Incremental Maximum Entropy phonotactics and the Shepard complexity hierarchy

Joe Pater
(joint work with Elliott Moreton, UNC Chapel Hill)

NECPhon, University of Maryland

October 6th, 2012
Summary

- Phonological learning experiments repeatedly find that difficulty of learning a pattern increases with the number of features on which it crucially depends.

- This pattern of results is predicted by an incremental MaxEnt phonotactics learner, as well as many other conceivable models.

- Shepard et al.'s 1961 study of non-linguistic concept learning provides a typology of 3-feature patterns on which model predictions differ.

- Moreton and Pertsova’s (2012) study of the full Shepard typology in phonotactic learning provides surprisingly good support for incremental MaxEnt.
Earlier empirical results

- Shepard *et al.* 1961 – a set defined by 3 binary features (e.g. color, shape, size) can be split 6 ways

<table>
<thead>
<tr>
<th>Easiest</th>
<th>Hardest</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="https://via.placeholder.com/150" alt="I" /></td>
<td><img src="https://via.placeholder.com/150" alt="VI" /></td>
</tr>
<tr>
<td><img src="https://via.placeholder.com/150" alt="II" /></td>
<td><img src="https://via.placeholder.com/150" alt="VI" /></td>
</tr>
<tr>
<td><img src="https://via.placeholder.com/150" alt="III" /></td>
<td><img src="https://via.placeholder.com/150" alt="VI" /></td>
</tr>
<tr>
<td><img src="https://via.placeholder.com/150" alt="IV" /></td>
<td><img src="https://via.placeholder.com/150" alt="VI" /></td>
</tr>
<tr>
<td><img src="https://via.placeholder.com/150" alt="V" /></td>
<td><img src="https://via.placeholder.com/150" alt="VI" /></td>
</tr>
</tbody>
</table>

- Using supervised learning experiments (training on both “IN” and “OUT”), Shepard *et al.* 1961 and subsequent concept learning studies find a consistent order of success/speed: $I > II > III, IV, V > VI$
Previous phonological studies have compared Type I with Type II, or Type II with Type VI, and have invariably replicated the $I > II$ and $II > VI$ orders (see Moreton and Pater 2012 LLC for a review, as well as for perhaps the first discussion of the relationship of these studies to the Shepard et al. 1961 et seq. literature).


<table>
<thead>
<tr>
<th>Easier (I)</th>
<th>Harder (II)</th>
<th>Easier (I)</th>
<th>Harder (II)</th>
<th>Easier (II)</th>
<th>Harder (VI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$</td>
<td>$t$</td>
<td>$m$</td>
<td>$b$</td>
<td>$pj$</td>
<td>$tj$</td>
</tr>
<tr>
<td>$k$</td>
<td></td>
<td>$n$</td>
<td>$k$</td>
<td>$p^hj$</td>
<td>$t^hj$</td>
</tr>
<tr>
<td>$b$</td>
<td>$d$</td>
<td>$f$</td>
<td></td>
<td>$pw$</td>
<td>$tw$</td>
</tr>
<tr>
<td>$g$</td>
<td></td>
<td>$z$</td>
<td></td>
<td>$p^hw$</td>
<td>$t^hw$</td>
</tr>
</tbody>
</table>

$[-\text{voice}]$ vs. $[\alpha \text{Cor}, \alpha \text{voice}]$  $[-\text{cont}]$ vs. $[+\text{nas}] \cup [+\text{cont}]$  $[\alpha \text{Lab}][-\alpha \text{Lab}] \text{ vs. } [\alpha \text{Lab} \neg \text{sg}] [-\alpha \text{Lab}] \cup [\alpha \text{Lab} \neg \text{sg}] [\alpha \text{Lab}]$
Incremental MaxEnt phonotactics

- As in Pater, Moreton, and Becker (2008: *BUCLD, NECPhon*) – see Moreton (2012: *JML*) Pater and Moreton (2012: *EFL*) for further details and references

- A simple toy example based roughly on Saffran and Thiessen – a universe of consonants:
  \[ \{[p], [t], [k], [b], [d], [g]\] 

