

# Statistical NLP Spring 2010



## Lecture 25: Diachronics

Dan Klein – UC Berkeley

# Evolution: Main Phenomena


The cover of Charles Darwin's book "The Origin of Species". The text on the cover includes: "ON THE ORIGIN OF SPECIES BY MEANS OF NATURAL SELECTION, OR THE PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE. BY CHARLES DARWIN, M.A., FELLOW OF THE ROYAL SOCIETY, LONDON, WITH AN APPENDIX, A THIRTIETH OF 'JOURNAL OF RESEARCHES INTO THE HISTORY AND GEOGRAPHY OF THE MANICOUAGAN RIVER SINCE THE YEAR 1841.' LONDON, JOHN MURRAY, ALBEMARLE STREET. 1859. The right of Translation is reserved."

### Mutations of sequences

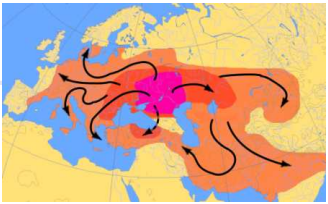
A diagram illustrating a mutation in a DNA sequence. At the top, a dinosaur icon is next to the sequence "C A T A C", where the 'T' is highlighted in blue. Below it, a bird icon is next to the sequence "C A G", where the 'G' is highlighted in purple. A vertical arrow labeled "Time" points downwards from the dinosaur to the bird, indicating the progression of time and the change in the sequence.

### Speciation

A diagram illustrating speciation. At the top, a single dinosaur icon is shown above a vertical line that splits into two diagonal lines leading to two bird icons. A vertical arrow labeled "Time" points downwards from the dinosaur to the birds, indicating the progression of time and the divergence of a single lineage into two distinct species.



# Tree of Languages



**Proto-Indo-European**

```

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

      IR --> Ind[Indic]
      IR --> Iran[Iranian]
      Ind --> Sans[Sanskrit]
      Sans --> Beng[Bengali]
      Sans --> Hindi[Hindi]
      Sans --> Urdu[Urdu]
      Sans --> Gujar[Gujarati]
      Iran --> Avestan[Avestan]
      Iran --> OP[Old Persian]
      OP --> MP[Middle Persian]
      MP --> Farsi[Farsi]
      MP --> Kurdish[Kurdish]

      H --> Greek[Greek]

      C --> Manx[Manx]
      C --> Irish[Irish]
      C --> Welsh[Welsh]
      C --> Scottish[Scottish]

      I --> Latin[Latin]
      Latin --> French[French]
      Latin --> Spanish[Spanish]
      Latin --> Portuguese[Portuguese]
      Latin --> Italian[Italian]
      Latin --> Rumanian[Rumanian]
      Latin --> Catalan[Catalan]


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

      G --> NG[North Germanic]
      G --> WG[West Germanic]
      NG --> ON[Old Norse]
      ON --> Norwegian[Norwegian]
      ON --> Icelandic[Icelandic]
      ON --> Swedish[Swedish]
      WG --> OFrisian[Old Frisian]
      OFrisian --> AngloFrisian[Anglo-Frisian]
      AngloFrisian --> OldEnglish[Old English]
      OldEnglish --> MiddleEnglish[Middle English]
      MiddleEnglish --> ModernEnglish[Modern English]
      OFrisian --> OldFrisian[Old Frisian]
      OldFrisian --> Frisian[Frisian]
      WG --> OldDutch[Old Dutch]
      OldDutch --> MiddleDutch[Middle Dutch]
      MiddleDutch --> Flemish[Flemish]
      MiddleDutch --> Dutch[Dutch]
      MiddleDutch --> Afrikaans[Afrikaans]
      WG --> OldHighGerman[Old High German]
      OldHighGerman --> MiddleHighGerman[Middle High German]
      MiddleHighGerman --> German[German]
      MiddleHighGerman --> Yiddish[Yiddish]
    
```

**Challenge: identify the phylogeny**

- 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>





# Statistical Inference Tasks


### Inputs

Modern Text

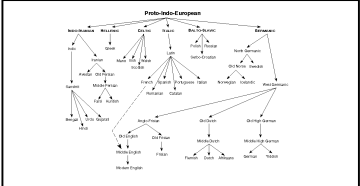
  
FR

  
IT

  
PT

  
ES

Phylogeny



### Outputs

focus


fuego
feu

Ancestral Word Forms

fuego
oeuf

huevo
feu

Cognate Groups / Translations

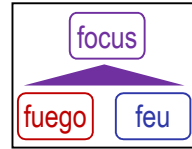


*les faits sont très clairs*

Grammatical Inference



## Outline



Ancestral  
Word Forms



Cognate Groups /  
Translations



Grammatical  
Inference



## Language Evolution: Sound Change

Latin

camera /kamera/

Deletion: /e/

Change: /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 [2009]

Appendix Probi [ca 300]

**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

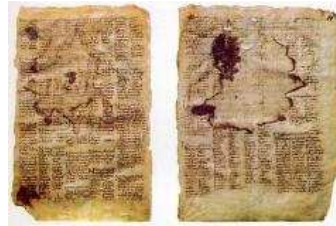
"Tonight" is the traditional version.

