Improved Reconstruction of Protolanguage Word Forms

Alexandre Bouchard-Côté Thomas L. Griffiths Dan Klein

Oceanic languages

Hawaiian

Samoan

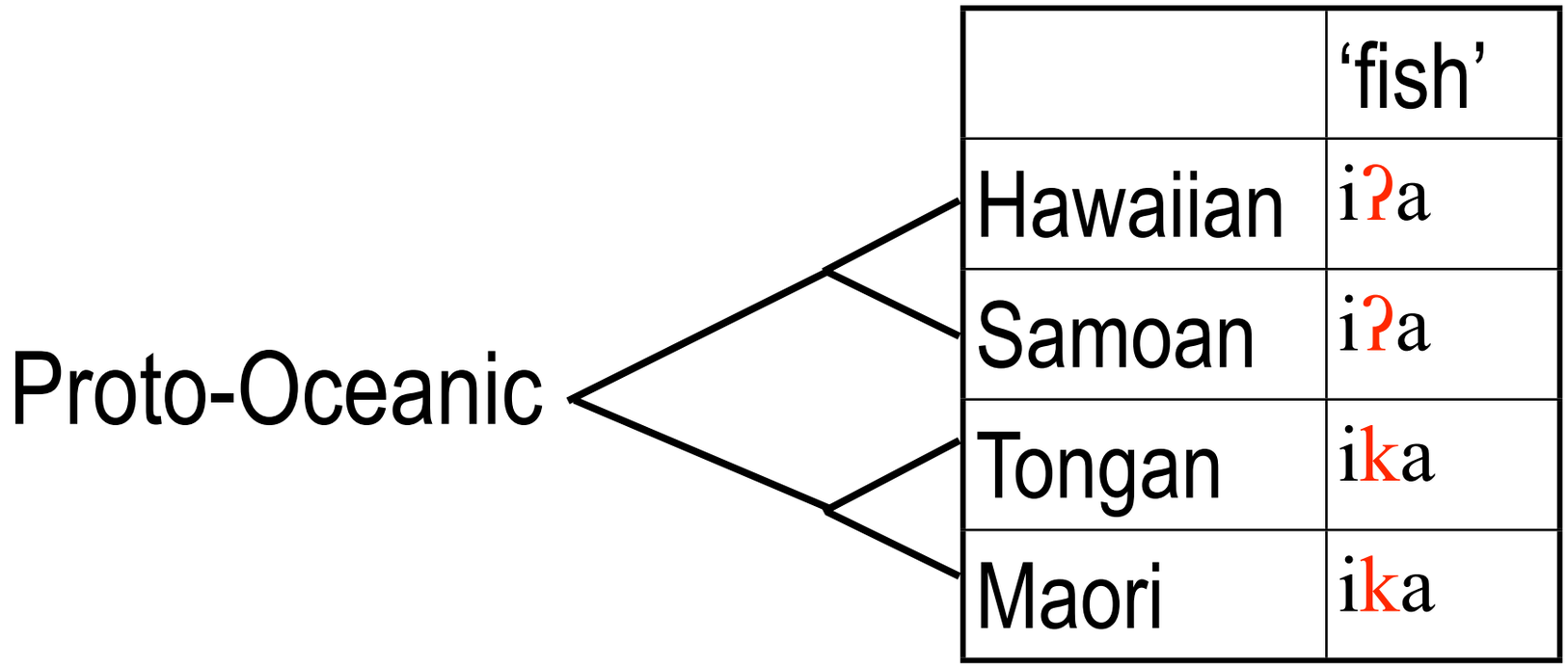
Tongan

Maori

Oceanic languages

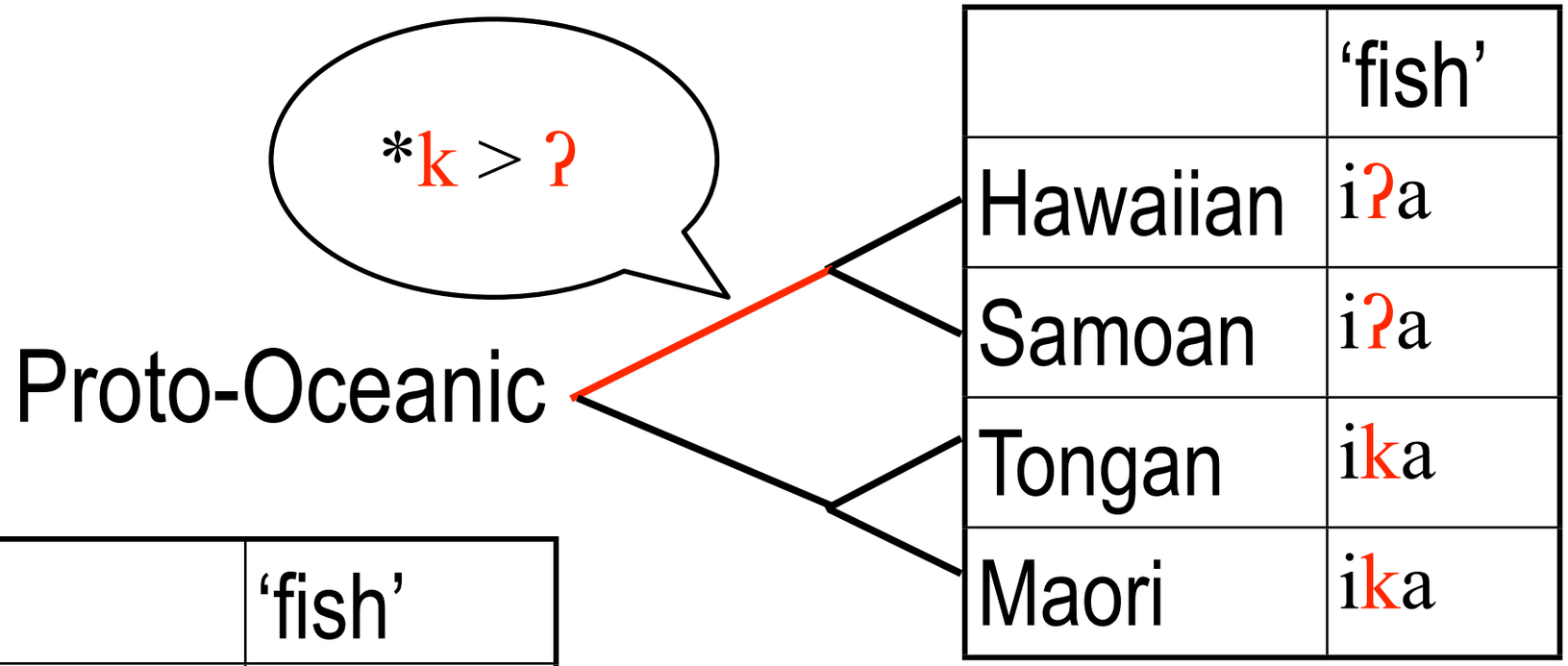
	‘fish’
Hawaiian	iʔa
Samoan	iʔa
Tongan	ika
Maori	ika

Shared ancestry



	‘fish’
Hawaiian	iʔa
Samoan	iʔa
Tongan	ika
Maori	ika

Shared ancestry



	‘fish’
Hawaiian	iʔa
Samoan	iʔa
Tongan	ika
Maori	ika

	‘fish’
POc	*ika

Shared ancestry

Proto-Oceanic

	‘fish’
Hawaiian	iʔa
Samoan	iʔa
Tongan	ika
Maori	ika

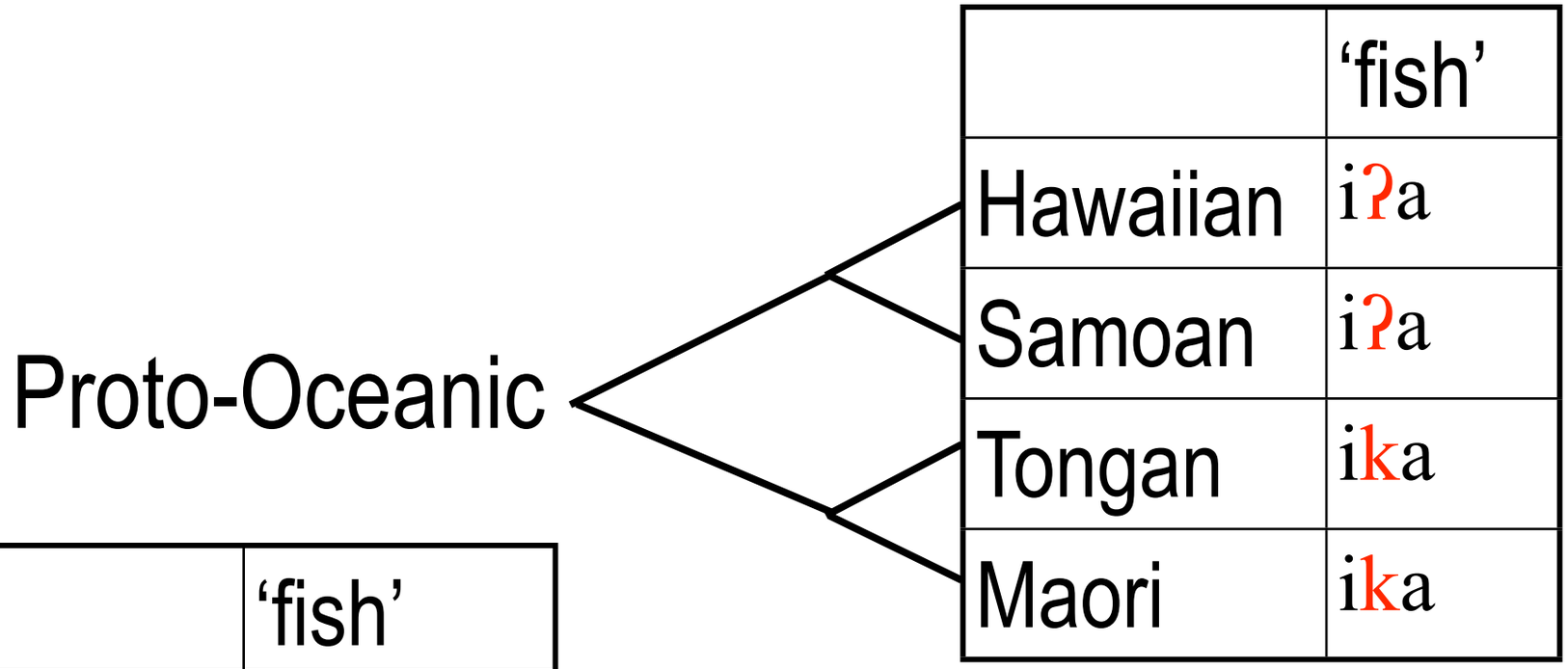
*ʔ > k

POc	*ika

or

	‘fish’
POc	*iʔa

Shared ancestry



	‘fish’
Hawaiian	iʔa
Samoan	iʔa
Tongan	ika
Maori	ika

	‘fish’
POc	*ika

or

	‘fish’
POc	*iʔa

University of California
C A L
N L P
Berkeley

Can we harness more languages?

