


Statistical NLP Spring 2009



Lecture 30: Diachronic Models

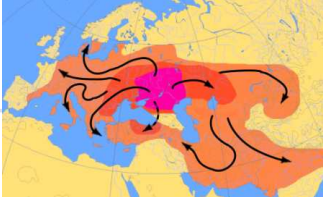
Dan Klein – UC Berkeley

Work with Alex Bouchard-Cote and
Tom Griffiths



University of
California
Berkeley

Tree of Languages



Proto-Indo-European

```

graph TD
    PIE[Proto-Indo-European] --> II[INDO-IRANIAN]
    PIE --> H[HELLENIC]
    PIE --> C[CELTIC]
    PIE --> I[ITALIC]
    PIE --> BS[BALTO-SLAVIC]
    PIE --> G[GERMANIC]

    II --> I1[Indic]
    II --> IR[Iranian]
    I1 --> S[Sanskrit]
    S --> B[Bengali]
    S --> H1[Hindi]
    S --> U[Urdu]
    S --> G1[Gujarati]
    IR --> A[Avestan]
    IR --> OP[Old Persian]
    OP --> F[Farsi]
    OP --> K[Kurdish]

    H --> G2[Greek]

    C --> M[Manx]
    C --> IR1[Irish]
    C --> W[Welsh]
    C --> S1[Scottish]

    I --> L[Latin]
    L --> F1[French]
    L --> Sp[Spanish]
    L --> P[Portuguese]
    L --> It[Italian]
    L --> R[Rumanian]
    L --> C1[Catalan]

    BS --> Po[Polish]
    BS --> Ru[Russian]
    BS --> SC[Serbo-Croatian]

    G --> NG[North Germanic]
    NG --> ON[Old Norse]
    ON --> Nw[Norwegian]
    ON --> I[Islandic]
    ON --> Sw[Swedish]

    G --> WG[West Germanic]
    WG --> AF[Anglo-Frisian]
    AF --> OE[Old English]
    OE --> ME[Middle English]
    ME --> ModE[Modern English]
    AF --> OF[Old Frisian]
    OF --> Fr[Frisian]

    WG --> OD[Old Dutch]
    OD --> MD[Middle Dutch]
    MD --> Fl[Flemish]
    MD --> Du[Dutch]
    MD --> Af[Afrikaans]

    WG --> OHG[Old High German]
    OHG --> MHG[Middle High German]
    MHG --> Ge[German]
    MHG --> Y[Yiddish]
    
```

Prepared by John Lynch, jlynch@cs.rutgers.edu

<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

Appendix Probi

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

"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

tonitru non tonotru

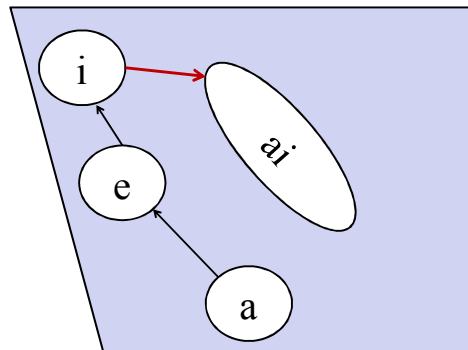
- Spelling (orthography) can reflect old pronunciation
- Corrections show when orthography hasn't kept up!



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

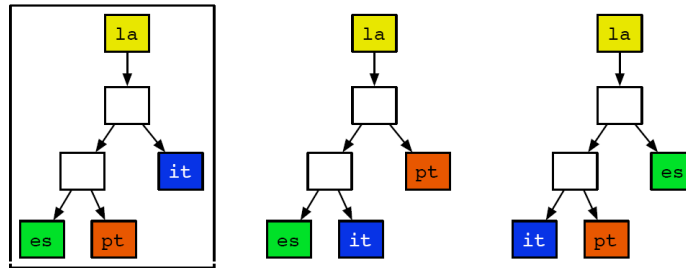


Synchronic (Comparative) Evidence

Gloss	Latin	Italian	Spanish	Portuguese
Word/verb	ver u m	ver o	ver o	ver u
Fruit	fructus	frutta	fruta	fruta
Laugh	ridere	ridere	reir	rir
Center	centr u m	centr o	centr o	centr o
August	aug u stus	ag o sto	ag o sto	ag o sto
Swim	natare	nuotare	nadar	nadar