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

Yun



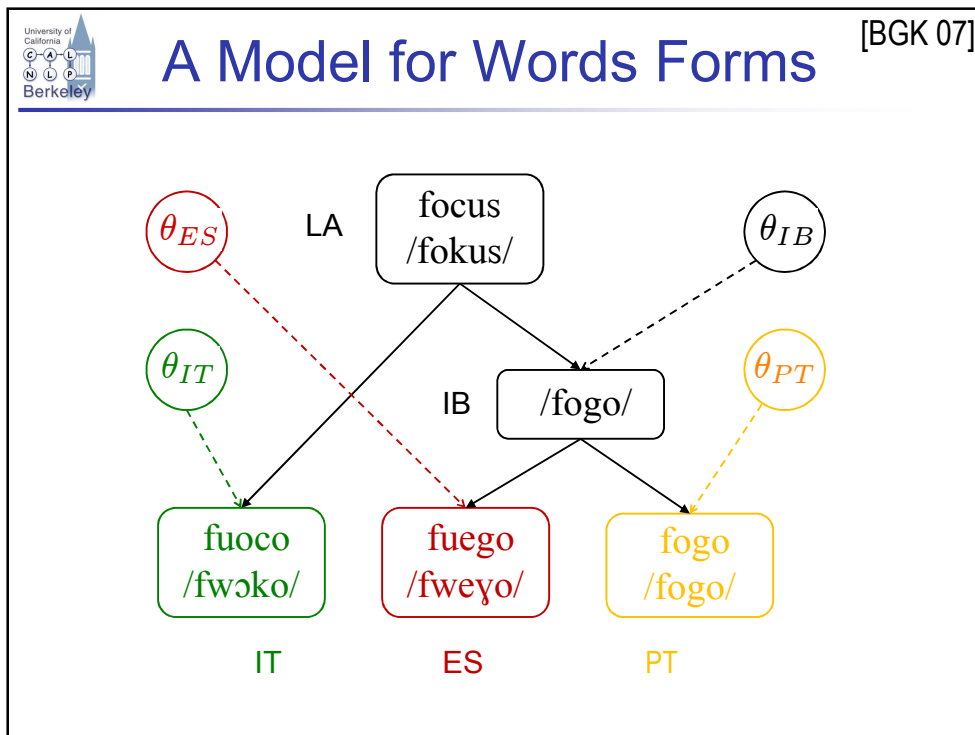
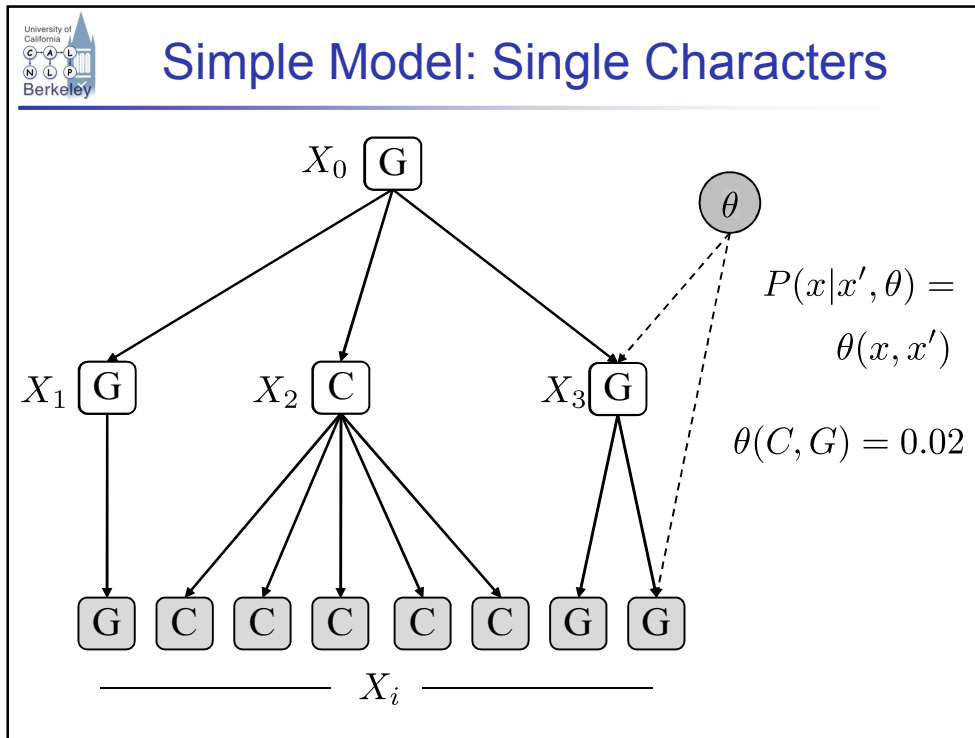
tonight not tonite

