A Probabilistic Approach to Diachronic Phonology

Alexandre Bouchard-Côté   Percy Liang

Tom Griffiths   Dan Klein

Languages evolve

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

⋮

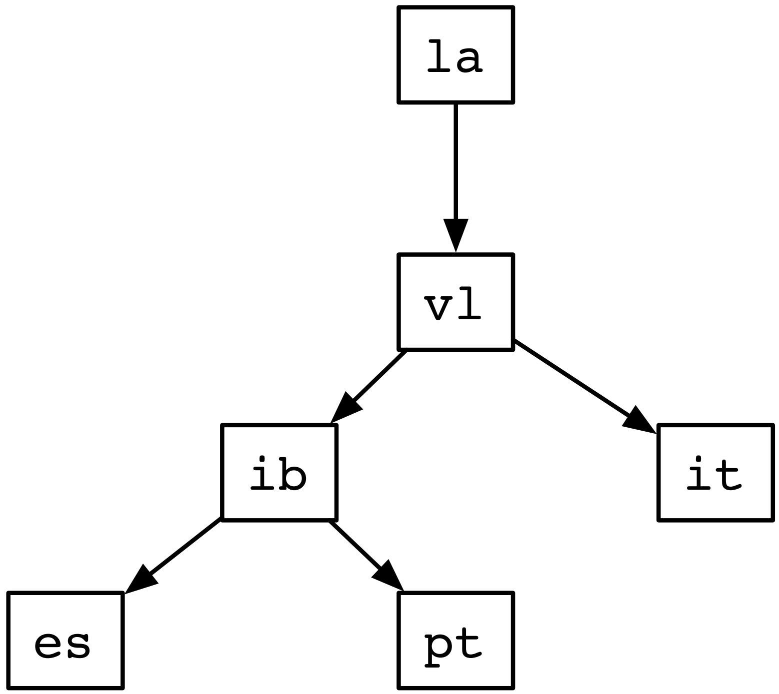
Language evolution

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

⋮

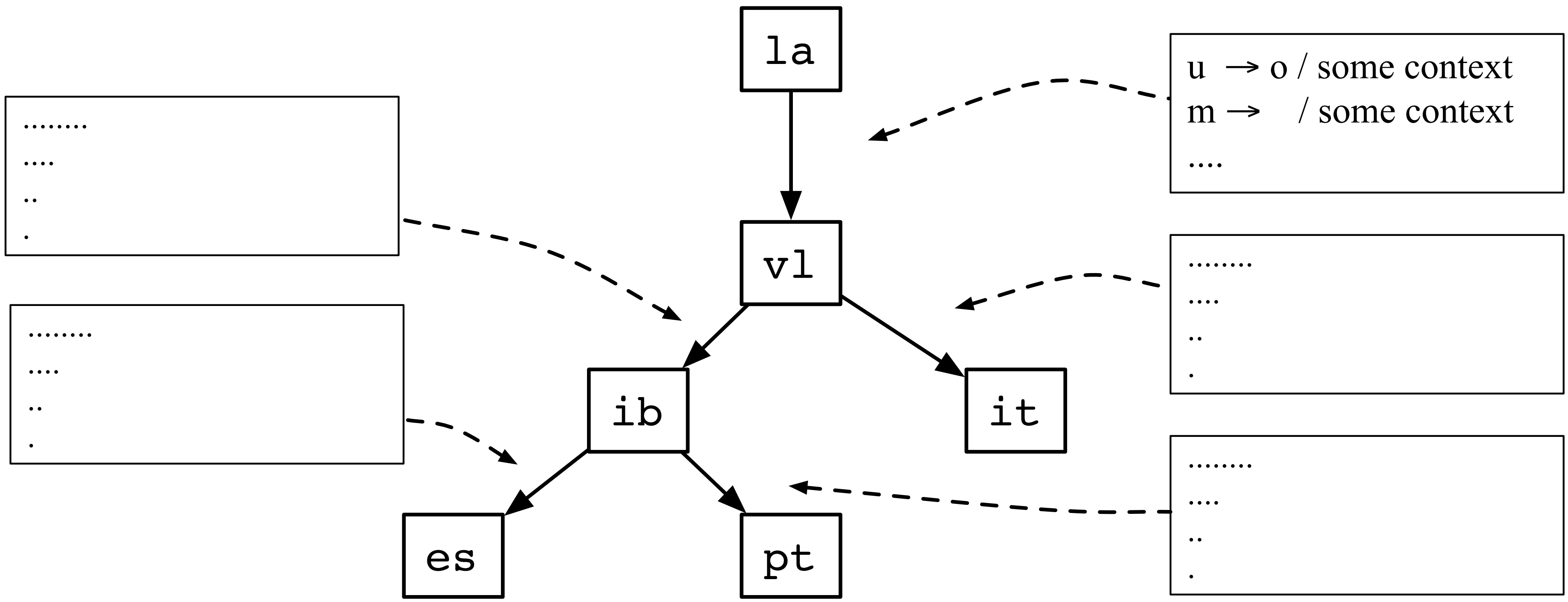
- Phonological rules more regular than morphological or syntactic ones
- basis of the *comparative method*

Example of a mutation process as seen by the comparative method



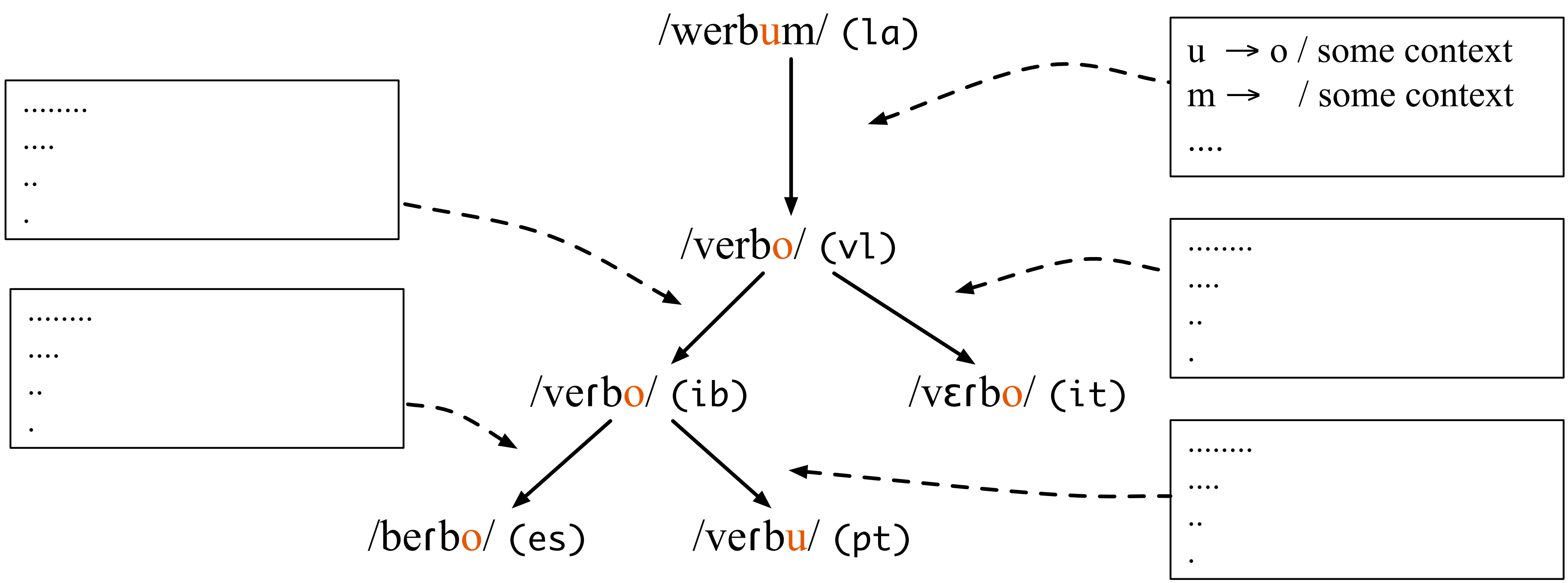
- `ib` : Proto-ibero Romance
- `vl` : Vulgar Latin

Example of a mutation process as seen by the comparative method



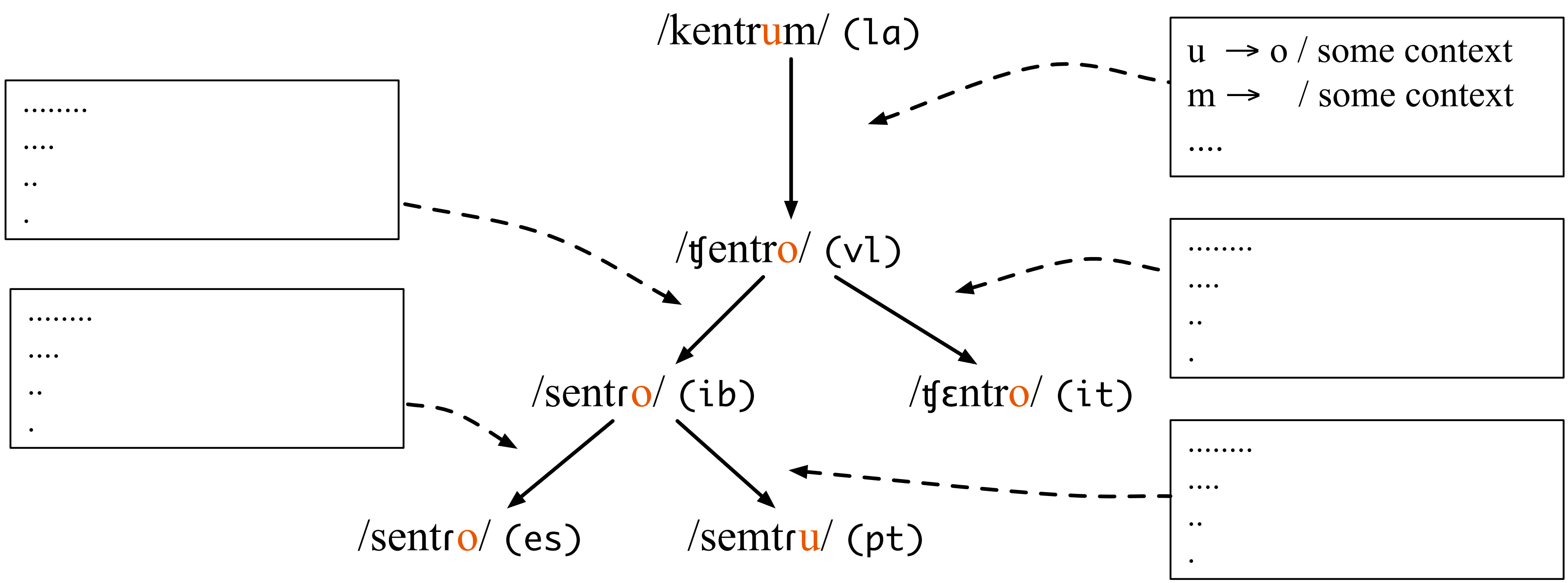
- Deterministic re-write rules at each branch
- Activated by some context