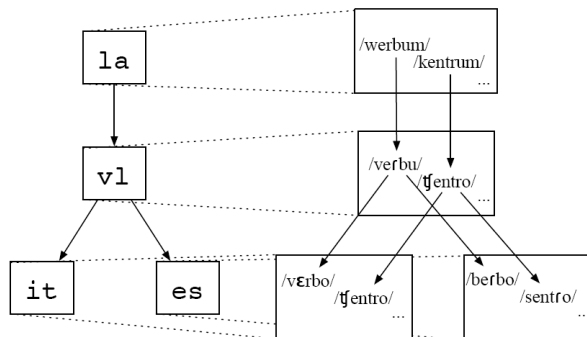


A Mini-Romance Phylogeny

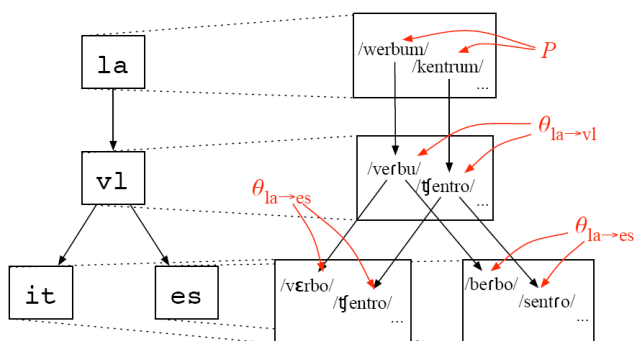




A Probabilistic Model

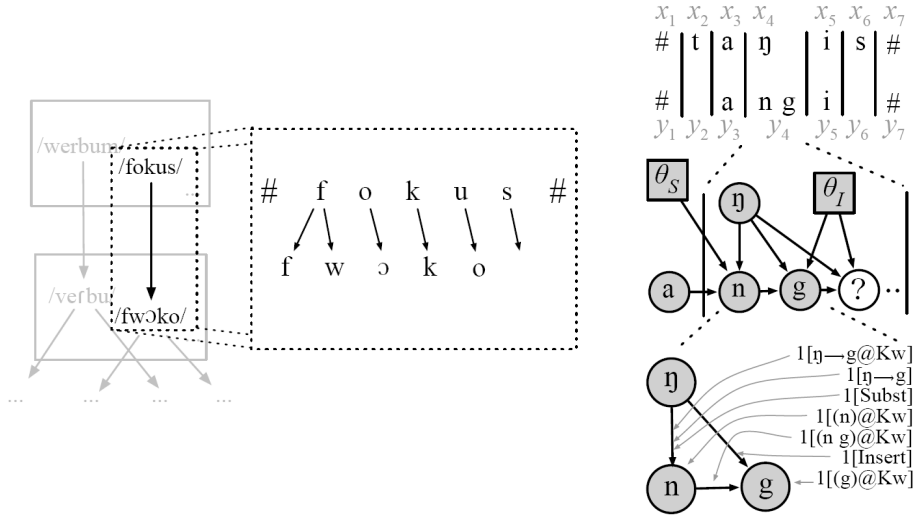


Model Parameters

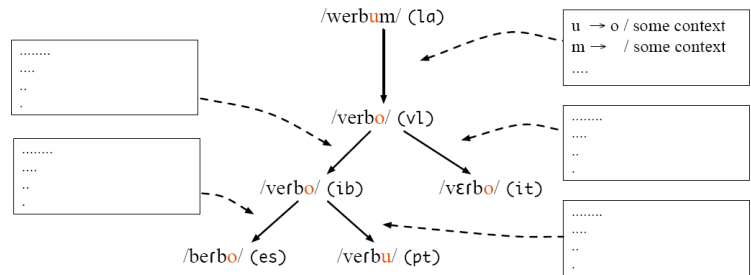




Local Mutation along Tree



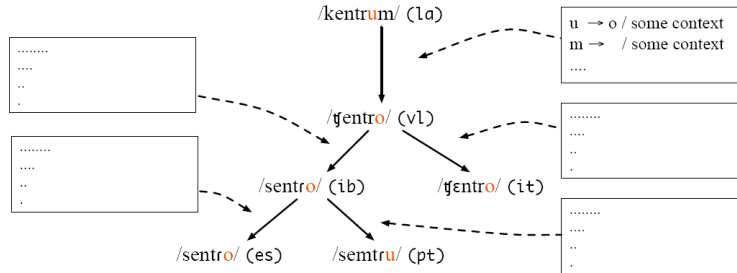
Ancient to Modern Forms



Gloss	Latin	Italian	Spanish	Portuguese
Word/verb	ver u m	ver o	ver o	ver u



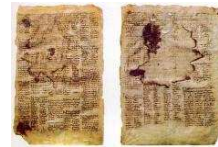
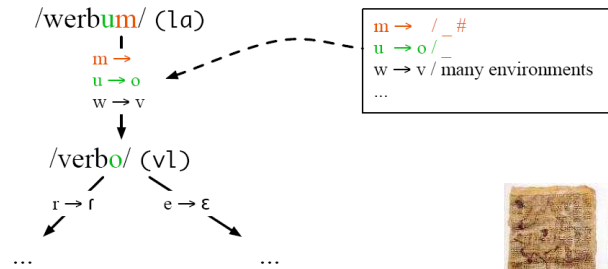
Ancient to Modern Forms



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



Learned Rules / Mutations

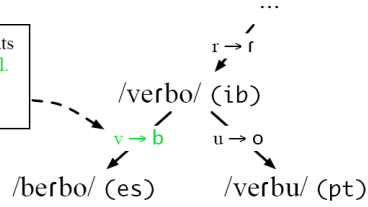