tonitru non tonotru



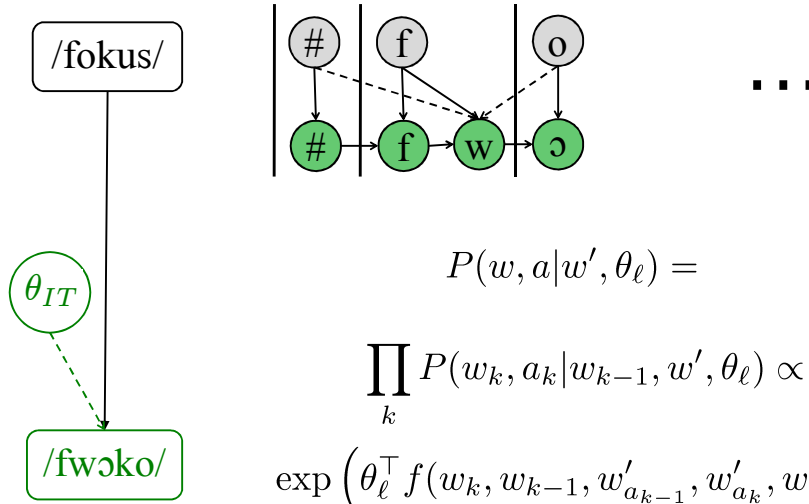
## Synchronic (Comparative) Evidence

Gloss	Latin	Italian	Spanish	Portuguese
Word/verb	ver <u>u</u> m	verbo	verbo	ver <u>u</u>
Fruit	fructus	frutta	fruta	fruta
Laugh	ridere	ridere	reir	rir
Center	centr <u>u</u> m	centro	centro	centro
August	aug <u>u</u> stus	agosto	agosto	agosto
Swim	natare	nuotare	nadar	nadar

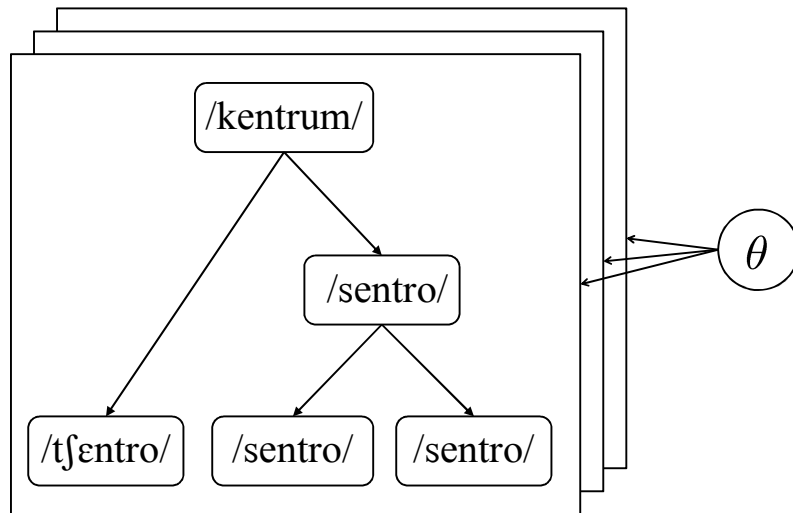




# Contextual Changes



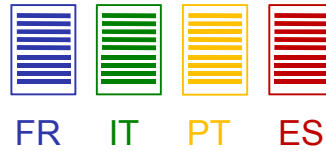
# Changes are Systematic





## Experimental Setup

- Data sets
  - Small: Romance
    - French, Italian, Portuguese, Spanish
    - 2344 words
    - Complete cognate sets
    - Target: (Vulgar) Latin
  - Large: Oceanic
    - 661 languages
    - 140K words
    - Incomplete cognate sets
    - Target: Proto-Oceanic [Blust, 1993]

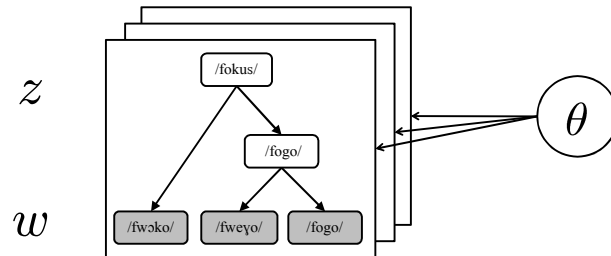


## Data: Romance

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



## Learning: Objective

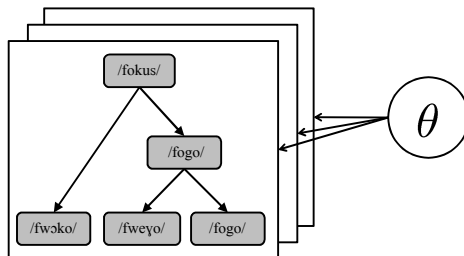


$$\max_{\theta} P(\theta | w_1 \dots w_L)$$

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

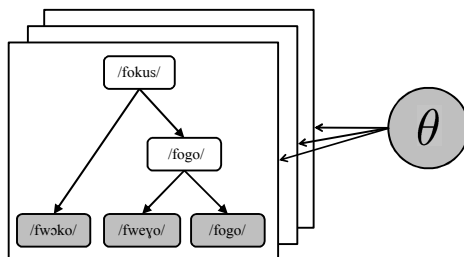


## Learning: EM



### ▪ 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

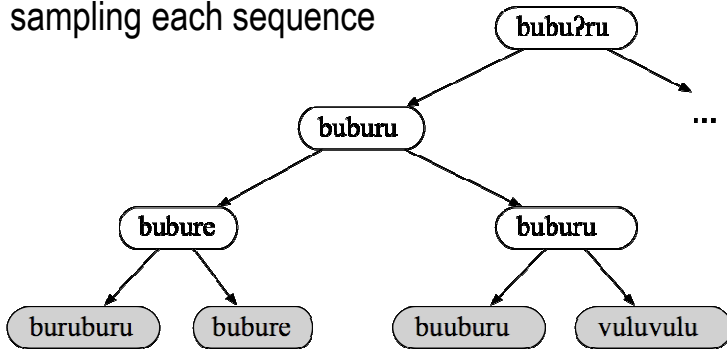




# Computing Expectations

[Holmes 01, BGK 07]

