Opportunities in quantitative historical linguistics

Eric Smith

work with
Tanmoy, Jon, Bill, Ian, Dan Hruschka, Logan Sutton, Hyejin Youn

collaborative with
Murray Gell-Mann, Ilia Peiros, George Starostin
Outline

• Brief history of languages at SFI

• The steps toward quantitative phylogenetics
  • Example I: lexical reconstruction
  • Example II: inferring models of word-order change

• New opportunities for empirical discovery
The Evolution of Human Languages Program at SFI

Goals: Deep reconstruction of language history and connection with genes and the archaeological record

Murray Gell-Mann

Ilia Peiros

George Starostin

Sergei Starostin
Deep reconstruction: the Nostratic hypothesis

- Coined by Holger Pedersen (1903)
- Modern form of the hypothesis by Vladislav Ilitch-Svitych and Aharon Dolgopolsky (1960s -- present)
- Estimated 12,000-15,000 BCE


http://en.wikipedia.org/wiki/Nostratic_languages
Methods: **lexicostatistics and glottochronology**

- Assign any sound map without penalty, but require regularity
- Exclude borrowed items in either language from consideration
- Identify fraction preserved cognates; convert to separation time (penalty)
- Attempt to fit separation times to an ultrametric structure (tree)

\[ P_{\text{Preserve}}(\text{word}) = e^{-t/\tau} \]

\[ \Delta t_{\text{sep}} = -\tau \log (\text{frac. preserved}) \]
EHL today

- The Dene-Caucasion hypothesis
- Borean? Relic of the last ice-age?
- Reconstructions of Dravidian, Khoisan, ...

- The EHL database and Starling

http://starling.rinet.ru/cgi-bin/response.cgi?root=config&morpho=0&basename=\data\alt\turcet&first=1
Concepts and steps in a quantitative phylogenetics

- Roles of Likelihood and Bayesian methods
  - Frequent-pattern versus rare-feature innovations
  - Bayes’s theorem and prior prejudice
  - Typological constraints and Bayesian priors
  - Information criteria and significance of parameters

- The likelihood part of a phylogenetic algorithm
  - Overall structure of sound and meaning change
  - Alignment, sound correspondence, and errors
  - Context discovery, detection of borrowings
  - Classification and reconstruction
Rare innovations versus clusters of common innovations

- Rare innovations: single features with \(~0\) probability to occur by chance *(go in Bayesian priors)*
  - Imply common descent or borrowing, even w/o mathematics
  - Only seen once: hard to assign probabilities from frequencies
  - Common in morpho-syntactic features
  - Useless for dating; do not support induction

- Common variations: *(estimate with likelihood)*
  - Examples: sound shift and meaning shift in core lexicon
  - Individually uninformative, but can assign probabilities from data
  - Require math to handle, but do support induction, and can be informative about dates if change processes are regular
Bayes’s Theorem and model comparison

• Represent both data and models with a joint probability

• Split joint probability into conditionals either of two ways

\[ P(m, d) = P(m \mid d) P(d) = P(d \mid m) P(m) \]

• Bayes’s theorem: priors and likelihoods

\[
P(m \mid d) = \frac{P(d \mid m) P(m)}{P(d)} \quad \text{Bayesian prior}
\]

\[
\text{Bayesian posterior} = \frac{P(d \mid m) P(m)}{\sum_{m'} P(d \mid m') P(m')}
\]
Typological constraints are a natural domain for priors (Ian’s lecture)

• Three roles for priors
  • Include frequency evidence from outside this sample
  • Include non-frequency evidence (rare innovations)
  • Represent out-of-field evidence (molecular phylogenies)

• On states
  • phoneme inventory, word order, ...
  • implicational relations (pronouns, time, color, aspect)

• On transitions
  • phoneme contexts, intermediate word-order states (NDO)
  • geometric models of phonology or semantics?
Akaike and Bayesian Information Criteria

\[ \text{AIC} \equiv -2 \log (L) + 2k \]

\[ \text{BIC} \equiv -2 \log (L) + k \log (n) \]

\[
P(n_1, \ldots n_K) = \prod_{i=1}^{K} p_{n_i}^{n_i} \left( \frac{N!}{n_1!, \ldots, n_K!} \right) = e^{-N D(\bar{n}/N \parallel \bar{p})}
\]

\[
D\left( \frac{\bar{n}}{N} \parallel \bar{p} \right) = \sum_{i=1}^{K} n_i \log \frac{n_i}{p_i}
\]

