Natural Language Processing

Diachronics

Dan Klein – UC Berkeley

Includes joint work with Alex Bouchard-Cote, Tom Griffiths, and David Hall
The Task
Lexical Reconstruction

<table>
<thead>
<tr>
<th>Latin</th>
<th>focus</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>French</th>
<th>Spanish</th>
<th>Italian</th>
<th>Portuguese</th>
</tr>
</thead>
<tbody>
<tr>
<td>feu</td>
<td>fuego</td>
<td>fuoco</td>
<td>fogo</td>
</tr>
</tbody>
</table>
We assume the phylogeny is known

- Much work in biology, e.g. work by Warnow, Felsenstein, Steele...
- Also in linguistics, e.g. Warnow et al., Gray and Atkinson...

http://andromeda.rutgers.edu/~jlynch/language.html
**Evolution through Sound Changes**

<table>
<thead>
<tr>
<th>Latin</th>
<th><strong>camera</strong> /kamera/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deletion</td>
<td>/e/, /a/</td>
</tr>
<tr>
<td>Change</td>
<td>/k/ .. /ʃ/ .. /ʃ/</td>
</tr>
<tr>
<td>Insertion</td>
<td>/b/</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>French</th>
<th><strong>chambre</strong> /ʃambʁ/</th>
</tr>
</thead>
</table>

Eng. camera from Latin, “camera obscura”

Eng. chamber from Old Fr. before the initial /t/ dropped
Changes are Systematic

camera /kamera/  
e → _

camra /kamra/

numerus /numerus/  
e → _

numrus /numrus/
Changes are Contextual

camera /kamera/

\[ e \rightarrow _{ } \]

\[ e \rightarrow _{ } / \text{after stress} \]

camra /kamra/
Changes Have Structure

_ → b

_ → b / m_r

_ → [stop x] / [nasal x]_r
Changes are Systematic

*English Great Vowel Shift (Simplified!)*

“time” = teem  →  “time” = taim
## English Great Vowel Shift

<table>
<thead>
<tr>
<th>Middle English</th>
<th>became</th>
<th>Early Modern English</th>
<th>became</th>
<th>Modern English</th>
</tr>
</thead>
<tbody>
<tr>
<td>[a:] 'name'</td>
<td></td>
<td>[ɛ:] [nɛ:m]</td>
<td></td>
<td>[ei] [nɛ:m]</td>
</tr>
<tr>
<td>[ɛ:] 'meat'</td>
<td></td>
<td>[e:] [me:t]</td>
<td></td>
<td>[i:] [mi:t]</td>
</tr>
<tr>
<td>[e:] 'meet'</td>
<td></td>
<td>[i:] [mi:t]</td>
<td></td>
<td>[i:] [mi:t]</td>
</tr>
<tr>
<td>[i:] 'ride'</td>
<td></td>
<td>[æi] [ræid]</td>
<td></td>
<td>[ai] [raid]</td>
</tr>
<tr>
<td>[ɔ:] 'boat'</td>
<td></td>
<td>[o:] [bo:t]</td>
<td></td>
<td>[ɔu/əu] (boot/boʊt)</td>
</tr>
<tr>
<td>[o:] 'boot'</td>
<td></td>
<td>[u:] [bu:t]</td>
<td></td>
<td>[u:] [bu:t]</td>
</tr>
</tbody>
</table>

Diachronic Evidence

Yahoo! Answers [ca 2000]
Resolved Question
Which is correct...tonight or tonite?
10 months ago
10 months ago

Best Answer - Chosen by Voters
"Tonight" is the traditional version.
If you'll observe, "tonite" is listed as a misspelling by the system here.
The use of "tonite" can probably be traced to the way that people make mistakes and they stick with a small group and then the use of it expands, making it become a use that people accept.
10 months ago

Appendix Probi [ca 300]

tonight not tonite

tonitru non tonotru
Synchronic (Comparative) Evidence

<table>
<thead>
<tr>
<th>Gloss</th>
<th>Latin</th>
<th>Italian</th>
<th>Spanish</th>
<th>Portuguese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word/verb</td>
<td>verbum</td>
<td>verbo</td>
<td>verbo</td>
<td>verbu</td>
</tr>
<tr>
<td>Fruit</td>
<td>fructus</td>
<td>frutta</td>
<td>fruta</td>
<td>fruta</td>
</tr>
<tr>
<td>Laugh</td>
<td>ridere</td>
<td>ridere</td>
<td>reir</td>
<td>rir</td>
</tr>
<tr>
<td>Center</td>
<td>centrum</td>
<td>centro</td>
<td>centro</td>
<td>centro</td>
</tr>
<tr>
<td>August</td>
<td>augustus</td>
<td>agosto</td>
<td>agosto</td>
<td>agosto</td>
</tr>
<tr>
<td>Swim</td>
<td>natare</td>
<td>nuotare</td>
<td>nadar</td>
<td>nadar</td>
</tr>
</tbody>
</table>