Standard approach, e.g. [Holmes 2001]:  
Gibbs sampling each sequence

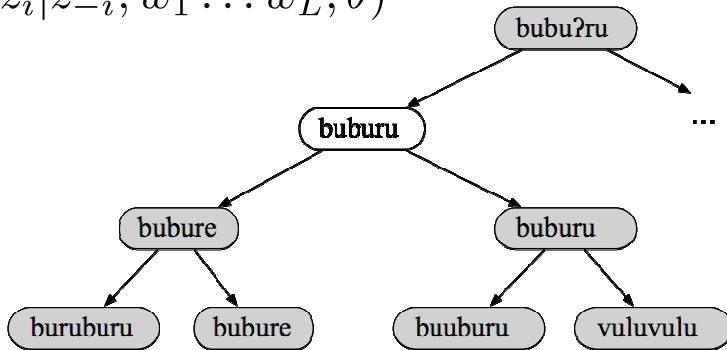


'grass'

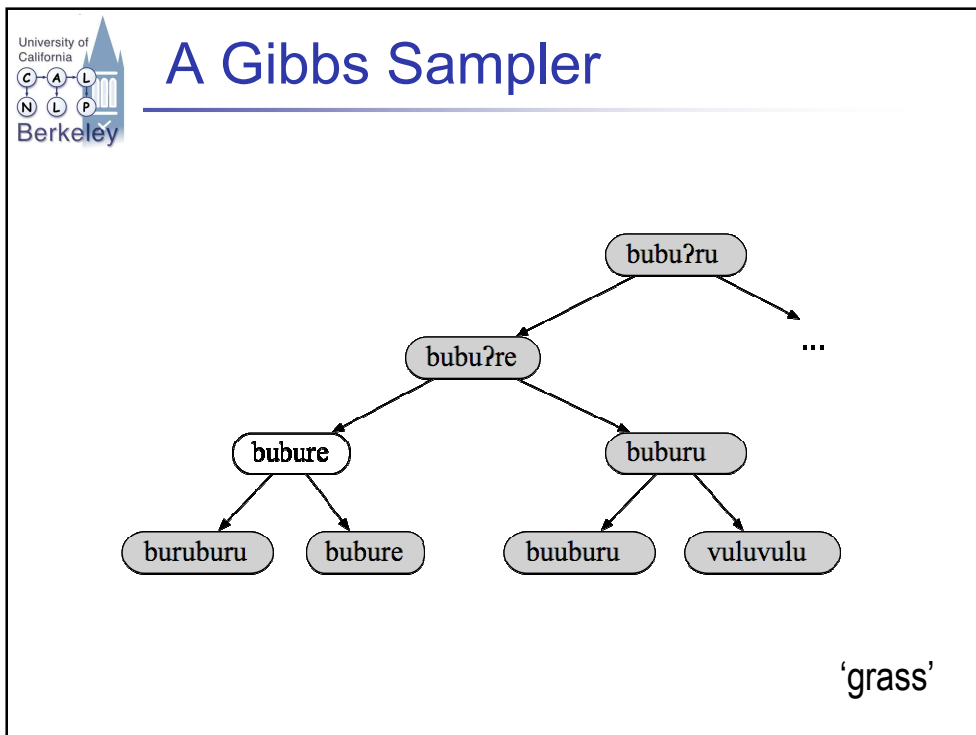
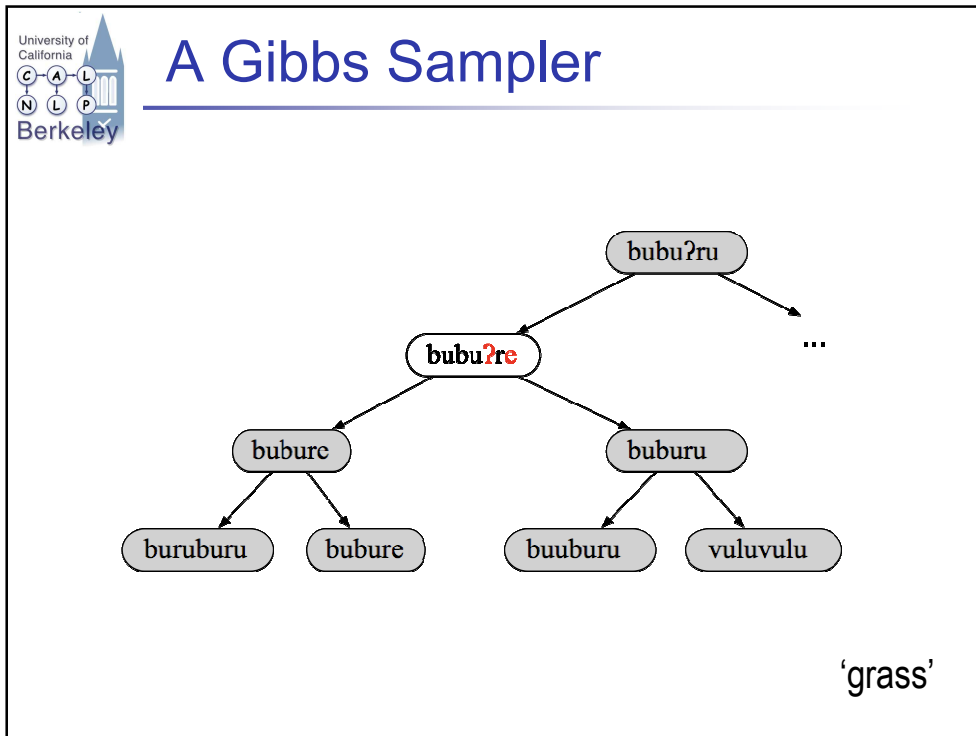


