Systemic Simplicity in Phonological Learning and Typology

Joe Pater
University of Massachusetts, Amherst

KNAW Academy Colloquium on Language Acquisition and Optimality Theory
Amsterdam, July 2\textsuperscript{nd}, 2009
“Systemic simplicity”

(1) The complexity (markedness) that a structure adds to a language depends on other aspects of the system.

For example:

(2) a. [g] adds less complexity to a language that contains [b] than to a language that has only [p]

b. [batá] adds less complexity to a language in which all other words are stressed on the final syllable than to a language in which they are stressed elsewhere

The account to be proposed here is a formalization of Martinet’s (1968: 483) explanation for cases like (2a) (what he called “featural economy” – see further Clements 2003 and references therein).
...each of [the features] being more frequent in speech, speakers will have more occasions to perceive and produce them, and they will get anchored sooner in the speech of children. A phoneme that is integrated into one of those bundles of proportional oppositions which we call ‘correlations’ will in principle be more stable than a non-integrated phoneme....

.../ɵ/ and /ð/ in English have been preserving for centuries their practically worthless opposition simply because they are perfectly integrated in the powerful correlation of voice....
The formal account part 1:

(4) Simple patterns learned more quickly by a gradual learning algorithm for Harmonic Grammar
(on HG-GLA see Boersma and Pater 2008, Jesney and Tessier NLLT; on HG see Smolensky and Legendre 2006 and Pater 2009; for the origins of this approach, see Rumelhart and McClelland 1986)

The formal account part 2:

(5) Learning biases have an effect on typology when incomplete learning is perpetuated and amplified through agent-based interaction
(see Boersma and Hamann 2008 and van Leussen 2008 for other applications of GLA+agent-based modeling; see Zuraw 2003 for an overview of agent-based modeling in phonology; for the origins of this approach, see Hare and Elman 1995)
1. Two cases of systemic simplicity

**Case 1:** The statistical dependence between [b] and [g] in the phoneme inventories of UPSID-92 (Maddieson and Precoda 1992; counts by John Kingston p.c.)

<table>
<thead>
<tr>
<th></th>
<th>[b]</th>
<th>no [b]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[g]</td>
<td>244 (O/E = 1.52)</td>
<td>11 (O/E = 0.12)</td>
</tr>
<tr>
<td>no [g]</td>
<td>43 (O/E = 0.34)</td>
<td>153 (O/E = 2.13)</td>
</tr>
</tbody>
</table>

\[ \chi^2 = 260, \text{ d.f.} = 1, \ p < 0.01 \]

Languages tend to have either both [b] and [g] or neither: it is especially unlikely for a language to have [g] without [b].

(7) Many languages have no sounds from the set [bdg], but if a language has one of them, it is likely to have all of them. These sounds are all [+voice]...having the full [bdg] set in a language maximizes the cross-classificatory effect of that distinctive feature.

Clements (2003) documents a number of other cases of feature economy, and provides a descriptive formalization.

Clements notes that feature economy has been very little studied. This is perhaps not surprising, given that phonological theory appears to provide no explanation for it.

Clements’ does not provide an account of why languages should tend towards economical use of features.
Case 2: Stress regularization

Words that disobey an otherwise general pattern of stress placement are often regularized in acquisition (e.g. Hochberg 1988, Fikkert 1994, Nouveau 2004), and in diachrony (e.g. Phillips 1998).

As Fidelholtz (1979) notes, languages differ in the rate of regularization by adult speakers.

(8) Polish: Generally penultimate, with few exceptions, frequent regularization

     English: Complicated pattern, many “exceptions”, infrequent regularization
Fidelholtz (1979: 58) points out that English and Polish would be analyzed identically in phonological theory:

(9) It appears to be a problem for linguistic theory that there is nothing in the formal description of Polish stress which would indicate that Polish is a “penultimate-stress’ language, as compared with the similar rules in English, which is essentially a free-stress language.