Kullback-Leibler divergence, or Relative Entropy

\[
\sum_{\tilde{n}} e^{-ND(\bar{n}/N \parallel \bar{p})} \delta\left( \frac{\tilde{n}}{N} - \bar{p} \right) = \rho(\bar{p})
\]

\[
\sum_{\tilde{n}} e^{-ND(\bar{n}/N \parallel \bar{p})} N \sum_{i} \tilde{p}_i \log \tilde{p}_i \approx N \sum_{i} \tilde{p}_i(\bar{p}) \log \tilde{p}_i(\bar{p}) + \frac{k}{2}
\]

\[
\sum_{\tilde{n}} e^{-ND(\bar{n}/N \parallel \bar{p})} \sum_{i} \tilde{n}_i(\bar{p}) \log \tilde{p}_i \approx N \sum_{i} \tilde{p}_i(\bar{p}) \log \tilde{p}_i(\bar{p}) - \sum_{\tilde{n}} e^{-ND(\bar{n}/N \parallel \bar{p})} N \frac{k}{2} \sum_{j=1}^{k} \left( \frac{\bar{p}(\tilde{n}/N) - \bar{p}(\bar{p})}{\tilde{p}_i(\bar{p})} \right)^2
\]

Maximize average likelihood of real samples over average models

\[
\tilde{p}_i \rightarrow \frac{\tilde{n}_i}{N} \quad \max_{\bar{p}} \log \mathcal{L}(\tilde{n}_1, \ldots, \tilde{n}_k) = N \sum_{j=1}^{k} \frac{\tilde{n}_j}{N} \log \frac{\tilde{n}_j}{N}
\]

(Distribution of max-likelihoods)

\[
\tilde{p}_i \approx N \sum_{i} \tilde{p}_i(\bar{p}) \log \tilde{p}_i(\bar{p}) - k \log \tilde{p}_i(\bar{p}) - k
\]

Unbiased sample estimator!
More detail on Akaike derivation

\[ L(\tilde{n}_1, \ldots, \tilde{n}_k) = \prod_{j=1}^{k} \tilde{p}_{\tilde{n}_j} \]

Likelihood of any data given a particular model

Average log-likelihood of actual data, from a model produced with fixed estimated parameters \( \tilde{p}_{\tilde{n}} \)

\[
\sum_{\tilde{n}} e^{-ND(\frac{\tilde{n}}{N} || \tilde{p})} \log L(\tilde{n} | \tilde{p}) = \sum_{i} \tilde{n}_i(\tilde{p}) \log \tilde{p}_i
\]

Now this averaged log-likelihood, averaged over estimated models; 2nd-order Taylor exp’n

\[
\sum_{\tilde{n}} e^{-ND(\frac{\tilde{n}}{N} || \tilde{p})} \sum_{i} \tilde{n}_i(\tilde{p}) \log \tilde{p}_i(\tilde{n}) \approx N \sum_{i} \tilde{p}_i(\tilde{p}) \log \tilde{p}_i(\tilde{p}) - \sum_{\tilde{n}} e^{-ND(\frac{\tilde{n}}{N} || \tilde{p})} N \frac{k}{2} \sum_{j=1}^{k} \left( \frac{\tilde{p}(\tilde{n}) - \tilde{p}(\tilde{p})}{\tilde{p}_i(\tilde{p})} \right)^2
\]

But we don’t have the ideal \( \tilde{p}_{\tilde{n}} \)-s; the best we can do is obtain an unbiased estimator from any single sample; for this we need to identify the bias typical of samples

\[
\sum_{\tilde{n}} e^{-ND(\frac{\tilde{n}}{N} || \tilde{p})} N \sum_{i} \tilde{p}_i(\tilde{n}) \log \tilde{p}_i \approx N \sum_{i} \tilde{p}_i(\tilde{p}) \log \tilde{p}_i(\tilde{p}) + \frac{k}{2}
\]

Use this to replace the ideal ML with an unbiased estimator from samples, get previous slide
Word lists are the starting point for lexical (= phonological / semantic) reconstruction

http://starling.rinet.ru/cgi-bin/response.cgi?root=config&morpho=0&basename=data\alt\turcet&first=1

