A Probabilistic Approach to Language Change

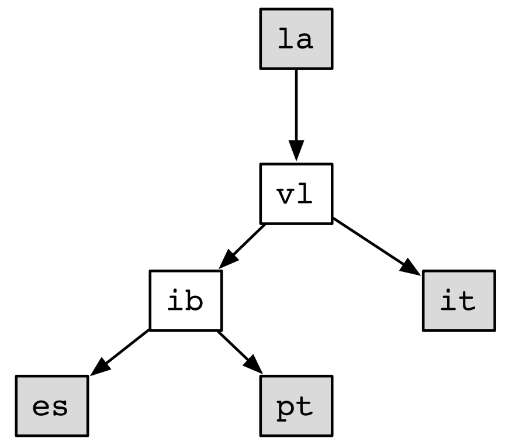
Alexandre Bouchard-Côté* Percy Liang* Thomas L. Griffiths† Dan Klein*

* Computer Science Division † Department of Psychology
University of California at Berkeley

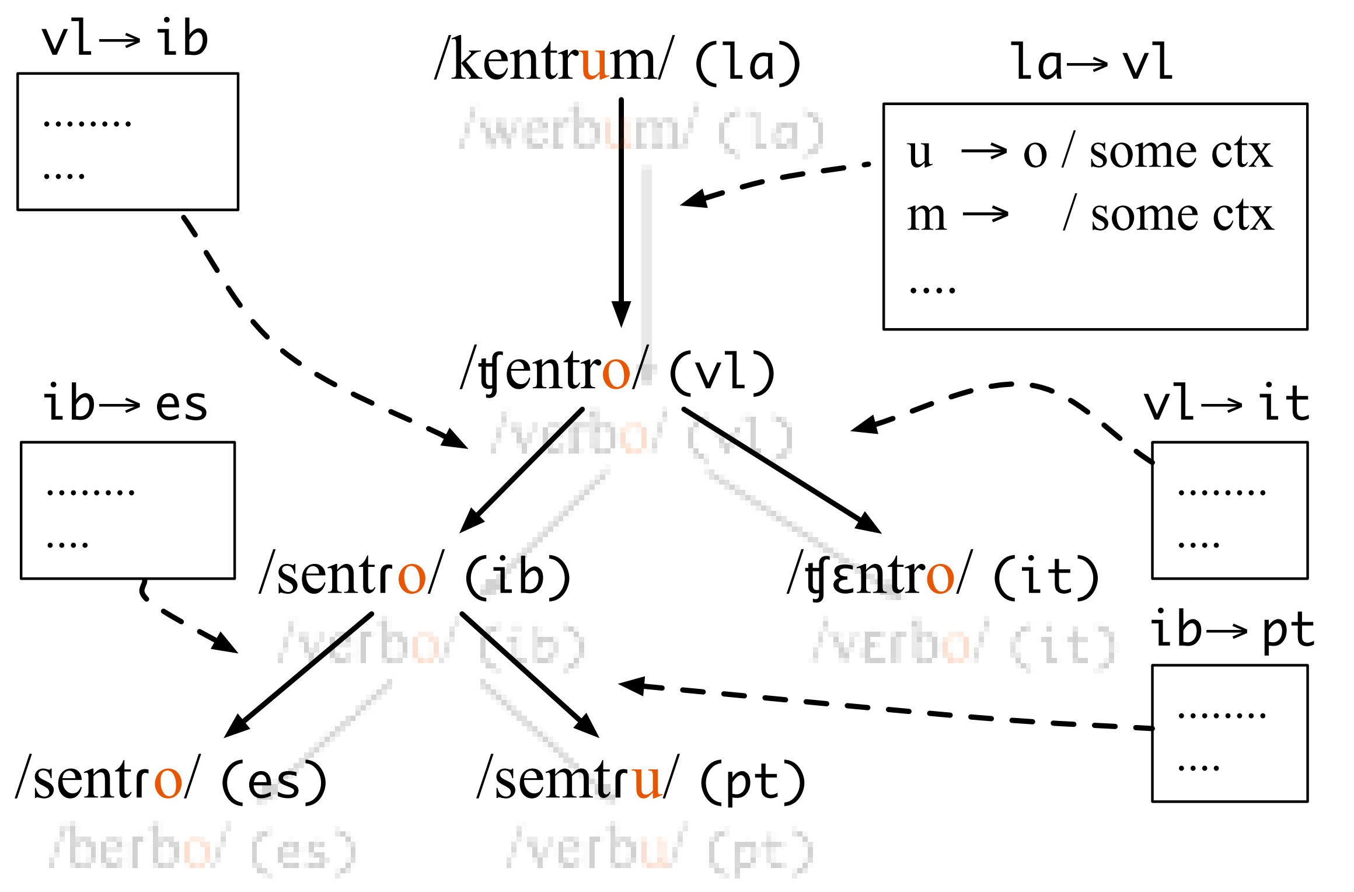
Gloss	Latin	Italian	Spanish	Portuguese
Word/verb	verbum	verbo	verbo	verbu
Fruit	fructus	frutta	fruta	fruta
Laugh	ridere	ridere	reir	rir
Center	centrum	centro	centro	centro
August	augustus	agosto	agosto	agosto
Swim	natare	nuotare	nadar	nadar
⋮				

