2001, Johnson et al.) and Harmony theory (Smolensky 1986)

- OT as a limiting case: higher-ranked constraints have infinitely stronger weights (sometimes realizable with \(\exp\) weighting)
2. Maxent model: Constraints

- Faithfulness (output-to-output)
  \[ F(k) \text{ violated by the change } k \rightarrow t \]
  \[ F(g) \text{ violated by the change } g \rightarrow d \]

- Markedness
  \[ M(C / V) \text{ violated by } C \text{ (k or g) followed by vowel with features } V \text{ ([±high, ±low])} \]

Features: i [\(+\text{high}, \text{−low}\)], e [\(−\text{high}, \text{−low}\)], a [\(−\text{high}, \text{+low}\)]
2. Maxent model: output probability

- Assume that experimental design restricts candidate set to two outputs
  - $y^{\text{pal}}$: palatalized ex. ki:\text{m\text{e}} ... tʃi:\text{m\text{e}}
  - $y^{\text{faith}}$: faithful ex. ki:\text{m\text{e}} ... ki:\text{m\text{e}}

- Probability of palatalizing stimulus $x$
  $$P(y^{\text{pal}}|x) = \frac{H(x,y^{\text{pal}})}{H(x,y^{\text{pal}})+H(x,y^{\text{faith}})}$$

where $H(y|x)$ is the harmony of $y$ given $x$ ...
2. Maxent model: harmony

- Harmony is an exp function of the sum of weighted ($\lambda_k$) constraint violations ($f_k$)

$$H(x, y) = \exp\left(-\sum_{i=1}^{K} \lambda_k f_k(x, y)\right)$$

- $H(x, y^{pal}) = \begin{cases} 
\lambda_{F(k)}, & \text{if } x \text{ begins with } k \\
\lambda_{F(g)}, & \text{if } x \text{ begins with } g
\end{cases}$

- $H(x, y^{faith}) = \sum \{ \lambda_{M(C/V)} | x \text{ begins with CV} \}$

Similarity and bias in phonological learning – p. 37/5
2. Maxent model: training

- Model trained on the same \((x^{(i)}, y^{(i)})\) pairs as the experiment participants (for \(D\) iterations)

- Objective \((\text{Lafferty et al. 2001, McCallum 2003})\)

\[
-D \sum_{i=1}^{N} \log P_{\Lambda}(y^{(i)}|x^{(i)}) \quad \text{(pseudo-likelihood)}
\]

\[
+ \sum_{k=1}^{K} \frac{(\lambda_k - \mu_k)^2}{2\sigma_k^2} \quad \text{(Gaussian prior/regularizer)}
\]

- Convex optimization problem \(\Rightarrow\) global min
2. Maxent model: substantive prior

- Given similarities calculated in (1), how can we incorporate Steriade’s law into the prior?
  \[ \sum_{k=1}^{K} \frac{(\lambda_k - \mu_k)^2}{2\sigma_k^2} \]  
  (from previous slide)

- Assign mean ranking value \( \mu = 0 \) to all M(C/V) constraints — consistent with L1

- Penalize deviation from \( \mu \) in proportion to the cost of changes that satisfy M(C/V)
  \[ \sigma_k \leftarrow \min \{ \text{biased-sim}(\text{velar}_k \rightarrow \text{pal} \mid V_k) \} \]
2. Maxent model: substantive prior

Example calculations

\[ M(k / [+high, -low]) \]  \[ \rightarrow \] *ki
\[ \sigma = \min \{ b(t\text{ʃ}i)\eta(ki,t\text{ʃ}i) \} = \min \{9.23^{-1}\} = 9.23^{-1} \]

\[ M(k / [-low]) \]  \[ \rightarrow \] *ki, ke
\[ \sigma = \min \{ b(t\text{ʃ}i)\eta(ki,t\text{ʃ}i), b(t\text{ʃ}e)\eta(ke,t\text{ʃ}e) \} = \min \{9.23^{-1}, 12.68^{-1}\} = 12.68^{-1} \]

\[ \Rightarrow \] Penalty for deviating from mean (0) weight is approximately 2 times greater for \[ M(k / [-low]) \]
2. Understanding the substantive bias

- All Markedness constraints considered here have a mean/default ranking values of 0. Set output-to-output Faithfulness much higher (10.0).