# A Gibbs Sampler

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

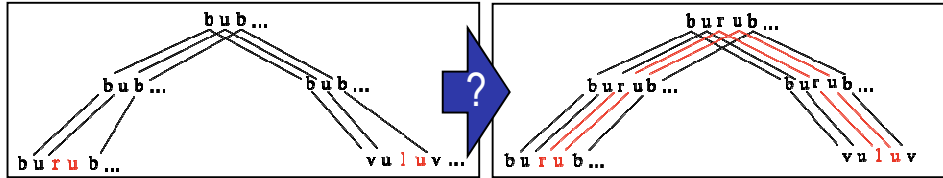


'grass'





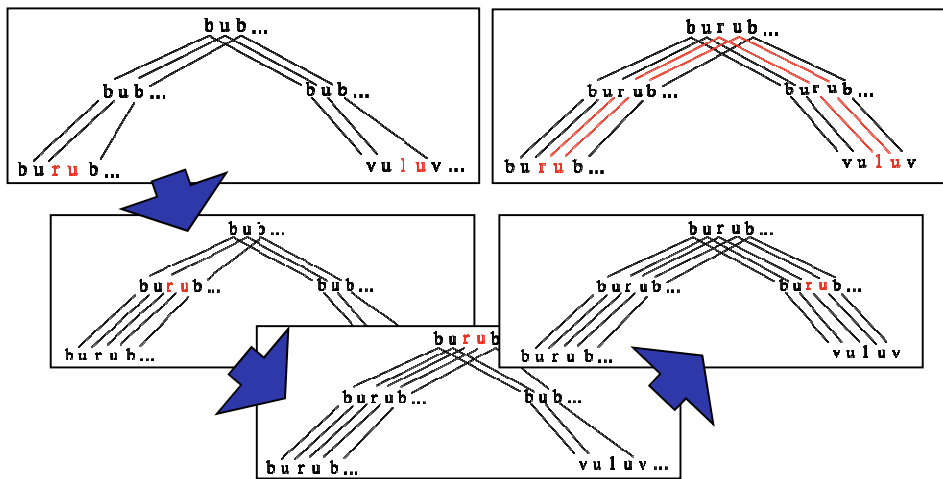
# Getting Stuck



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



# Getting Stuck



University of California  
C A L  
N L P  
Berkeley

[BGK 08]

## Solution: Vertical Slices

Single Sequence Resampling

Ancestry Resampling

University of California  
C A L  
N L P  
Berkeley

## Details: Defining “Slices”

The sampling domains (kernels) are indexed by contiguous subsequences (*anchors*) of the observed leaf sequences

Correct construction  $\text{section}(G)$  is non-trivial but very efficient



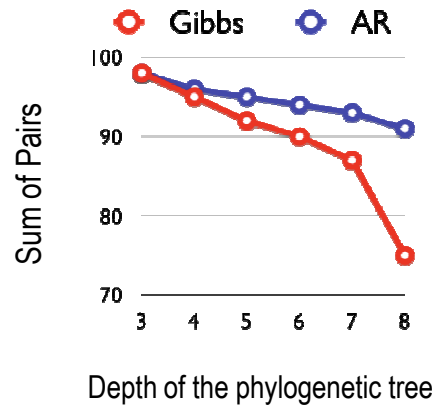
## Results: Alignment Efficiency

Is ancestry resampling faster than basic Gibbs?