I believe that Fidelholtz’s problem remains to be completely resolved.
One sometimes finds statements that exceptionality adds complexity to a grammar, as in Pater (to appear):

(10) ...languages maintain a certain degree of homogeneity, for example by regularizing exceptional patterns over time. This would receive a relatively natural explanation in indexation theory, if one assumes that constraint clones and indices add complexity to the system that is sometimes eliminated over the course of time.

As far as I know, diachronic loss of exceptional lexical specification has never been formalized in generative phonology, under any theory of exceptionality.
2. Complexity leads to slower learning

We’ll first see that systemically complex forms are learned relatively slowly under standard assumptions about gradual learning in OT/HG (Boersma 1998, Boersma and Pater 2008)

I assume “Exponential Noisy HG” as a grammar model:

(11) a. Numerically weighted constraints, as in HG
    b. Noise added to generate variation and realistic learning curves, as in stochastic OT
    c. Weights are exponentiated before evaluation to keep them strictly positive

For learning, I assume HG-GLA, the adaptation to HG of perceptron/delta rule/stochastic gradient descent (in the current examples, it is also identical to Boersma’s OT-GLA)
The initial state in simulation 1 (mean weights and harmony, and candidates probability with noise 0.2)

\[
\begin{array}{c|ccc|c}
\text{/ga/} & *\text{Voice} & *[g] & \text{Ident-Voice} \\
& e^1 = 2.72 & e^1 = 2.72 & e^0 = 1 \\
0 [ga] & -1 & -1 & -5.44 \\
1 [ka] & & -1 & -1 \\
\end{array}
\]

The learning rate (plasticity) in the simulations was 0.01. After one exposure to /ga/ → [ga], the grammar would be:

\[
\begin{array}{c|ccc|c}
\text{/ga/} & *\text{Voice} & *[g] & \text{Ident-Voice} \\
& e^{0.99} = 2.69 & e^{0.99} = 2.69 & e^{0.01} = 1.01 \\
0 [ga] & -1 & -1 & -5.38 \\
1 [ka] & & -1 & -1.01 \\
\end{array}
\]
Learners were exposed to one of two languages:

(14) \textit{bdg}

/ba/ \ [ba]  
-da/ \ [da]  
/ga/ \ [ga]

(15) \textit{ptg}

/pa/ \ [pa]  
/ta/ \ [ta]  
/ga/ \ [ga]

The following chart graphs \%correct [ga], for each of the two languages, at intervals of 50 pieces of learning data.
Correct [g] over time

% correct

n learning data
Recall the initial state:

\[
(16) \quad \begin{array}{|c|c|c|c|}
\hline
/ga/ & *\text{Voice} & *[g] & \text{Ident-Voice} \\
\hline
& e^1 = 2.72 & e^1 = 2.72 & e^0 = 1 \\
\hline
0 \ [ga] & -1 & -1 & -5.44 \\
1 \ [ka] & & -1 & -1 \\
\hline
\end{array}
\]

After one exposure to /ba/ → [ba], the grammar is updated as follows:

\[
(17) \quad \begin{array}{|c|c|c|c|}
\hline
/ga/ & *\text{Voice} & *[g] & \text{Ident-Voice} \\
\hline
& e^{0.99} = 2.69 & e^1 = 2.72 & e^{0.01} = 1.01 \\
\hline
0 \ [ga] & -1 & -1 & -5.41 \\
1 \ [ka] & & -1 & -1.01 \\
\hline
\end{array}
\]

Learning on /ba/ → [ba] generalizes to /ga/ → [ga] through the change in values of general *VOICE and IDENT-VOICE.
Simulation 2 used a very simple set of stress constraints.

(18) STRESS-INITIAL (violated by word-final stress)
    STRESS-FINAL (violated by word-initial stress)
    STRESS-INITIAL-L (violated by final stress on word L)
    STRESS-FINAL-L (violated by final stress on word L)

All constraints started at weight zero.

