Feature economy and iterated grammar learning

Joe Pater    Robert Staubs

University of Massachusetts Amherst

21st Manchester Phonology Meeting
Overview

- Feature economy: an *unsolvable* problem in standard phonological theory
- Featural simplicity and ease of learning with an incremental MaxEnt model
- Feature economy and contrast in the output of iterated learning
The challenge of feature economy

- A simple example of feature economy  
  (J. Kingston p.c., based on Madiesson and Precoda 1992)

<table>
<thead>
<tr>
<th></th>
<th>[b]</th>
<th>no [b]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[g]</td>
<td>244</td>
<td>11</td>
</tr>
<tr>
<td>no [g]</td>
<td>43</td>
<td>153</td>
</tr>
</tbody>
</table>

\( \chi^2 = 260, \ d.f. \ = 1, \ p < 0.01 \)

- Languages tend to have either both [b] and [g] or neither: it is especially unlikely for a language to have [g] without [b].
- More generally, a segment is more likely if its feature values are shared by other segments.
- In other words, languages tend toward feature economy (Martinet 1968; Clements 2003)
First difficulty - feature economy is a property of systems, not of individual representations or derivations.

How do we express the dependency of [b] on [g] and vice versa?

Standard phonological theories, be they rule- or constraint-based, do not provide a formal mechanism to express such systemic dependencies.
The challenge of feature economy - cont’d

- Second difficulty - feature economy is a *tendency*
- Languages with [p k g] or [p k b] are rare, not unattested
- Standard phonological theories deal only with typological absolutes, not probabilities
The challenge of feature economy - cont’d

- Our claim - we do not need a new kind of phonological grammar to deal with feature economy
- Instead, we incorporate learning into typological explanation (as in fact suggested by Martinet 1968)
- We’ll first show how featurally simple systems are learned more quickly by an incremental MaxEnt learner with conjunctive constraint schema
- We’ll then show how this learning bias can probabilistically affect typology, using iterated learning/agent-based modeling
Featural simplicity and learning bias

- Simplicity bias in the learning of inventories in this space:

<table>
<thead>
<tr>
<th></th>
<th>Voiced</th>
<th>Voiceless</th>
<th>Aspirated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labial</td>
<td>b</td>
<td>p</td>
<td>pʰ</td>
</tr>
<tr>
<td>Coronal</td>
<td>d</td>
<td>t</td>
<td>tʰ</td>
</tr>
<tr>
<td>Dorsal</td>
<td>g</td>
<td>k</td>
<td>kʰ</td>
</tr>
</tbody>
</table>

- One laryngeal feature language: [b d g]
- Two laryngeal feature language: [b t g]
- Three laryngeal feature language: [b t kʰ]
Featural simplicity and learning bias - cont’d

- “General” constraints target each feature (lab, cor, dor, vce, vcl, asp)
- “Specific” constraints target each conjunction (e.g. lab∧vce = [b])
- Constraints can have negative or positive weight
- Probability of a representation proportional to exp(Harmony) (as in Hayes and Wilson 2008)
- Weights are adjusted incrementally to move probability onto the observed forms
- All as in Pater and Moreton 2012
Probability of observed data by epoch

- bdg
- btg
- btK
[bdg] constraint weights by epoch

weight

epoch

vce

g
b
k
d
K
p
P
t
T
asp
vcl
[btg] constraint weights by epoch

- vce
- t
- g
- b
- vcl
- K
- P
- T
- k
- p
- d
- asp
Iterated learning models give an idealized view of language change: agents learn languages from one another in succession. (Hare and Elman 1995, see overviews in Zuraw 2003, Wedel 2011)

Probabilistic biases of learning are reflected in the predicted frequency of resulting languages.
We use just two agents in our iterated learning simulations.

Both agents learn from random data for 100 iterations.

"Production"

1. Agent is chosen at random.
2. That agent chooses a meaning, produces a pronunciation according to its (Maximum Entropy) grammar.
Pairs of words with three potentially pronunciations each:

<table>
<thead>
<tr>
<th>Meaning</th>
<th>Pronunciations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meaning 1</td>
<td>bi</td>
</tr>
<tr>
<td>Meaning 2</td>
<td>pi</td>
</tr>
<tr>
<td>Meaning 3</td>
<td>di</td>
</tr>
<tr>
<td>Meaning 4</td>
<td>ti</td>
</tr>
<tr>
<td>Meaning 5</td>
<td>gi</td>
</tr>
<tr>
<td>Meaning 6</td>
<td>ki</td>
</tr>
</tbody>
</table>

General and specific constraints on consonantal place and laryngeal features, as in the earlier simulation.

Constraints demanding mappings between meaning and sound (e.g. M1 → [pi]).
The agent that is “listening” does not know the intended meaning—it must be inferred.

We use a version of Robust Interpretive Parsing (Tesar & Smolensky, 2000).

The agent chooses the meaning which would be most likely to yield the given pronunciation according to its own grammar.

Learner uses perceived meaning to produce a pronunciation.
Learner’s grammar is updated if its pronunciation does not match its teacher’s.

One step of iterated learning

Teacher Production: \[ M1 \rightarrow \begin{array}{c} a \\ b \end{array} \]

Interpretation:

\[ a \rightarrow M1 \]

\[ M1 \rightarrow M2 \]

Learner Production:

\[ M2 \rightarrow \begin{array}{c} a \\ b \end{array} \]

Update: Output is not \( a \) \( \Rightarrow \) \( M2 \rightarrow a \uparrow \), \( M2 \rightarrow b \downarrow \)
Emergent effects

Resulting languages...

1. Have contrasting pronunciations for meanings.
2. Use near-categorical pronunciations.
3. Distinguish meanings in an economical way.
Homophony avoidance: one meaning per sound

Out of 10,000 runs 9,979 distinguished all 6 meanings.