	‘fish’
Hawaiian	iʔa
Samoan	iʔa
Tongan	ika
Maori	ika
Geser	ikan
Rapanui	ika
Nukuoro	iga
Niue	ika

Welcome to Oceanic Park

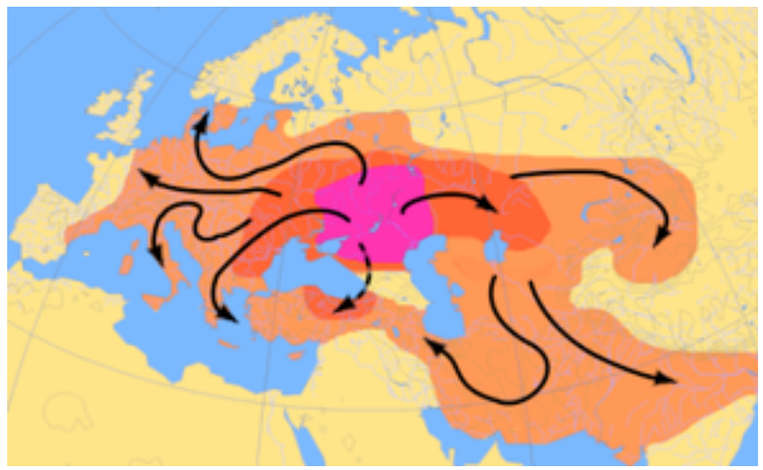
Outline:

- Motivation
- Computational model
- Learning and inference
- Experiments on Proto-Oceanic

Why reconstruct?

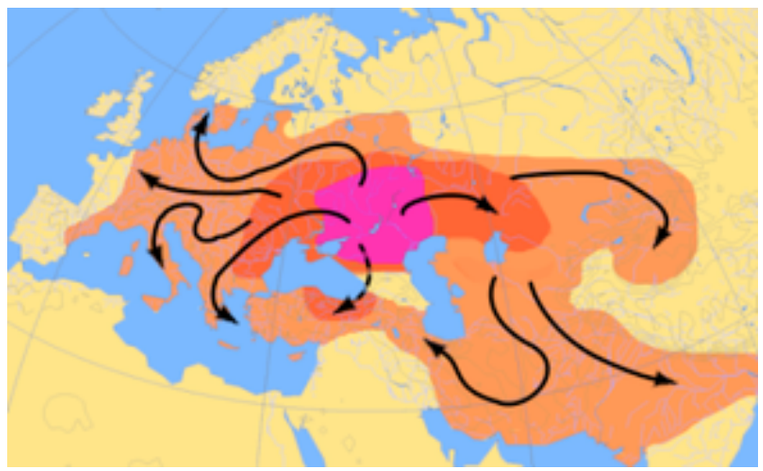
- Can answer a large number of questions about our past

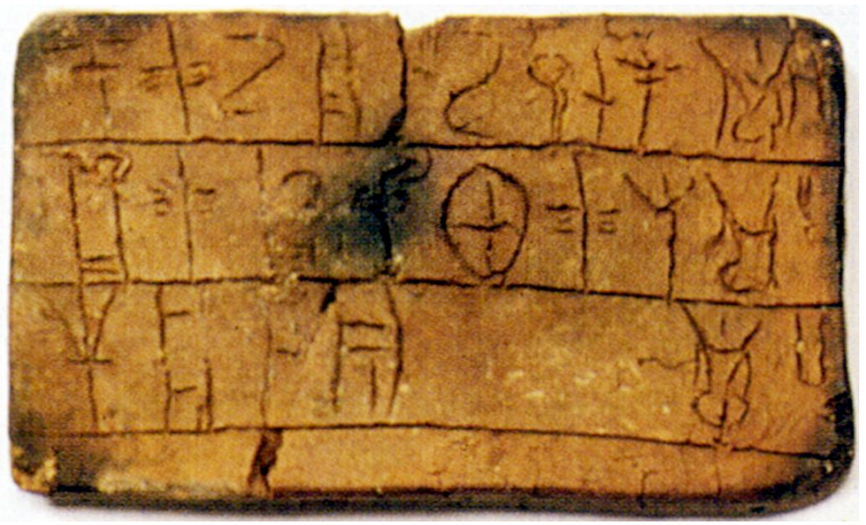
  - Learn about ancient populations' migrations



Why reconstruct?

- Can answer a large number of questions about our past
  - Learn about ancient populations' migrations
  - Decipherment of ancient scripts





How linguists do reconstruction

- Direct diachronic evidence, sometimes

How linguists do reconstruction

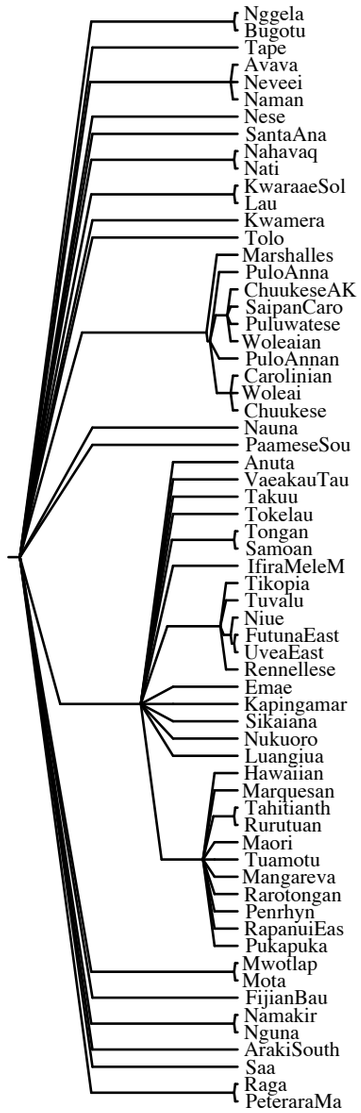
- Direct diachronic evidence, sometimes

	‘fish’	‘fear’
Hawaiian	iʔa	makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	manavahē
Maori	ika	mataku

- Often not available (prehistorical cultures)
  - The comparative method
  - Unsupervised setup

Computational Model

Input



+

	‘fish’	‘fear’
Hawaiian	iʔa	makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	
Maori	ika	mataku

…

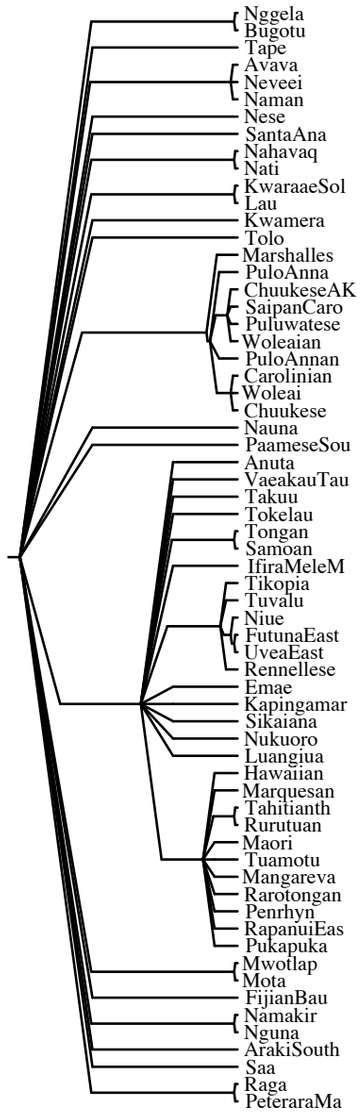
⋮

512 languages X 6856 cognate sets

IPA format

Density: 60K entries

Input



+

	‘fish’	‘fear’
Hawaiian	iʔa	makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	
Maori	ika	mataku

…

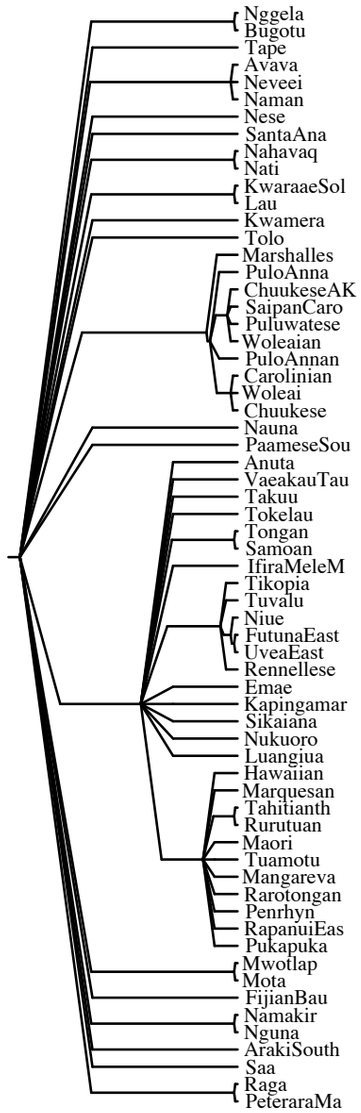
⋮

512 languages X 6856 cognate sets

IPA format

Density: 60K entries

Input



+

	‘fish’	‘fear’
Hawaiian	iʔa	makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	
Maori	Missing data	taku

