Network models of sound change

Dan Hruschka
Santa Fe Institute
dhrusch@santafe.edu

in collaboration with Tanmoy Bhattacharya, Eric Smith & Jon Wilkins

“Statistical Inference for Complex Networks”
December 2008
“Egg” in 24 Turkic languages

- jumurtGa
- simit
- umurxa
The Plan

**Question**—What theories of sound change best account for observed diversity in related languages?

- Brief introduction—concepts, units
- Specify one model in a likelihood framework
- Discuss plans for specifying other models and raise issues of inference and model selection
### Vowels

<table>
<thead>
<tr>
<th>Front</th>
<th>Central</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>e</td>
<td>u</td>
</tr>
<tr>
<td>/i/</td>
<td>/e/</td>
<td>/u/</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td>High</td>
</tr>
<tr>
<td>Lowered-high</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Higher-mid</td>
<td></td>
<td>Mid</td>
</tr>
<tr>
<td>Mid</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raised-low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Tones:**
- Extra high
- High
- Mid
- Low
- Extra low
- Rising
- Falling

### Consonants

<table>
<thead>
<tr>
<th>Labial</th>
<th>Coronal</th>
<th>Dorsal</th>
<th>Guttural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilabial</td>
<td>Labiodental</td>
<td>Labial-palatal</td>
<td>Labial-velar</td>
</tr>
<tr>
<td>Dental</td>
<td>Alveolar</td>
<td>Postalveolar</td>
<td>Retracted</td>
</tr>
<tr>
<td>Palatal</td>
<td>Velar</td>
<td>Velar-velar</td>
<td>Uvular</td>
</tr>
<tr>
<td>Velar-velar</td>
<td>Pharyngeal</td>
<td>Epiglottal</td>
<td>Glottal</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obstructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop</td>
</tr>
<tr>
<td>Implosive</td>
</tr>
<tr>
<td>Fricative</td>
</tr>
<tr>
<td>Affricate</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Click</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ø</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Liquids</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trill</td>
</tr>
<tr>
<td>Tap</td>
</tr>
<tr>
<td>Flap</td>
</tr>
<tr>
<td>Lat. Fricative</td>
</tr>
<tr>
<td>Lat. Approximant</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nasal</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Approximant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ø</td>
</tr>
</tbody>
</table>
Many Theories

• No expectation of what changes to what
• Regularity, but no expectation of what changes to what
• Random walk through feature space
• Lenition ($t \rightarrow ts \rightarrow s \rightarrow h$)
• Articulatory Reduction
• Other accounts
  • “An eft” becomes “a newt”
  • Sound symbolism—'stamp', stomp', 'tamp', 'tromp', 'tramp'
  • Constraints on language design (e.g. optimality theory)
An Opportunity

• “Although no comprehensive study of sound change that would allow us to distinguish common from uncommon innovations has ever been undertaken, historical linguists have acquired a sense of what kinds of change are likely to occur” (Blust 2004)
Common Methods

• **Sociolinguistics**—Study dialectical variation and change over short periods of time

• **Psycholinguistics**—Study variation in production and perception

**Advantage**
- Close view of change in action

**Disadvantage**
- Limited to frequent or fast changes
Historical Linguistics

Study extant between-language variation and ancient recorded languages (Greek, Sanskrit)

Advantages
- Longer time scales

Disadvantages
- Can’t see process
- Requires some way of making inferences about generative models

Jacob and Wilhelm Grimm (1855)
A Model

**Tarvag**

1. Align words
2. Estimate probabilities of sound change
3. Estimate probabilities of ancestral sounds
4. Identify cognates
5. Infer historical relationships
Likelihood Function

\[ p(L_1, L_2, \ldots, L_N \mid p(A), p(s \mid A), \text{alignment}) = \prod \sum \prod \prod p(A) \prod p(s \mid A) \]

\( p(s \mid A) = \) probability of observed sound given an ancestor
\( p(A) = \) probability of ancestral sound

Some assumptions were required to get here
Estimation Strategy

\[ P(\text{Observed sound}|\text{Ancestral sound}) \]

\[ P(\text{Ancestral sound}) \]

Find alignment that maximizes \( L \)
• Word lists (350 to 1400 words each) for 29 Turkic languages
• “Phonetic” transcriptions

• Advantages
  – Shallow time depth

• Weaknesses
  – Idealization
  – Synchronic
Finding ML estimates

- **100 words**
  - 9635 observations
  - 24 hours
  - loglikelihood = -8239
  - Ancestral sounds—15 consonants and 9 vowels with probability $> 10^{-17}$

- **200 words**
  - 16127 Observations
  - 118 hours
  - loglikelihood = -16903
  - Ancestral sounds—18 consonants and 9 vowels with probability $> 10^{-17}$
Distribution of Probabilities in the Sound Change Matrix
Vowels

Consonants (Velar, Uvular, Labial)

Consonants (Palatal, Alveolar, Postalveolar)

Vowels
Comparing 2 samples’ estimates

Estimated Sound Change Probabilities

Estimated Ancestral Probabilities
Recap

• We have ML estimates for data given model
• They fit qualitative observations made by many linguists
• They are consistent for different samples
Questions

• How to compare with other models?
  – Feature-based changes
  – Natural classes (Bouchard-Cote et al. 2007)

• We will always have misspecification. What to do?
  – Historical relationships
  – Regularity of sound change (Foot-Pied, Five-Pent)
  – Idiosyncratic, but common kinds of change (“A newt” & “an eft”)

• Is there a way to estimate faster?