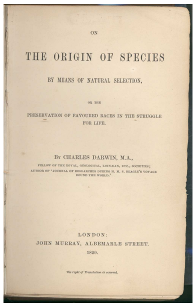
Statistical NLP
Spring 2010

Lecture 25: Diachronics

Dan Klein – UC Berkeley

Evolution: Main Phenomena



ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

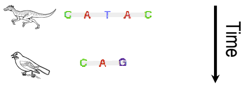
By CHARLES DARWIN, M.A.,

FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;
AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE ROUND THE WORLD.'

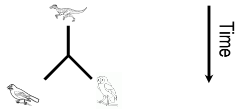
LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

The right of Translation is reserved.

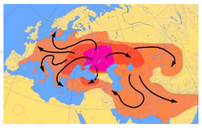
Mutations of sequences

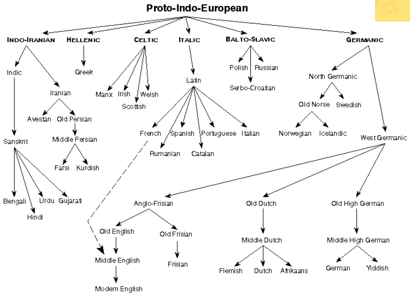


Speciation



Tree of Languages





Challenge: identify the phylogeny

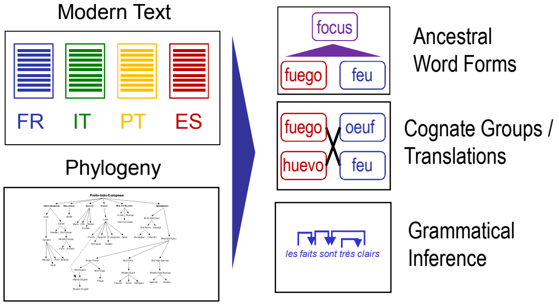
- Much work in biology, e.g. work by Warnow, Felsenstein, Steele…
- Also in linguistics, e.g. Warnow et al., Gray and Atkinson…

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

Statistical Inference Tasks

Inputs

Outputs



Modern Text

FR IT PT ES

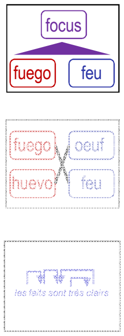
Phylogeny

Ancestral Word Forms

Cognate Groups / Translations

Grammatical Inference

Outline



Ancestral Word Forms

Cognate Groups / Translations

Grammatical Inference

Language Evolution: Sound Change

Latin camera /kamera/

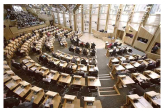
Deletion: /e/

Change: /k/ .. /tʃ/ .. /ʃ/

Insertion: /b/

French chambre /ʃambʁ/

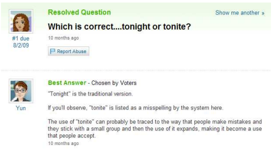
Eng. camera from Latin, "camera obscura"



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

Diachronic Evidence

Yahoo! Answers [2009]



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.

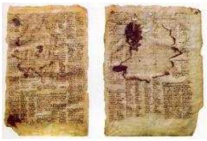
If you'll observe, "tonite" is listed as a misspelling by the system here.

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 [ca 300]

