Three Correlates of the Typological Frequency of Quantity-Insensitive Stress Systems

Max Bane and Jason Riggle
{bane, jriggle}@uchicago.edu

Department of Linguistics / Chicago Language Modeling Lab
University of Chicago
http://clml.uchicago.edu

June 19, 2008
SIGMORPHON 2008
The Ohio State University
A characteristic property of Language

- The distribution of linguistic properties is very uneven typologically.

Examples

- Sound inventories
  - 921 distinct phonemes found in a sample of 451 languages; average language uses only about 30 (Maddieson 1984).
  - Some sounds extremely common (≈universal): [m], [k]; others extremely rare: [口中], [œ]

- Stress patterns
  - 26 distinct QI stress patterns in a sample of 306 languages (Heinz 2007).
  - But over 60% of the languages use one of just 3 patterns.

- Morphosyntactic, semantic properties
Previous typological research

- The goal of most typological studies is to construct a theory or model that predicts:
  - As many attested patterns/languages as possible, and
  - As few unattested patterns/languages as possible
- The “inclusion-exclusion” criterion. (cf. precision/recall)
- Few attempt the additional (harder) task of predicting:
  - The typological frequencies of attested patterns
- (Though see, e.g., Liljencrants & Lindblom 1972, de Boer 2000, and others)
Focus on the typology of quantity-insensitive (QI) stress systems, as collected by Bailey (1995), Gordon (2002), and combined by Heinz (2007).

Question: to what extent can an existing optimality theoretic model of QI stress, developed by Gordon (2002) to meet the inclusion-exclusion criterion, predict the frequencies of stress patterns?
Focus on the typology of quantity-insensitive (QI) stress systems, as collected by Bailey (1995), Gordon (2002), and combined by Heinz (2007).

Question: to what extent can an existing optimality theoretic model of QI stress, developed by Gordon (2002) to meet the inclusion-exclusion criterion, predict the frequencies of stress patterns?

Answer: frequency is surprisingly well predicted by three things:

- the $n$-gram entropy of a stress pattern,
- its “confusability”,
- the number of constraint rankings consistent with it
Assumptions, definitions

- Stress pattern
  - Any “culminative” accentual system; there is one most prominent accentual unit.
  - Any given unit may bear primary or secondary stress; exactly one primary stressed unit per prosodic word.
- I assume that the stress-bearing unit is the syllable.
- Quantity-insensitive stress pattern
  - The assignment of stresses to a word’s syllables depends only on the number of syllables, not on the contents of the syllables.
Examples

- Notation: $\sigma =$ unstressed syllable, $\acute{\sigma} =$ primary stressed syllable, $\grave{\sigma} =$ secondary stressed syllable.
- Penultimate primary stress (Mohawk, Albanian, ...)
  - 2 syl. word: $\acute{\sigma}\sigma$
  - 3 syl. word: $\sigma\acute{\sigma}\sigma$
  - 4 syl. word: $\sigma\sigma\acute{\sigma}\sigma$
  - 5 syl. word: $\sigma\sigma\sigma\acute{\sigma}\sigma$
  - ...
- Even-numbered from right, leftmost primary (Malakmalak)
  - 2 syl. word: $\acute{\sigma}\sigma$
  - 3 syl. word: $\sigma\acute{\sigma}\sigma$
  - 4 syl. word: $\acute{\sigma}\sigma\grave{\sigma}\sigma$
  - 5 syl. word: $\sigma\acute{\sigma}\sigma\grave{\sigma}\sigma$
  - 6 syl. word: $\acute{\sigma}\sigma\grave{\sigma}\sigma\grave{\sigma}\sigma$
  - ...
The stress typology

- Heinz’s (2007) dissertation:
  - Combines two previous typologies by Bailey (1995) and Gordon (2002), collected from other studies and primary source grammars.
  - Samples a total of 422 languages with stress, 306 of which have quantity-insensitive systems.
  - Sample chosen to balance representation of major language stocks.
  - Caveats:
    - Only “dominant” stress patterns represented.
    - Lexical exceptions, subregularities not included.
    - Some may contribute multiple stress patterns. E.g., Lenakel nouns vs Lenakel verbs — counts as two “languages.”
Between the 306 languages with QI stress, there are 26 distinct stress patterns, distributed as follows:
The stress typology

- A very skewed distribution (Zipfian; Gauss-Newton: $R^2 = .809, p < 0.001$).
- The top three most common patterns, together representing a majority of languages surveyed:
  - Fixed final stress (24.2% of systems)
  - Fixed initial stress (22.5% of systems)
  - Fixed penultimate stress (19.6% of systems)
- Of $N = 306$ sampled languages, $n_1 = 13$ have patterns attested only once.
- Good-Turing estimate (Good 1953):
  - We should reserve about $n_1 / N = 4.3\%$ of probability/frequency-mass for unseen patterns.
  - $\Rightarrow$ a fairly exhaustive sample.
Gordon’s (2002) model

