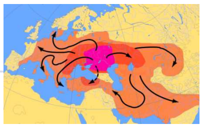
Statistical NLP
Spring 2009

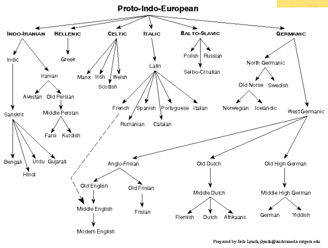
Lecture 30: Diachronic Models

Dan Klein – UC Berkeley

Work with Alex Bouchard-Cote and
Tom Griffiths

Tree of Languages





http://andromeda.rutgers.edu/~jlynch/language.html

Language Evolution

Latin camera /kamera/

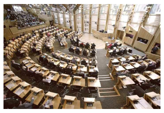
Deletion: /e/

Change of place: /k/ .. /tʃ/ .. /ʃ/

Insertion: /b/

French chambre /ʃambʁ/

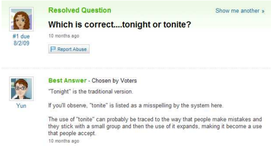
Eng. camera from Latin, "camera obscura"



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

Diachronic Evidence

Yahoo! Answers



Resolved Question

Show me another »

Which is correct....tonight or tonite?

#1 due 8/2/09

10 months ago

Report Abuse

Best Answer - Chosen by Voters

Yun

"Tonight" is the traditional version.

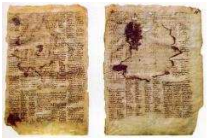
If you'll observe, "tonite" is listed as a misspelling by the system here.

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

tonight not tonite

Appendix Probi



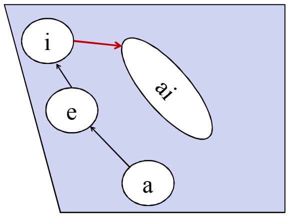
tonitru non tonotru

- Spelling (orthography) can reflect old pronunciation
- Corrections show when orthography hasn't kept up!

Example: Great Vowel Shift

(Simplified!)

"time" = teem ➡ "time" = taim



This is why the letter "i" is spoken as "ee" by many other languages, etc.

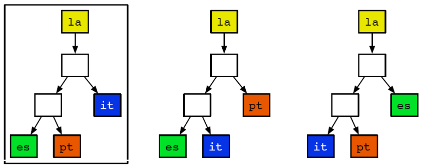
Where's It Going?

- Language isn't going anywhere in particular
- In fact, it's basically going everywhere
  - Over time, languages drift around
  - Related languages diverge
  - Eventually, results say more about the human language system than about history [Griffiths and Kalish 2007]
- Examples of tradeoffs
  - More consonant clusters vs. more syllables
  - More morphology vs. more rigid word order
  - Stress vs. tones vs. vowel variety

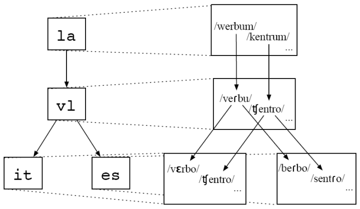
Synchronic (Comparative) Evidence

Gloss	Latin	Italian	Spanish	Portuguese
Word/verb	verbum	verbo	verbo	verbu
Fruit	fructus	frutta	fruta	fruta
Laugh	ridere	ridere	reir	rir
Center	centrum	centro	centro	centro
August	augustus	agosto	agosto	agosto
Swim	natare	nuotare	nadar	nadar