…

⋮

512 languages X 6856 cognate sets

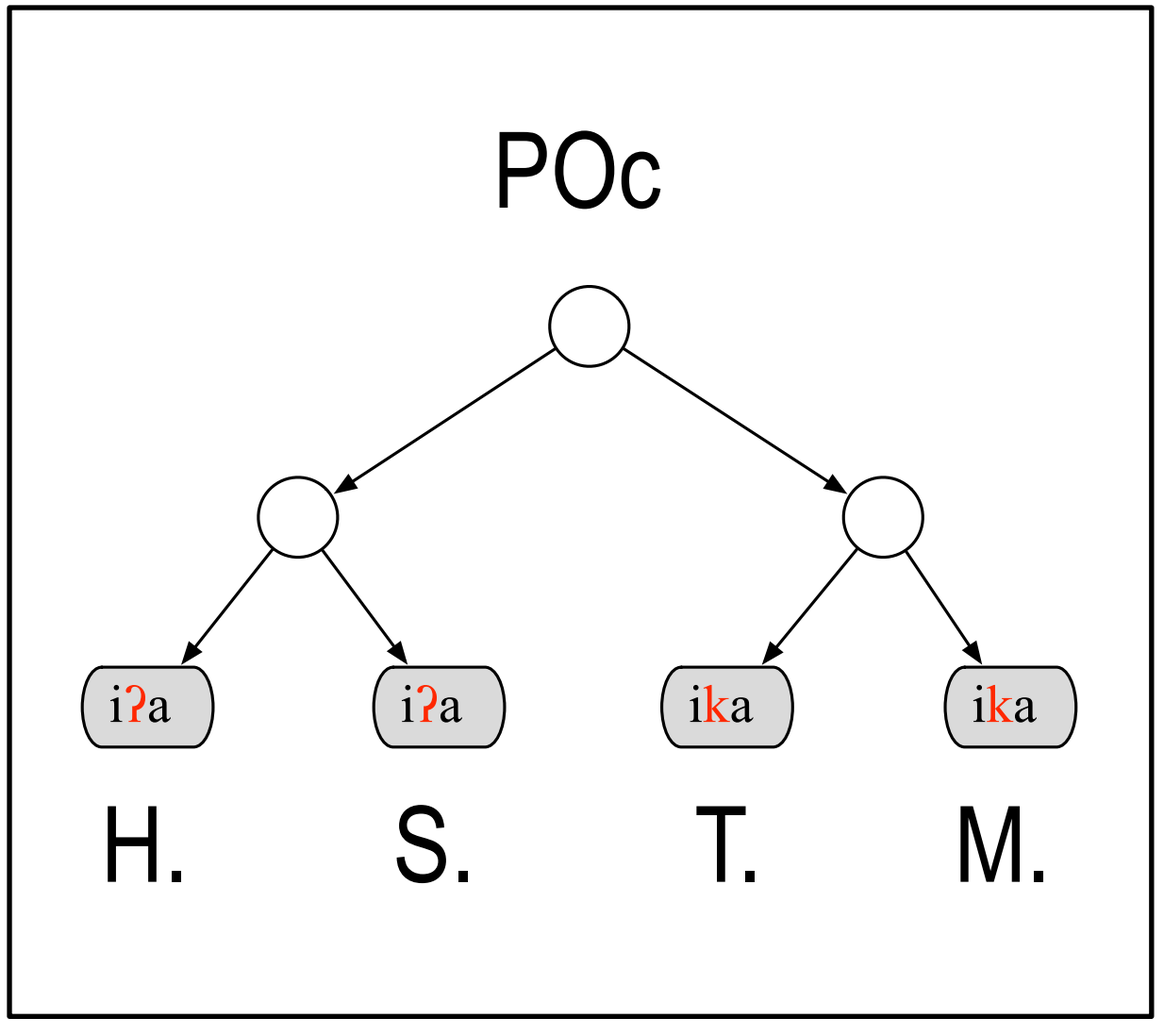
IPA format

Density: 60K entries

Graphical model

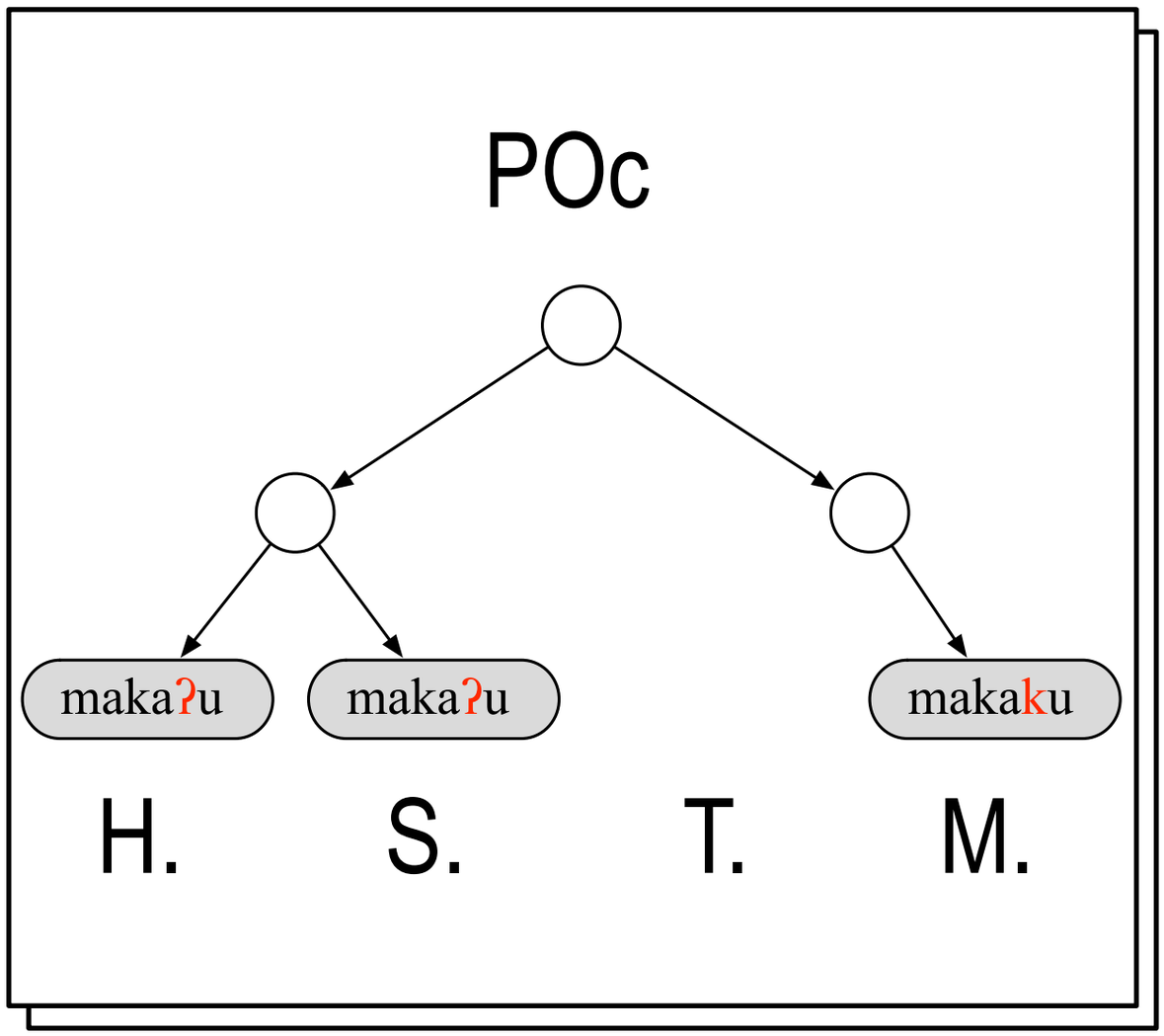
	‘fish’	‘fear’
Hawaiian	iʔa	makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	
Maori	ika	mataku

Graphical model



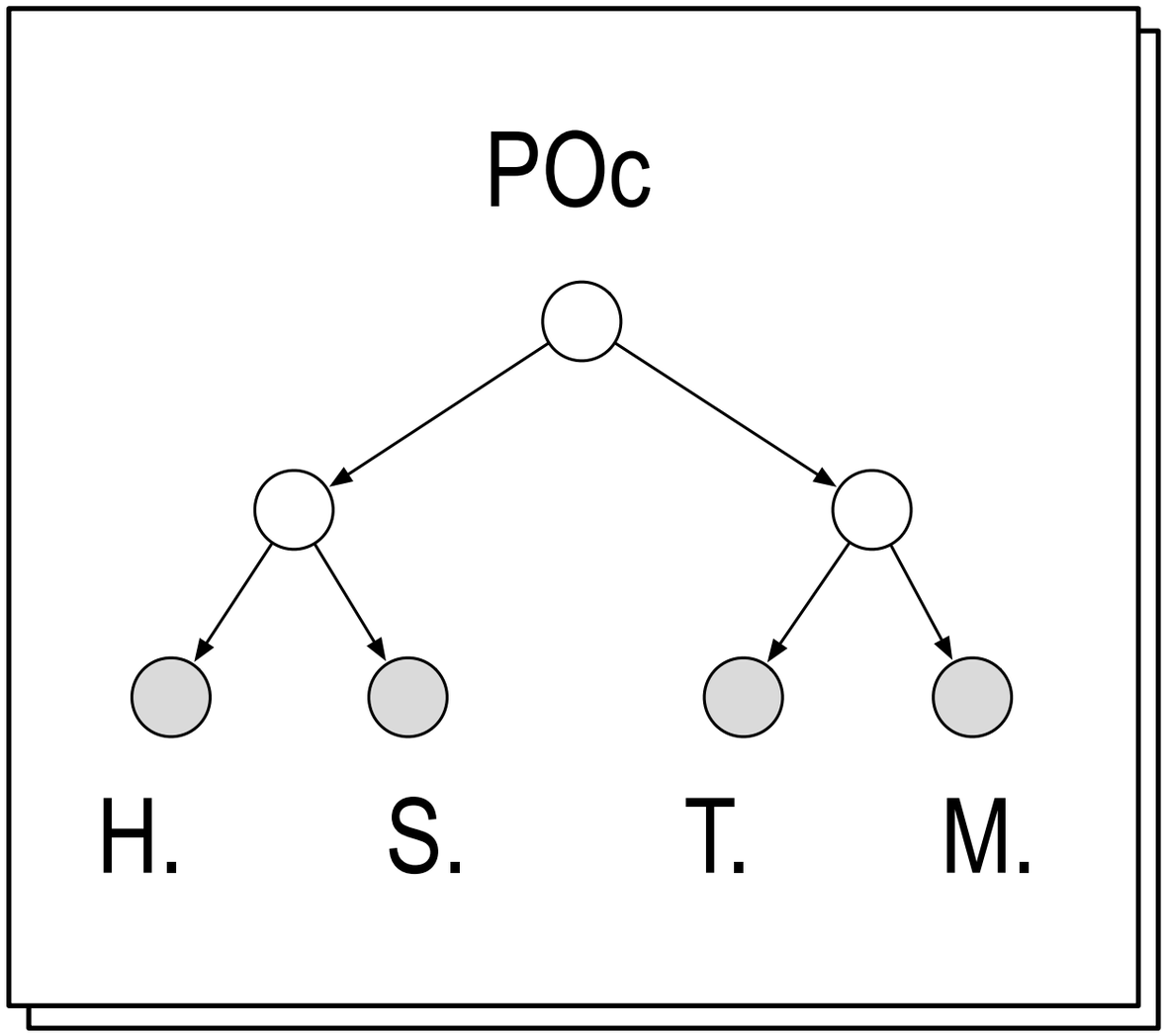
	‘fish’	‘fear’
Hawaiian	iʔa	makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	
Maori	ika	mataku