- Two “languages” with subsets of the consonants:
  “Simple” = \{[p], [t], [k]\}
  “Complex” = \{[p], [d], [k]\}
Model of grammar: Hayes and Wilson’s (2008) Maximum Entropy phonotactic grammar, which defines a probability distribution over the space of possible representations (usually words) using weighted constraints (on the intellectual history of MaxEnt in linguistics and machine learning, see Johnson 2007: NECPhon I)

Probability is proportional to the exponential of a representation’s Harmony, or weighted sum of constraint scores

A conjunctive constraint schema: Constraints target each feature value and each conjunction

[+Voice] targets [b, d, g]
[+Voice] & [Coronal] targets [d]
A weighting of constraints yielding a probability distribution over our consonantal universe - note that the behavior of a constraint depends on the polarity of its weight value.

<table>
<thead>
<tr>
<th>Consonant</th>
<th>[+Vce]</th>
<th>[-Vce]</th>
<th>[+Vce] ∧ [+Cor]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[b]</td>
<td>1</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>[d]</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>[g]</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[p]</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>[t]</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>[k]</td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

\[ H = -4, p < .001 \]
\[ H = 4, p = .25 \]
\[ H = -4, p < .001 \]
\[ H = 4, p = .25 \]
\[ H = 4, p = .25 \]
\[ H = 4, p = .25 \]
The Perceptron Update/Delta Rule applied to MaxEnt phonotactics:

- A single learning datum is sampled from the target distribution (Target)
- A single datum is sampled from the distribution defined by the learner’s grammar (Grammar)
- The difference is taken between the vectors of constraint scores (Target - Grammar), and the resultant difference vector is scaled by the learning rate and then added to the vector of constraint weights to get the updated values
- Probability shifted more quickly onto observed forms in the *ptk* case (initial weights zero, rate 0.01)
The “simplicity bias” in this model is due to the structure of the constraint set, and the nature of the update rule – see Moreton (2012: *JML*) for formal analysis.

In this case:

- [+Voice] and [-Voice] separate [p, t, k] from [b, d, g]: [-Voice] quickly gains positive weight and [+Voice] quickly gains negative weight, shifting probability to [p, t, k]

- No single constraint separates [p, d, k] from [b, t, g] – the general voice constraints in fact cause over-generalization – higher probability on [t] than [d] in early stages
We run simulations on the Shepard types by using a constraint set that includes constraints rewarding each single feature value, and all two- and three-feature conjunctions.

The resulting predicted order:

\[ I > III, IV > II, V > VI \]

The \( I > II > VI \) sub-order matches previous work on phonological learning, as well as the supervised concept learning results.

The \( III, IV > II \) sub-order contradicts the supervised concept learning results, and also any expectation that difficulty must increase with the number of crucial features.
The incremental MaxEnt model is essentially equivalent to Gluck and Bower’s (1988) Configural Cue Model of concept learning.

The Configural Cue Model has either been revised (Nosofsky et al. 1994), or abandoned in subsequent concept learning work (see Moreton and Pater 2012 and Pater and Moreton 2012 for references).

But there are many differences between non-linguistic concept learning studies, and phonological learning experiments...
To understand the predictions that incremental MaxEnt makes for the full Shepard hierarchy, we can consider a cube representation of the three dimensions: here top/bottom = black/white, back/front = big/small, left/right = circle/triangle.

Examples are shown for patterns defined over the three features color (black vs white), shape (circle vs triangle), and size (large vs small):

- **Type I**: Only color matters
- **Type II**: Only color and shape matter
- **Types III–V**: All three features matter but some subsets can be described with fewer key features (e.g., white triangles)
- **Type VI**: Every subset requires all three features

Non-linguistic pattern learning studies:
- Supervised learning: See a stimulus, guess whether it's positive or negative, get feedback. Typically, give the same results: I < II < {III, IV, V} < VI (Shepard et al. 1991; Neisser and Weene 1998; Nosofsky et al. 1999; Feldman 2019; Love 2020; Smith et al. 2021).
We have already discussed $I > II$ in terms of the single-feature constraints that separate IN and OUT stimuli in Type $I$, and the absence of such constraints in Type $II$.

On the cubes, these are the faces that join only black dots/corners, or white dots/corners.

The $III, IV > II, V$ ordering is due to the operation of two-feature constraints that target only IN or only OUT stimuli, and can thus shift probability onto only IN or only off OUT.