Key idea: changes occur uniformly across the lexicon
The Data
The Data

- **Data sets**
  - **Small: Romance**
    - French, Italian, Portuguese, Spanish
    - 2344 words
    - Complete cognate sets
    - Target: (Vulgar) Latin
  FR  IT  PT  ES
The Data

- **Data sets**
  - **Small: Romance**
    - French, Italian, Portuguese, Spanish
    - 2344 words
    - Complete cognate sets
    - Target: (Vulgar) Latin

  - **Large: Austronesian**
    - 637 languages
    - 140K words
    - Incomplete cognate sets
    - Target: Proto-Austronesian
Austronesian
### Austronesian Examples

**Word: bird**

Entries for "bird":

<table>
<thead>
<tr>
<th>ID</th>
<th>Language</th>
<th>Item</th>
<th>Annotation</th>
<th>Cognacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>34274</td>
<td>Banggai (W.dialect)</td>
<td>manu-manuk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34275</td>
<td>Banggi</td>
<td>bohed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34276</td>
<td>Banoni</td>
<td>manughu</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>34277</td>
<td>Bantik</td>
<td>manu?</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>34278</td>
<td>Gayo</td>
<td>manuk</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>34279</td>
<td>Gedaged</td>
<td>ma</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>34280</td>
<td>Geser</td>
<td>manuk</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>34281</td>
<td>Ghari</td>
<td>manu</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>34282</td>
<td>Gimán</td>
<td>manik</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>34283</td>
<td>Fijian (Bau)</td>
<td>manumanu</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>34284</td>
<td>Gorontalo (Hulondalo)</td>
<td>buuruŋi</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>34285</td>
<td>Hanunóo</td>
<td>manúk</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>34286</td>
<td>Bima</td>
<td>nasl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34287</td>
<td>Bintulu</td>
<td>manuk</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>34288</td>
<td>Bobot</td>
<td>ohas</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

From the Austronesian Basic Vocabulary Database
The Model
Simple Model: Single Characters

\[
P(x|x', \theta) = \theta(x, x')
\]

\[
\theta(C, G) = 0.02
\]

[cf. Felsenstein 81]
Changes are Systematic
Parameters are Branch-Specific

[Note: The diagram illustrates the parameters for different branches, focusing on the phonetic aspects of words in various languages.]
Edits are Contextual, Structured

\[ P(w, a | w', \theta_\ell) = \prod_{k} P(w_k, a_k | w_{k-1}, w', \theta_\ell) \propto \exp \left( \theta_\ell^T f(w_k, w_{k-1}, w'_{a_{k-1}}, w'_{a_k}, w'_{a_{k+1}}) \right) \]
Inference
Learning: Objective

\[
\max_{\theta,z} P(\theta, z | w_1 \ldots w_L)
\]
Learning: EM

- **M-Step**
  - Find parameters which fit (expected) sound change counts
  - Easy: gradient ascent on theta

- **E-Step**
  - Find (expected) change counts given parameters
  - Hard: variables are string-valued
Computing Expectations

Standard approach, e.g. [Holmes 2001]:
Gibbs sampling each sequence

[Holmes 01, Bouchard-Cote, Griffiths, Klein 07]
A Gibbs Sampler

\[ P(z_i \mid z_{-i}, w_1 \ldots w_L, \theta) \]
A Gibbs Sampler

‘grass’
A Gibbs Sampler

bubure

buruburu  bubure

bubure

buburu

buuburu  vuluvulu

‘grass’
Getting Stuck

How could we jump to a state where the liquids /r/ and /l/ have a common ancestor?
Getting Stuck
Efficient Sampling: Vertical Slices

- Single Sequence Resampling

- Ancestry Resampling

[Bouchard-Cote, Griffiths, Klein, 08]
Results
### Results: Romance

<table>
<thead>
<tr>
<th>Gloss</th>
<th>Latin</th>
<th>Italian</th>
<th>Spanish</th>
<th>Portuguese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word/verb</td>
<td><em>verb</em>um</td>
<td><em>verbo</em></td>
<td><em>verbo</em></td>
<td><em>verbu</em></td>
</tr>
<tr>
<td>Fruit</td>
<td><em>fructus</em></td>
<td><em>frutta</em></td>
<td><em>fruta</em></td>
<td><em>fruta</em></td>
</tr>
<tr>
<td>Laugh</td>
<td><em>ridere</em></td>
<td><em>ridere</em></td>
<td><em>reir</em></td>
<td><em>rir</em></td>
</tr>
<tr>
<td>Center</td>
<td><em>centrum</em></td>
<td><em>centro</em></td>
<td><em>centro</em></td>
<td><em>centro</em></td>
</tr>
<tr>
<td>August</td>
<td><em>augustus</em></td>
<td><em>agosto</em></td>
<td><em>agosto</em></td>
<td><em>agosto</em></td>
</tr>
<tr>
<td>Swim</td>
<td><em>natare</em></td>
<td><em>nuotare</em></td>
<td><em>nadar</em></td>
<td><em>nadar</em></td>
</tr>
</tbody>
</table>
Learned Rules / Mutations