Example of a mutation process as seen by the comparative method



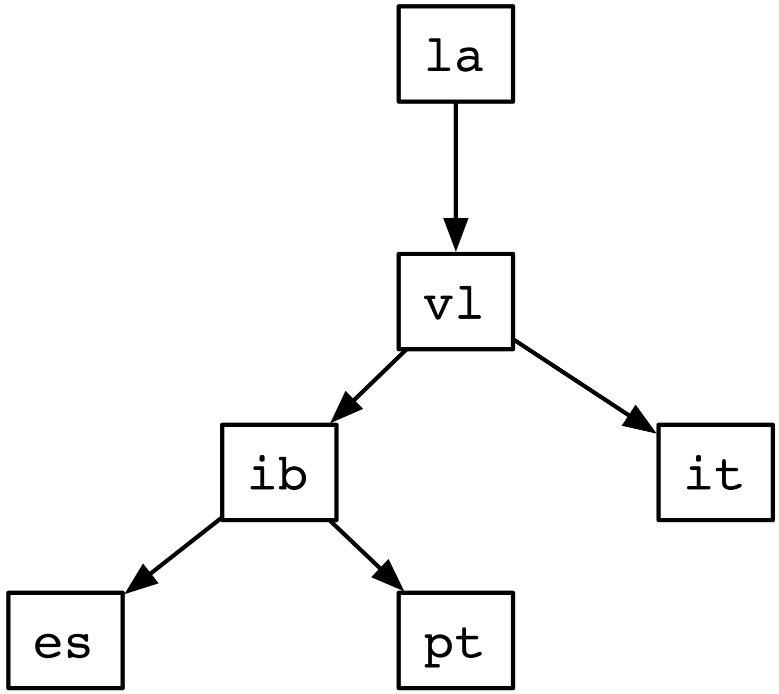
Gloss	Latin	Italian	Spanish	Portuguese
Word/verb	verbum	verbo	verbo	verbu

Example of a mutation process as seen by the comparative method



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

Example of a mutation process as seen by the comparative method

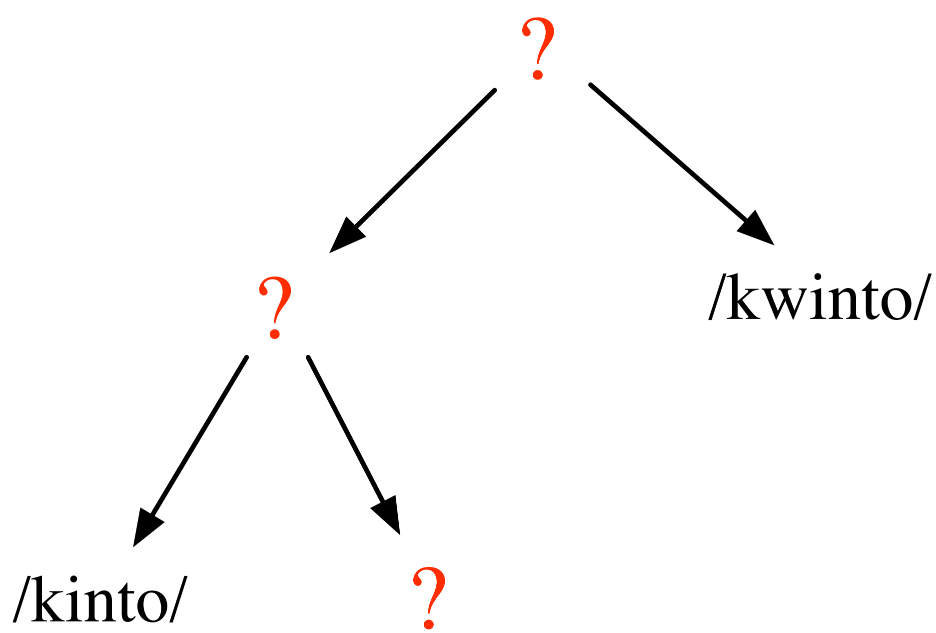


- In practice, the ancient words and/or the evolutionary tree are unknown
- Methodology: manually inspecting the data

Our work:

- A probabilistic model that captures phonological aspects of language change.

- Many usages:

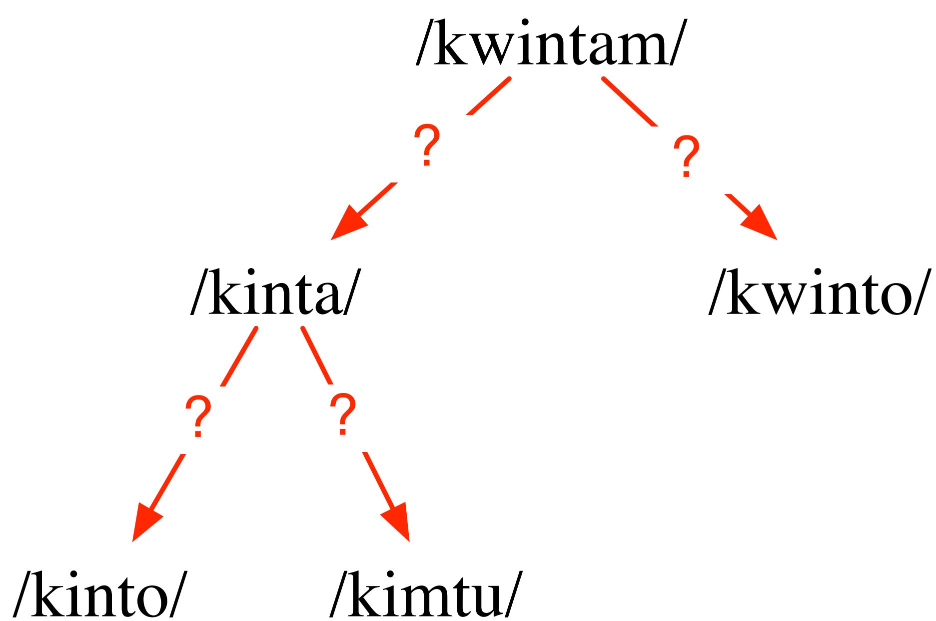


Reconstruction of word forms (ancient and modern)

Our work:

- A probabilistic model that captures phonological aspects of language change.

- Many usages:

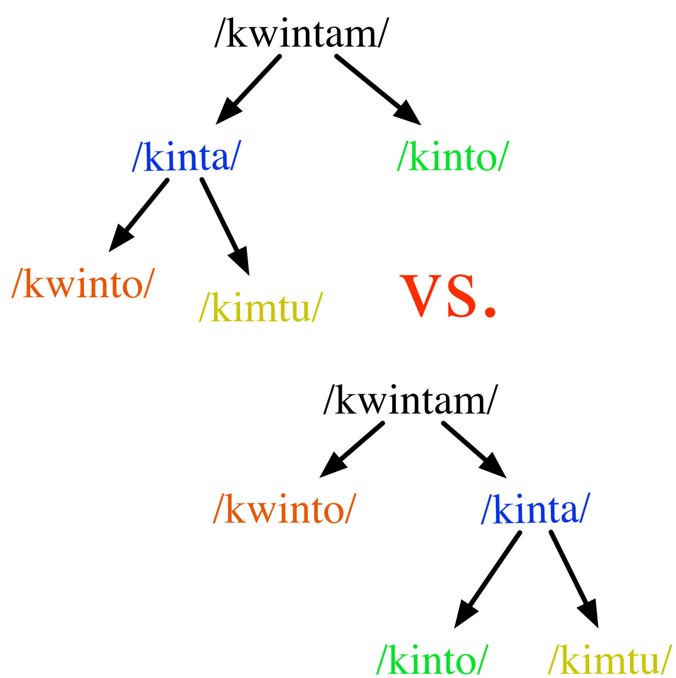


Inference of phonological rules

Our work:

- A probabilistic model that captures phonological aspects of language change.

- Many usages:



Selection of phylogenies

Our work:

- A probabilistic model that captures phonological aspects of language change.