coluber non coluber
 passim non passi



Learned Rules / Mutations

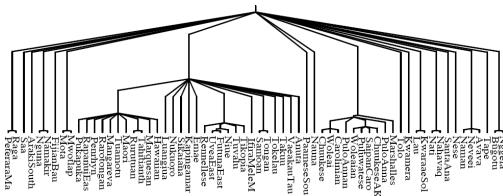
u → o / many environments
 v → b / *init. or intervocal.*
 t → te / ALV_#
 ...



Oceanic Languages



Proto-Oceanic



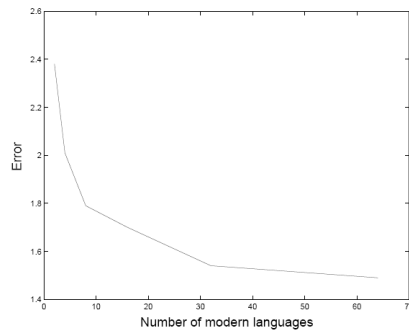


Oceanic Data

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

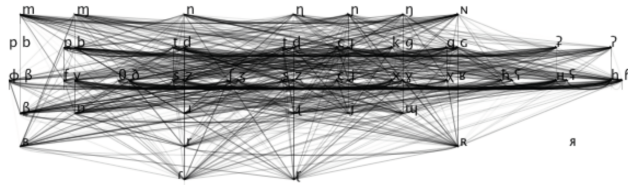


Condition	Edit dist.
Full system	1.87
-FAITHFULNESS	2.02
-MARKEDNESS	2.18
-Sharing	1.99
-Topology	2.06

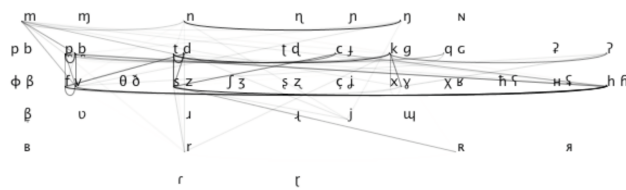


Learned Phonological Shifts

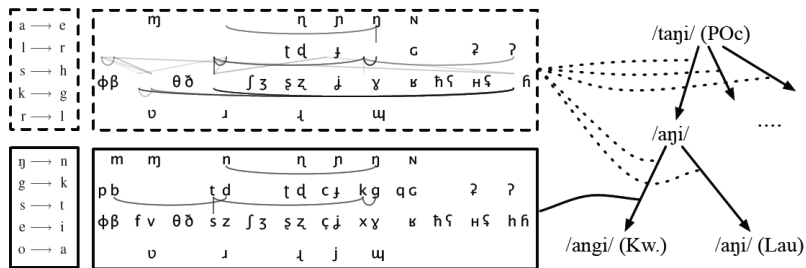
Prior Weights



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!