There are three meaning pairs with three pronunciations for each.

Full contrast is thus expected only \( \frac{8}{27} = 29.63\% \) of the time.
Categorical results: one sound per meaning

With three candidates and initial zero weights, starting probabilities are 33.33%.

In 9,979 of the runs all 6 meanings had a candidate with probability > 50%.
Economy

One way to measure economy is the number of distinct patterns of contrast used.

For example:

- Voiced vs. voiceless suffices to distinguish all 3 pairs. *(Maximally economical.)*

- Voiced vs. voiceless, voiceless vs. aspirated, voiced vs. aspirated could be used. *(Minimally economical.)*
<table>
<thead>
<tr>
<th>Count</th>
<th>1 contrast</th>
<th>2 contrasts</th>
<th>3 contrasts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation</td>
<td>32.14%</td>
<td>57.96%</td>
<td>09.69%</td>
</tr>
<tr>
<td>Chance</td>
<td>11.11%</td>
<td>66.66%</td>
<td>22.22%</td>
</tr>
</tbody>
</table>

$n = 10,000$
Explaining contrast and categorical results

A tendency toward a categorical choice of sound for a meaning occurs when agents learn from one another.

The interpretation step of iterated learning is important to both effects.

A learner must decide which meaning was intended in order to update its grammar.

Use of Robust Interpretive Parsing (or something similar) pushes learners to choose categorical, distinct pronunciations.
RIP is based on preferences of the following type:
Pronunciation $x$ is more likely to originate from meaning $y$

Two logically possible situations

C I Meaning 1 preferred for pronunciation $a$.
Meaning 2 preferred for $b$ and $c$.

C II Meaning 1 not preferred for any pronunciation.
Meaning 2 preferred for $a$, $b$, and $c$. 
Case I

| Prefers | | 1: a, 2: b, c |
|---------|------------------|
| Hear    | a                | b                | c    |
| Interpret | 1                | 2                | 2    |
| Rewards | 1→a              | 2→b              | 2→c  |
| Penalizes | 1→b, 1→c        | 2→a, 2→c        | 2→a, 2→b |
| Prefers | 1: a↑, 2: b↑, c↑ | 1: a↑, 2: b↑, c↓ | 1: a↑, 2: b↓, c↑ |

a is interpreted as Meaning 1. This is reinforced. Other pronunciations become more preferred for Meaning 2.

b or c interpreted as Meaning 2. This is reinforced.

⇒ Pronunciation of Meaning 1 becomes fixed and distinct from Meaning 2’s pronunciations.


**Case II**

<table>
<thead>
<tr>
<th></th>
<th>1: $\emptyset$, 2: $a, b, c$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hear</strong></td>
<td>$a$</td>
</tr>
<tr>
<td><strong>Interpret</strong></td>
<td>2</td>
</tr>
<tr>
<td><strong>Rewards</strong></td>
<td>$2 \rightarrow a$</td>
</tr>
<tr>
<td><strong>Penalizes</strong></td>
<td>$2 \rightarrow b, 2 \rightarrow c$</td>
</tr>
<tr>
<td><strong>Prefers</strong></td>
<td>$2: a \uparrow, b \downarrow, c \downarrow$</td>
</tr>
</tbody>
</table>

Preference for interpreting one pronunciation as Meaning 2 decreases to the point where it’s better interpreted as Meaning 1.

This is Case I.
If homophony is not possible across candidate sets, meanings are distinguished much less often.

For example, if the pronunciations of two words differ in vowels their consonants are not under pressure to be distinct.

In such a simulation only 1,364 pairs were distinguished by their consonants (compare with 29,937).
Feature economy in iterated learning

More economical systems rely on more general constraints, as shown earlier.

These constraints allow faster updates in probability in iterated learning.

A pronunciation is more likely to gain advantage – or to capitalize on chance – if it is helped by a general constraint.

The perception/production process can work across meaning pairs through general constraints.
## Reduction to chance without general constraints

<table>
<thead>
<tr>
<th>Count</th>
<th>1 contrast</th>
<th>2 contrasts</th>
<th>3 contrasts</th>
</tr>
</thead>
<tbody>
<tr>
<td>General constraints</td>
<td>32.14%</td>
<td>57.96%</td>
<td>09.69%</td>
</tr>
<tr>
<td>No general constraints</td>
<td>10.93%</td>
<td>65.99%</td>
<td>22.67%</td>
</tr>
<tr>
<td>Chance</td>
<td>11.11%</td>
<td>66.66%</td>
<td>22.22%</td>
</tr>
</tbody>
</table>

\[ n = 10,000 \]
Feature economy and contrast emerge in the output of iterated learning without any principles specifically demanding economy or contrast.

Boersma and Hamann (2008) demonstrate that dispersion can also emerge from agent interaction

Suggests that phonological grammars might not need to evaluate whole systems (cf. especially Flemming’s 1995 constraints on dispersion and contrast)
Further work

Mackie and Mielke (2011) show that some feature economy effects emerge without distinctive features - what is the relationship between our models?

An account of non-economical features - e.g. Kingston 2005 points out that place is not particularly economical across phonation types.
Acknowledgments

This research was supported by grant BCS-0813829 from the National Science Foundation to the University of Massachusetts Amherst and an NSF Graduate Research Fellowship to Robert Staubs. Thanks to John Kingston and Elliott Moreton for discussion.
Use of the same production and perception grammar means that “ties” are broken sooner.

In Case I this is the ambiguous updates from $b$ and $c$.

In Case II it holds for all three pronunciations.

Eventually one will dominate by random chance. Biasing production probabilities helps this happen faster.