- Phonological rules more regular than morphological or syntactic ones

- Basis of the **comparative method:**



- `la` : Classical Latin
- `vl` : ``Vulgar Latin"
- `ib` : ``Proto-ibero Romance"

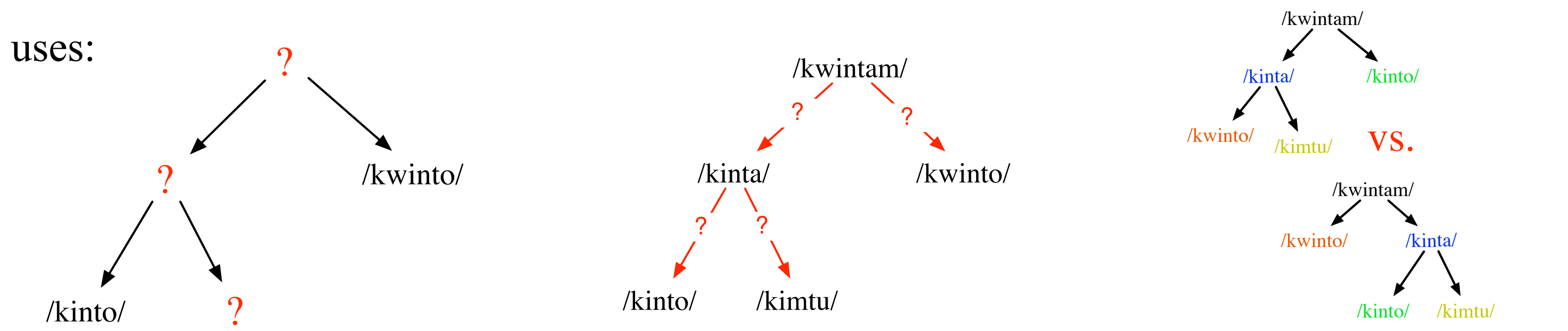


- In practice, the ancient words and/or the evolutionary tree are unknown

- Methodology: manually inspecting the data

- **Our work: A probabilistic model that captures phonological aspects of language change**

- Many uses:



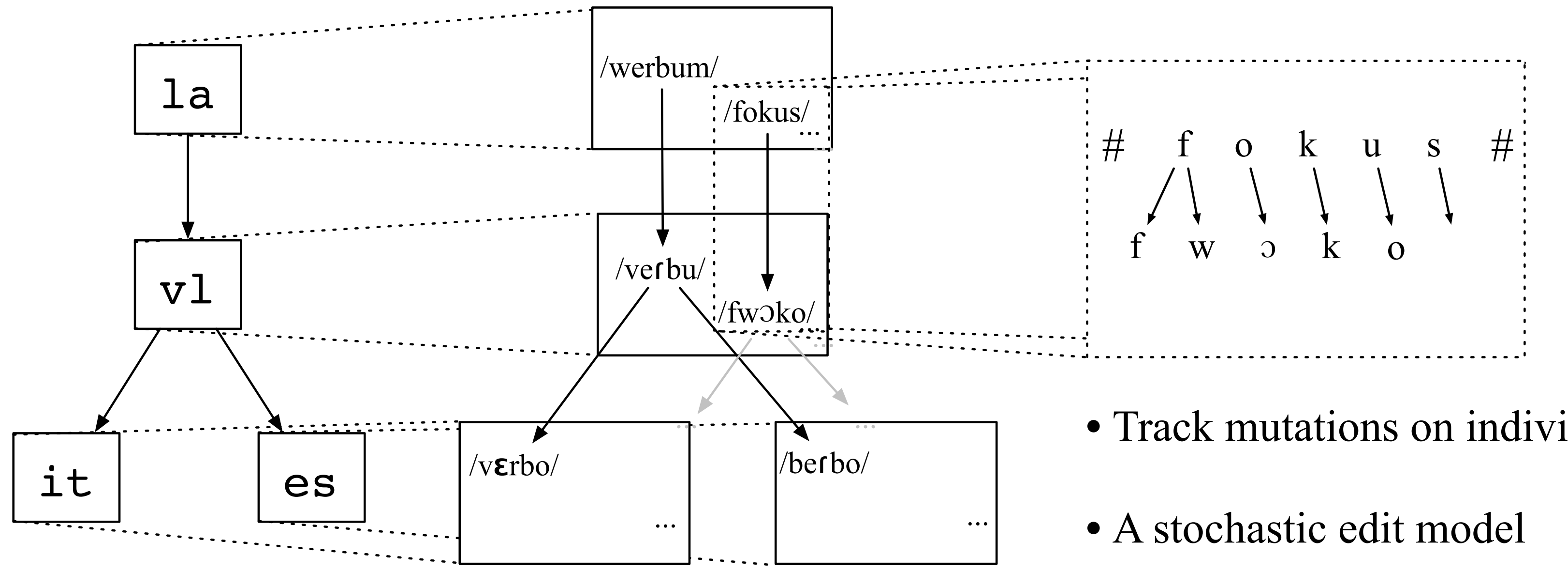
Reconstruction of word forms (ancient and modern)

Inference of phonological rules

Selection of phylogenies

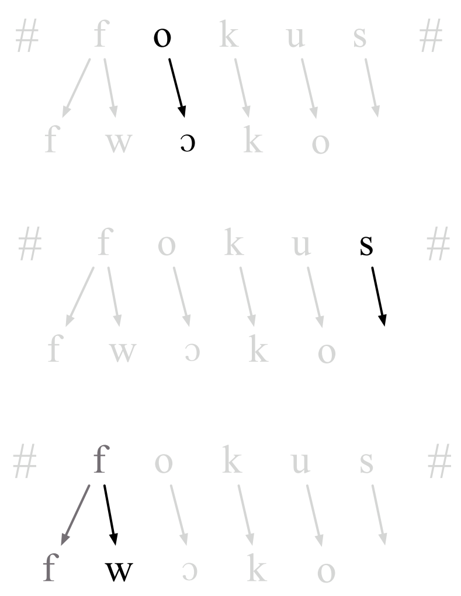
- An inference procedure and experiments on all three applications
- A loglinear parametrization of the edit model

- **Model:** assume for now that the tree topology is known

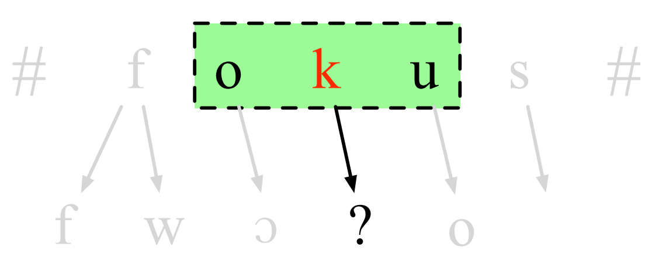


- Track mutations on individual words
- A stochastic edit model

- Types of operations:

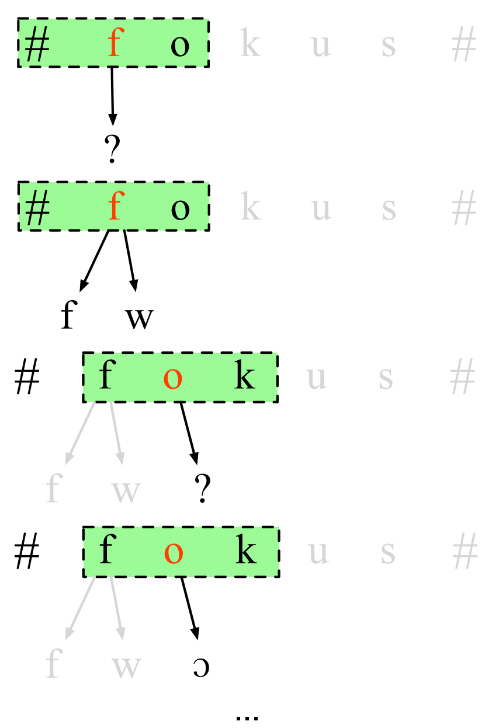


- Context:

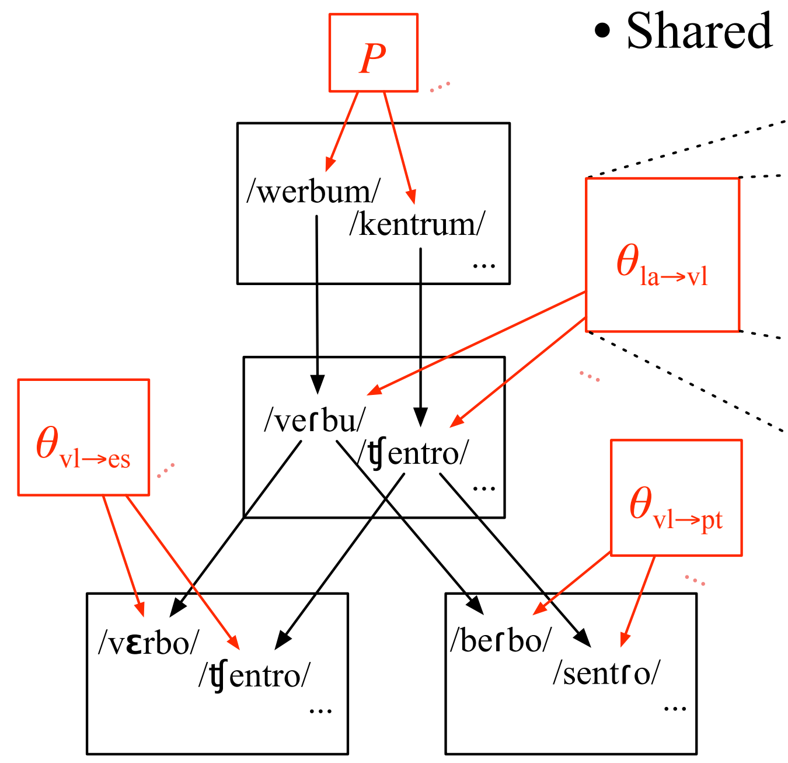


Distribution over operation conditioned on features of the adjacent phonemes (locally normalized)

- Example:



- Edit parameters: one set of parameters $\theta_{A\to B}$ for each edge A→B in the tree



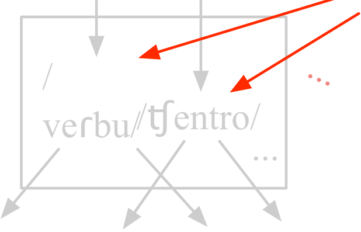
- Shared across all word forms evolving along this edge

context	operation	ℙ(operation|context)
u m #	deletion	0.1
u m #	substitution to /m/	0.8
u m #	substitution to /b/	0.1
a c b	deletion	0.8
a c b	insertion of c	0.1
⋮	⋮	⋮

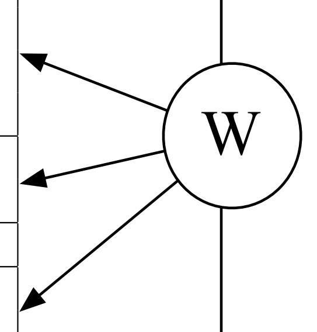
⋮

- Sparsity problems
- No single grouping of contexts is satisfactory

- Prior:



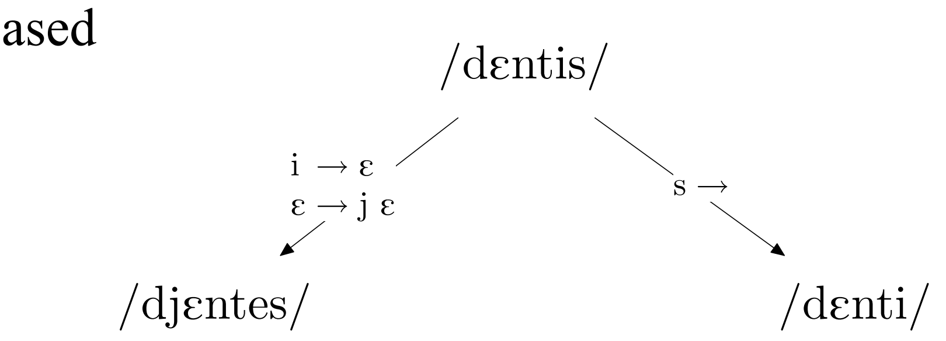
context	operation	IS-INSERT	IS-SUB	m → / _ #	IS-SELF-SUB	v → / intervocalic	...	ℙ(...)
u m #	deletion	0*-1.5	0*-.5	1*1.5	0*1.2	0*1.3	...	0.8
u m #	substitution to /m/	0*-1.5	1*-.5	0*1.5	1*1.2	0*1.3	...	0.1
u m #	substitution to /b/	0*-1.5	1*-.5	0*1.5	0*1.2	0*1.3	...	0.02
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮
i m #	deletion	0*-1.5	1*-.5	1*1.5	0*1.2	0*1.3	...	0.02
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮
a v i	substitution to /b/	0*-1.5	1*-.5	0*1.5	0*1.2	1*1.3	...	0.9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮

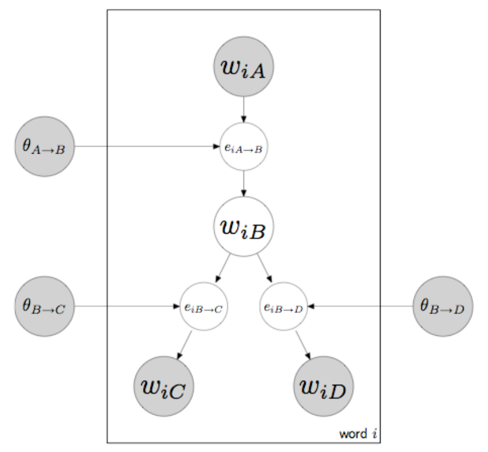


- A log-linear model
- Standard L_2 regularization
- Features:
  - Type of operation
  - Various context granularities

- **Inference**: stochastic EM **(e**xact E step is intractable)

- We an approximate E step based on Gibbs sampling



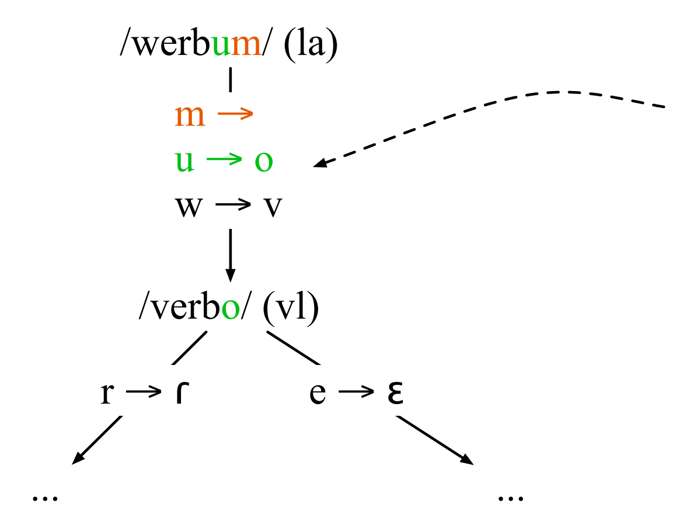


- **Experiments**

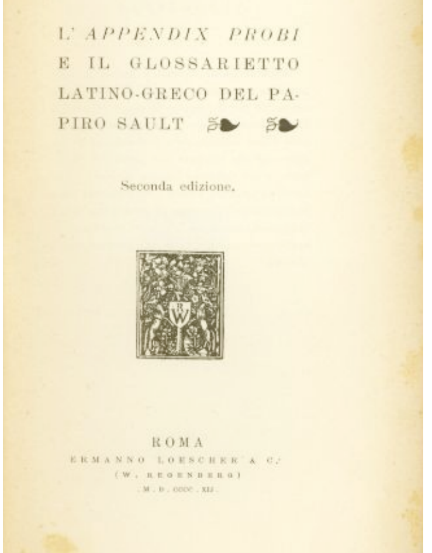
Task 1: reconstruction of Latin given all of the Spanish and Italian words, and some of the Latin words

Model	Baseline	Model	Improvement
Dirichlet	3.59	3.33	7%
Log-linear (0)	3.59	3.21	11%
Log-linear (0,1)	3.59	3.14	12%
Log-linear (0,1,2)	3.59	3.10	14%

