Natural Language Processing

Diachronics

Dan Klein – UC Berkeley

Includes joint work with Alex Bouchard-Cote, Tom Griffiths, and David Hall

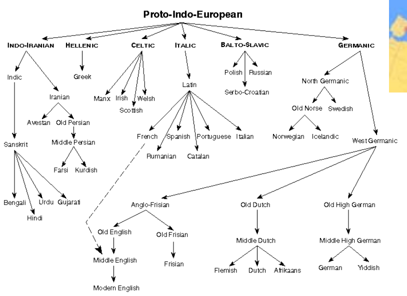
The Task

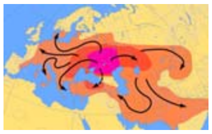
Lexical Reconstruction

Latin
focus

French	Spanish	Italian	Portuguese
feu	fuego	fuoco	fogo

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/ .. /tȍ̈/ .. /ȍ̈/

Insertion: /b/

French chambre /ȍ̈ambЙ/

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 → _

e → _ / after stress

camra /kamra/

Changes Have Structure

camra /kamra/

_ → b

_ → b / m_r

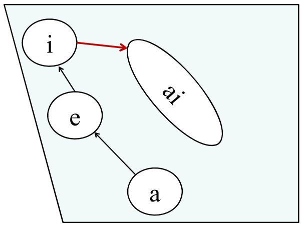
_ → [stop x] / [nasal x]_r

cambra /kambra/

Changes are Systematic

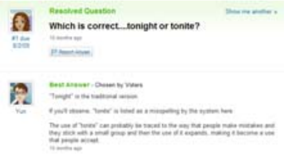
English Great Vowel Shift (Simplified!)

"time" = teem ➡ "time" = taim



Diachronic Evidence

Yahoo! Answers [ca 2000]



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

tonight not tonite

Appendix Probi [ca 300]

tonitru non tonotru

Synchronic (Comparative) Evidence

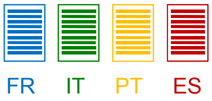
Gloss	Latin	Italian	Spanish	Portuguese
Word/verb	verbum	verbo	verbo	verbu
Fruit	fructus	frutta	fruta	fruta
Laugh	ridere	ridere	reir	rir
Center	centrum	centro	centro	centro
August	augustus	agosto	agosto	agosto
Swim	natare	nuotare	nadar	nadar

Key idea: changes occur uniformly across the lexicon

The Data

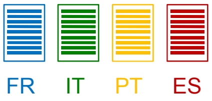
The Data

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

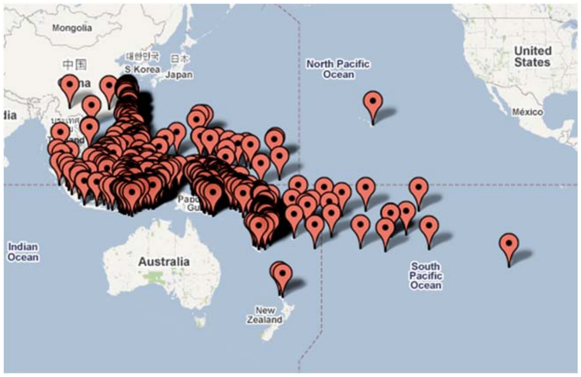


The Data

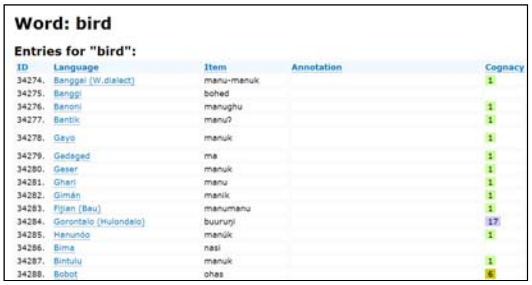
- Data sets
  - Small: Romance
    - French, Italian, Portuguese, Spanish
    - 2344 words
    - Complete cognate sets
    - Target: (Vulgar) Latin
  - Large: Austronesian
    - 637 languages
    - 140K words
    - Incomplete cognate sets
    - Target: Proto-Austronesian