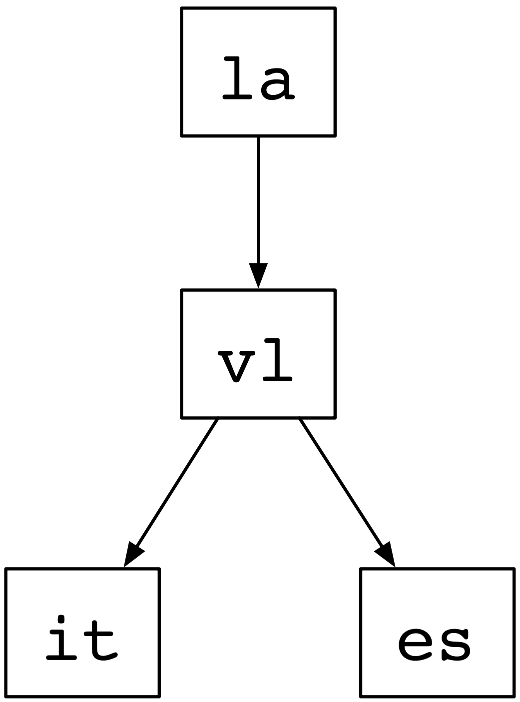
- Many usages:
  - Reconstruction of word forms (ancient and modern)
  - Inference of phonological rules
  - Selection of phylogenies

- An inference procedure and experiments on all three applications

- A new task and evaluation framework

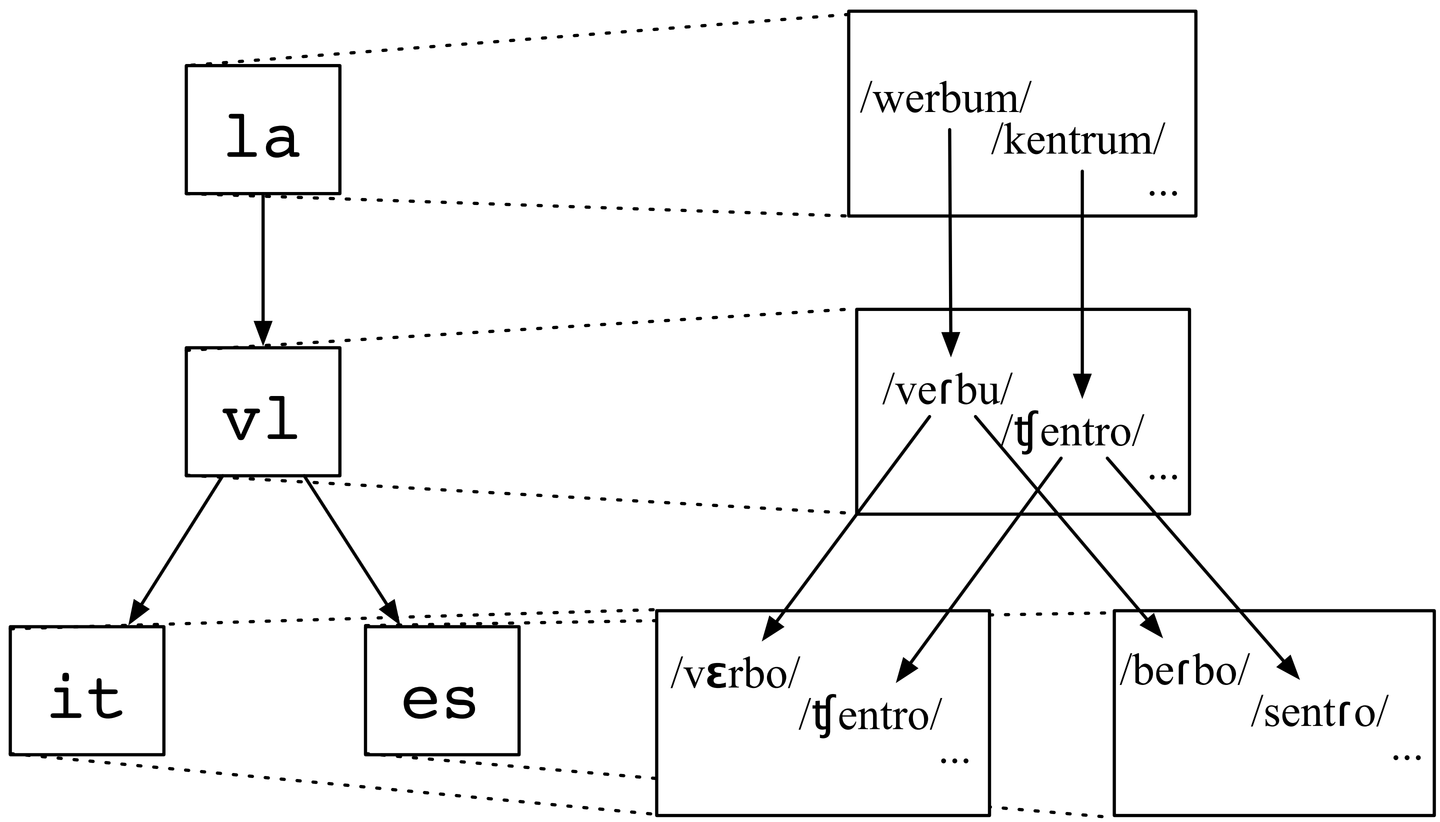
The model

Big picture



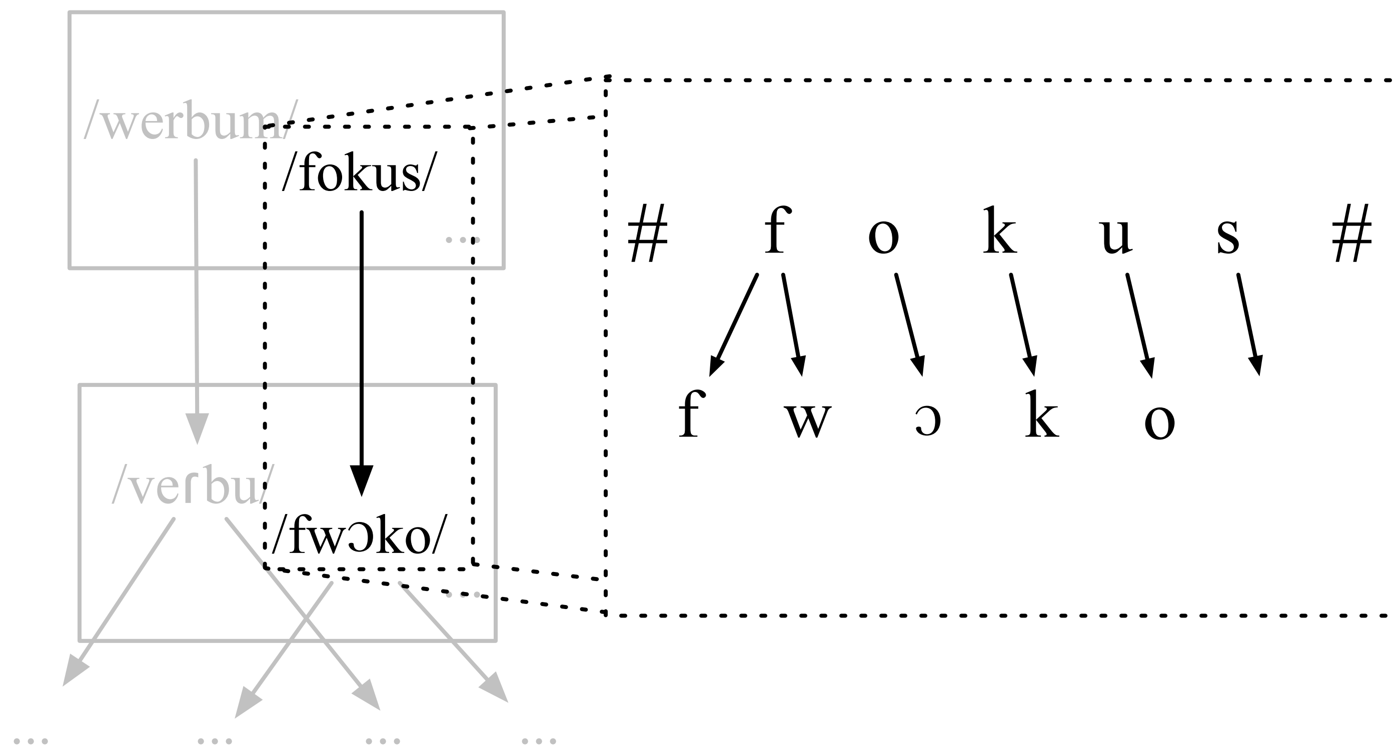
- Assume for now that the tree topology is known

Big picture



- Assume for now that the tree topology is known
- Track individual words

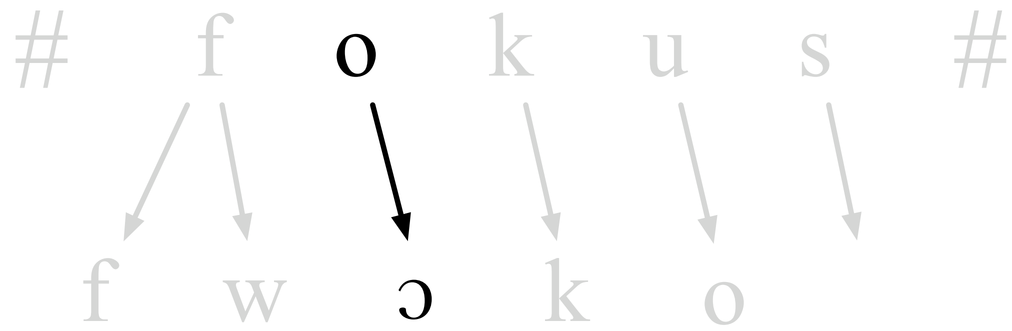
Stochastic edit model



- Let's look at how a single words evolve along one of the edges of the tree

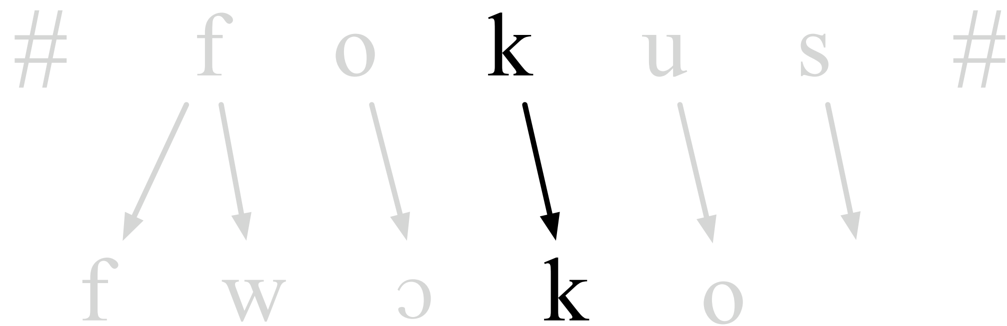
- Mutation of Latin *FOCUS* (/fokus/) into Italian *fuoco* (/fwɔko/) (fire)