Hypothesis: Larger gains for deeper trees

Setup: Fixed wall time

Synthetic data, same parameters

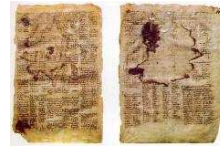
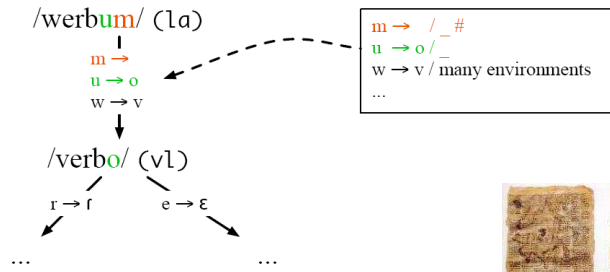


## Results: Romance

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



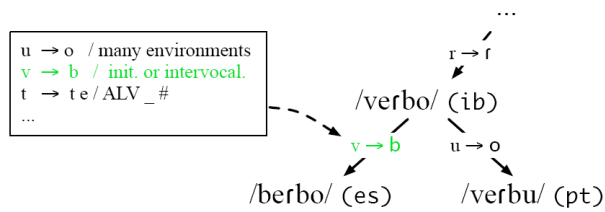
# Learned Rules / Mutations



coluber    non coluber  
passim    non passi



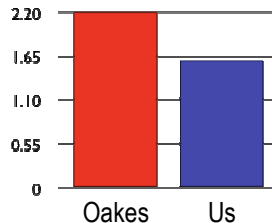
# Learned Rules / Mutations





# Comparison to Other Methods

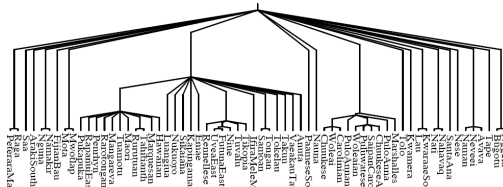
- Evaluation metric: edit distance from a reconstruction made by a linguist (lower is better)
- Comparison to system from [Oakes, 2000]
  - Uses exact inference and deterministic rules
  - Reconstruction of Proto-Malayo-Javanic cf [Nothofer, 1975]



# Data: Oceanic



## Proto-Oceanic





## Data: Oceanic

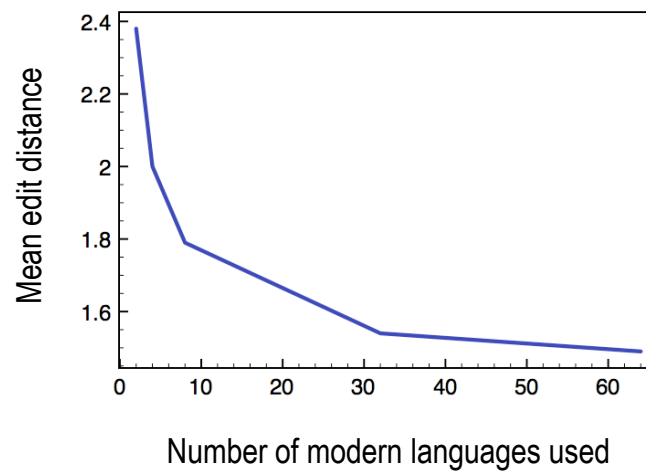
Gloss	Hawai'ian	Maori	Samoan	Tongan
'break'	haki	whati	fati	fasi
'house'	hale	whare	fale	fale
'yam'	uhi	uhi	ufi	ufi
'woman'	wahine	wahine	fafine	fefine
'moon'	mahina	mahina	masina	mahina

<http://language.psy.auckland.ac.nz/austronesian/research.php>



## Result: More Languages Help

Distance from [Blust, 1993] Reconstructions









## Regularity and Functional Load

In a language, some pairs of sounds are more contrastive than others (higher functional load)

**Example:** English “p”/“b” versus “t”/“th”

“p”/“b”: pot/dot, pin/din, dress/press,  
pew/dew, ...

“t”/“th”: thin/tin



## Functional Load: Timeline

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

**Previous research within linguistics:** “FLH does not seem to be supported by the data” [King, 67]

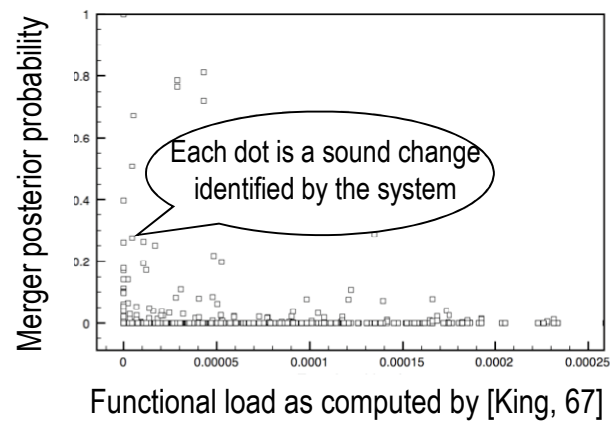
**Caveat:** only four languages were used in King’s study [Hockett 67; Surandran et al., 06]