Austronesian



Austronesian Examples



Word: bird

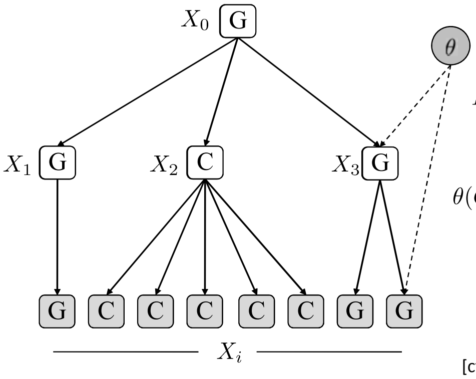
Entries for "bird":

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

From the Austronesian Basic Vocabulary Database

The Model

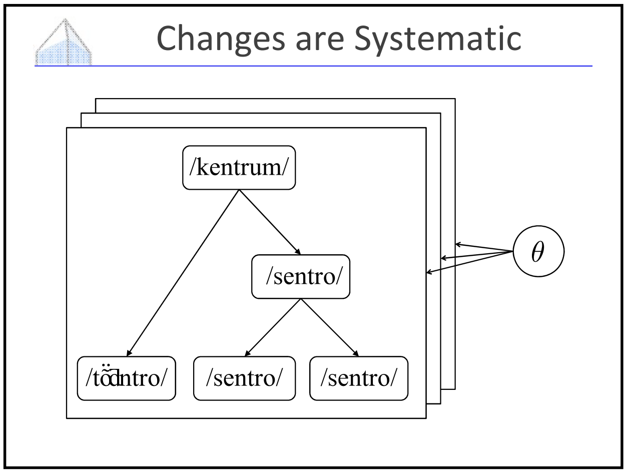
Simple Model: Single Characters



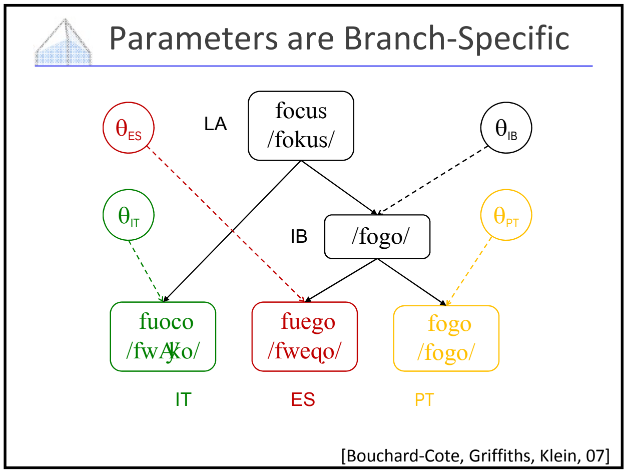
$$P(x|x',\theta) = \theta(x,x')$$

$$\theta(C,G) = 0.02$$

[cf. Felsenstein 81]

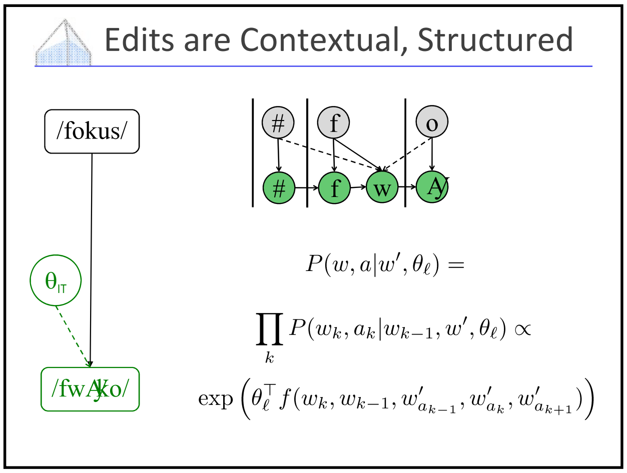


Changes are Systematic



Parameters are Branch-Specific

[Bouchard-Cote, Griffiths, Klein, 07]