Task 2: inference of phonological rules



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


L' APPENDIX PROBI
E IL GLOSSARIETTO
LATINO-GRECO DEL PA-
PIRO SAULT

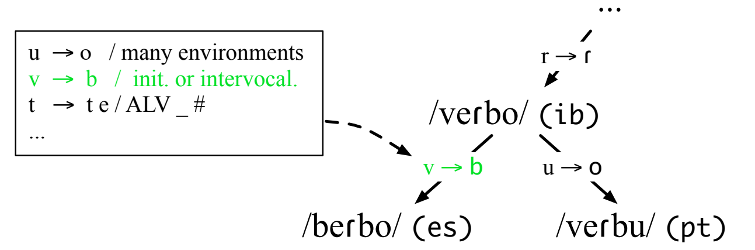
Seconda edizione.

ROMA
ERMANNO LOESCHER & C.
(W. REGENBERG)
M. D. CCCC. XII.

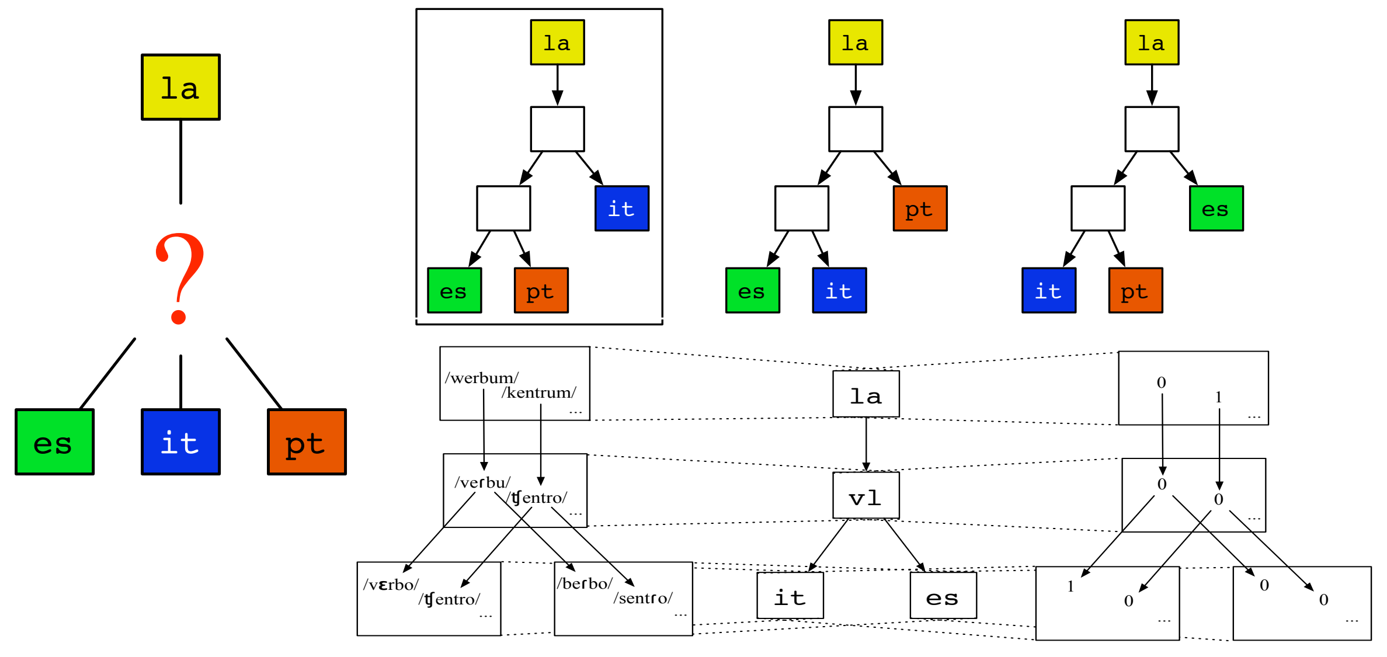
- Comparison with *Appendix Probi:*

coluber	non colober
passim	non passi

- /v/ to /b/fortition
- /s/ to /z/voicing in Italian



Task 3: Selection of phylogenies



Conclusion and future work:

- A probabilistic approach to diachronic phonology
- Log-linear prior yields better reconstructions; interesting connection with stochastic optimality theory
- Enables reconstruction of ancient and modern word forms, phonological rules and tree topologies