- Gordon presents a model of QI stress that aims to include as many attested languages, and exclude as many unattested, as possible.
- Optimality Theoretic.
  - 12 constraints (plus 1 “meta” constraint) on the assignment of stress to syllables.
  - $12! = 479,001,600$ possible QI stress grammars.
<table>
<thead>
<tr>
<th>Constraint(s)</th>
<th>Penalizes...</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ALIGNEDGE</strong></td>
<td>each edge of the word with no stress</td>
</tr>
<tr>
<td><strong>ALIGN(x₁, L/R)</strong></td>
<td>each stressed syllable for each other syllable between it and the left/right edge</td>
</tr>
<tr>
<td><strong>ALIGN(x₂, L/R)</strong></td>
<td>each primary stressed syllable for each secondary stressed syllable between it and the left/right edge</td>
</tr>
<tr>
<td><strong>NONFINALITY</strong></td>
<td>the last syllable if it is stressed</td>
</tr>
<tr>
<td><strong>LAPSE</strong></td>
<td>each adjacent pair of unstressed syllables</td>
</tr>
<tr>
<td><strong>CLASH</strong></td>
<td>each adjacent pair of stressed syllables</td>
</tr>
<tr>
<td><strong>EXTLAPSE</strong></td>
<td>each occurrence of three consecutive unstressed syllables</td>
</tr>
<tr>
<td><strong>LAPSELEFT/RIGHT</strong></td>
<td>the left/right-most syllable if more than one unstressed syllable is between it and the left/right edge</td>
</tr>
<tr>
<td><strong>EXTLAPSERIGHT</strong></td>
<td>the right-most syllable if more than two unstressed syllables are between it and the right edge</td>
</tr>
</tbody>
</table>
Multiple constraint-rankings may, on the surface, specify the same QI stress pattern.

The 12! possible constraint-rankings (i.e., grammars) correspond to only 152 distinct QI stress patterns (looking at words up to 8 syllables long).
Gordon’s (2002) model

- Multiple constraint-rankings may, on the surface, specify the same QI stress pattern.
- The 12! possible constraint-rankings (i.e., grammars) correspond to only 152 distinct QI stress patterns (looking at words up to 8 syllables long).
- 24 of the 26 attested patterns are predicted by Gordon’s model.
  - Undergenerates by $26 - 24 = 2$ patterns (those of Bhojpuri and Içuã Tupi)
  - Overgenerates by $152 - 24 = 128$ patterns.
- Gordon’s choice of constraints offers a reasonable current baseline for predicting QI stress typology.
What about frequency?

- Gordon gives a recipe for constructing QI stress grammars
  - Capable of describing 152 distinct QI patterns
  - Usual interpretation: these are the QI stress patterns predicted to be possible in human languages.
- Gordon makes no claims about how frequent these possible stress patterns should be expected to occur.
Our claim

Even though Gordon’s goal is only to design a set of constraints that (nearly) satisfy the inclusion-exclusion criterion, his model can also be used to predict attestation and frequency by computing three quantities:

- $n$-gram entropy,
- “confusability vectors”,
- the “ranking-volume” of each predicted language.
Motivation

- Hypothesized principle of least effort or simplicity
- Predicts that cross-linguistically frequent patterns should be simpler (according to some metric) than rare ones
- Predicted patterns (like the 152 predicted by Gordon) should be more likely to be attested if they are less complex

We find evidence consistent with both predictions for QI stress

- According to at least one information theoretic definition of complexity
Compute conditional transition probabilities between the symbols (σ, ́σ, ̀σ,#) of each word in a stress pattern (2–8 syllables), given the previous $n – 1$ symbols.

Use these to calculate the Shannon entropy of the pattern.

- The number of bits necessary to efficiently describe the pattern, according to the $n$-gram probability model.

Patterns where it’s easy to predict next symbol based on previous $n – 1$ symbols: low entropy = less “complex”.

