Grammatical Correlates of Cross-linguistic Frequency: Quantity-Insensitive Stress

Max Bane
bane@uchicago.edu

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

January 23, 2009
Council on Advanced Studies
Workshop on Language, Cognition, and Computation
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: [ʁ], [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: how do the typological predictions of three different models of QI stress relate to the set of attested systems and their cross-linguistic frequencies?
This talk

Focus on the typology of quantity-insensitive (QI) stress systems, as collected by Bailey (1995), Gordon (2002), and combined by Heinz (2007).

Question: how do the typological predictions of three different models of QI stress relate to the set of attested systems and their cross-linguistic frequencies?

Answer: attestedness and frequency are correlated with three things:

1. the $n$-gram entropy of a stress pattern,
2. its “confusability” with other predicted patterns (for at least one model),
3. the number of parameter settings that specify it in the models.
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:** $\sigma = \text{unstressed syllable}, \dot{\sigma} = \text{primary stressed syllable}, \ddot{\sigma} = \text{secondary stressed syllable}.$

- **Penultimate primary stress (Mohawk, Albanian, ...)**
  - 2 syl. word: $\dot{\sigma}\sigma$
  - 3 syl. word: $\sigma\dot{\sigma}\sigma$
  - 4 syl. word: $\sigma\sigma\dot{\sigma}\sigma$
  - 5 syl. word: $\sigma\sigma\sigma\dot{\sigma}\sigma$
  - ... 

- **Even-numbered from right, leftmost primary (Malakmalak)**
  - 2 syl. word: $\dot{\sigma}\sigma$
  - 3 syl. word: $\sigma\dot{\sigma}\sigma$
  - 4 syl. word: $\dot{\sigma}\sigma\dot{\sigma}\sigma$
  - 5 syl. word: $\sigma\dot{\sigma}\sigma\dot{\sigma}\sigma$
  - 6 syl. word: $\dot{\sigma}\sigma\dot{\sigma}\sigma\dot{\sigma}\sigma$
  - ...
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.”
The stress typology

Between the 306 languages with QI stress, there are 26 distinct stress patterns, distributed as follows:

Frequencies of Attested Stress Patterns

<table>
<thead>
<tr>
<th>Stress Pattern</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>L076</td>
<td>0.00</td>
</tr>
<tr>
<td>L118</td>
<td>0.05</td>
</tr>
<tr>
<td>L004</td>
<td>0.10</td>
</tr>
<tr>
<td>L132</td>
<td>0.15</td>
</tr>
<tr>
<td>L110</td>
<td>0.20</td>
</tr>
<tr>
<td>L044</td>
<td>0.25</td>
</tr>
<tr>
<td>L008</td>
<td></td>
</tr>
<tr>
<td>L022</td>
<td></td>
</tr>
<tr>
<td>L143</td>
<td></td>
</tr>
<tr>
<td>L065</td>
<td></td>
</tr>
<tr>
<td>L054</td>
<td></td>
</tr>
<tr>
<td>L095</td>
<td></td>
</tr>
<tr>
<td>L040</td>
<td></td>
</tr>
<tr>
<td>L077</td>
<td></td>
</tr>
<tr>
<td>L071</td>
<td></td>
</tr>
<tr>
<td>L033</td>
<td></td>
</tr>
<tr>
<td>L113</td>
<td></td>
</tr>
<tr>
<td>L037</td>
<td></td>
</tr>
<tr>
<td>LXX1</td>
<td></td>
</tr>
<tr>
<td>L047</td>
<td></td>
</tr>
<tr>
<td>L042</td>
<td></td>
</tr>
<tr>
<td>L089</td>
<td></td>
</tr>
<tr>
<td>LXX2</td>
<td></td>
</tr>
<tr>
<td>LXX3</td>
<td></td>
</tr>
<tr>
<td>L082</td>
<td></td>
</tr>
<tr>
<td>L084</td>
<td></td>
</tr>
</tbody>
</table>
The stress typology

- A very skewed distribution (power law? Gauss-Newton regression to Zipf’s law: $R^2 = .809, p < 0.001$).
- 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.
Modeling QI stress: Overview

- Three models of QI stress
  - Gordon’s (2002) OT model
  - A dynamic linear model
  - A weighted (“harmonic grammar”) version of Gordon’s model

- All three overgenerate and undergenerate
  - “Overgeneration” = predicting unattested patterns
  - “Undergeneration” = not predicting attested patterns
  - Just worrying about overgeneration today