<table>
<thead>
<tr>
<th>Proto-Turkic</th>
<th>Altaic etymology</th>
<th>Meaning</th>
<th>Russian meaning</th>
<th>Old Turkic</th>
<th>Karakhanid</th>
<th>Turkish</th>
<th>Tatar</th>
<th>Middle Turkic</th>
<th>Uzbek</th>
<th>Uighur</th>
<th>Sary-Yughur</th>
<th>Azerbaidzhan</th>
<th>Tur.</th>
</tr>
</thead>
<tbody>
<tr>
<td>*Ab</td>
<td>Altaic etymology</td>
<td>hunt, chase</td>
<td>охота</td>
<td>ab (Orkh.), av (OUygh.)</td>
<td>av (MK)</td>
<td>av</td>
<td>aw</td>
<td>aw (Pav. C.)</td>
<td>ов</td>
<td>av, dial. ө</td>
<td>ов</td>
<td>ов</td>
<td>ав</td>
</tr>
<tr>
<td>*ab-</td>
<td>Altaic etymology</td>
<td>to crowd, come together</td>
<td>собираться, встречаться</td>
<td>av- (OUygh.)</td>
<td>av- (MK, KB)</td>
<td>uč</td>
<td>avuč</td>
<td>ee</td>
<td>эwię</td>
<td>oč</td>
<td>oš</td>
<td>ovuč</td>
<td></td>
</tr>
<tr>
<td>*abuč</td>
<td>Altaic etymology</td>
<td>handful</td>
<td>пригоршня</td>
<td>avuč</td>
<td>avuč</td>
<td>abuša (MK, KB)</td>
<td>avuč</td>
<td>abuša (MA, Sangl., Bop. Bal.)</td>
<td>abuša, abuša (Sib.)</td>
<td>abuša, abuša (Abush., Sangl.)</td>
<td>abuša, abuša</td>
<td>abuša, abuša</td>
<td>abuša, abuša</td>
</tr>
<tr>
<td>*Abuč-ka</td>
<td>Altaic etymology</td>
<td>1 husband, old man 2 foster-mother 3 elder sister 4 uncle</td>
<td>1 муж, старик 2 кормилица 3 старшая сестра 4 дядя</td>
<td>avičya, avičya 1, avičya 2 (OUygh.)</td>
<td>avičya</td>
<td>abuša, abuša 1 dial. (Sib.)</td>
<td>abuša, abuša 1 dial. (Sib.)</td>
<td>abuša, abuša 4 (Abush., Sangl.)</td>
<td>abuša, abuša 4</td>
<td>abuša, abuša 4</td>
<td>abuša, abuša 4</td>
<td>abuša, abuša 4</td>
<td>abuša, abuša 4</td>
</tr>
</tbody>
</table>

Turkic etymology:
Objects that must be modeled are joint histories of sound and meaning for a collection of words

- n.b., word forms are attested; meanings are indirectly inferred, and often ambiguous
- easy to trace a form; but inadequate to infer history from forms alone
Representing sound and meaning “innovation” in the comparative method (Bill’s lecture)

- Suppose that some stable meaning categories can be identified
- Identify primary words for each meaning
- Try to exclude “borrowed” terms; suppose that what is left has been transmitted through vertical descent
- Identify systematic sound relations and try to infer historical sound changes
- Associate semantic innovations with in-language substitutions within meaning categories
Preserved meanings suggest sound maps

(Numbers give an example in which we can treat primary meanings as language-universal and historically relatively stable)

Use preservation of meaning to infer regular relations of sounds: here, e.g., thr <> tr

(Innovation)
Sound maps help identify meaning shifts.

- PIE: dhus, ane
  - k\textsuperscript{w}en

- PG
  - deuza
  - rai\textsuperscript{z}az
  - \(\chi\)un-da

- Lat
  - animal
  - canis

- Ger
  - tier
  - reh
  - hund

- Eng
  - animal
  - deer
  - dog
  - hound

- It
  - animale
  - cane

- Fr
  - animal
  - chien
The “alignment problem”: what to compare w/ what?