On the cubes, these are vertices that connect only black dots, only white ones.
Gluck and Bower (1988) discuss the importance of what they term “partially valid cues” – we call them “valid edges” or “valid constraints”.

An easier pattern has more valid edges/constraints.

Note that the valid edges connecting only OUT stimuli mirror those connecting only IN.
In terms of the IN-preferring valid constraints, Type II and V have only 2 – here Type II has the “white triangle” and “black circle” constraints.

Types III and IV have 3 – here Type IV has “white triangle”, “small white” and “small triangle”
Moreton and Pertsova (2012: *NELS*) – phonology with all six Shepard types

Follows the basic methodology of Moreton (2008: *Phonology*; 2012: *JML*) – unsupervised phonotactics, with only IN items in training

Participants told they will learn to pronounce words in an artificial language, and then be tested on ability to recognize words in that language

Listen to and repeat aloud 32 positive stimuli 4 times over, then hear 32 pairs of new stimuli (one IN, one OUT) and try to identify the “word of the language”
For each subject, 3 binary features are randomly chosen from a set of 8 features of the CVCV stimuli (MBROLA synthesized)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Stimulus segment</th>
<th>Consonants</th>
<th>Vowels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$C_1$ $V_1$ $C_2$ $V_2$</td>
<td>$k$ $t$ $g$ $d$</td>
<td>$æ$ $ç$ $i$ $u$</td>
</tr>
<tr>
<td>voiced</td>
<td>± ±</td>
<td>- - + +</td>
<td></td>
</tr>
<tr>
<td>Coronal</td>
<td>± ±</td>
<td>- + - +</td>
<td></td>
</tr>
<tr>
<td>high</td>
<td>± ±</td>
<td>- - + +</td>
<td></td>
</tr>
<tr>
<td>back</td>
<td>± ±</td>
<td>- + - +</td>
<td></td>
</tr>
</tbody>
</table>

These are used to define a Shepard type - for example:

$T_I$ \quad $IN = [+Vce]C_1$ \quad $OUT = [-Vce]C_1$

$T_2$ $IN = [+Vce]C_1&[+High]V_2$ or $[-Vce]C_1&[-High]V_2$
Preliminary results ($n = 139$) – individuals and means
There are a number of potentially confounding factors – we include 3 in our regression model (see Moreton 2012: JML on the first two):

- “Correct first” – a response bias
- “Reduplication”
  1 if the correct choice has the same syllable twice, -1 if the incorrect one does, 0 if neither does
- “All Same Segment”
  Are all the features on one segment?
  1 for Type \(I\), 0 for Types \(III-VI\)
  1 or 0 for Type \(II\)
- Mixed effects logit model with Type as a single six-level variable with Type II as the intercept (random intercept for Subject; only fixed effects shown)

| Coefficient         | Estimate | SE   | Pr(>|z|) |
|---------------------|----------|------|----------|
| Intercept           | 0.185    | 0.132| 0.162    |
| Type I              | 0.126    | 0.247| 0.610    |
| Type III            | 0.392    | 0.169| 0.020    |
| Type IV             | 0.607    | 0.170| < 0.001  |
| Type V              | 0.272    | 0.168| 0.105    |
| Type VI             | -0.108   | 0.167| 0.518    |
| AllSameSeg          | 0.660    | 0.255| < 0.001  |
| Redup               | -0.785   | 0.103| < 0.001  |
| CorrFirst           | 0.240    | 0.065| < 0.001  |
Type II is significantly harder than III and IV, rather than significantly easier than III and IV – matches predictions of the incremental MaxEnt model.

Type I is no easier than Type II when AllSameSeg is controlled (means 0.73 vs. 0.70) – this is unexpected.

Tukey pairwise comparisons also show Type IV was harder than V, which also match the predictions of the learning model.

The comparisons between Type I and Types III – VI are hard to interpret; they compare the advantage of Type I over Type II, controlled for AllSameSegment, with the advantage of Types III – VI over Type II, also controlled for AllSameSegment, but the relevant values of AllSameSegment are different.
The incremental MaxEnt model makes predictions about the relative probabilities of correct choices across types of IN vs. OUT test pairs.