**Our work:** we reexamined the question with two orders of magnitude more data [BGK, *under review*]



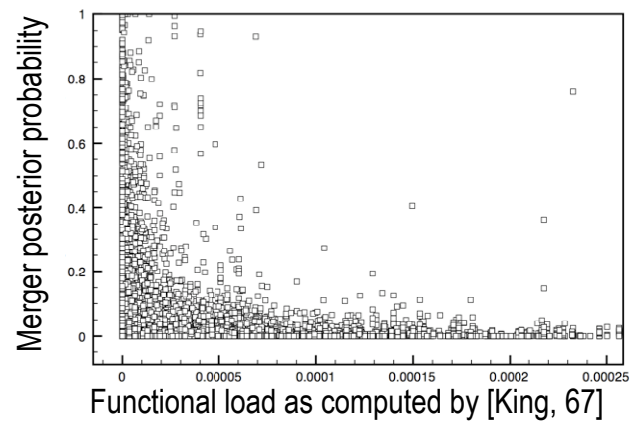
## Regularity and Functional Load

**Data:** only 4 languages from the Austronesian data



## Regularity and Functional Load

**Data:** all 637 languages from the Austronesian data

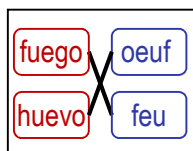




# Outline



Ancestral  
Word Forms



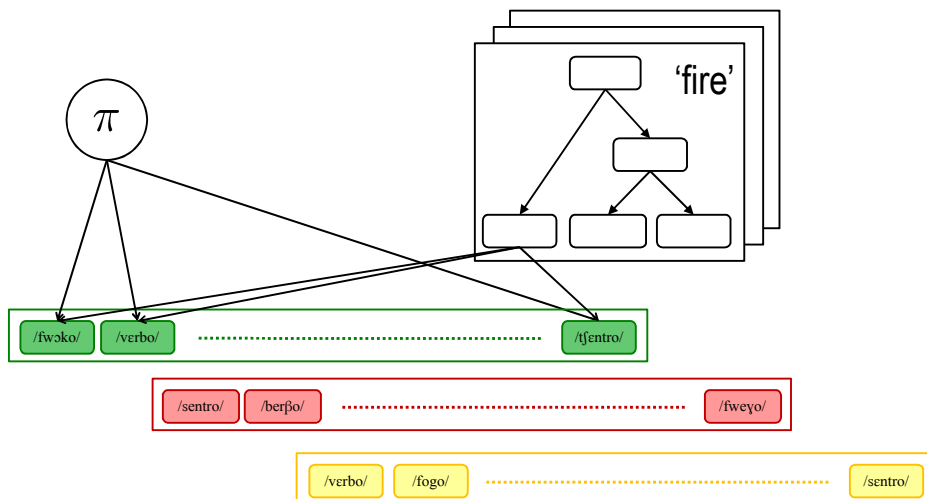
Cognate Groups /  
Translations



Grammatical  
Inference

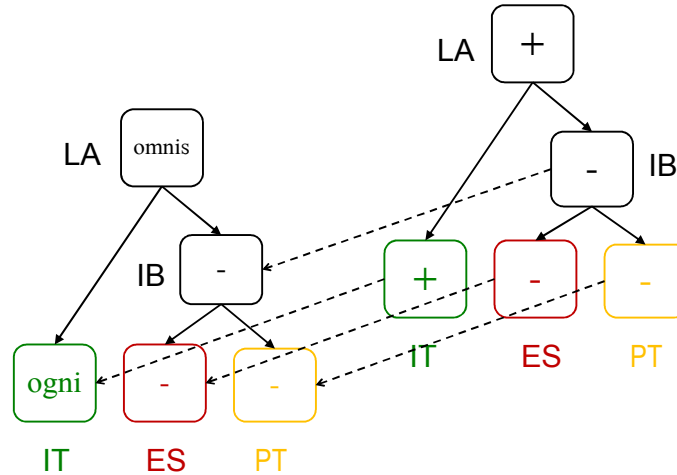


# Cognate Groups



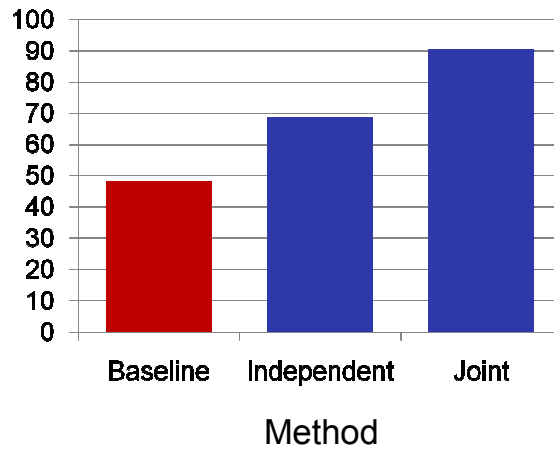


## Model: Cognate Survival



## Results: Grouping Accuracy

Fraction of Words Correctly Grouped



[Hall and Klein, in submission]



## Semantics: Matching Meanings

EN

day

Occurs with:

"night"

"sun"

"week"

tag

EN

Occurs with:

"name"

"label"

"along"

DE

tag

Occurs with:

"nacht"

"sonne"

"woche"