Graphical model



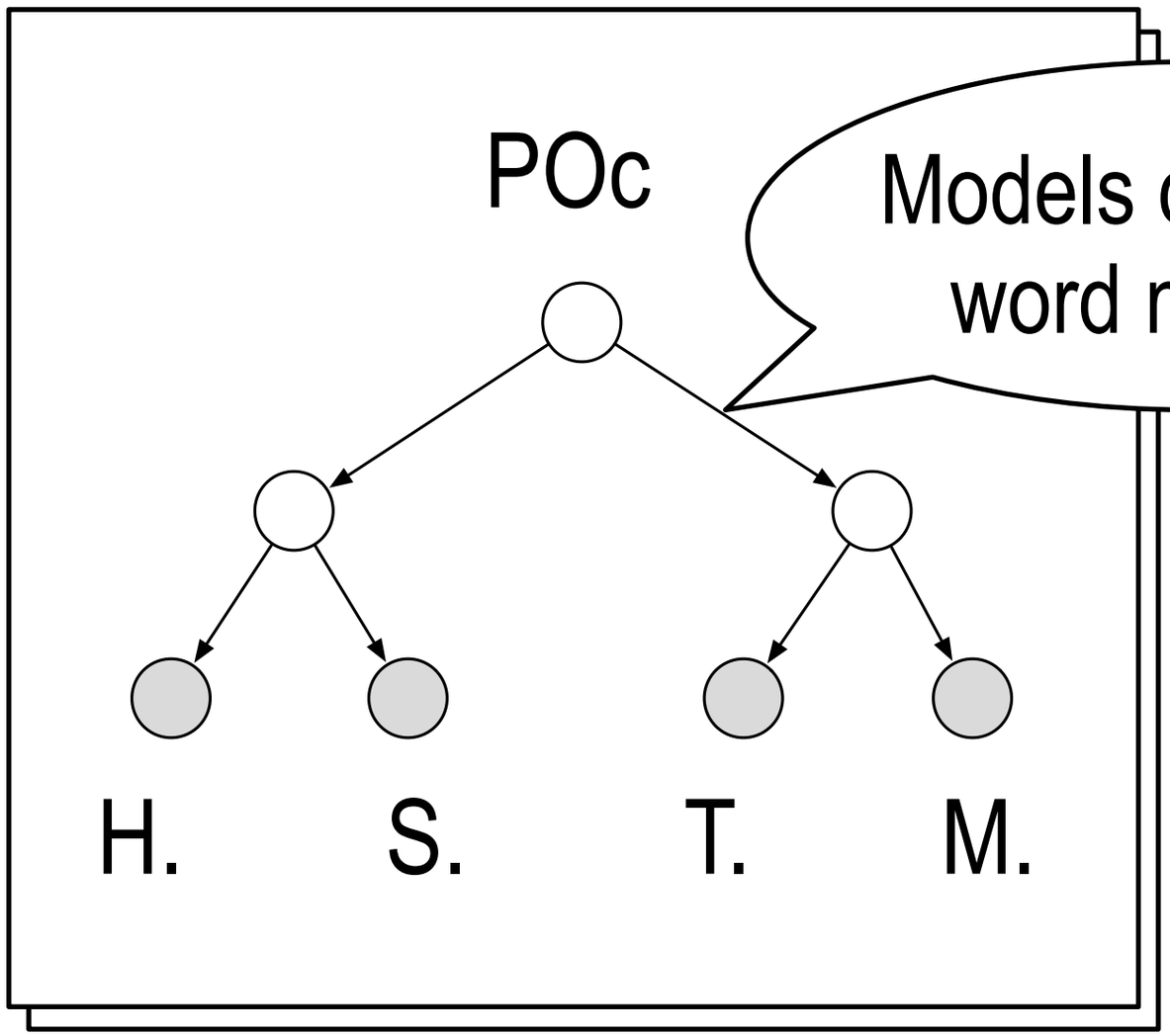
	‘fish’	‘fear’
Hawaiian	iʔa	makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	
Maori	ika	mataku

Graphical model



	‘fish’	‘fear’
Hawaiian	iʔa	makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	
Maori	ika	mataku

Graphical model



	fish’	‘fear’
Hawaiian	iʔa	makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	
Maori	ika	mataku

University of California
C A L
N L P
Berkeley

Modeling string mutation

- What kind of string mutations need to be captured?
  - Substitution

    *k > ?

Modeling string mutation

- What kind of string mutations need to be captured?
  - Substitution

    *k > ʔ

  - Insertion (and deletion)

    *h > wh

	‘break’
Hawaiian	haki
Samoan	fati
Tongan	fasi
Maori	whati

Modeling string mutation

- What kind of string mutations need to be captured?
  - Substitution

    *k > ʔ

  - Insertion (and deletion)

    *h > wh

  - Context

    *h > wh / # _

	‘break’
Hawaiian	haki
Samoan	fati
Tongan	fasi
Maori	whati

Modeling string mutation

- What kind of string mutations need to be captured?
  - Substitution

    *k > ʔ

  - Insertion (and deletion)

    *h > wh

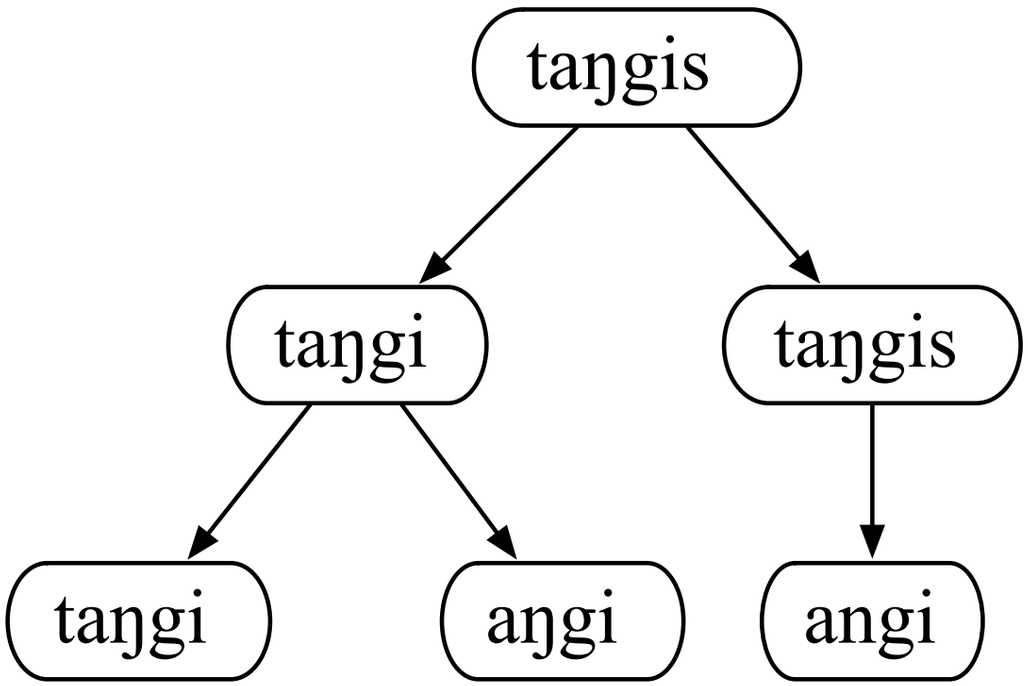
  - Context

    *h > wh / # _

	‘break’	‘aloha’
Hawaiian	haki	aloha
Samoan	fati	alofa
Tongan	fasi	ʔalofa
Maori	whati	aroha