The candidate sets consisted of initially and finally stressed words (bisyllables); the learning data were evenly randomly sampled over the words from the following lexica:

(19) One-final
    9 words with initial stress, 1 with final stress

    All-final
    10 words with final stress
This sort of effect should generalize to other markedness constraints in a specific-to-general relationship, for instance:

(20) **HEAD-INITIAL** (violated by a non-head-initial phrase)
    
    **HEAD-FINAL** (violated by a non-head-final phrase)
    
    **HEAD-INITIAL-XP** (violated by a non-head-initial phrase of category X)
    
    **HEAD-FINAL-XP** (violated by a non-head-final phrase of category X)

*Data?*

In a third language, initial and final stress were equally represented in the learning data:

(21) **Five-final**
    
    Five initially stressed words, five finally stressed ones
Correct final stress over time

% correct

n learning data

one-final
five-final
3. Agent-based modeling amplifies simplification

While simulations like the preceding may be useful in modeling learning data, they do not in themselves explain typological skews or diachronic changes to simplicity. Although a complex pattern is learned more slowly, given enough data, it will be learned completely.

We’ll now see that incorporating interaction between learners (agent-based modeling) can lead to the amplification of the probability of a simplified form.

(22) i. Learners exposed to some number of learning data from the parent distribution (“Childhood”)
    
    ii. Learners then act as “teachers” to one another (“Adolescence”)
The basic learning parameters were the same as in the previous simulations:

(23) Exponential Noisy HG, noise = 0.2
    Learning rate = 0.01

Phase 1:

(24) Two learners each supplied with a different random sequence of 200 pieces of learning data.

Phase 2 (10,000 iterations):

(25) a. One grammar is picked at random and produces a random input-output pair
    b. The other grammar is updated based on that learning datum
Agents after interaction

Number of learners

Percentage correct [g]

0- 10- 20- 30- 40- 50- 60- 70- 80- 90-

ptg
bdg

0 1 2 3 4 5 6 7

0 0 0 0 0 0 0 0 0 0
Typical accuracy on [g] after 200 pieces of data:

\[(26) \quad \text{bdg} \quad 60\% \text{ correct} \]
\[(\text{ptg} \quad 6\% \text{ correct}) \]

In the post-adolescent results shown above, our \textit{bdg} speakers tended to produce $> 90\%$ correct.

The interaction between agents tends to amplify an existing skew (tends toward reduction of error)

In the stress simulations, the initial learning phase consisted of 400 pieces of data. Typical scores at that point:

\[(27) \quad \text{One-final} \quad 40\% \text{ correct final stress} \]
\[(\text{Five-final} \quad 70\% \text{ correct final stress}) \]
\[(\text{All-final} \quad 98\% \text{ correct final stress}) \]
In these simulations, final stress is more stable in a language in which all words are stressed on the final syllable than in a language in which 10% are.

The stability of final stress in a language with 50% of the words stressed on the final syllable is in between the two.

These simulations thus mirror the apparent correlation between proportion of regular words in the lexicon and probability of regularization.

Diachronic studies of stress regularization find that word frequency negatively correlates with probability of regularization (on English, see Phillips 1984, Sonderegger and Niyogi 2009).

A fourth language:

(28) **One-final\*five**

One word with final stress, with 5 times frequency
5. Conclusions

As far as I know, this is the only fully explicit account of phenomena termed feature economy, and the only generative account of stress regularization (where “generative” means the crucial use of a phonological grammar).

In both cases, systemic simplification comes from the workings of a general constraint (on voicing, on stress placement). The effect “emerges” from the workings of the learning algorithm and agent-based interaction. Other cases with similarly structured constraint sets should fall to similar treatments.

This approach can also be combined with a constraint induction algorithm to yield an account of the ease of learning “simple generalizations” (Pater, Moreton, and Becker BUCLD 2008)
Other accounts of stress regularization in child speech:

(29) a. Analogical (Eddington 2000)
    b. Exemplar-based (Gillis, Daelemans and Durieux 2000)

Account of diachronic shifts in stress:

(30) Dynamical systems (Sonderegger and Niyogi 2009)

With further development of the agent-based GLA model, we should be able to compare it to these alternatives.

One reason to be optimistic:

(31) The theory of grammar (OT/HG) has been applied to a wide range of phenomena; when coupled with this approach to learning, we get a wide range of cases of emergent systemic simplification