Stochastic edit model: operations



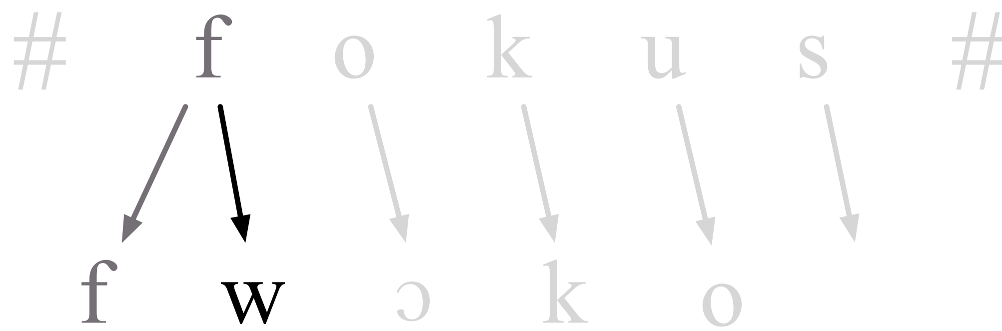
- Substitution

Stochastic edit model: operations



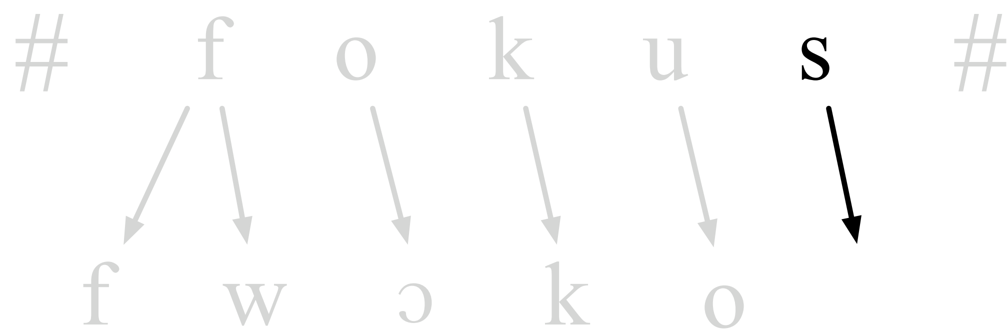
- Substitution (incl. self-substitution)

Stochastic edit model: operations



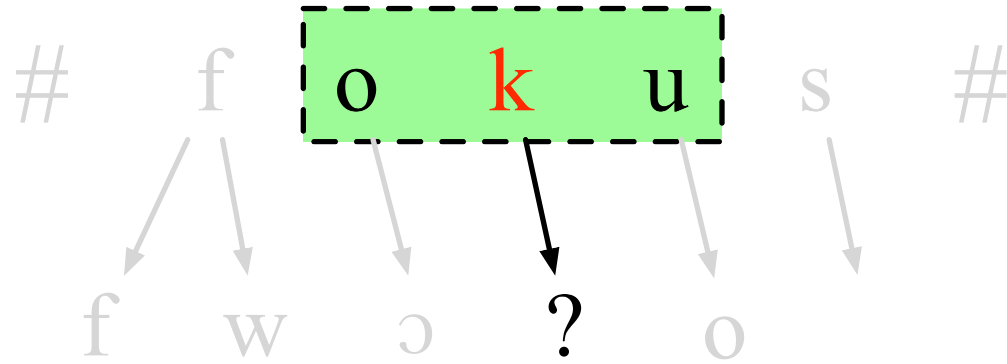
- Substitution (incl. self-substitution)
- Insertion

Stochastic edit model: operations



- Substitution (incl. self-substitution)
- Insertion
- Deletion

Stochastic edit model: context

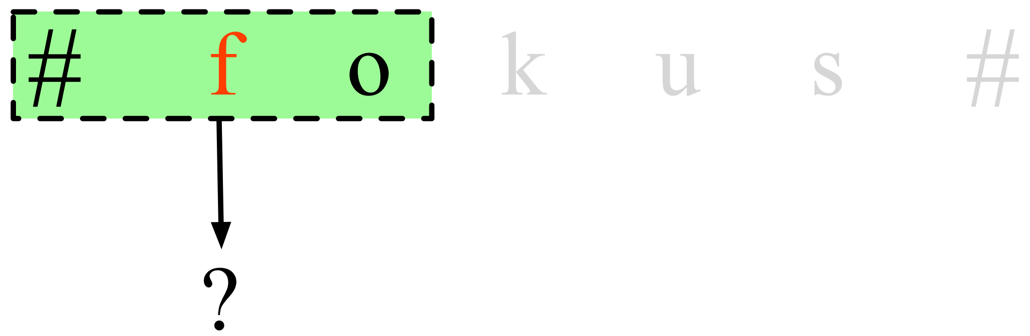


- Distribution over operations conditioned on adjacent phonemes

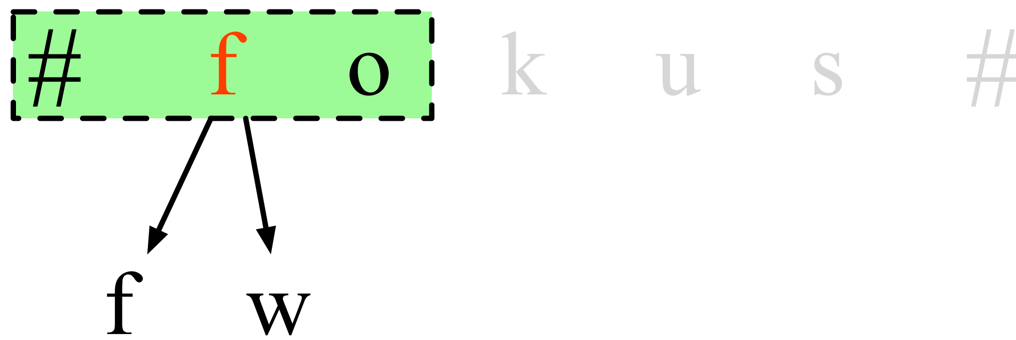
Stochastic edit model: generation process

f o k u s

Stochastic edit model: generation process

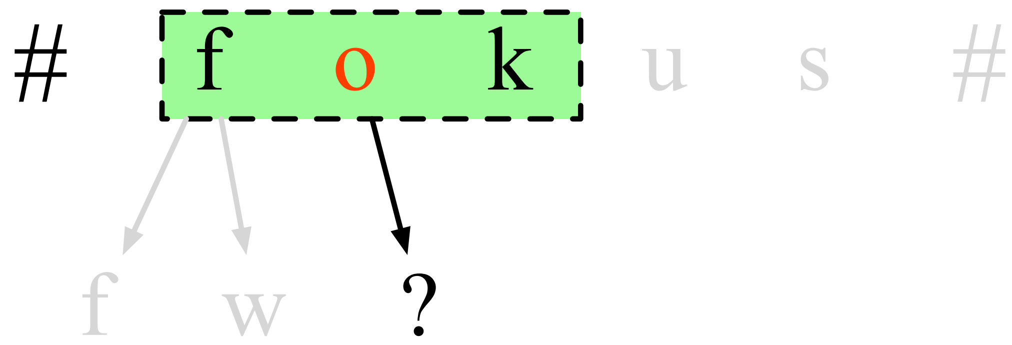


Stochastic edit model: generation process



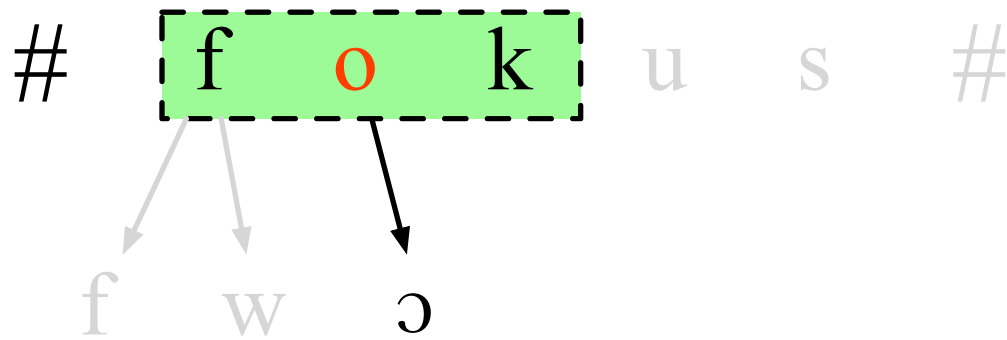
- $\mathbb{P}(\mathsf{f} \longrightarrow \mathsf{f}\ \mathsf{w}\ /\ \#\ \_\ \mathsf{V}) = 0.05$

Stochastic edit model: generation process



- $\mathbb{P}(\mathsf{f} \rightarrow \mathsf{f\ w}\ /\ \#\ \_\ \mathsf{V}) = 0.05$

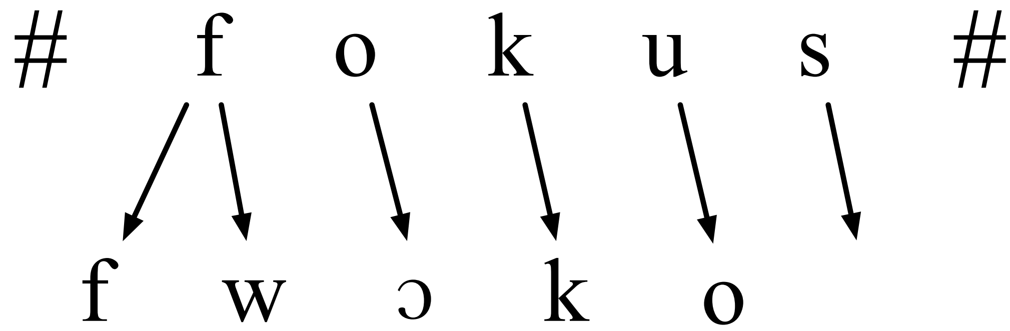
Stochastic edit model: generation process



- $\mathbb{P}(\mathsf{f} \longrightarrow \mathsf{f\ w}\ /\ \#\ \_\ \mathsf{V}) = 0.05$

- $\mathbb{P}(\mathsf{o} \longrightarrow \text{ɔ}\ /\ \mathsf{C}\ \_\ \mathsf{V}) = 0.1$