- Within the predictions of each model:
  - What separates the attested patterns from the unattested?
  - What separates the cross-linguistically frequent from the infrequent?
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.
  - A grammar is a ranking of the constraints.
  - For each word length ($n = 2, \ldots, 8$ syllables), choose the stress assignment ($\in \{\sigma, \dot{\sigma}, \ddot{\sigma}\}^n$) that best satisfies the highest ranked constraints.
<table>
<thead>
<tr>
<th>Constraint(s)</th>
<th>Penalizes...</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALIGNEDGE</td>
<td>each edge of the word with no stress</td>
</tr>
<tr>
<td>ALIGN($x_1$, L/R)</td>
<td>each stressed syllable for each other syllable between it and the left/right edge</td>
</tr>
<tr>
<td>ALIGN($x_2$, L/R)</td>
<td>each primary stressed syllable for each secondary stressed syllable between it and the left/right edge</td>
</tr>
<tr>
<td>NONFINALITY</td>
<td>the last syllable if it is stressed</td>
</tr>
<tr>
<td>*LAPSE</td>
<td>each adjacent pair of unstressed syllables</td>
</tr>
<tr>
<td>*CLASH</td>
<td>each adjacent pair of stressed syllables</td>
</tr>
<tr>
<td>*EXTLAPSE</td>
<td>each occurrence of three consecutive unstressed syllables</td>
</tr>
<tr>
<td>*LAPSELEFT/RIGHT</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>*EXTLAPSERIGHT</td>
<td>the right-most syllable if more than two unstressed syllables are between it and the right edge.</td>
</tr>
</tbody>
</table>
Gordon’s (2002) model

- **Meta constraint:**
  - Only one of $\text{ALIGN}(x_2, L)$, $\text{ALIGN}(x_2, R)$ is active at once.
  - $\Rightarrow$ two sub-models: one without $\text{ALIGN}(x_2, L)$, one without $\text{ALIGN}(x_2, R)$
  - Reason: preserves an apparent universal property of stress systems: secondary stress always appears to one side (left or right) of primary stress, without vascillating back and forth across word lengths.
Gordon’s (2002) model

- **Typological predictions**
  - Each sub-model allows 11! possible grammars.
  - $2 \cdot 11! = 79,833,600$
  - Multiple constraint-rankings may, on the surface, specify the same QI stress pattern.
  - Each sub-model gives some number of possible patterns.
  - Typological predictions = union of sub-models’ possible patterns.
Gordon’s (2002) model

- The $2 \cdot 11!$ possible constraint-rankings (grammars) correspond to only 152 distinct QI stress patterns (looking at words up to 8 syllables long).
Gordon’s (2002) model

- The $2 \cdot 11!$ possible constraint-rankings (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.
- Has the tightest inclusion/exclusion fit of the three models.
A dynamic linear model

- Treat stress assignment as a continuous “wave-like” phenomenon.
- “Connectionist” model
  - Each syllable corresponds to a node in a linear network.
  - A continuous quantity of “activation” “flows” from node to node, according to linear inhibition/amplification parameters, until equilibrium is reached.
  - At equilibrium, nodes with more activation than neighbors correspond to stressed syllables.
A dynamic linear model

\[ u_i^t = \text{activation of node } i \text{ at time } t. \]

Parameters:
- \( \alpha, \beta = \text{real-valued left/right inhibition constants} \)
- \( P_1, P_{-1}, P_{-2} = \text{starting activations of initial, final, penultimate syllables} \)
- \( S = \text{boolean whether left- or rightmost stressed syllable is primary} \)
A dynamic linear model

Network evolution:

\[ u_i^{t+1} = u_i^1 + \alpha u_{i+1}^t + \beta u_{i-1}^t \]

Prince (1993) shows conditions for convergence and gives a closed-form solution for it.

Typological predictions

- Can be computed approximately by discrete sampling of parameter space.
- Parameters vary freely with granularity 0.2, \( \alpha \) and \( \beta \) must be inhibitory \((\in [0, -1])\).
A dynamic linear model

\[ P_1 = P_{-1} = P_{-2} = 1.0, \] colored regions represent distinct stress patterns.
A dynamic linear model

Results
- Generates 1,470 distinct QI stress systems
- 14 attested
- Undergenerates by 12, overgenerates by 1,456.

Caveats
- Preliminary implementation
- Other parameterizations, granularities possible
- Currently cannot generate stress clash!
A weighted constraints model

- Constraint-based like OT, but with constraint weightings rather than strict rankings.
- Allows “additive” effects
  - E.g., many violations of low-weighted constraint can overwhelm action of higher-weighted constraint.

Parameterization
- For $k$ constraints, a grammar $\vec{w}$ is a $k$-vector ($\in \mathbb{R}_+^k$) of weights.
- Each candidate output ($\in \{\sigma, \dot{\sigma}, \ddot{\sigma}\}^n$) gives a vector $\vec{v}$ of constraint violations ($\in \mathbb{Z}_+^k$).
- Optimal candidate minimizes $\vec{w} \cdot \vec{v}$. 
As a starting point, we’ll consider an HG model using the same constraints as Gordon’s (2002) OT model.

- Same meta-constraint that only one of the $\text{ALIGN}(x_2, \text{L/R})$ constraints be active at once.