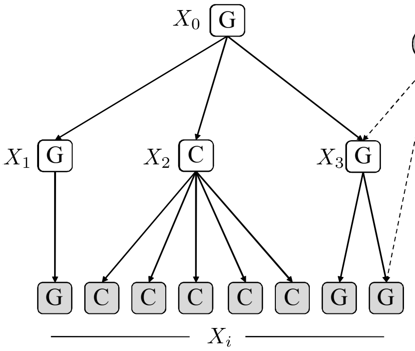


tonitru non tonotru

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

Simple Model: Single Characters

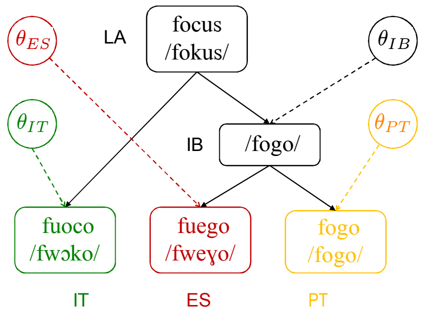


$$P(x|x',\theta) = \theta(x,x')$$

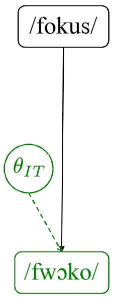
$$\theta(C,G) = 0.02$$

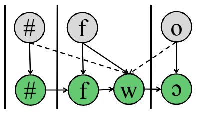
A Model for Words Forms

[BGK 07]



Contextual Changes





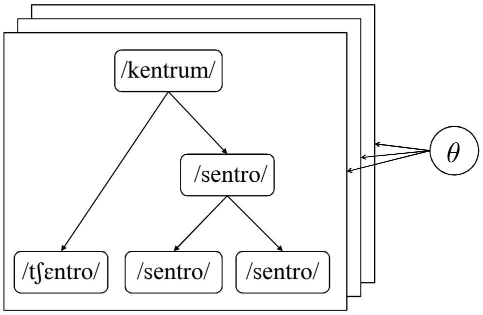
. . .

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

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

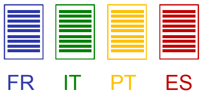
$$\exp\left(\theta_\ell^\top f(w_k,w_{k-1},w'_{a_{k-1}},w'_{a_k},w'_{a_{k+1}})\right)$$

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]

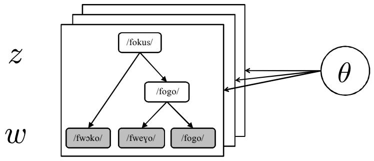


Mongolia
S Korea
Japan
North Pacific Ocean
United States
México
Indian Ocean
Australia
New Zealand
South Pacific Ocean

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

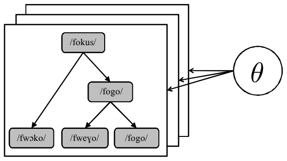
Learning: Objective



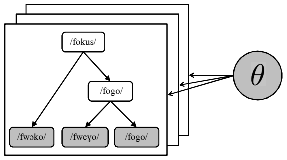
$$\max_{\theta} P(\theta | w_1 \dots w_L)$$

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

Learning: EM



- M-Step
  - Find parameters which fit (expected) sound change counts
  - Easy: gradient ascent on theta

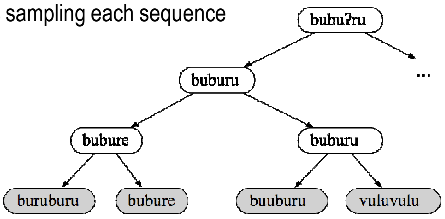


- E-Step
  - Find (expected) change counts given parameters
  - Hard: variables are string-valued

Computing Expectations

[Holmes 01, BGK 07]

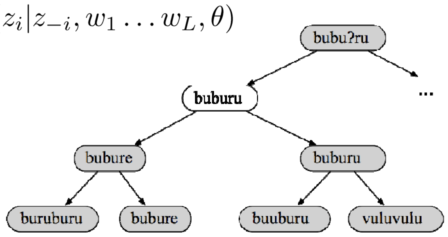
Standard approach, e.g. [Holmes 2001]:
Gibbs sampling each sequence



'grass'

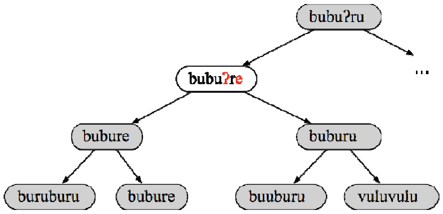
A Gibbs Sampler

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



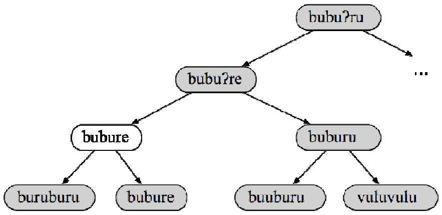
'grass'