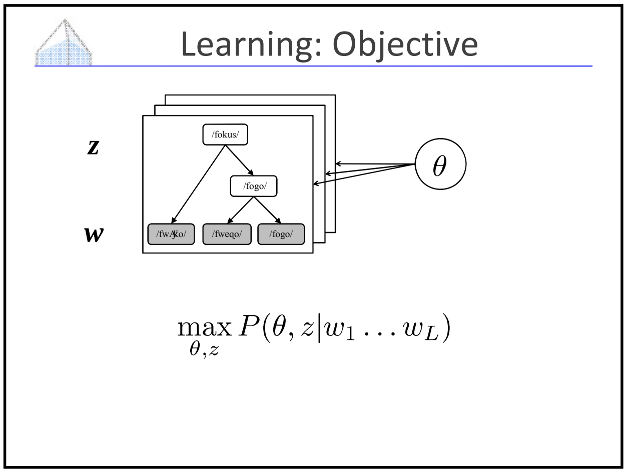
Edits are Contextual, Structured

$$P(w, a|w', \theta_\ell) =$$

$$\prod_k P(w_k, a_k|w_{k-1}, w', \theta_\ell) \propto$$

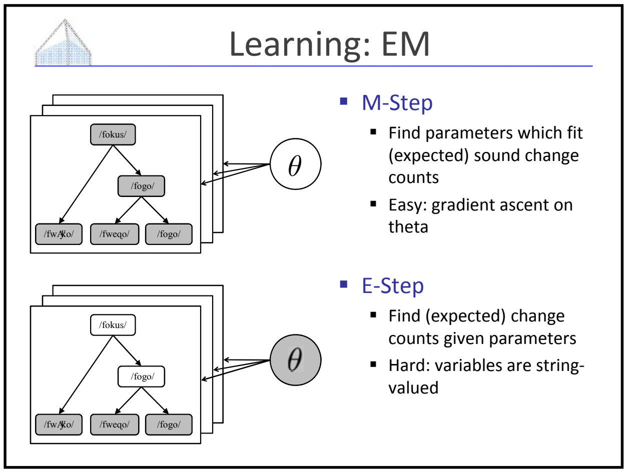
$$\exp\left(\theta_\ell^\top f(w_k, w_{k-1}, w'_{a_{k-1}}, w'_{a_k}, w'_{a_{k+1}})\right)$$

Inference



Learning: Objective

$$\max_{\theta, z} P(\theta, z|w_1 \ldots w_L)$$

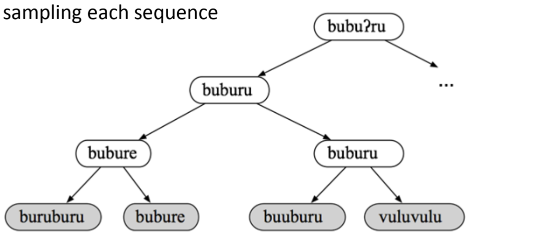


Learning: EM

- 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

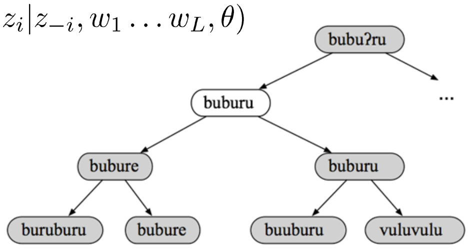


'grass'

[Holmes 01, Bouchard-Cote, Griffiths, Klein 07]

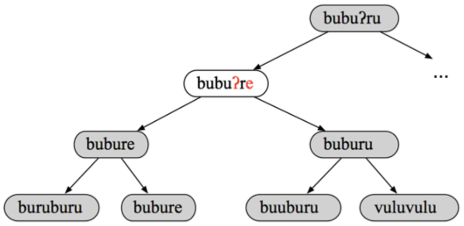
A Gibbs Sampler

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



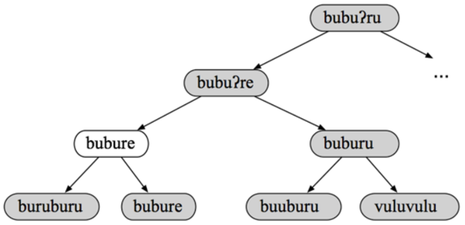
'grass'