## Outline



Ancestral  
Word Forms



Cognate Groups /  
Translations

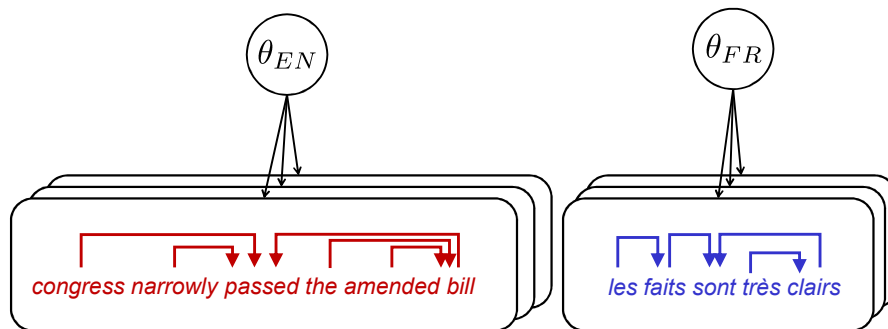


Grammatical  
Inference



## Grammar Induction

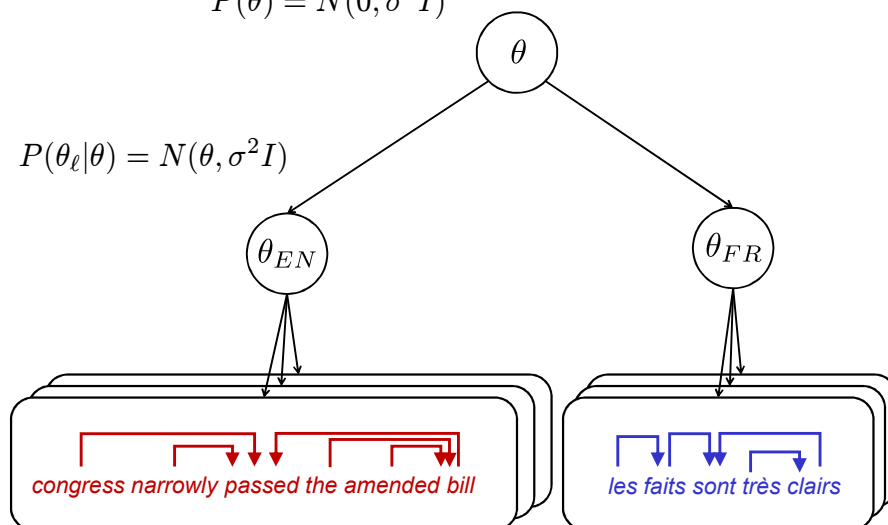
**Task:** Given sentences, infer grammar  
(and parse tree structures)

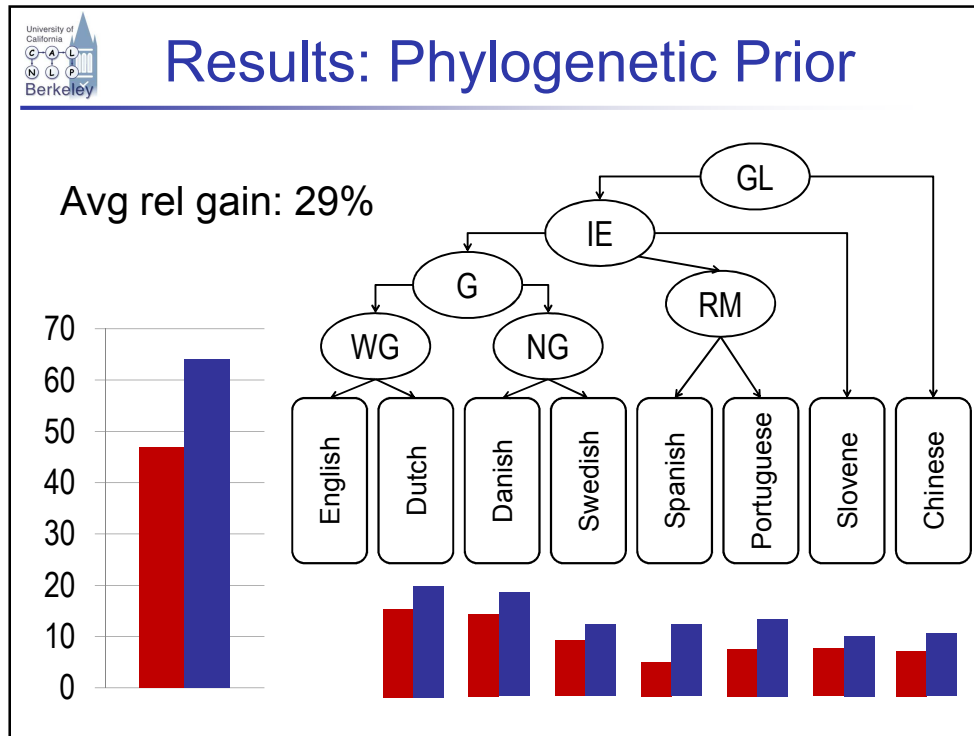


## Shared Prior

$$P(\theta) = N(0, \sigma^2 I)$$

$$P(\theta_\ell | \theta) = N(\theta, \sigma^2 I)$$





University of California Berkeley

## Conclusion

- Phylogeny-structured models can:
  - Accurately reconstruct ancestral words
  - Give evidence to open linguistic debates
  - Detect translations from form and context
  - Improve language learning algorithms
- Lots of questions still open:
  - Can we get better phylogenies using these high-res models?
  - What do these models have to say about the very earliest languages? Proto-world?



Thank you!



[nlp.cs.berkeley.edu](http://nlp.cs.berkeley.edu)