A Gibbs Sampler



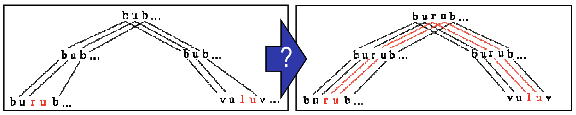
'grass'

A Gibbs Sampler



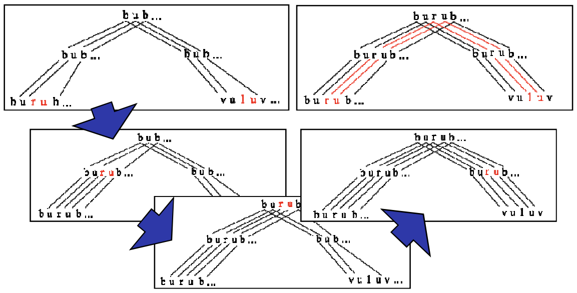
'grass'

Getting Stuck



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

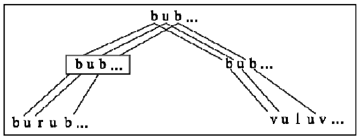
Getting Stuck



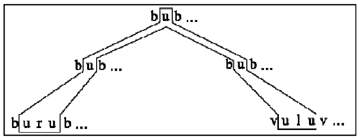
Solution: Vertical Slices

[BGK 08]

Single Sequence Resampling

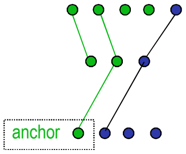


Ancestry Resampling



Details: Defining "Slices"

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



Correct construction section(G) is non-trivial but very efficient

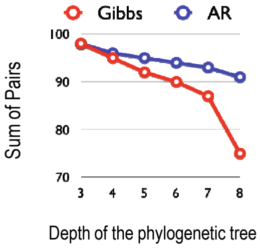
Results: Alignment Efficiency

Is ancestry resampling faster than basic Gibbs?

Hypothesis: Larger gains for deeper trees

Setup: Fixed wall time

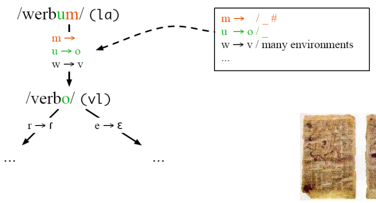
Synthetic data, same parameters



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



/werbum/ (la)

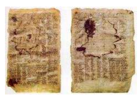
m →
u → o
w → v

/verbo/ (vl)

r → ɾ
e → ɛ

...
...

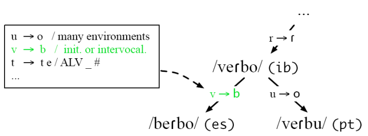
m → / _#
u → o / _
w → v / many environments
...



coluber non colober

passim non passi

Learned Rules / Mutations



u → o / many environments
v → b / init. or intervocal.
t → t e / ALV _ #
...

...

r → ɾ

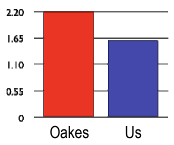
/veɾbo/ (ib)

v → b
u → o

/beɾbo/ (es)
/veɾbu/ (pt)

Comparison to Other Methods

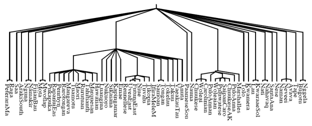
- Evaluation metric: edit distance from a reconstruction made by a linguist (lower is better)
- Comparison to system from [Oakes, 2000]
  - Uses exact inference and deterministic rules
  - Reconstruction of Proto-Malayo-Javanic cf [Nothefer, 1975]