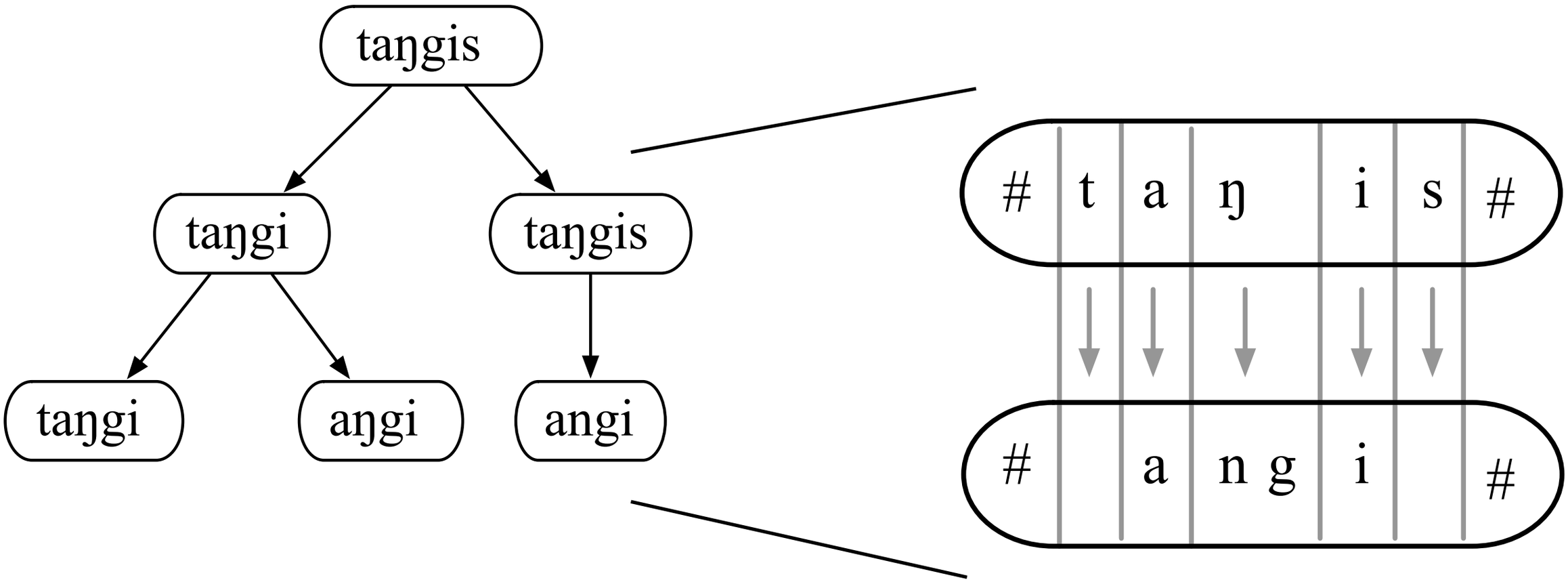
NOT: arowha

String transducer



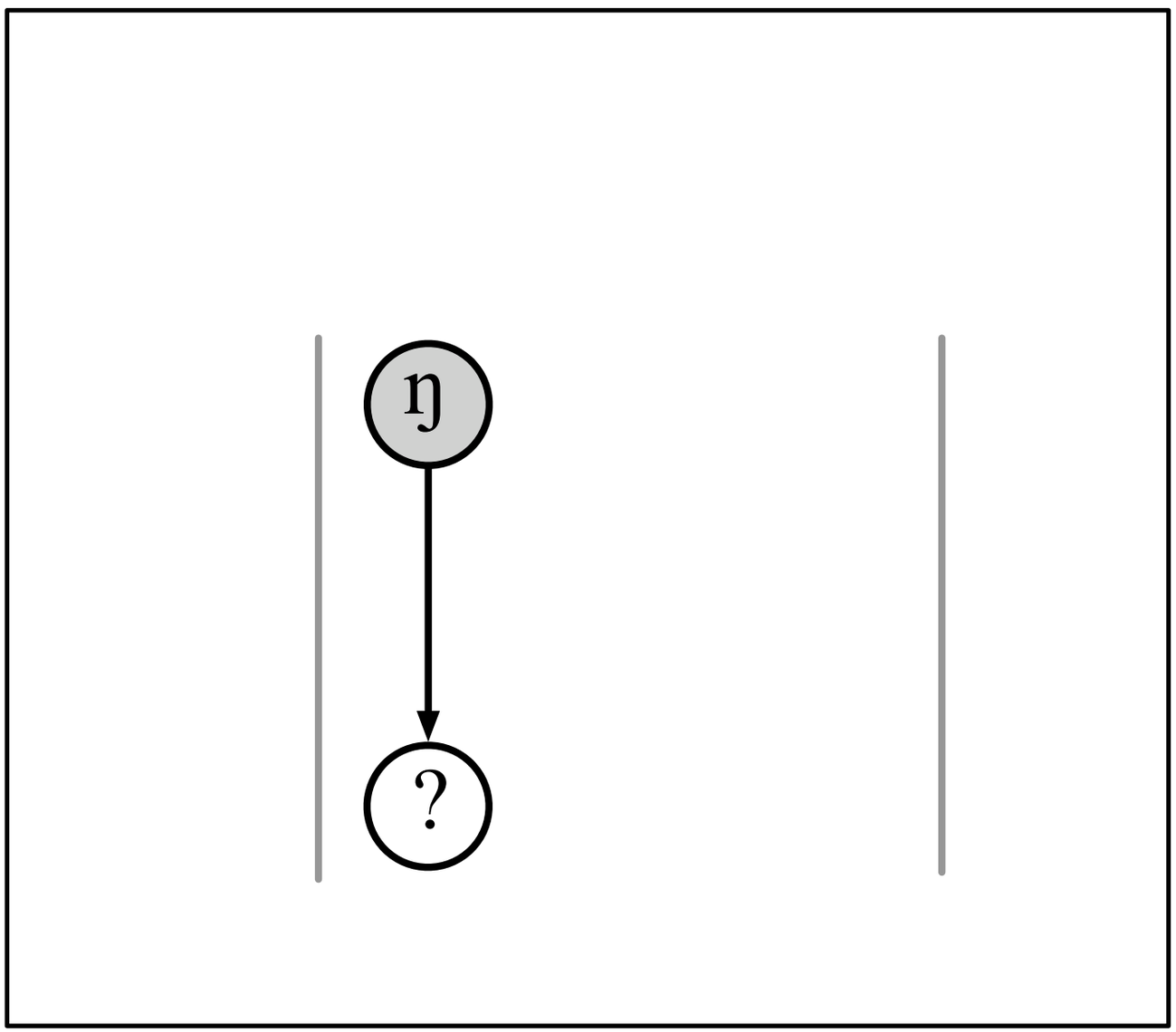
‘to cry’

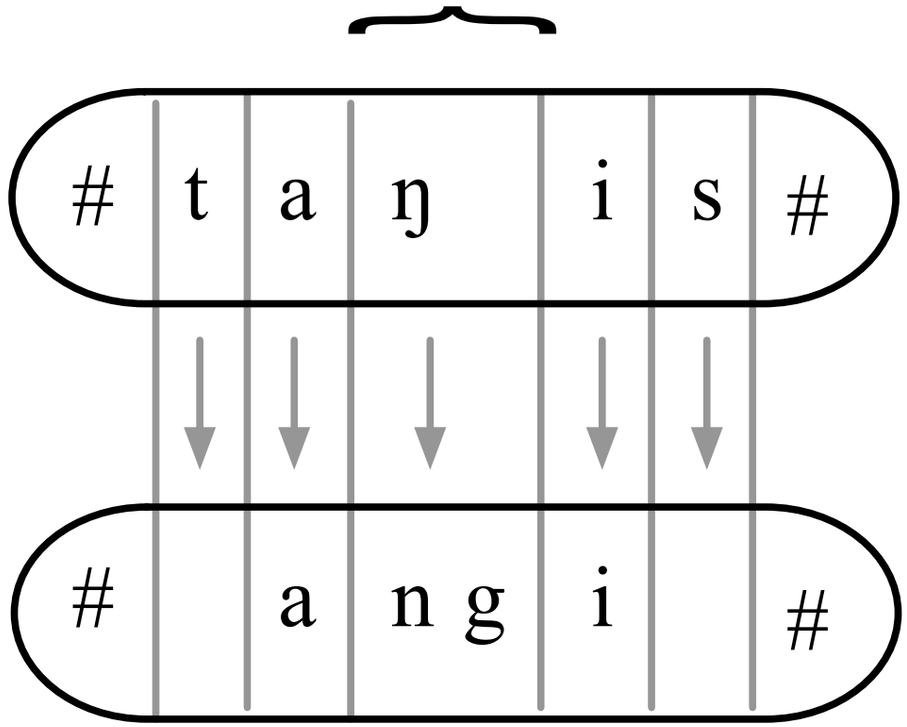
String transducer



'to cry'