- Constraints with 0 weight have no effect on the output probabilities, hence the default output is faithful (no palatalization).

- Markedness constraints that compel more perceptually costly palatalization receive greater penalties for deviating from 0 weight.
2. Prior density by $\sigma$

- $sd = 0.11$
- $sd = 0.08$
- $sd = 0.01$
- $sd = 0.05$
- $sd = 0.02$
- $sd = 0.01$
2. Prior penalty by $\sigma$ and weight ($\lambda$)
2. Unbiased alternative

- In the substantively-biased model, all Faith constraints have a relatively large $\sigma (.01)$

- An unbiased version of the model is easy to construct: simply assign the same $\sigma = .01$ to all Markedness constraints as well

- The next section compares the predictions of the biased and unbiased models with the results of Experiment 1 and 2
3. Experiment 1: High

biased
\[ r = .910 \ (82.8\%) \]

cf. unbiased
\[ r = .913 \ (83.4\%) \]
3. Experiment 1: Mid

biased

$r = .859 (73.8\%)$

cf. unbiased

$r = .550 (30.3\%)$
3. Bias and asymmetric generalization

- High condition. The prior $\sigma$ values of $M(k/ [+hi, -lo])$ and $F(k)$ are approx. equal, therefore the constraints “meet in the middle”
  \[ M' < M(k/ [+hi, -lo]) \approx F(k) \]

- Mid condition. But the prior $\sigma$ value of $M(k/ [-hi, -lo])$ is much smaller than that of $F(k)$, so the latter gets “dragged down”
  \[ M' \approx M(k/ [-hi, -lo]) \approx F(k) \]

- Cf. unbiased model never generalizes
3. Experiment 2: Voiceless

biased

\[ r = 0.807 \ (65.1\%) \]

cf. unbiased

\[ r = 0.811 \ (65.8\%) \]
3. Experiment 2: Voiced

biased

\[ r = .920 \ (84.6\%) \]

cf. unbiased

\[ r = .875 \ (76.6\%) \]
Discussion and conclusions
Evidence for substantive bias

- Substantively-biased model, but not unbiased model, accounts for the asymmetric generalization in Exp 1

- Difference in the Mid condition — the only one showing significant generalization — is stark (approx. 30% of variance)

- Modeling of Exp 2 results shows how behavior depends on prior \textit{and} constraints
Discussion: hard vs soft bias

- Substantively-biased phonology does not impose hard restrictions on phono patterns.

- Given sufficient high-quality input, even $g \rightarrow \widehat{d}3 \_ \ a$ is learnable (“overwhelming the prior”).

- This mitigates the empirical problems — attested random rules — that plague more rigid phonetically based proposals.
Discussion: types of generalization

- Why do we observe generalization on the \textit{context}, but not on the \textit{focus}?

- Similar results (for k g) found in speech-error experiments of Goldrick 2004

- Plausibly due to an architectural feature motivated, on average, by economy changes can be identified (listed) with few features \textit{must} generalize to compactly describe contexts
Discussion: relationship to OT

- Empirical difference between (stochastic) OT and maxent w.r.t. harmonic bounding

- These experiments sidestep the issue: candidates violate constraints at most once

- Assuming harmonically bounded candidates always lose, can we formulate stochastic OT learning as a convex optimization problem?
Summary

- What is the relationship between phonetics and phonology?
  
  *Knowledge of phonetics (here, perception) biases the symbolic system in favor of states that are functionally motivated*

- Substantive bias can be formalized with standard methods from psych and ling

- The nature of how learners **generalize** from (possibly quite limited) exposure is the central theoretical issue of phonology