Data: Oceanic

Australia
North Pacific Ocean
United States
South Pacific Ocean
Indian Ocean

Proto-Oceanic



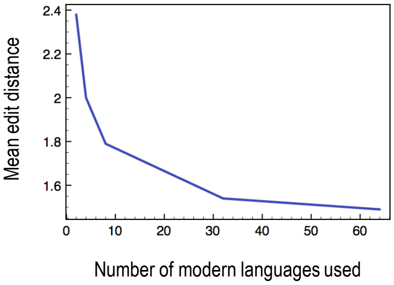
Data: Oceanic

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

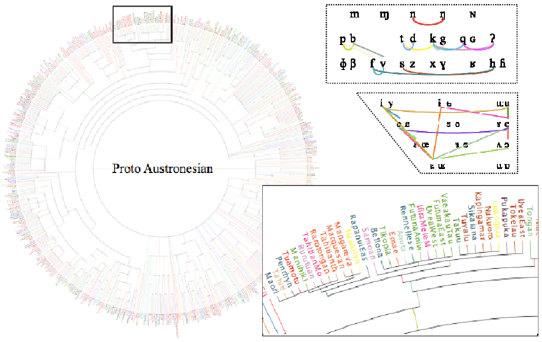
http://language.psy.auckland.ac.nz/austronesian/research.php

Result: More Languages Help

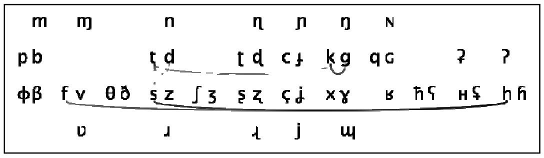
Distance from [Blust, 1993] Reconstructions



Results: Large Scale



Visualization: Learned universals



m ɱ n ɳ ɲ ŋ ɴ

p b t d ʈ ɖ c ɟ k g q ɢ ʡ ʔ

ɸ β f v θ ð s z ʃ ʒ ʂ ʐ ç ʝ x ɣ ʁ ħ ʕ ʜ ʢ h ɦ

ʋ ɹ ɻ j ɰ

*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"/"b" versus "t"/"th"

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

"t"/"th": thin/tin

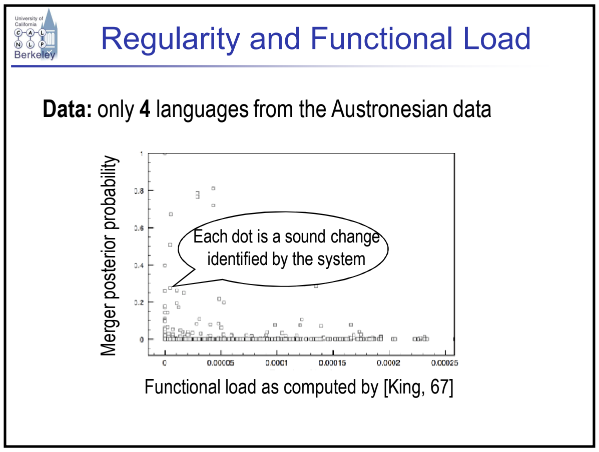
Functional Load: Timeline

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 [Hocket 67; Surandran et al., 06]

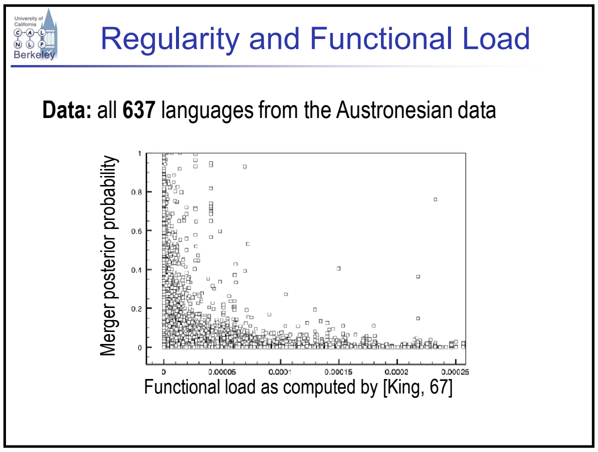
Our work: we reexamined the question with two orders of magnitude more data [BGK, *under review*]