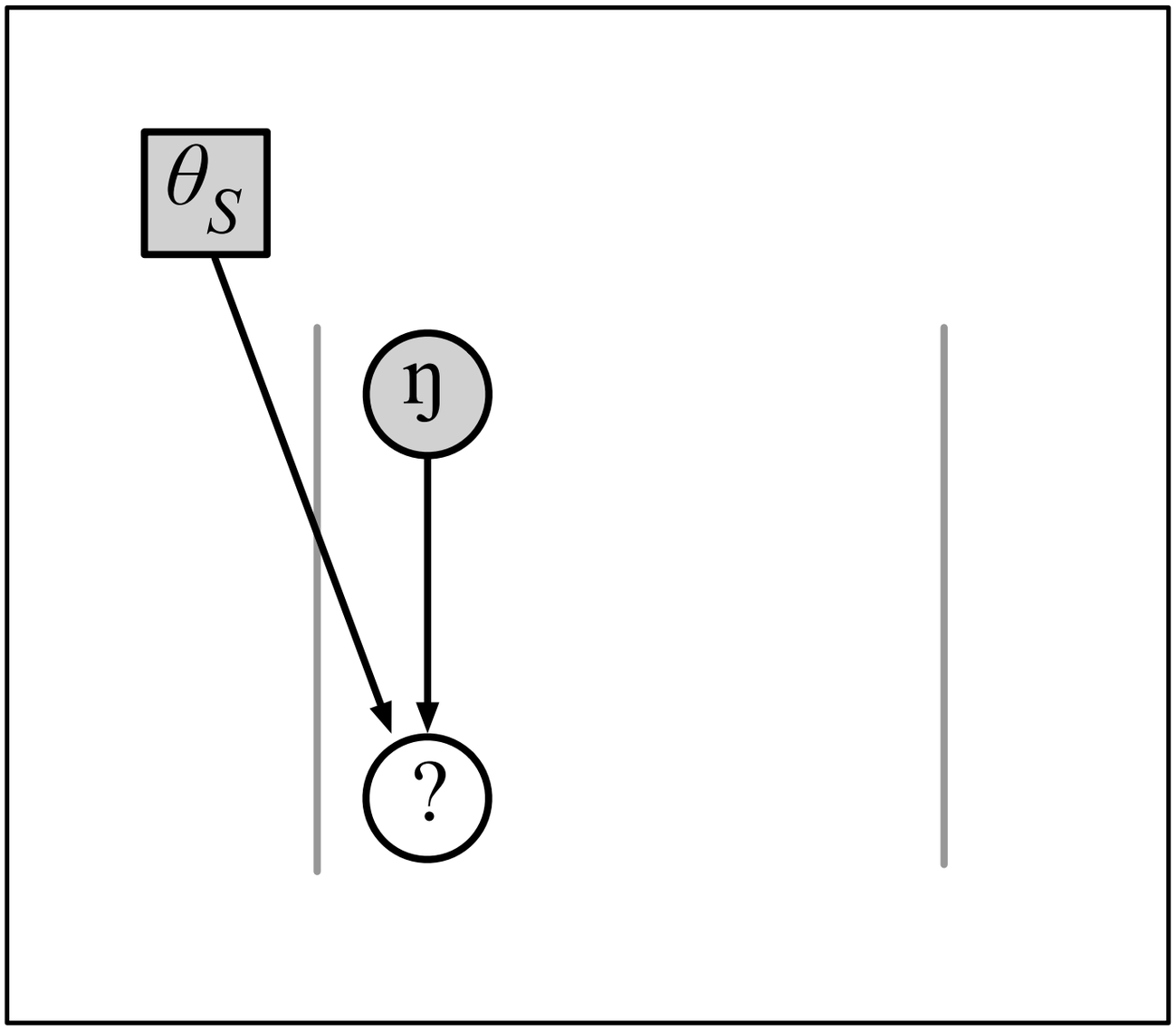
String transducer

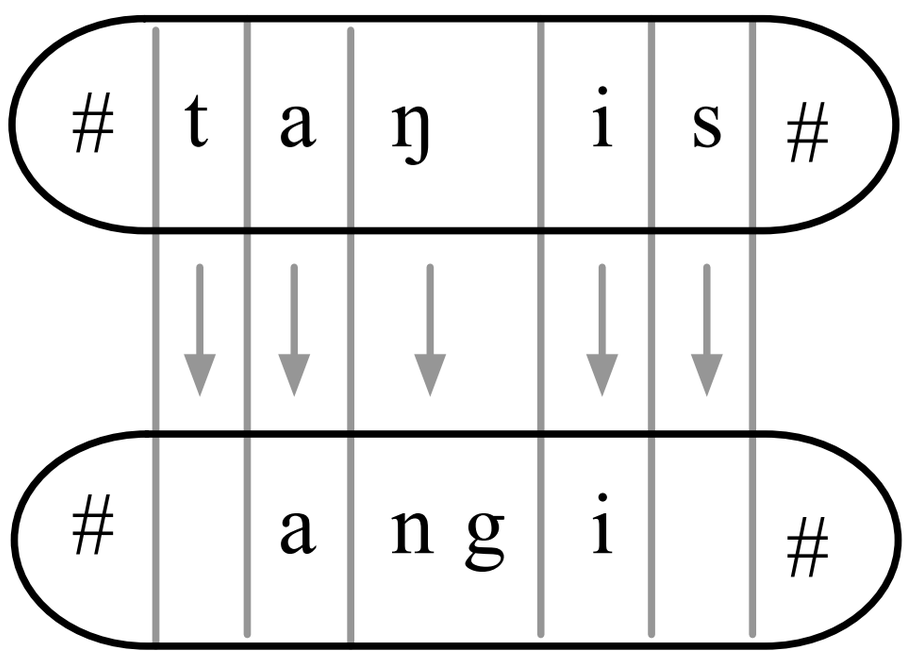




String transducer

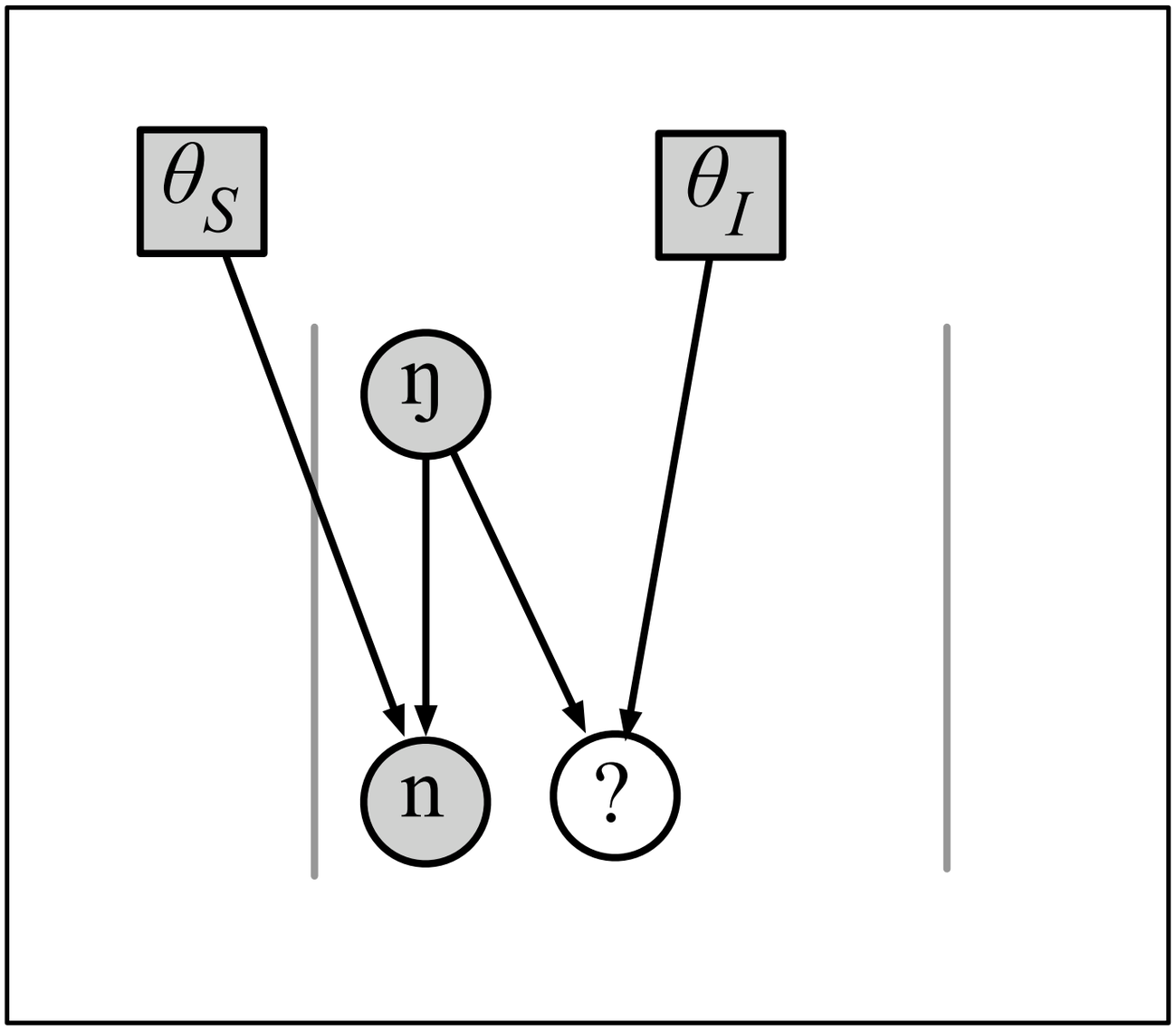
θ_S: Substitution/Deletion Parameters

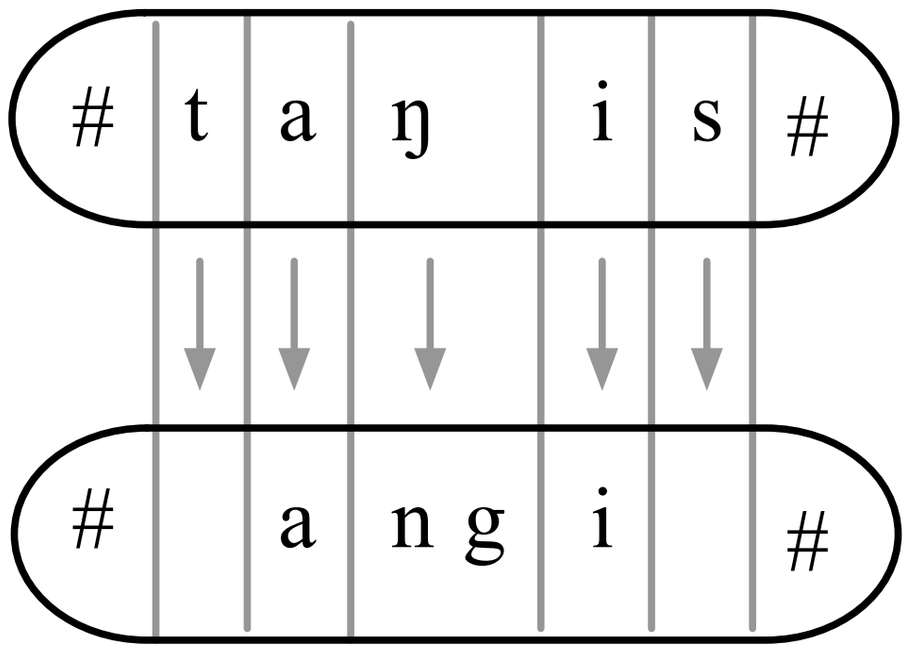




String transducer

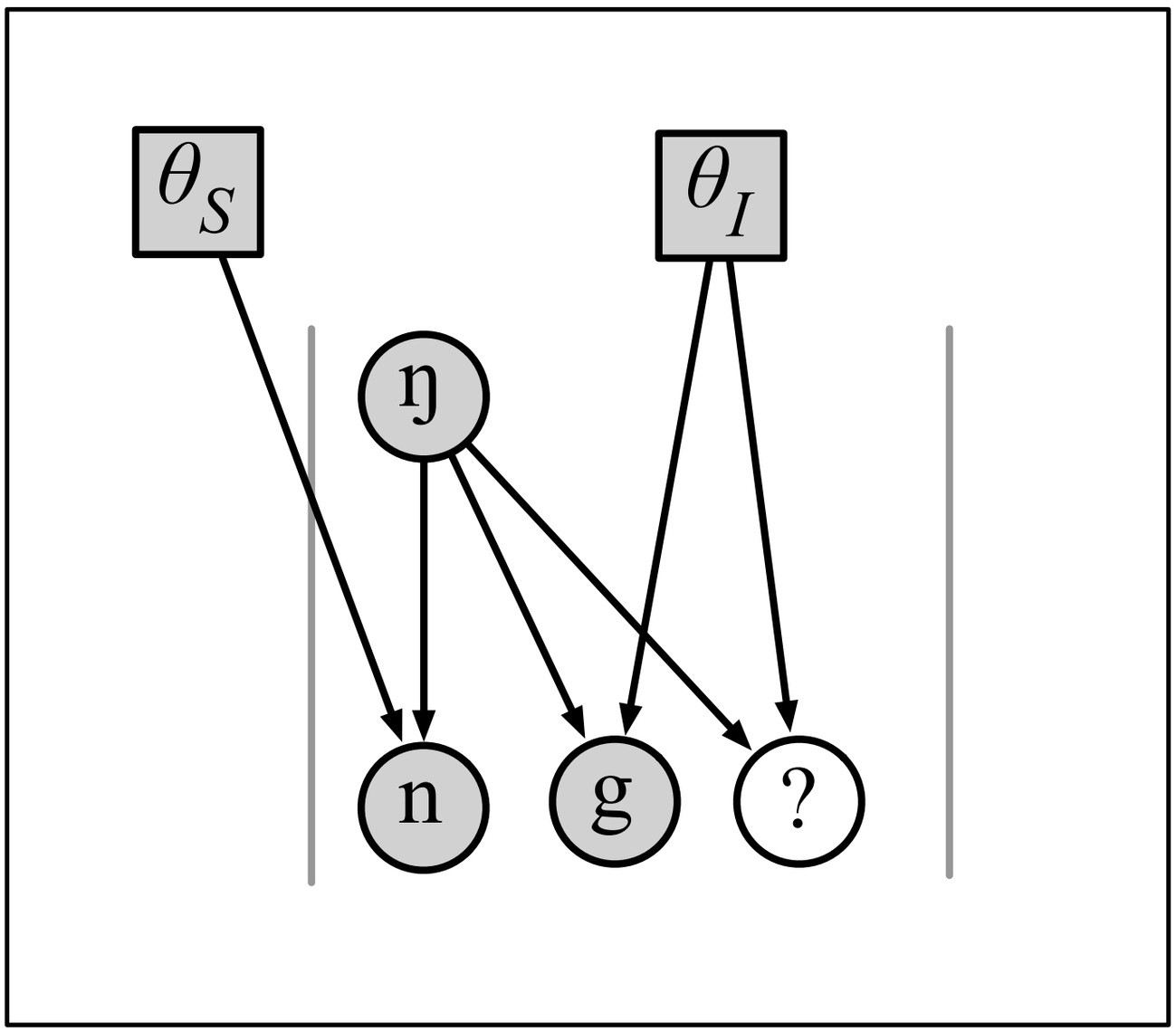