A Gibbs Sampler



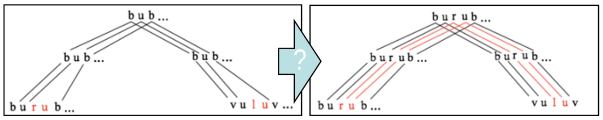
'grass'

A Gibbs Sampler



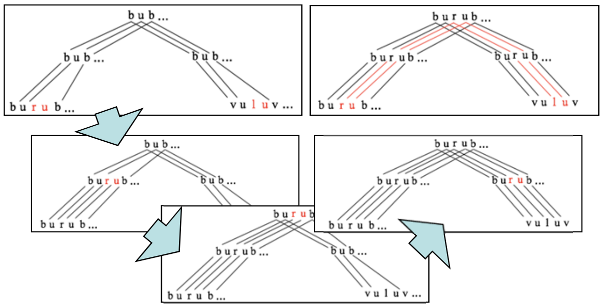
'grass'

Getting Stuck



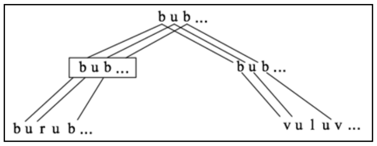
How could we jump to a state where the liquids /r/ and /l/ have a common ancestor?

Getting Stuck

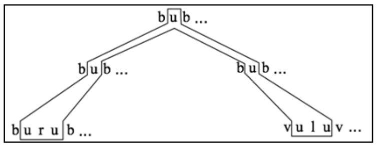


Efficient Sampling: Vertical Slices

Single Sequence Resampling



Ancestry Resampling



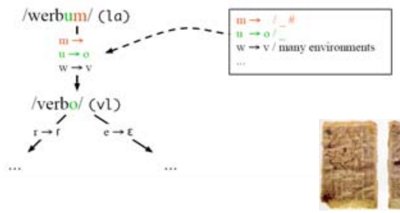
[Bouchard-Cote, Griffiths, Klein, 08]