Typological predictions

- Exact computation using finite-state methods (Bane & Riggle in preparation).
- 36,846 distinct stress systems possible (!)
- 25 of them attested
- Undergenerates by 1 (Biangai), overgenerates by 36,821.
Motivation

- Hypothesized principle of least effort or simplicity
- Predicts that cross-linguistically frequent patterns should be simpler (according to some metric) than rare ones
- Patterns predicted by a model 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 $(\sigma, \dot{\sigma}, \ddot{\sigma}, #)$ 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”
Attestation: Bigram Entropy

Bigram Entropy of OT Typology

Bigram Entropy of DLM Typology

Bigram Entropy of HG Typology

Significant ($p < 0.05$): DLM, HG.
Attestation: Trigram Entropy

Trigram Entropy of OT Typology

Trigram Entropy of DLM Typology

Trigram Entropy of HG Typology

Significant ($p < 0.05$): OT, DLM, HG.
Attestation: 4-gram Entropy

4-gram Entropy of OT Typology

4-gram Entropy of DLM Typology

4-gram Entropy of HG Typology

Significant ($p < 0.05$): DLM, HG.
Typological frequency of attestation:
- High frequency (above median) patterns have significantly lower trigram entropy than low frequency (below median) patterns. $U = 51.5, \ p = 0.0428$
Confusability

- Some patterns are very similar to each other, or confusable
  - Must observe very long forms (i.e., many syllables) in order to distinguish them from each other
- Example: Albanian and Malakmalak
  - 2 syl. word: ʼσσ vs ʼσσ
  - 3 syl. word: σʼσσ vs σʼσσ
  - 4 syl. word: σσʼσσ vs σʼσσʼσσ
  - 5 syl. word: σσσʼσσ vs σσʼσσσʼσσ
  - 6 syl. word: σσσσʼσσ vs σσʼσσσσʼσσ
- 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 predicted pattern in each model.

- Ex: Albanian’s (fixed penultimate primary) stress pattern:
  - \langle 101, 39, 10, 0, 0, 0, 0 \rangle (in Gordon’s OT model)
  - Confusability 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
Difference is significant only in OT

\[ U = 1005.5, \quad p < 0.001 \]
Confusability Sums

- Confusability sums are not significantly different for attested vs unattested patterns.
- But, within attested patterns, frequency of attestation significantly correlated with linear combination of confusability sums and pivots ($p < 0.05$, $R^2 = 0.271$).
  - Only among Gordon’s (OT) predicted languages.
Within each model, each predicted pattern can be associated with the number of model parameter settings that produce it.

For OT:
- The number of constraint rankings that generate the pattern.
- Efficiently computable by an algorithm worked out by Riggle (2008).
Parameter-volume ($\rho$-volume)

- For DLMs:
  - Volume of the region in the $\{\alpha, \beta, P_1, P_{-1}, P_{-2}, S\}$-space that produces the pattern.
  - Approximated by discrete search of the space.

- For HG:
  - Volume of the region in the constraint weighting space ($R^k_+$) that produces the pattern.
  - Can be computed exactly, but NP-hard
  - . . . still in progress (need better computer!)
Majority of the parameter volume is concentrated in a few stress patterns.

Power law, similar to cross-linguistic frequency (nearly Zipfian, $p < 0.001$, $R^2 = 0.968$ according to Gauss-Newton)
What can $p$-volume predict?

- The $p$-volumes of the stress patterns generated by Gordon’s OT model prove to correlate well with:
  - Which predicted patterns are actually attested, and
  - The frequencies of the attested patterns.
OT $p$-volume correlates with attestation

R–Volumes of Predicted Stress Patterns
(152 patterns)

R–Volumes of Predicted Stress Patterns
(152 patterns; attested are blue)
The \( p \)-volumes of attested stress patterns are significantly greater than those of the unattested.

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

- Frequency is more robustly predicted by a nonlinear function of the logarithm of the $p$-volume.
- Exponential regression: $R^2 = 0.704$, $p < 0.001$
DLM $p$-volume correlates with attestation.
Discussion

- Surprising, since these correlates depend entirely on the set of languages generated by each model.
- p-volume findings might suggest learning model in which...
  - Learners randomly select a grammar consistent with the evidence they’ve seen so far.
    - ⇒ learners will tend to choose languages with frequency proportional to p-volume
  - Or where learners select from the languages consistent with their observations according to p-volume.
- In progress: predicting the typological frequencies with iterated learning simulations.
In probabilistic and information-theoretic theories, we often need to quantify the prior probability of a grammar.

Often not obvious

- How is one constraint-ordering, or one vector in $\mathbb{R}_+^k$ more probable than another?

$p$-volume results suggest an answer:

- $P(g) \propto pvol(L(g))$
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, detectably).
especially if those same systematic differences also relate to cross-linguistic frequency.