θ_I: Insertion Parameters

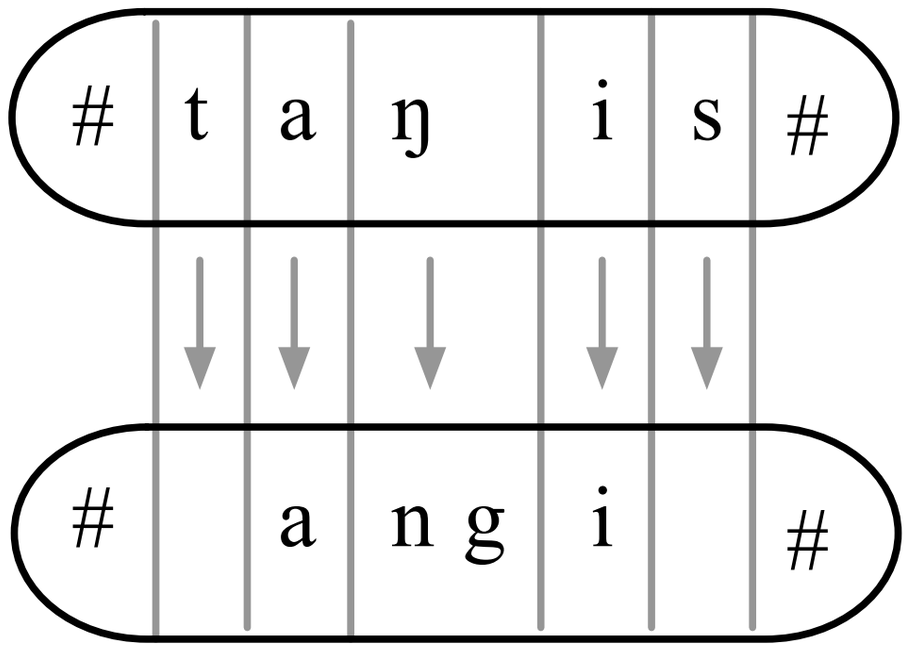




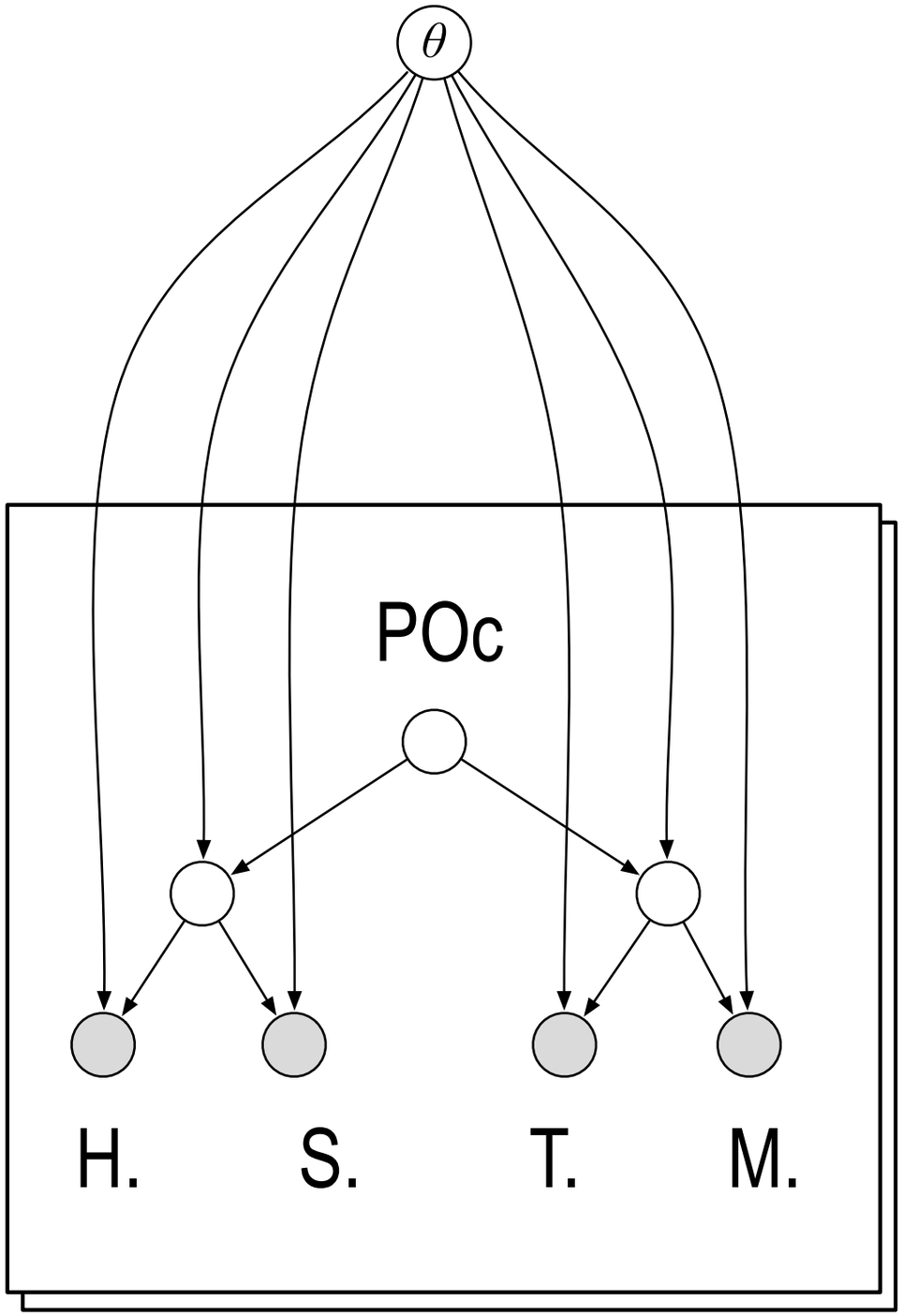
University of California
C A L
N L P
Berkeley

String transducer





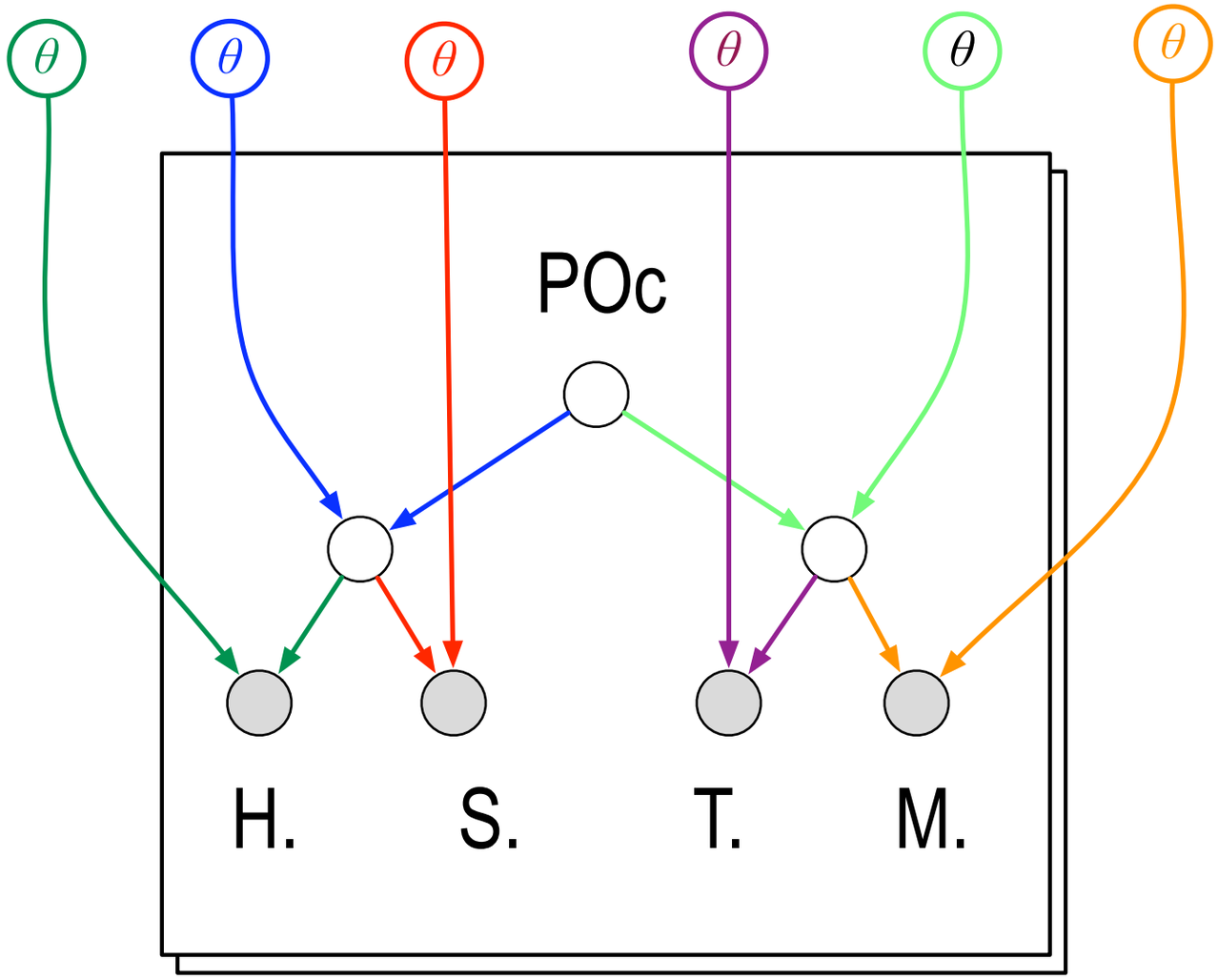
Parameters



- Global?
  - Cannot explicitly represent sound changes!