Results

Results: Romance

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

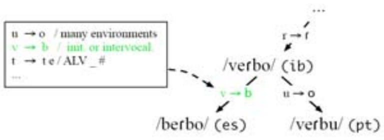
Learned Rules / Mutations



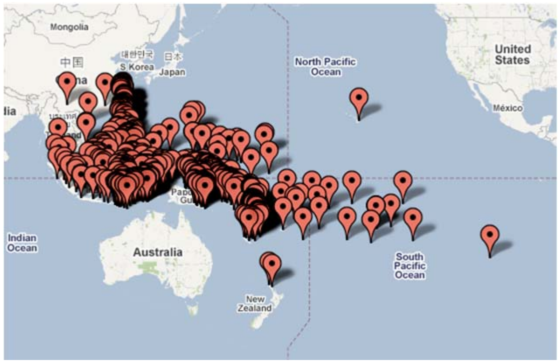
coluber non colober

passim non passi

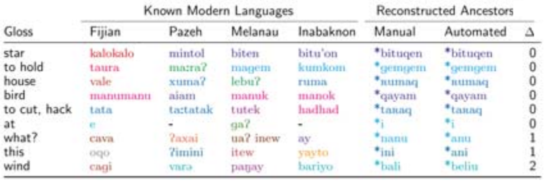
Learned Rules / Mutations



Results: Austronesian



Examples: Austronesian

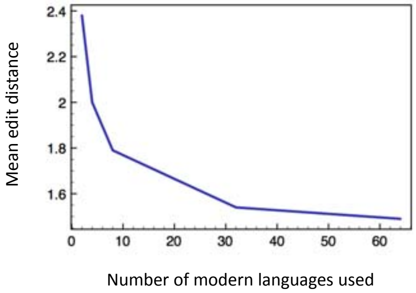


	Known Modern Languages				Reconstructed Ancestors		
Gloss	Fijian	Pazeh	Melanau	Inabaknon	Manual	Automated	Δ
star	kalokalo	mintol	biten	bitu'on	*bituqen	*bituqen	0
to hold	taura	ma:raʔ	magem	kumkom	*gemgem	*gemgem	0
house	vale	xumaʔ	lebuʔ	ruma	*ʀumaq	*ʀumaq	0
bird	manumanu	aiam	manuk	manok	*qayam	*qayam	0
to cut, hack	tata	ta:tatak	tutek	hadhad	*taʀaq	*taʀaq	0
at	e	-	gaʔ	-	*i	*i	0
what?	cava	ʔaxai	uaʔ inew	ay	*nanu	*anu	1
this	oqo	ʔimini	itew	yayto	*ini	*ani	1
wind	cagi	varə	paŋay	bariyo	*bali	*beliu	2

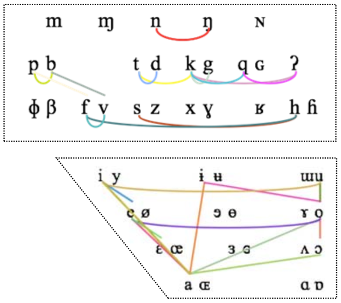
[Bouchard-Cote, Hall, Griffiths, Klein, 13]

Result: More Languages Help

Distance from Blust [1993] Reconstructions



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

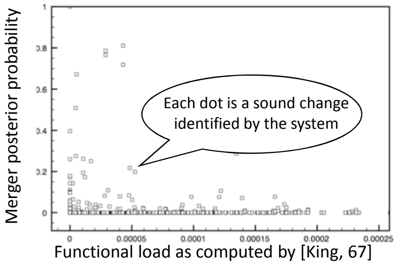
1955: Functional Load Hypothesis (FLH): Sound changes are less frequent when they merge phonemes with high functional load [Martinet, 55]

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

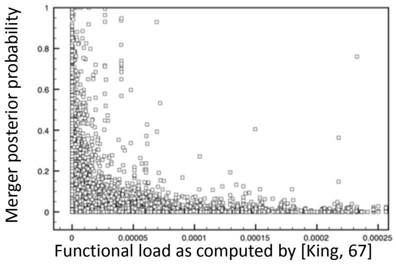
Data: only 4 languages from the Austronesian data



Each dot is a sound change identified by the system

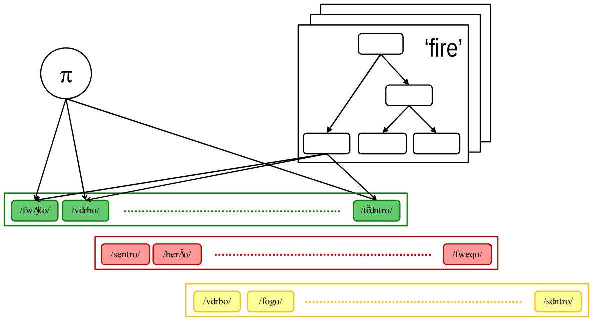
Regularity and Functional Load

Data: all 637 languages from the Austronesian data



Extensions

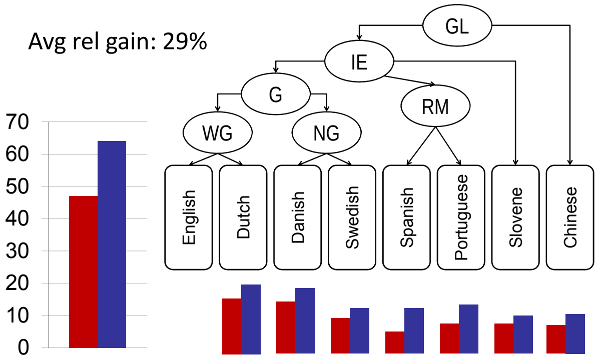
Cognate Detection



[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]