Stochastic edit model: generation process



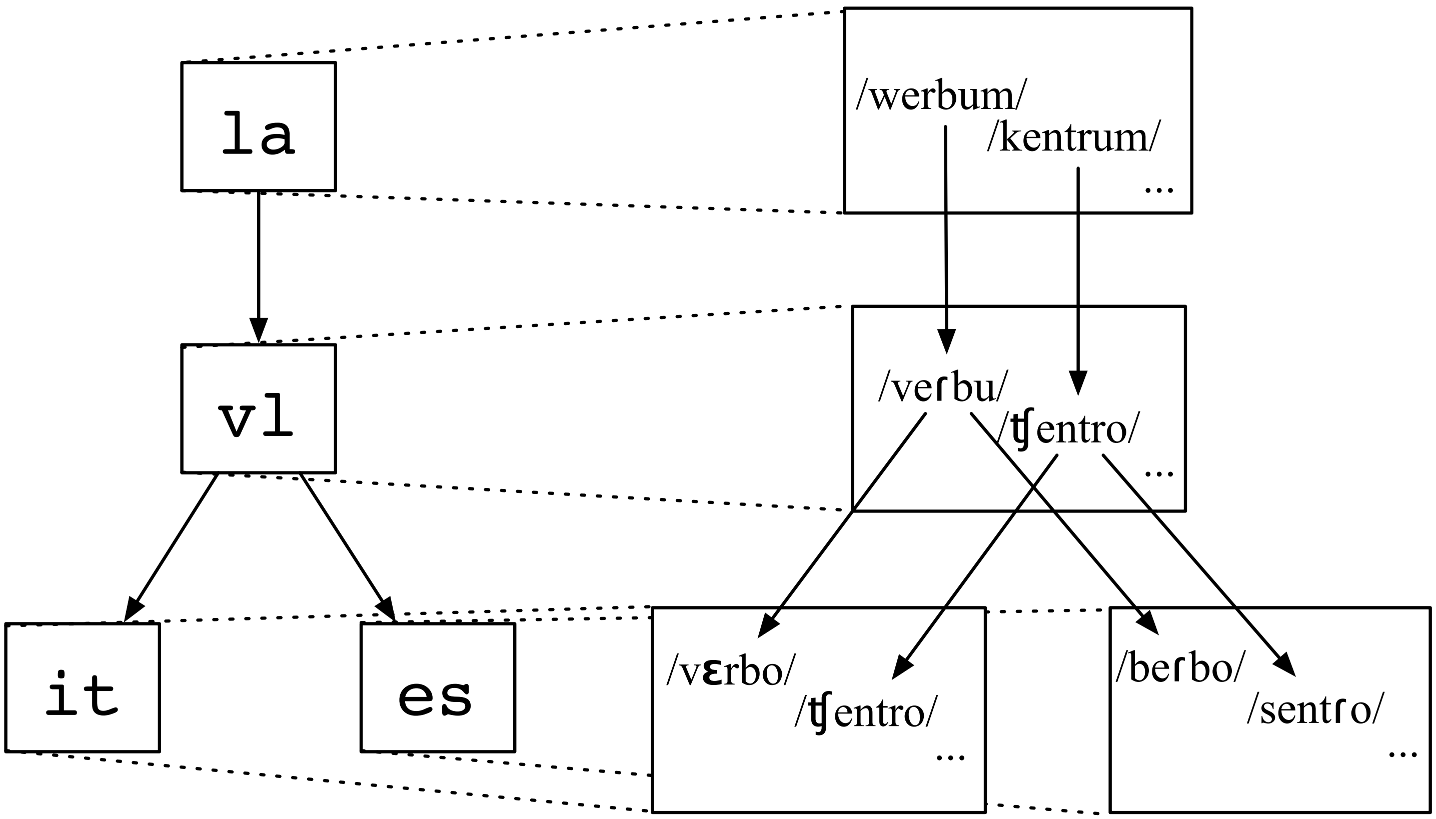
- $\mathbb{P}(\mathsf{f} \rightarrow \mathsf{f\ w}\ /\ \#\ \_\ \mathsf{V}) = 0.05$

- $\mathbb{P}(\mathsf{o} \rightarrow \text{ɔ}\ /\ \mathsf{C}\ \_\ \mathsf{V}) = 0.1$

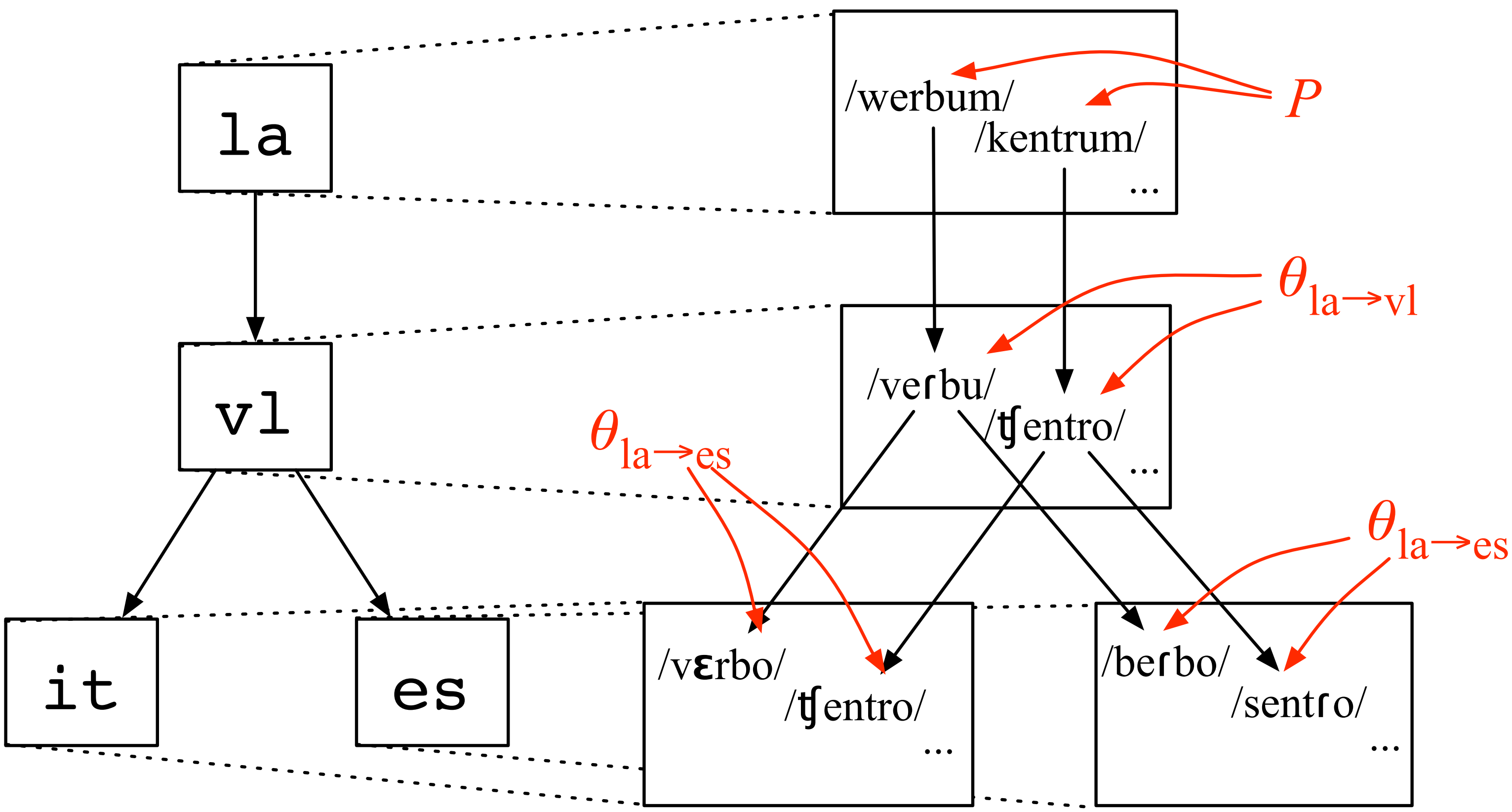
- $\ldots$

- $\mathbb{P}(/\text{fokus}/ \rightarrow /\text{fwɔko}/)) = 0.05 \times 0.1 \times \cdots$

Edit parameters

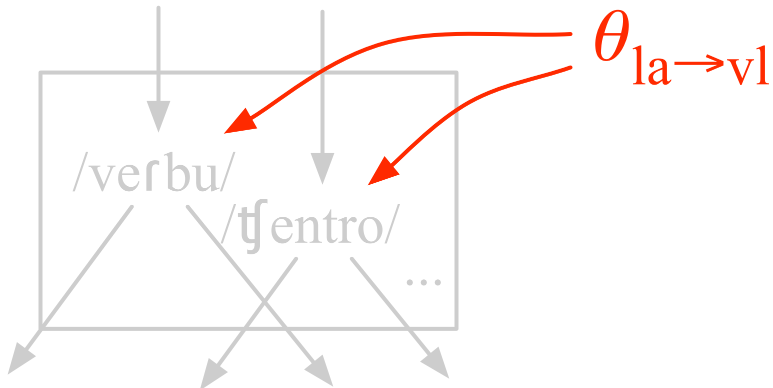


Edit parameters



- One set of parameter $\theta_{A \to B}$ for each edge $A \to B$ in the tree
- Shared across all word forms evolving along this edge

Edit parameters



- $\theta_{A\to B}$ specifies $\mathbb{P}(\mathrm{operation}|\mathrm{context})$

context	operation	$\mathbb{P}(\mathrm{operation}\|\mathrm{context})$
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
⋮	⋮	⋮

⋮

Distribution on the edit parameters

- Too many parameters
- Addressed by:
  - Sparsity prior: independent Dirichlet priors (one for each context)
  - Group context distributions. Example:

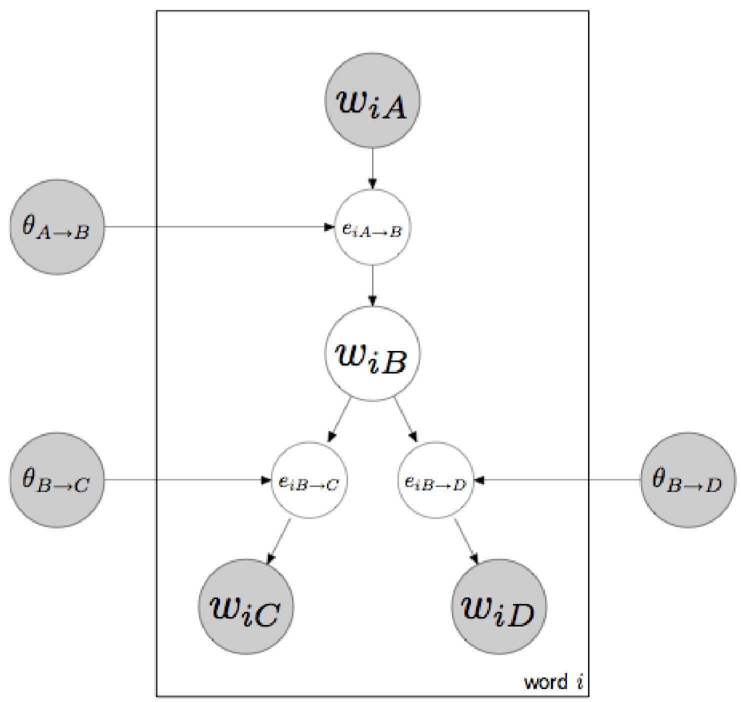
context	operation	$\mathbb{P}(\text{operation}\|\text{context})$
V m #	deletion	0.1
V m #	substitution to /a/	0.8
V m #	substitution to /b/	0.1
V c C	deletion	0.8
V c C	insertion of c	0.1
⋮	⋮	⋮

⋮

Inference and experiments

Inference: EM

- Exact E step is intractable
  - We use a stochastic E step based on Gibbs sampling
- E: fix the edit parameters, resample the derivations
- M: update the edit parameters from expected edit counts



Automatic extraction of a Romance corpus

$$
\begin{tikzcd}
\text{Wiktionary} \arrow[r] & \text{XML dump} \arrow[dr] & & \\
\text{Bible} \arrow[r] & \text{Align.} \arrow[r] & \text{Closure} \arrow[r] & \text{Cognate detector} \\
\text{Europarl} \arrow[r] & \text{Align.} \arrow[ur] & &
\end{tikzcd}
$$

- Noisier than manually curated cognate lists
- More data available
- Our model overcomes this noise

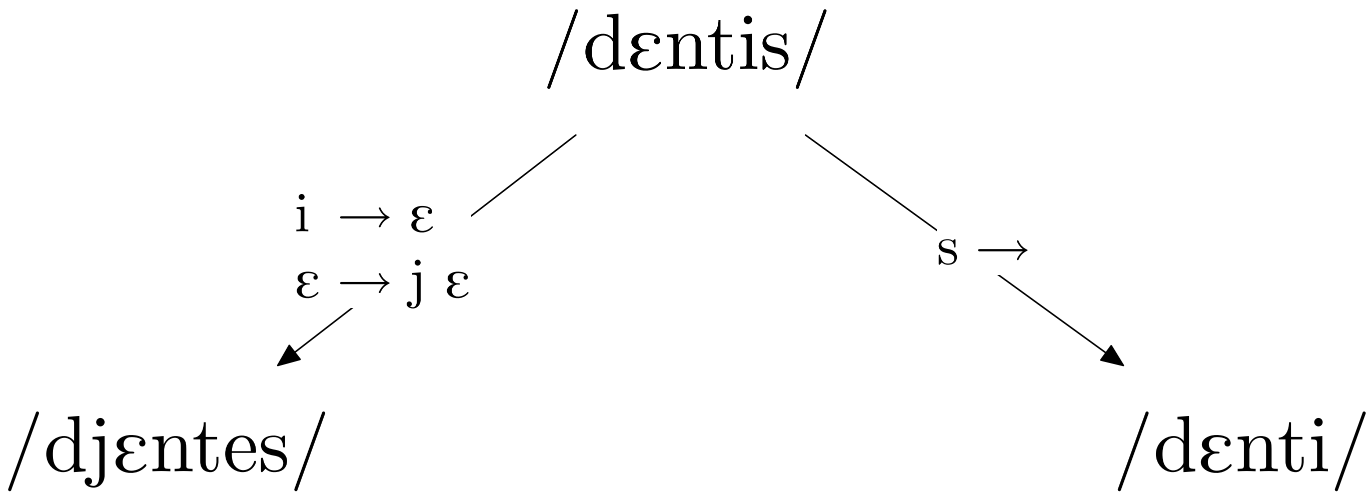
Data available online:
`http://nlp.cs.berkeley.edu/pages/historical.html`

Reconstruction of ancient word forms

- Task: reconstruction of Latin given all of the Spanish and Italian words, and some of the Latin words
- Evaluation: uniform cost edit distance on held-out data
- Baseline: pick one of the modern languages at random

Reconstruction of ancient word forms

- Task: reconstruction of Latin given all of the Spanish and Italian words, and some of the Latin words
- Example: “teeth”, nearly correctly reconstructed

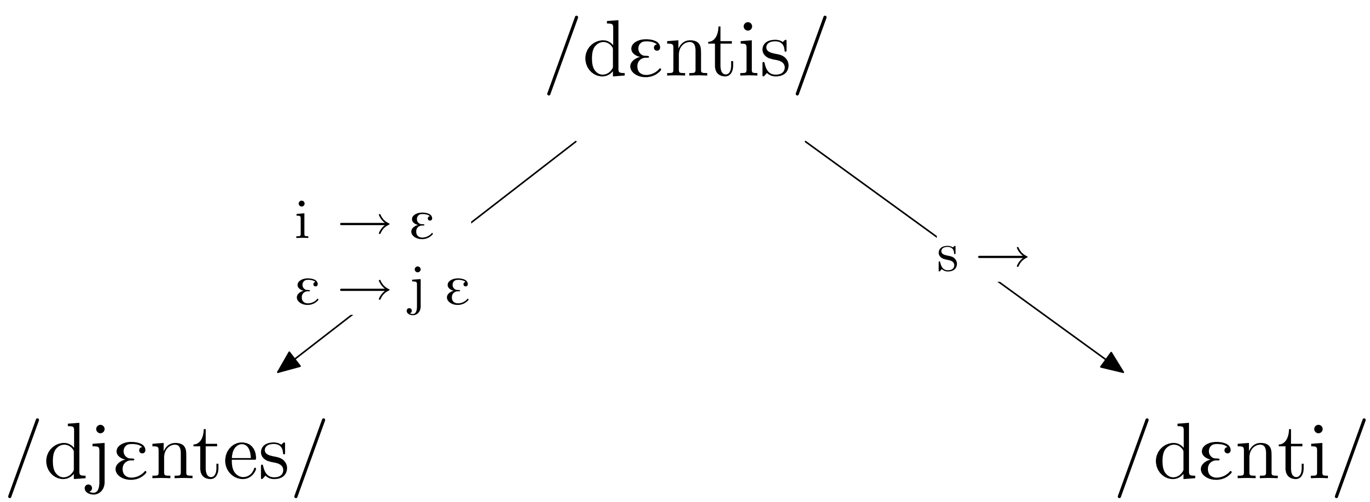


- Numbers:

Language	Baseline	Model	Improvement
Latin	2.84	2.34	9%

Reconstruction of word forms

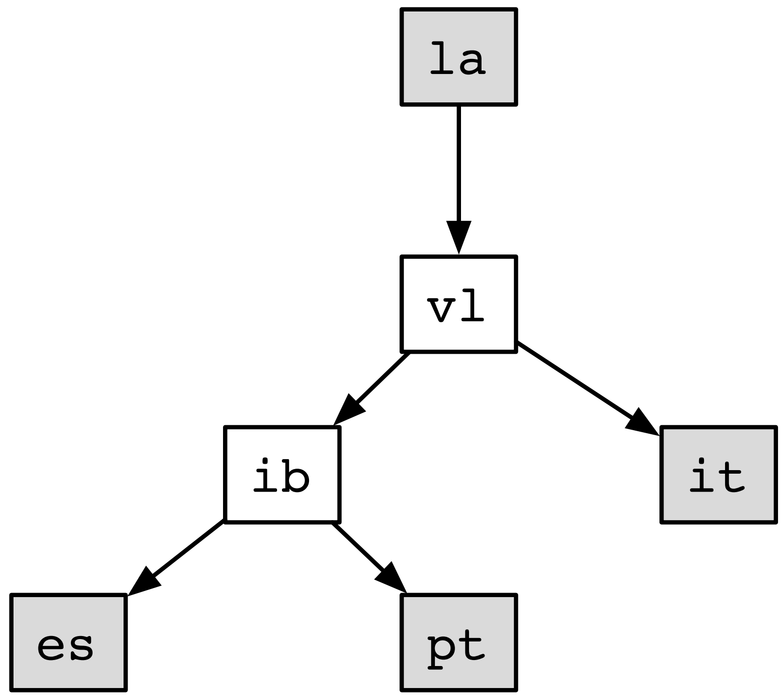
- Evaluation: uniform cost edit distance on held-out data
- Baseline: pick one of the modern languages at random
- Example: “teeth”, nearly correctly reconstructed



- Numbers:

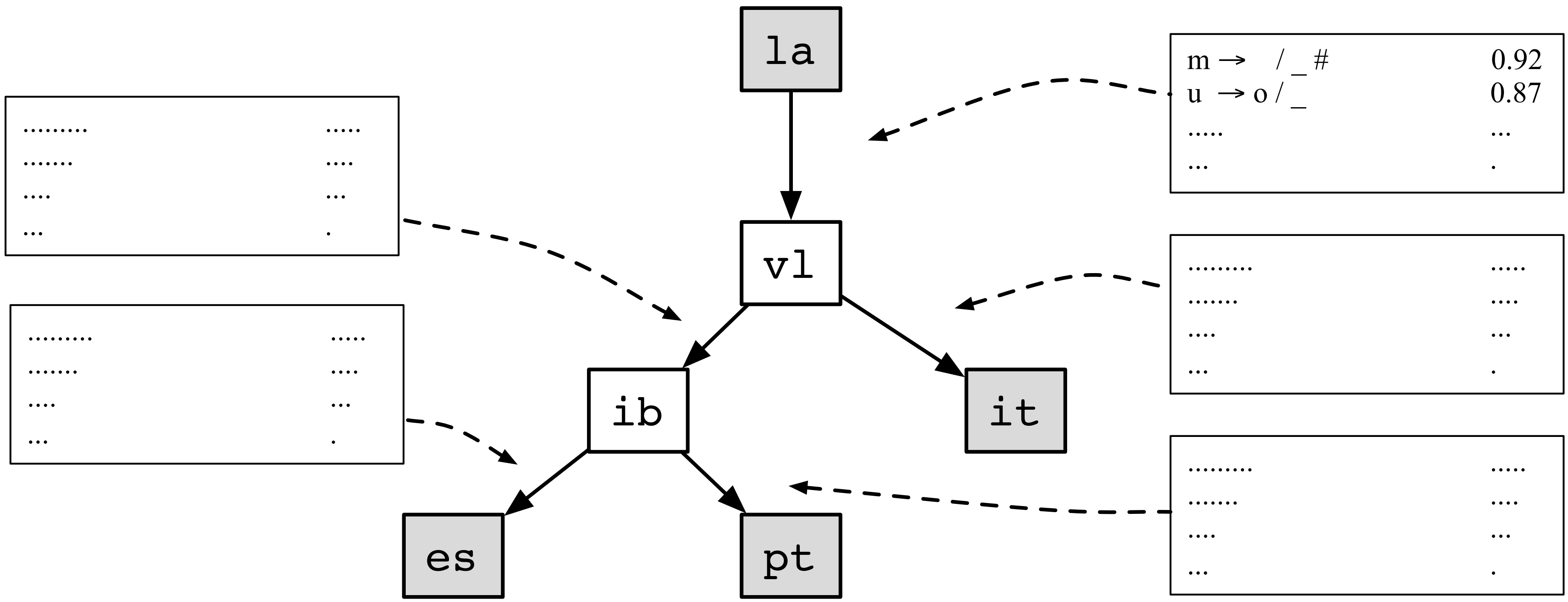
Language	Baseline	Model	Improvement
Latin	2.84	2.34	9%
Spanish	3.59	3.21	11%

Inference of phonological rules



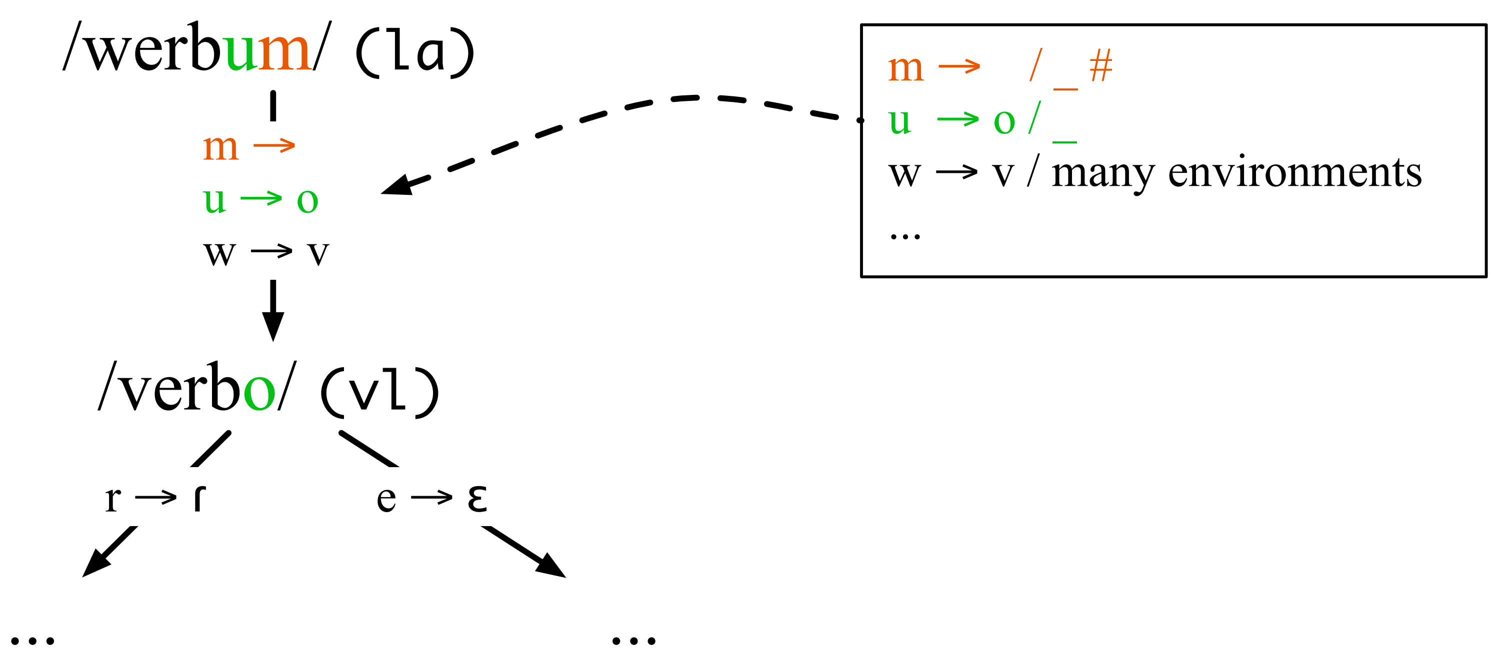
- `ib` : Proto-ibero Romance
- `vl` : Vulgar Latin

Inference of phonological rules



- Reconstruct the internal nodes
- Focus on the rules used most often during the last E step

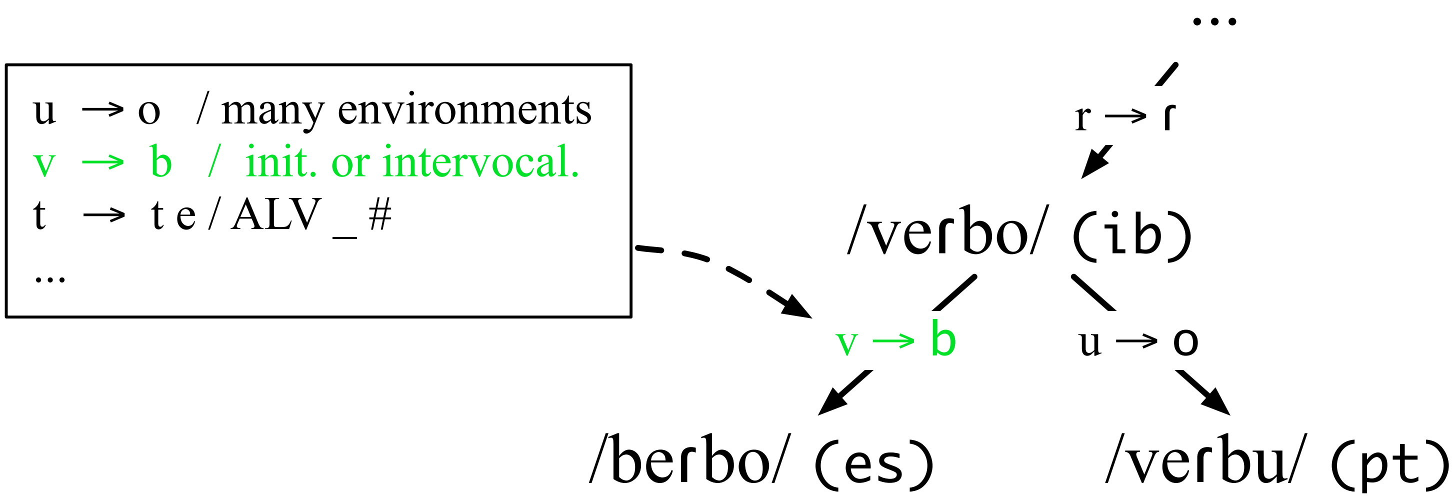
Hypothesized derivation for “word” along with top rules



- Comparison with historical evidence: the *Appendix Probi*

coluber	non colober
passim	non passi

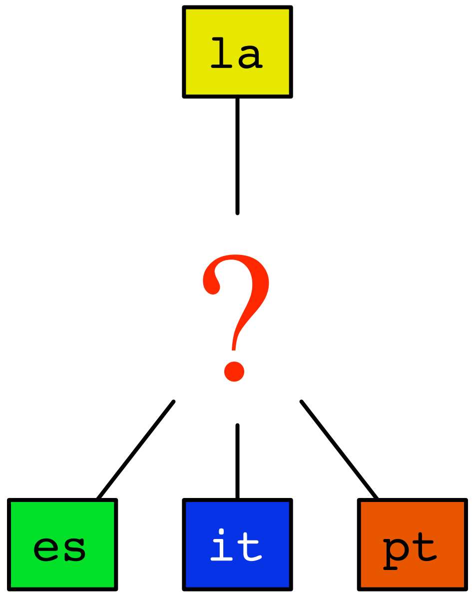
Hypothesized derivation for “word” along with top rules