Etymology 4 “belly”

Etymology 5 “big, high”

Etymology 12: “Breast, nipple”
Maximum-likelihood estimation of history and process (phonological only, here)

- Suppose we have proposed an alignment of positions in the daughter languages
- Propose phoneme assignments to aligned positions in the ancestor (with probabilities)
- Estimate regular correspondence of ancestor to daughter phonemes (w/ or w/o probabilities)
- Estimate random violations (with probabilities)
Sound correspondences among the languages

Typology I: sound relations and features inferred from sound changes
Context dependence: an Artificial Intelligence problem

<table>
<thead>
<tr>
<th>Stage</th>
<th>katt</th>
<th>keaff</th>
<th>kinn</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>I</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>II</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>III</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage</th>
<th>ka:ra</th>
<th>flo:s</th>
<th>flo:ses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>I</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>II</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>III</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Sound change can be regular, but not at single-phoneme level
- Conditioning contexts can be lost; must be guessed
- AIC and BIC can be used to judge guesses

From R.L. Trask, “Language change”
Vowel harmony in the Turkic language family: a predictive supra-segmental context

Front/back agreement in roots and affixes

nom  a book
inek  a cow
nomnar  books
inekter  cows
Identification of borrowings: words that do not fit the system pattern

Receiver Operating Characteristic curve
New conceptual domain: mathematical likelihood modeling of semantic shift

- Phonological and semantic constraints interact with polysemy and synonymy to structure sound and meaning change

- Semantic categories, split, join, and move in some “space” which we do not know
Phylogeny and reconstruction

- Modern glottochronology applied to expert linguists’ judgments of cognate classifications
- Presence/absence data format modeled after genes
- *Not yet a model of processes of sound and meaning change*
Maslova: how much can you do with incomplete reconstructions of the past?

Two quantities: stationary frequency typical number of pairs

\[ \bar{f} = \frac{p}{p+q} \]

\[ \bar{h} = (p + q)^2 [2 - (p + q)] 2 \bar{f} (1 - \bar{f}) \]

Regression model for observed frequencies estimate \((p/q)\):

\[ f_i = \bar{f} + \epsilon_i \]

Fit of number of pairs against mean, in children of a common family ancestor:

put bounds on \((p+q), p/q\)

\[ h_i - \bar{h} = (f_i - \bar{f}) 2 (p + q) (1 - 2 \bar{f}) + \epsilon_i \]
Three partitions of standard word order

\[ f(OV) \approx 0.53 \]
\[ \frac{1}{p + q} \approx 24\text{ky} \]

\[ f(SO) \approx 0.96 \]
\[ \frac{1}{p + q} \leq 11\text{ky} \]
\[ p/q \geq 15 \]

\[ f(SV) \approx 0.86 \]
\[ \frac{1}{p + q} \in (10\text{ky}, 30\text{ky}) \]
\[ p/q \geq 3 \]
New opportunities for quantitative characterization of regularities

• Cross-linguistic regularities in frequency of use, and relations to rates of change
• Punctuated equilibrium and correlations of the "clock" of language change with culture
• Polysemy, synonymy, and semantics
• Full speaker-corpus archives (Norquist et al.), formant-based analysis (Labov), ...
Typology II: frequency of use and rates of change

Frequency of word-use predicts rates of lexical evolution throughout Indo-European history
Mark Pagel, Quentin D. Atkinson & Andrew Meade

Quantifying the evolutionary dynamics of language
Erez Lieberman, Jean-Baptiste Michel, Joe Jackson, Tina Tang & Martin A. Nowak

a. English
b. Spanish
c. Russian
d. Greek

Mean rate (per 10^10 yr)

Frequency of word (number of times per 10^6 words)

Number of irregular verbs

Old English
Middle English
Modern English

Old to Modern English
Middle to Modern English

Regularization rate

-0.48
-0.51

Frequency
Linguistic punctuated equilibrium: do cultural phenomena like splitting drive language divergence?

Languages Evolve in Punctuational Bursts

Quentin D. Atkinson, Andrew Meade, Chris Venditti, Simon J. Greenhill, Mark Pagel

1 FEBRUARY 2008 VOL 319 SCIENCE www.sciencemag.org
Toward a likelihood model of semantic shift: Inferring polysemy with English as a meta-language

<table>
<thead>
<tr>
<th>Meta-language</th>
<th>L1</th>
<th>L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>xun</td>
<td>w1</td>
</tr>
<tr>
<td>day</td>
<td></td>
<td>w2</td>
</tr>
<tr>
<td>moon</td>
<td>ai</td>
<td>w3</td>
</tr>
<tr>
<td>month</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

```
Sun
1
Day
Moon
2
Month
```
Network of polysemes in 81 diverse languages

Data and Graph courtesy Bill Croft, Logan Sutton, and Hyejin Youn
Summary comments

• Much linguistics traditionally modeled on logic; Can we root historical linguistics in probability?

• Language evolution is not molecular evolution; Same principles lead to different models

• New and refined quantitative signatures: Both methodological and conceptual opportunities