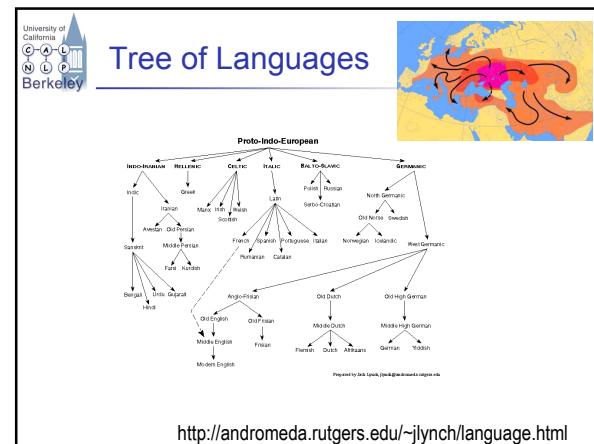


Statistical NLP
Spring 2009



Lecture 30: Diachronic Models
Dan Klein – UC Berkeley
Work with Alex Bouchard-Cote and Tom Griffiths



Language Evolution

Latin **camera /kamera/**
Deletion: /e/
Change of place: /k/ .. /tʃ/ .. /ʃ/
Insertion: /b/
French **chambre /ʃambr/**
Eng. camera from Latin, "camera obscura"

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


Diachronic Evidence

Yahoo! Answers
Resolved Question
Which is correct....tonight or tonite?
Show me another x
12 months ago
Report Abuse

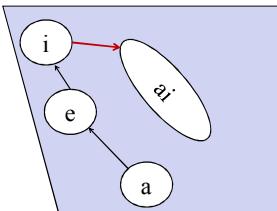
Appendix Probi

tonight not tonite 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

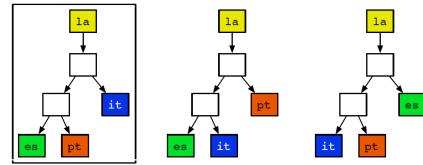


Synchronic (Comparative) Evidence

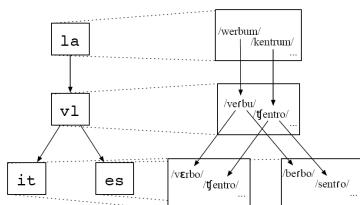
Gloss	Latin	Italian	Spanish	Portuguese
Word/verb	verbum	verbo	verbo	verbu
Fruit	fructus	frutta	fruta	fruta
Laugh	ridere	ridere	reir	rir
Center	centrum	centro	centro	centro
August	augustus	agosto	agosto	agosto
Swim	natare	nuotare	nadar	nadar



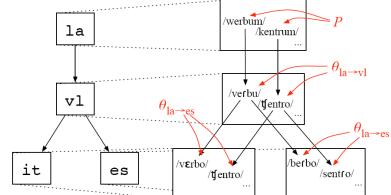
A Mini-Romance Phylogeny



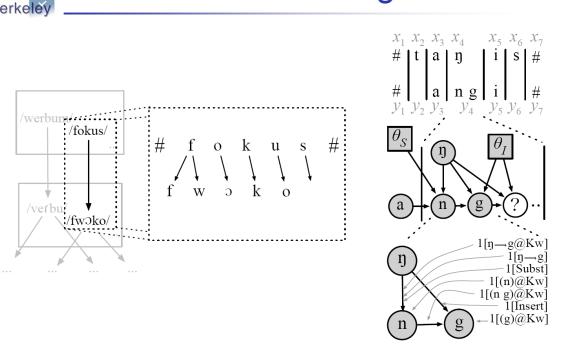
A Probabilistic Model



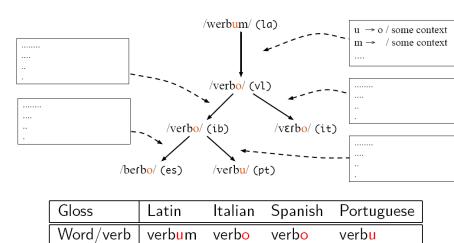
Model Parameters



Local Mutation along Tree



Ancient to Modern Forms



Ancient to Modern Forms

The diagram illustrates the historical development of the word 'centrum'. It starts with the Latin form /kentrum/ (1a), which undergoes mutations to /sentro/ (1b) and /sentru/ (1c). These then lead to further mutations: /sentro/ (1b) leads to /sentro/ (1d) and /sentru/ (1e); /sentru/ (1c) leads to /sentro/ (1f). Finally, these lead to modern forms in Italian, Spanish, and Portuguese: centro, centro, and centro respectively. A table below shows the gloss, Latin, Italian, Spanish, and Portuguese equivalents.

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

Learned Rules / Mutations

This section shows learned rules for mutations. It starts with the Latin form /werbum/ (1a), which undergoes mutations to /verbo/ (1b) and then to /verbu/ (1c). The mutations are defined by rules: m → /-#/, u → o / -/, w → v, and e → ε. Below the diagram, two ancient manuscript fragments are shown, with the reconstructed words coluber and non colober, passim and non passi.

Learned Rules / Mutations

This section shows learned rules for mutations. It starts with the Latin form /berbo/ (1a), which undergoes mutations to /verbo/ (1b) and then to /verbu/ (1c). The mutations are defined by rules: u → o / many environments, v → b / init. or intervocal., t → t e / ALV _ #, and ...

Oceanic Languages

This section discusses Oceanic languages. It includes a map of the Pacific region with red dots indicating language distribution and a phylogenetic tree titled 'Proto-Oceanic' showing the relationships between various language families.

Oceanic Data

This section presents Oceanic data. A table compares words across five language families: Hawaiian, Maori, Samoan, Tongan, and ProtoOceanic. The words listed are 'break', 'house', 'yam', 'woman', and 'moon'.

Gloss	Hawaiian	Maori	Samoan	Tongan	ProtoOceanic
'break'	haki	whati	fati	fasi	*fati
'house'	hale	whare	fale	fale	*fale
'yam'	uhī	uhī	ufī	ufī	*ufī
'woman'	wahine	wahine	fafine	fefine	*wafine
'moon'	mahina	mahina	masina	mahina	*masiana

POc Reconstruction Results

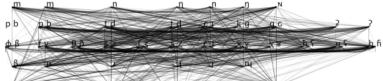
This section presents results of Proto-Oceanic reconstruction. On the left, a graph shows Error (Y-axis, 1.0 to 2.5) versus Number of modern languages (X-axis, 0 to 70). On the right, a table compares different conditions based on edit distance:

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

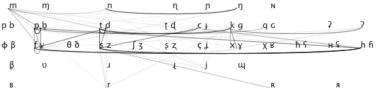


Learned Phonological Shifts

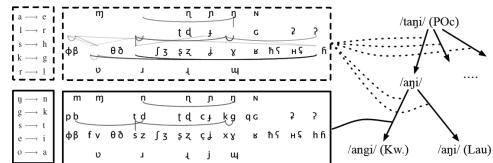
Prior Weights



Posterior Weights



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!