Regularity and Functional Load

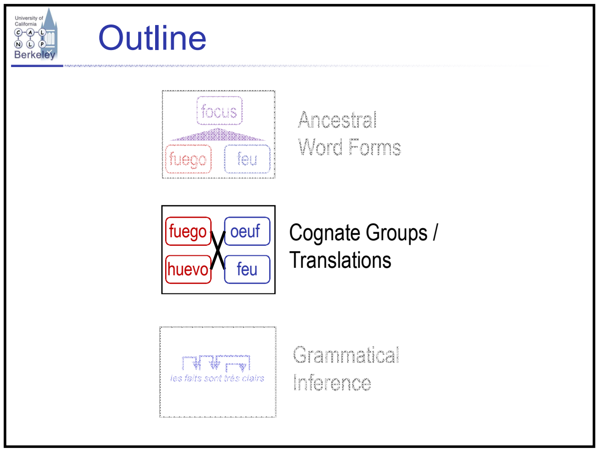
Data: only **4** languages from the Austronesian data

Each dot is a sound change identified by the system



Regularity and Functional Load

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

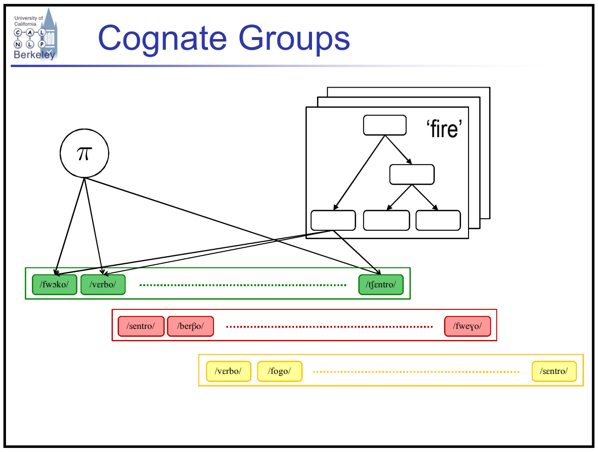


Outline

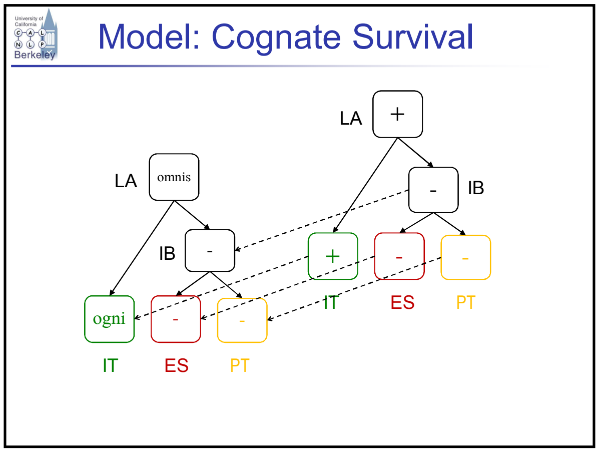
Ancestral Word Forms

Cognate Groups / Translations

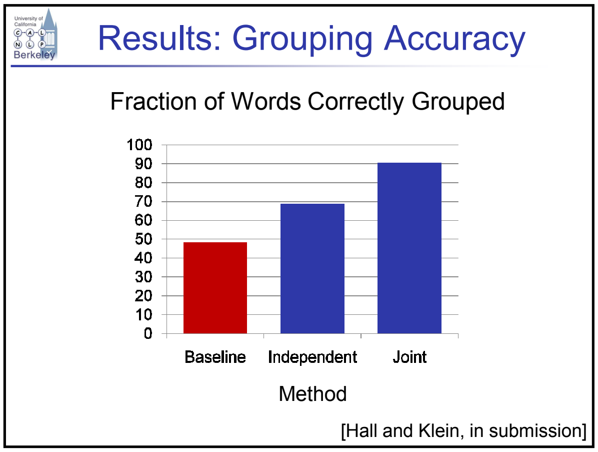
Grammatical Inference



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"