Back to the valid edges – the following figure shows how many valid two-feature constraints target each of the 4 types of IN stimulus within each of the Shepard types (OUT stimuli are symmetrical).

In our earlier Type IV example, small white triangles were targeted by all three two-feature constraints.
For a given IN - OUT pair, there is a possible range of 0 - 6 valid two-feature constraints that separate the stimuli

Only Type IV can have 6, Type VI always has 0

When we run a regression including the number of valid constraints as a factor, it is a highly significant predictor of correctness, and the Types become largely irrelevant (Type 6 is better than its lack of valid constraints would predict)

The next figure shows log-odds of correct choice the number of edge constraints supporting correct over incorrect choice for each stimulus type in each category
- X-axis shows number of relevant valid constraints

![Graph](attachment:image.png)

- Log-odds of correct choice
- Edge constraints supporting correct over incorrect choice

<table>
<thead>
<tr>
<th>X-axis shows number of relevant valid constraints</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>V</td>
</tr>
<tr>
<td>I</td>
<td>IV</td>
</tr>
<tr>
<td>II</td>
<td>III</td>
</tr>
<tr>
<td>III</td>
<td>IV</td>
</tr>
<tr>
<td>IV</td>
<td>III</td>
</tr>
<tr>
<td>V</td>
<td>IV</td>
</tr>
<tr>
<td>VI</td>
<td></td>
</tr>
</tbody>
</table>

Edge constraints supporting correct over incorrect choice
An implementation of the learning model, with Luce choice probabilities converted to log-odds probability correct, and human data as rings.
Conclusions

- The results of the experiment provide surprisingly good support for the predictions of the incremental MaxEnt model with a conjunctive constraint schema.

- The ordering of Types II - V match those predictions, rather than the pattern shown in supervised concept learning, or a “more features is harder” prediction for II vs. III - V.

- The individual stimulus pairs also match the prediction of the model: correctness increases with the number of valid constraints supporting the choice.
Discussion: Relation to concept learning

- These results provide support for a model of concept learning that has long been abandoned, Gluck and Bower’s (1988) configural cue model, a single layer feedforward network, whose “cues” or “features” are set up like our “constraints”

- They also raise the question of why our results are different than those from supervised concept learning (our $IV > II$ was found in unsupervised learning by Love 2002)

- Given the great number of differences between the tasks, an answer requires new experiments

- Preliminary results of Moreton and Pertsova’s “cake” experiment show a similar Type order in unsupervised learning of analogous non-linguistic patterns
Discussion: Modeling phonological knowledge

- Prior research using a MaxEnt model of phonotactics has shown it to be useful in modeling native speaker judgments (Hayes and Wilson 2008, Daland et al. 2011, Kager and Pater 2012)
- We have shown that by implementing it with an incremental learning algorithm, we can generate predictions for phonotactic learning
- Along with refining our current model to make it better fit the data, it seems particularly promising to fit priors to the data in batch MaxEnt or Bayesian models (see related work by Culbertson and Smolensky)
Discussion: Further experiments

- Since our learning model is probabilistic, it can be straightforwardly be applied to cases in which the learning data is probabilistic; see work in progress by Robert Staubs.

- Learning of phonological alternations in some sense supervised (e.g. observation of plural [-s] in some environment implies not [-z]) – raises interesting modeling and experimental questions.

- Relationship of lab learning to natural acquisition? Preliminary ERP data from an experiment with Lisa Sanders finds a P600 signature for lab-learned phonotactic violations, as has been found for naturally learned phonotactics.
Discussion: Modeling phonological typology

- Phonological typology shows skews to featurally simple inventories and patterns (e.g. Clements 2003)

- Are these the result of biases in phonological learning? (Martinet 1968 and Bach and Harms 1972)

- We are currently working with iterated learning models to try to match typology; some initial results are presented in Pater and Moreton (2012)

- We also would like use modeling to explore the hypothesis that the role of phonetic substance in typological patterning can be reduced to what Moreton (2008) call channel bias.
Acknowledgments

This research was supported by grant BCS-0813829 from the National Science Foundation to the University of Massachusetts Amherst.

See acknowledgments from M&P (2012) and P&M (2012) for thanks to those who contributed to this work; thanks to the UMass Amherst computational phonology group for feedback on an earlier version of this presentation.