Trigram entropy ($n = 3$) correlates with pattern attestation and frequency...
Trigram Entropy

- Attestedness of Gordon’s 152 predicted patterns
- Mann-Whitney $U$-test (two tailed): $U = 1196$, $p = 0.0106$
Typological frequency of attestation:

- High frequency (above median) patterns have significantly lower trigram entropy than low frequency (below median) patterns. $U = 51.5$, $p = 0.0428$
Some patterns are very similar to each other, or confusible

- Must observe very long forms (i.e., many syllables) in order to distinguish them from each other

Example: Albanian and Malakmalak

- 2 syl. word: $\acute{o}\sigma$ vs $\acute{o}\sigma$
- 3 syl. word: $\sigma\acute{o}\sigma$ vs $\sigma\acute{o}\sigma$
- 4 syl. word: $\sigma\sigma\acute{o}\sigma$ vs $\acute{o}\sigma\sigma\sigma$
- 5 syl. word: $\sigma\sigma\sigma\sigma\sigma$ vs $\sigma\sigma\sigma\sigma\sigma$
- 6 syl. word: $\sigma\sigma\sigma\sigma\sigma\sigma$ vs $\sigma\sigma\sigma\sigma\sigma\sigma$

Identical stress assignments for two and three syllable words

- Must see words of 4+ syllables to tell them apart
Confusability Vectors

- Potential factor in learnability of stress patterns. Patterns that are easily identified at short word-lengths from among the competing possibilities may be more faithfully acquired by learners, and thus more “typologically stable”:
  - more likely to be attested, more frequently attested, or both
- Test: construct **confusability vector** for each of Gordon’s 152 predicted patterns
- Ex: Albanian’s (fixed penultimate primary) stress pattern:
  - \(\langle 101, 39, 10, 0, 0, 0, 0 \rangle\)
  - Confusable with 101 other patterns at 2 syllables, with 39 at 3 syllables, with 10 at 4 syllables, with none at 5+ syllables
- **Confusability sum**: sum all the numbers in the vector
- **Confusability pivot**: number of syllables at which pattern is uniquely identifiable
Attested patterns have significantly lower confusability pivots than $(U = 1005.5, \ p < 0.001)$; but no significant difference for confusability sums.

Frequency of attestation significantly predicted by linear combination of confusability sums and pivots $(p < 0.05, \ R^2 = 0.271)$. 
Let $\text{CON}$ be a set of constraints.

Let $L$ be a language generated by some grammar over (i.e., ranking of) $\text{CON}$; in this case, a QI stress pattern.

Then define the ranking volume ($r$-volume) to be the number of rankings of $\text{CON}$ that generate $L$.

$r$-volumes can be computed efficiently using Prince’s (2002) Elementary Ranking Conditions and a recursive algorithm described by Riggle (2008).

To find the $r$-volumes of the QI stress patterns predicted by Gordon’s model:

- Implemented Gordon’s constraints as finite state transducers, giving a finite state OT model (Riggle 2004),
- Generated the factorial typology of stress systems implied by the constraints, replicating and confirming Gordon’s findings,
- Computed the $r$-volume of each of the 152 languages predicted by the model.
Majority of the ranking volume is concentrated in a few stress patterns.

Almost perfectly Zipfian distribution
\( p < 0.001, R^2 = 0.968 \) according to Gauss-Newton
What can $r$-volume predict?

The $r$-volumes of the stress patterns generated by Gordon’s model prove to correlate well with:

- Which predicted patterns are actually attested, and
- The frequencies of the attested patterns.
r-volume correlates with attestation
The $r$-volumes of attested stress patterns are significantly greater than those of the unattested.

Mann-Whitney $U$ test: $U = 2113.5$, $\rho = 71.2\%$, $p < 0.01$
$r$-volume correlates with frequency

- Linear regression: Frequency $\propto r$-volume
- $R^2 = 0.712, p < 0.001$
- But overly sensitive to outliers (Q-Q plot nonlinearity, significant Cook’s distance).
Frequency is more robustly predicted by a nonlinear function of the logarithm of the $r$-volume.

Exponential regression: $R^2 = 0.704$, $p < 0.001$
Surprising, since these correlates depend entirely on the set of languages generated by Gordon’s model.

R-volume findings are consistent with an OT learning model in which...

- Learners randomly select a grammar consistent with the evidence they’ve seen so far.
  - \( \Rightarrow \) learners will tend to choose languages with frequency proportional to R-volume

- Or where learners select from the languages consistent with their observations according to R-volume.

In progress: predicting the typological frequencies with iterated learning simulations.
Discussion

- A different perspective on typological models
  - One usually wants overgeneration to be as small as possible (inclusion-exclusion)
  - And as non-“pathological” as possible...i.e., similar to what is attested.
A different perspective on typological models
- One usually wants overgeneration to be as small as possible (inclusion-exclusion)
- And as non-“pathological” as possible. . . i.e., similar to what *is* attested.

But maybe large amounts of overgeneration are not a problem
- as long as unattested languages are systematically, detectably different from attested.
- i.e., as long as overgeneration *is* pathological (systematically).
Future/current work

- Other models/analyses of QI stress and whether these correlates are still significant
- What are the crucial properties of a constraint set that determine the distribution of $r$-volumes?
  - Anttila’s (2006) T-orders?
- How does Gordon’s set of predicted stress patterns compare to random sets of patterns in terms of these correlates?
- Other correlates of frequency?