\[ /\text{werbum} / (\text{la}) \]
- \( m \rightarrow /_\# \)
- \( u \rightarrow o / _\) \)
- \( w \rightarrow v / \text{many environments} \)
- ...  

\[ /\text{verbo} / (\text{vl}) \]
- \( r \rightarrow f \)
- \( e \rightarrow \varepsilon \)
- ...  

\[ \text{coluber} \quad \text{non colober} \]
\[ \text{passim} \quad \text{non passi} \]
Learned Rules / Mutations

u → o / many environments
v → b / init. or intervocal.
t → the / ALV_#

/verbo/ (ib)
  v → b
  u → o

/berbo/ (es) /verbu/ (pt)

r → r

...
Results: Austronesian
### Examples: Austronesian

<table>
<thead>
<tr>
<th>Gloss</th>
<th>Fijian</th>
<th>Pazeh</th>
<th>Melanau</th>
<th>Inabaknon</th>
<th>Reconstructed Ancestors</th>
</tr>
</thead>
<tbody>
<tr>
<td>star</td>
<td>kalokalo</td>
<td>mintol</td>
<td>biten</td>
<td>bitu’on</td>
<td>*bituqen</td>
</tr>
<tr>
<td>to hold</td>
<td>taura</td>
<td>ma:ra?</td>
<td>magem</td>
<td>kumkom</td>
<td>*gemgem</td>
</tr>
<tr>
<td>house</td>
<td>vale</td>
<td>xuma?</td>
<td>lebu?</td>
<td>ruma</td>
<td>*rumaq</td>
</tr>
<tr>
<td>bird</td>
<td>manumanu</td>
<td>aiam</td>
<td>manuk</td>
<td>manok</td>
<td>*qayam</td>
</tr>
<tr>
<td>to cut, hack</td>
<td>tata</td>
<td>ta:tatak</td>
<td>tutek</td>
<td>hadhad</td>
<td>*taraq</td>
</tr>
<tr>
<td>at</td>
<td>e</td>
<td>-</td>
<td>ga?</td>
<td>-</td>
<td>*i</td>
</tr>
<tr>
<td>what?</td>
<td>cava</td>
<td>?axai</td>
<td>ua? inew</td>
<td>ay</td>
<td>*nanu</td>
</tr>
<tr>
<td>this</td>
<td>oqo</td>
<td>?imini</td>
<td>itew</td>
<td>yayto</td>
<td>*ini</td>
</tr>
<tr>
<td>wind</td>
<td>cagi</td>
<td>varə</td>
<td>paŋay</td>
<td>bariyo</td>
<td>*bali</td>
</tr>
</tbody>
</table>

[Bouchard-Cote, Hall, Griffiths, Klein, 13]
Result: More Languages Help

Distance from Blust [1993] Reconstructions

![Graph showing the relationship between number of modern languages used and mean edit distance.](image-url)
*The model did not have features encoding natural classes
Regularity and Functional Load

In a language, some pairs of sounds are more contrastive than others (higher functional load)

Example: English p/d versus t/th

High Load: p/d: pot/dot, pin/din
dress/press, pew/dew, ...

Low Load: th/t: thin/tin
1955: Functional Load Hypothesis (FLH): Sound changes are less frequent when they merge phonemes with high functional load [Martinet, 55]

1967: Previous research within linguistics: “FLH does not seem to be supported by the data” [King, 67] (Based on 4 languages as noted by [Hocket, 67; Surandran et al., 06])

Our approach: we reexamined the question with two orders of magnitude more data [Bouchard-Cote, Hall, Griffiths, Klein, 13]
Regularity and Functional Load

Data: only 4 languages from the Austronesian data

Each dot is a sound change identified by the system

Functional load as computed by [King, 67]
Regularity and Functional Load

Data: all 637 languages from the Austronesian data

Functional load as computed by [King, 67]
Extensions
Cognate Detection

[Hall and Klein, 11]
Grammar Induction

Avg rel gain: 29%

[Berg-Kirkpatrick and Klein, 07]
Language Diversity

Why are the languages of the world so similar?

Universal grammar answer: Hardware constraints

Common source answer: Not much time has passed

[Rafferty, Griffiths, and Klein, 09]