Natural Language Processing

Diachronics

Dan Klein – UC Berkeley

Includes joint work with Alex Bouchard-Cote, Tom Griffiths, and David Hall
The Task
<table>
<thead>
<tr>
<th>Latin</th>
<th>focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>Spanish</td>
</tr>
<tr>
<td>feu</td>
<td>fuego</td>
</tr>
</tbody>
</table>
Tree of Languages

- 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

Latin

**camera /kamera/**

Deletion: /e/, /a/

Change: /k/ .. /tõ/ .. /õ/

Insertion: /b/

French

**chambre /ɔmbʁ̩/**

Eng. camera from Latin, “camera obscura”

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

<table>
<thead>
<tr>
<th>camera /kamera/</th>
<th>numerus /numerus/</th>
</tr>
</thead>
<tbody>
<tr>
<td>e → _</td>
<td>e → _</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>camra /kamra/</th>
<th>numrus /numrus/</th>
</tr>
</thead>
</table>
Changes are Contextual

camera /kamera/

e → _

e → _ / after stress

camra /kamra/
Changes Have Structure

\[
\begin{align*}
camra & /kamra/ \\
_ & \rightarrow b \\
_ & \rightarrow b / m_r \\
_ & \rightarrow [\text{stop } x] / [\text{nasal } x]_r \\
cambra & /kambra/
\end{align*}
\]
Changes are Systematic

*English Great Vowel Shift (Simplified!)*

“time” = teem  ➔  “time” = taim
Diachronic Evidence

Yahoo! Answers [ca 2000] 

Resolved Question
Which is correct...tonight or tonite?
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>nature</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

---

**FR** | **IT** | **PT** | **ES**
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>Benggai (W. dialect)</td>
<td>manu-manuk</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>34275</td>
<td>Benggi</td>
<td>bohed</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>34276</td>
<td>Banoni</td>
<td>manughu</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>34277</td>
<td>Bantik</td>
<td>manuʔ</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>34278</td>
<td>Gayo</td>
<td>manuk</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>34279</td>
<td>Gedaged</td>
<td>ma</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>34280</td>
<td>Geser</td>
<td>manuk</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>34281</td>
<td>Ghari</td>
<td>manu</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>34282</td>
<td>Gimēn</td>
<td>manik</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>34283</td>
<td>Fijian (Bau)</td>
<td>manumanu</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>34284</td>
<td>Gorontalo (Hulondalo)</td>
<td>buuruni</td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>34285</td>
<td>Hanunüdo</td>
<td>manúk</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>34286</td>
<td>Bima</td>
<td>nasi</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>34287</td>
<td>Bintulu</td>
<td>manuk</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>34288</td>
<td>Bobot</td>
<td>ohas</td>
<td></td>
<td>6</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

[Image of a diagram showing the relationship between different languages and their focus on words like "fuoco", "fuego", and "fogo".]

[Text: [Bouchard-Cote, Griffiths, Klein, 07]]
Edits are Contextual, Structured

\[ P(w, a | w', \theta_\ell) = \prod_k P(w_k, a_k | w_{k-1}, w', \theta_\ell) \propto \exp \left( \theta_\ell^T f(w_k, w_{k-1}, w'_{a_k-1}, w'_a, w'_{a_k+1}) \right) \]
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

\[ P(z_i | z_{-i}, w_1 \ldots w_L, \theta) \]

'grass'
A Gibbs Sampler

'grass'
A Gibbs Sampler

'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>verbum</td>
<td>verbo</td>
<td>verbo</td>
<td>verbo</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>
Learned Rules / Mutations

/werbum/ (la)

m →
u → o
w → v

/vero/ (vl)

r → r
e → ε

m → / _ #
u → o / _
w → v / many environments

...  ...  ...  ...

coluber  non colober
passim  non passi
Learned Rules / Mutations

\[
\begin{align*}
  u & \rightarrow o \quad / \text{many environments} \\
  v & \rightarrow b \quad / \text{init. or intervocal.} \\
  t & \rightarrow te / ALV_{-}# \\
\end{align*}
\]

\[
\begin{align*}
  r & \rightarrow f \\
  v & \rightarrow b \\
  u & \rightarrow o \\
  /verb/ (ib) \\
  /berbo/ (es) & \quad /verb/ (pt)
\end{align*}
\]
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>
<th>Manual</th>
<th>Automated</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>star</td>
<td>kalokalo</td>
<td>mintol</td>
<td>biten</td>
<td>bitu'on</td>
<td>*bituqen</td>
<td>*bituqen</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>to hold</td>
<td>taura</td>
<td>maza?</td>
<td>magem</td>
<td>kumkom</td>
<td>*gemgem</td>
<td>*gemgem</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>house</td>
<td>vale</td>
<td>xuma?</td>
<td>lebu?</td>
<td>ruma</td>
<td>*rumaq</td>
<td>*rumaq</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>bird</td>
<td>manumanu</td>
<td>aiam</td>
<td>manuk</td>
<td>manok</td>
<td>*qayam</td>
<td>*qayam</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>to cut, hack</td>
<td>tata</td>
<td>tatak</td>
<td>tutek</td>
<td>hadhad</td>
<td>*taraq</td>
<td>*taraq</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>at</td>
<td>e</td>
<td>-</td>
<td>ga?</td>
<td>-</td>
<td>*i</td>
<td>*i</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>what?</td>
<td>cava</td>
<td>axai</td>
<td>uai? inew</td>
<td>ay</td>
<td>*nanu</td>
<td>*nanu</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>this</td>
<td>oqo</td>
<td>imini</td>
<td>itew</td>
<td>yayto</td>
<td>*ani</td>
<td>*ani</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>wind</td>
<td>cagi</td>
<td>varə</td>
<td>panyay</td>
<td>bariyo</td>
<td>*bali</td>
<td>*beleru</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

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

Distance from Blust [1993] Reconstructions

Mean edit distance vs. Number of modern languages used
Visualization: Learned Universals

*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: t/th:  thin/tin
Functional Load: Timeline

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
Regularity and Functional Load

Data: all 637 languages from the Austronesian data

[Graph showing the relationship between merger posterior probability and functional load as computed by King, 67]
Extensions
Cognate Detection

\[ \pi \]

\[ /\text{fwe}/ \]
\[ /\text{fogo}/ \]
\[ /\text{fw}/ \]
\[ /\text{sentro}/ \]
\[ /\text{vdrbo}/ \]
\[ /\text{vetro}/ \]
\[ /\text{fweqo}/ \]
\[ /\text{vdrbo}/ \]
\[ /\text{vdrbo}/ \]
\[ /\text{vdrbo}/ \]
\[ /\text{sentro}/ \]

[Hall and Klein, 11]
Grammar Induction

Avg rel gain: 29%

[Graph showing language relationships with bar chart showing average relative gain of 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]