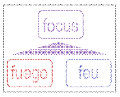
DE tag

Occurs with:
"nacht"
"sonne"
"woche"

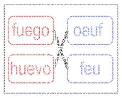
tag EN

Occurs with:
"name"
"label"
"along"

Outline



Ancestral Word Forms

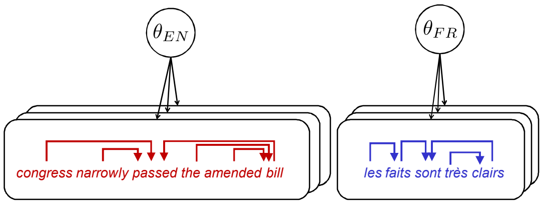


Cognate Groups / Translations

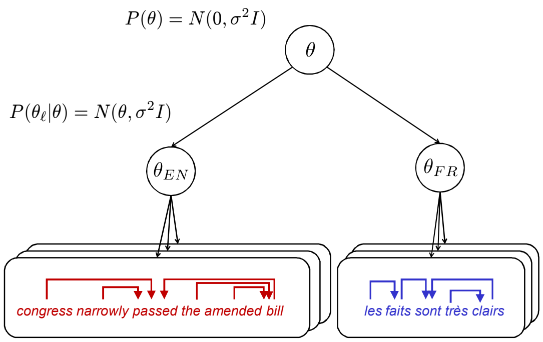
Grammatical Inference

Grammar Induction

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

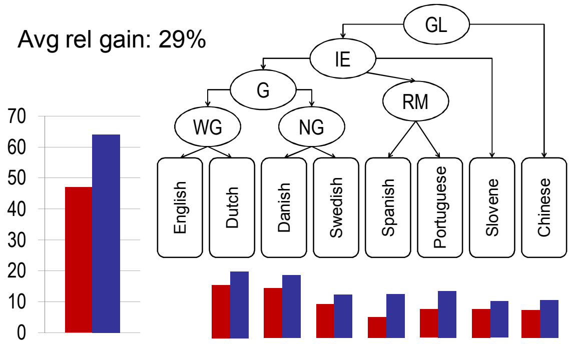

θ_{EN}
θ_{FR}
congress narrowly passed the amended bill
les faits sont très clairs

Shared Prior


$P(\theta) = N(0, \sigma^2 I)$
θ
$P(\theta_\ell|\theta) = N(\theta, \sigma^2 I)$
θ_{EN}
θ_{FR}
congress narrowly passed the amended bill
les faits sont très clairs

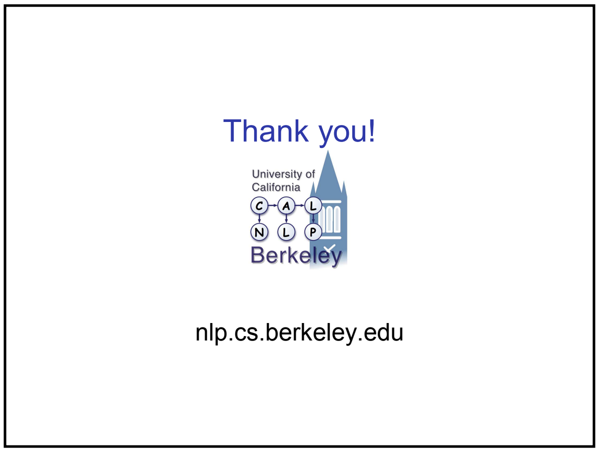
Results: Phylogenetic Prior

Avg rel gain: 29%



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