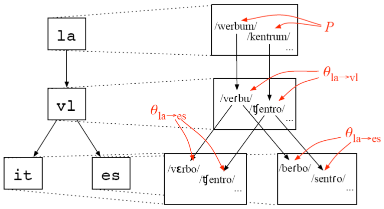
A Mini-Romance Phylogeny



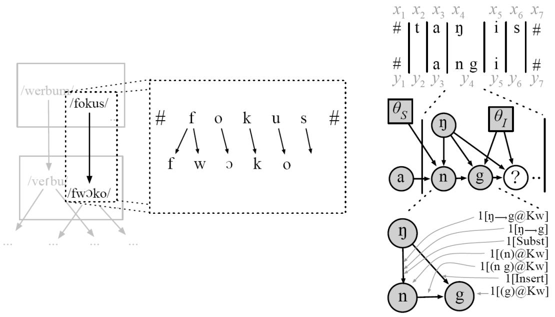
A Probablistic Model



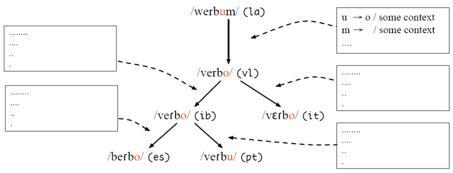
Model Parameters



Local Mutation along Tree

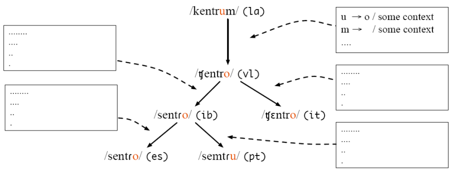


Ancient to Modern Forms



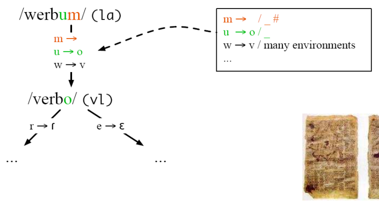
Gloss	Latin	Italian	Spanish	Portuguese
Word/verb	verbum	verbo	verbo	verbu

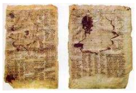
Ancient to Modern Forms



Gloss	Latin	Italian	Spanish	Portuguese
Word/verb	verbum	verbo	verbo	verbu
Center	centrum	centro	centro	centro

Learned Rules / Mutations

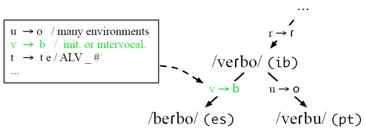




coluber non colober

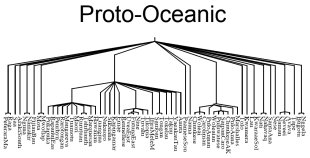
passim non passi

Learned Rules / Mutations



Oceanic Languages

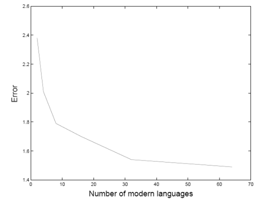
Russia
Mongolia
Japan
India
Pakistan
North Pacific Ocean
Indian Ocean
Australia
Google
Map data ©2008 Europa Tec



Oceanic Data

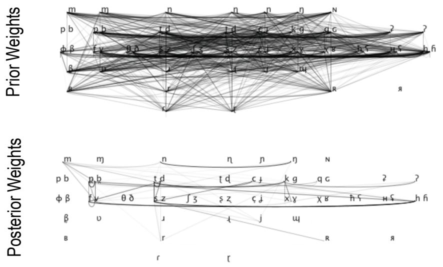
Gloss	Hawai'ian	Maori	Samoan	Tongan	ProtoOceanic
'break'	haki	whati	fati	fasi	*fati
'house'	hale	whare	fale	fale	*fale
'yam'	uhi	uhi	ufi	ufi	*ufi
'woman'	wahine	wahine	fafine	fefine	*wafine
'moon'	mahina	mahina	masina	mahina	*masiana

POc Reconstruction Results

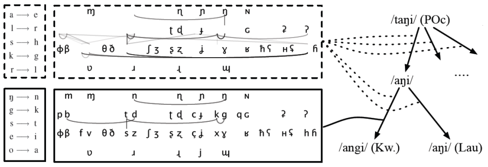


Condition	Edit dist.
Full system	1.87
-FAITHFULNESS	2.02
-MARKEDNESS	2.18
-Sharing	1.99
-Topology	2.06

Learned Phonological Shifts



Example Parameters



Conclusion

- Languages undergo evolutionary processes
- Can model as regular edits along a tree
- Using modern forms ONLY:
  - We can determine the historical phylogeny
  - We can reconstruct ancient forms (though inherently less accurate for older forms)
- A lot still left to do!

Thank You!