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

Selection of phylogenies

Inference of topology



Example of previous approaches

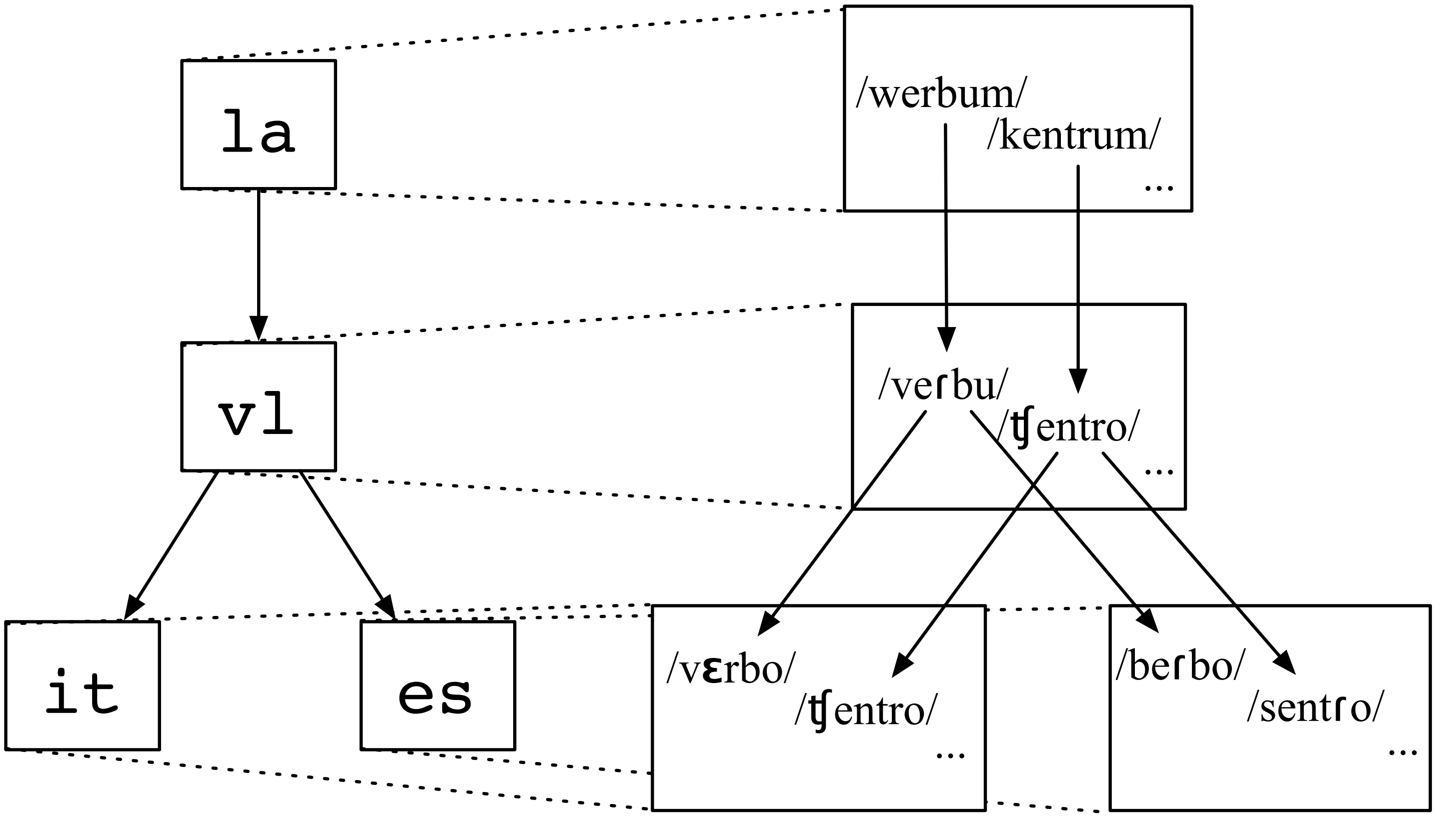
- Gray and Atkinson, 2003
- Coarse encoding:

Latin	mandere (to chew)
French	manger
Italian	mangiare
Latin	comedere (to consume)
Spanish	comer
Portuguese	comer

Meaning	Eat		$\cdots$
Cognate set	1	2	$\cdots$
Latin	1	1	$\cdots$
French	1	0	$\cdots$
Italian	1	0	$\cdots$
Spanish	0	1	$\cdots$
Portuguese	0	1	$\cdots$

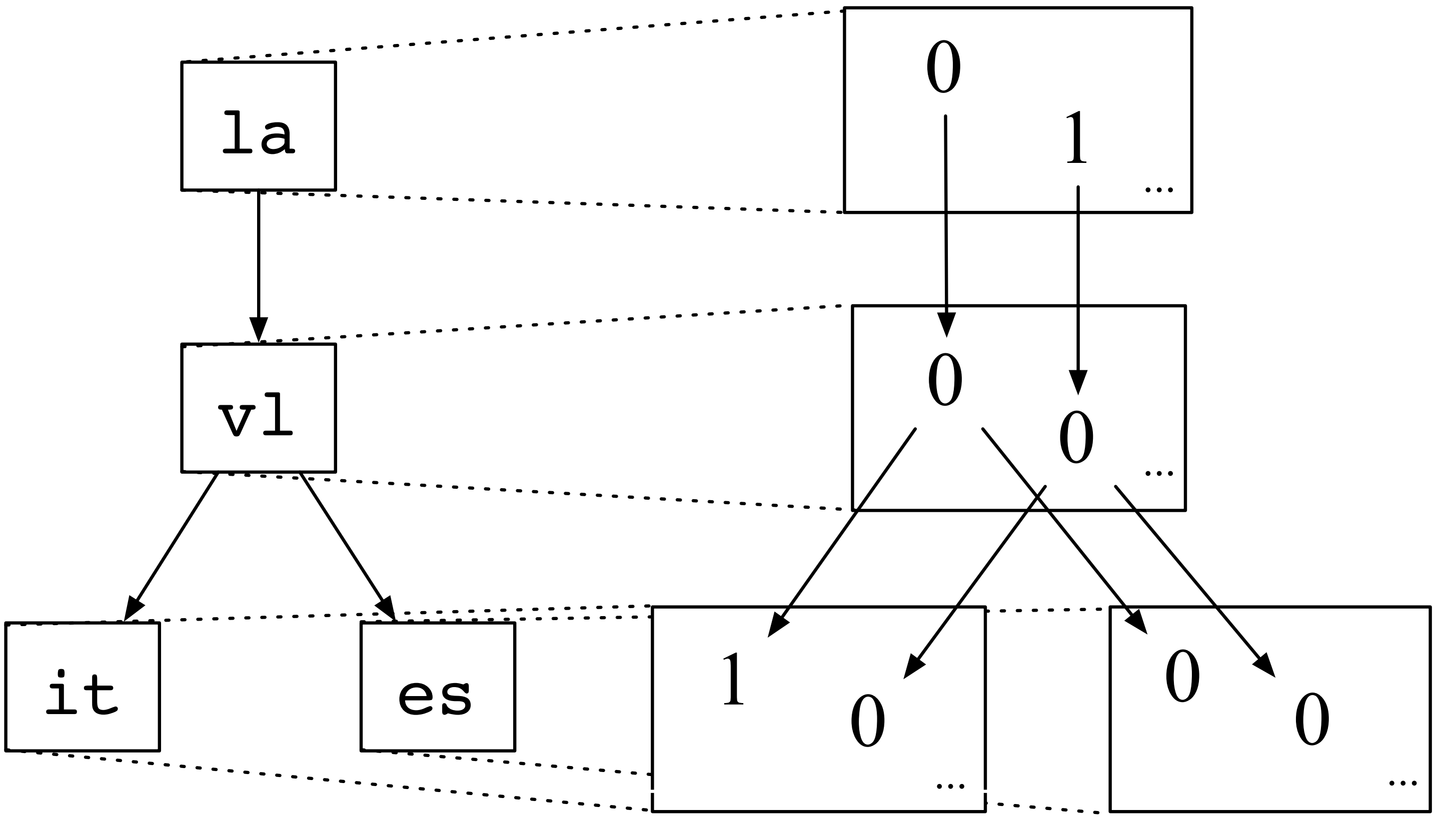
- These characters evolve independently in their model
- Lots of information discarded

Comparison



Our samples look like this

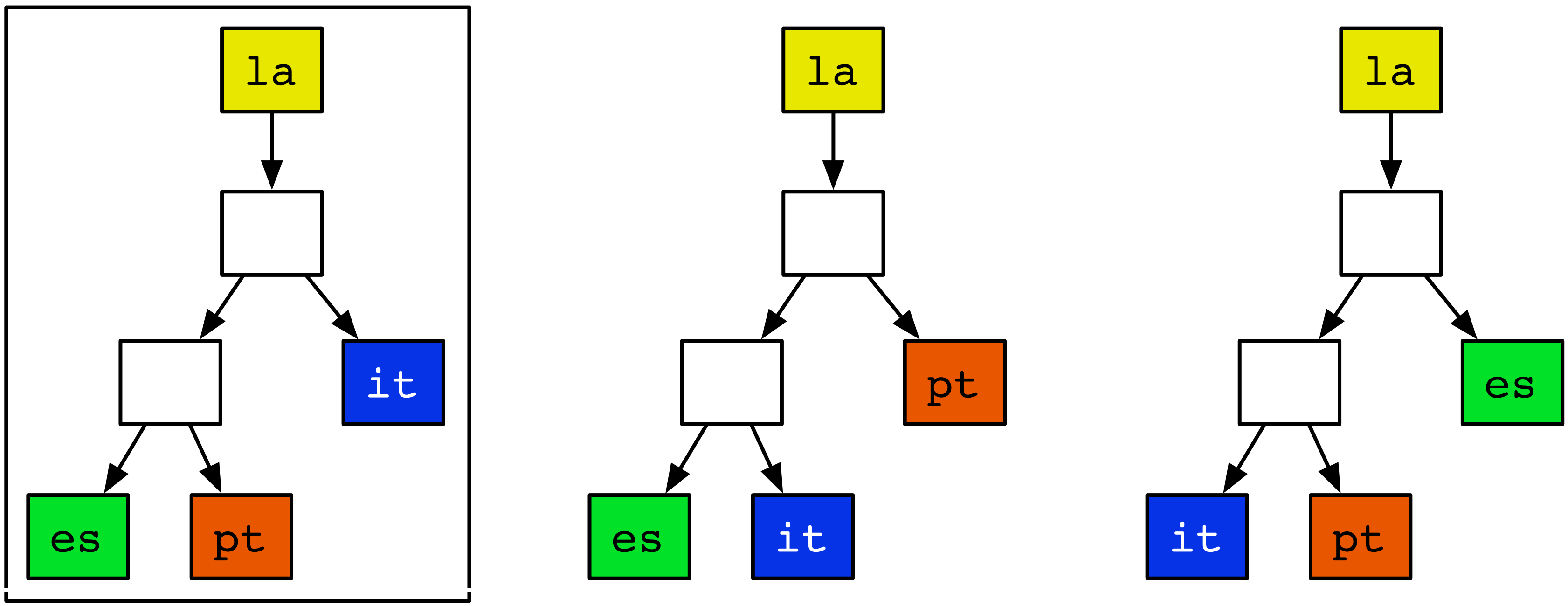
Comparison



Atkinson's

What we did

- Present good vs. bad topologies and compute the likelihood ratio



- this can be turned into a full topology inference algorithm using the quartet method [Erdos et al., 1996]

Conclusion

- Introduced a probabilistic approach to diachronic phonology
- Enables reconstruction of ancient and modern word forms, phonological rules and tree topologies
- Future work:
  - We are scaling it up to larger phylogenies
  - We are working on an extension using a log-linear parametrization of the contexts, reminiscent of stochastic OT
- Data available online: `http://nlp.cs.berkeley.edu/pages/historical.html`