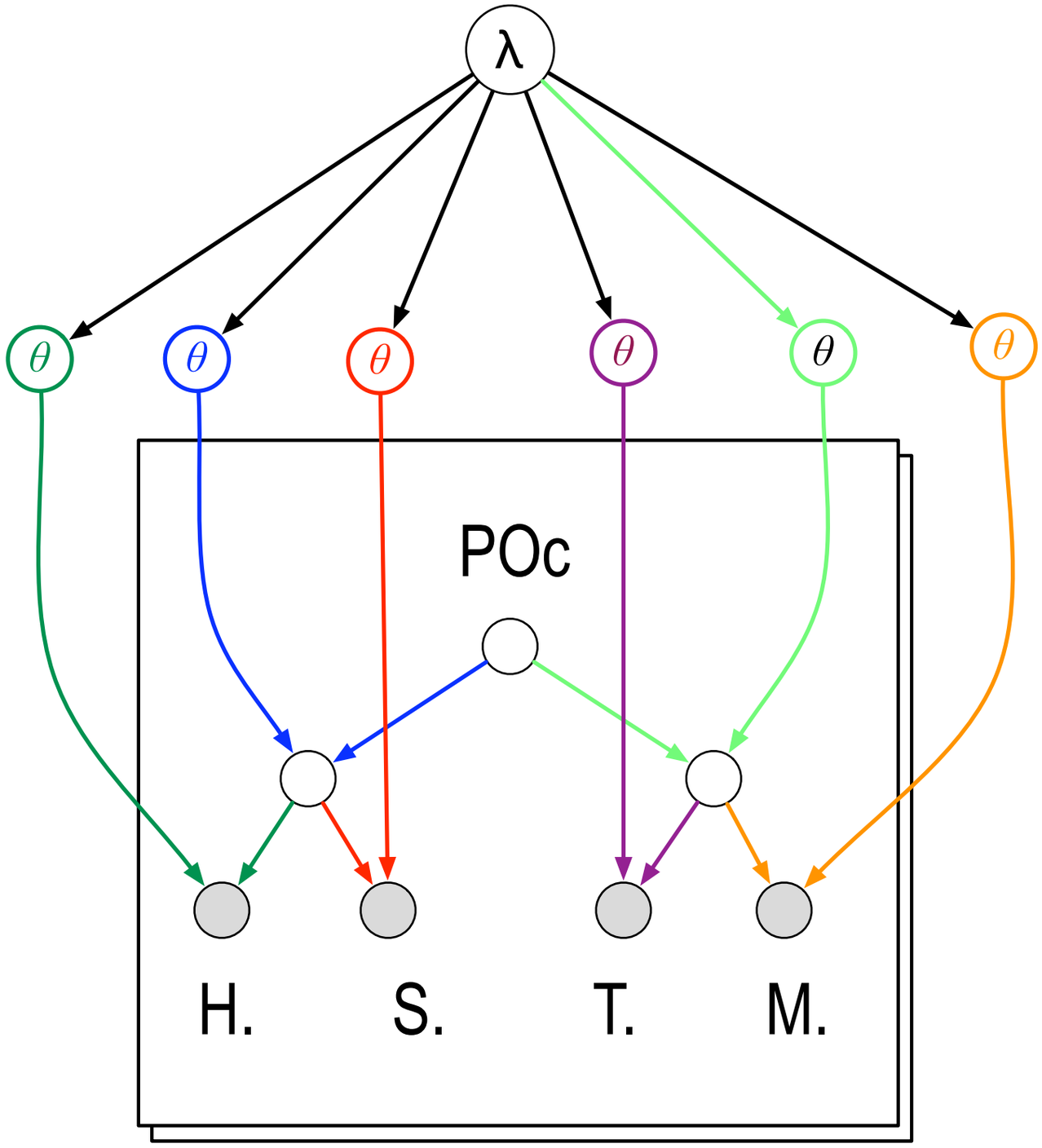
$$\theta = \theta_S \ \& \ \theta_I$$

Parameters



- Global?
  - Cannot explicitly represent sound changes!
- Branch-specific
  - Parameter proliferation!

Parameters



- Global?
  - Cannot explicitly represent sound changes!
- Branch-specific
  - Parameter proliferation!
- Solution:
  - Learning cross-linguistic trends

Cross-linguistic trends

- Some sound changes are unlikely cross-linguistically:
  - Velar stop to vowel: $k > a$

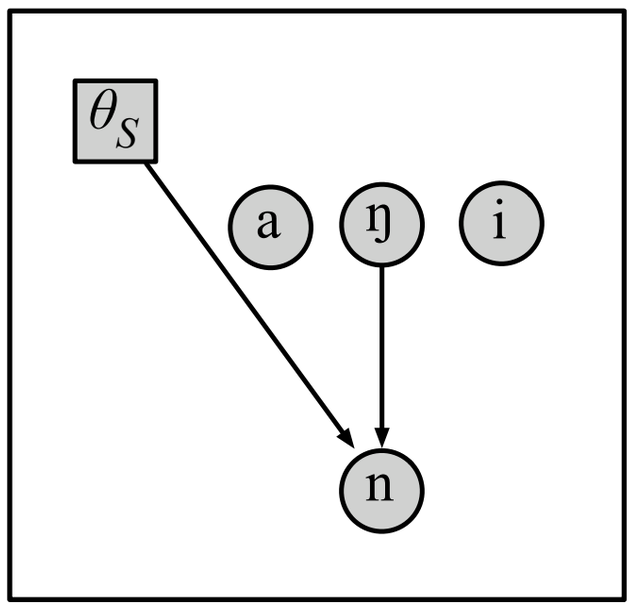
Cross-linguistic trends

- Some sound changes are unlikely cross-linguistically:
  - Velar stop to vowel: k > a

- Some sound changes are frequent cross-linguistically:
  - Consonant place change: k > ʔ
  - Debuccalization: f > h
  - Identity (faithfulness): $x > x$

Learning cross-linguistic trends

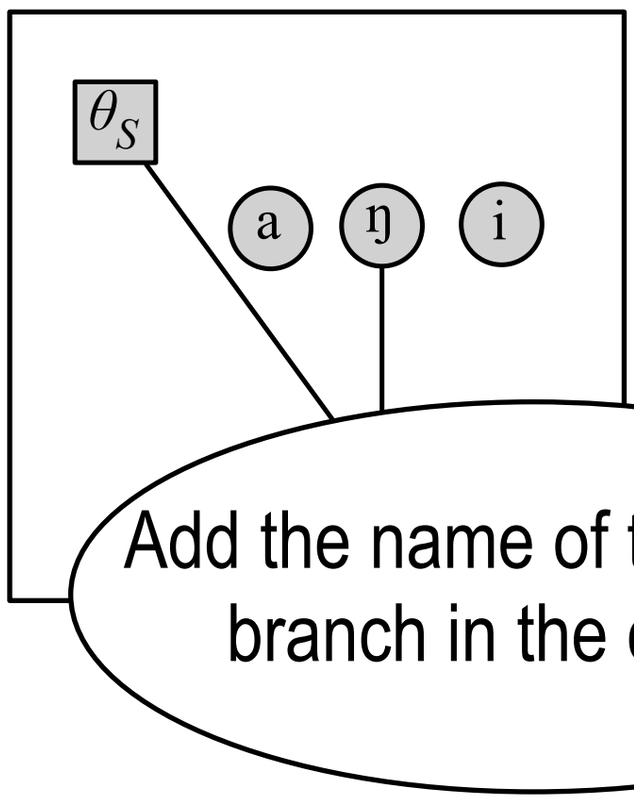
- How to learn these universals: express the transducer parameters as the output of a log-linear model



$$\theta_S \left(\begin{array}{l} \blacktriangleright\ \text{ŋ} > \text{n} \\ \blacktriangleright\ \text{mutation to Maori} \end{array} \right) \propto \exp\{ \langle \lambda, f \rangle \}$$

Learning cross-linguistic trends

- How to learn these universals: express the transducer parameters as the output of a log-linear model

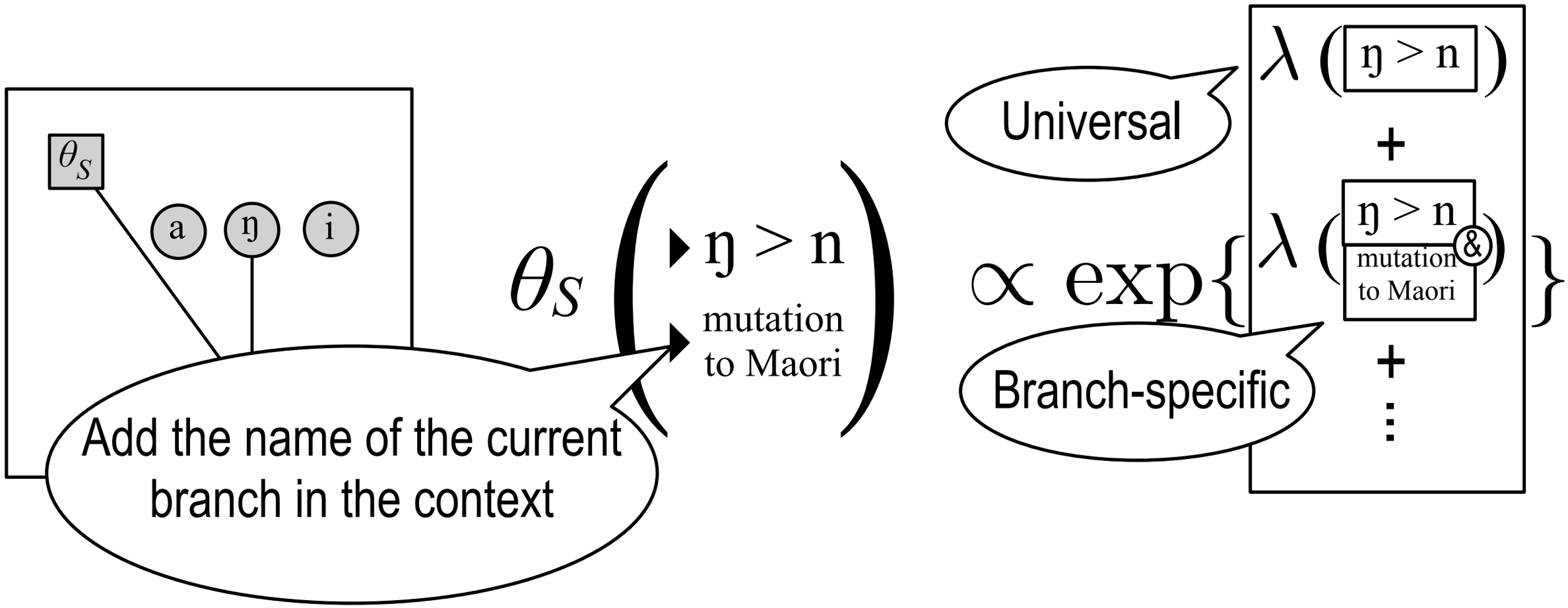


$$\theta_S\left(\blacktriangleright\ \eta > n \quad \text{mutation to Maori}\right) \propto \exp\{\ \langle \lambda, f \rangle\ \}$$

University of California
C A L
N L P
Berkeley

Learning cross-linguistic trends

- How to learn these universals: express the transducer parameters as the output of a log-linear model
- Universals ignore the name of the current branch



University of California
C A L
N L P
Berkeley

Second improvement

- Response to a concrete problem: sound changes are *not* exceptionless in real data

University of California
CALNLP
Berkeley

Second improvement

- Response to a concrete problem: sound changes are *not* exceptionless in real data
- Example: tension between a sound change and a morphological paradigm

Second improvement

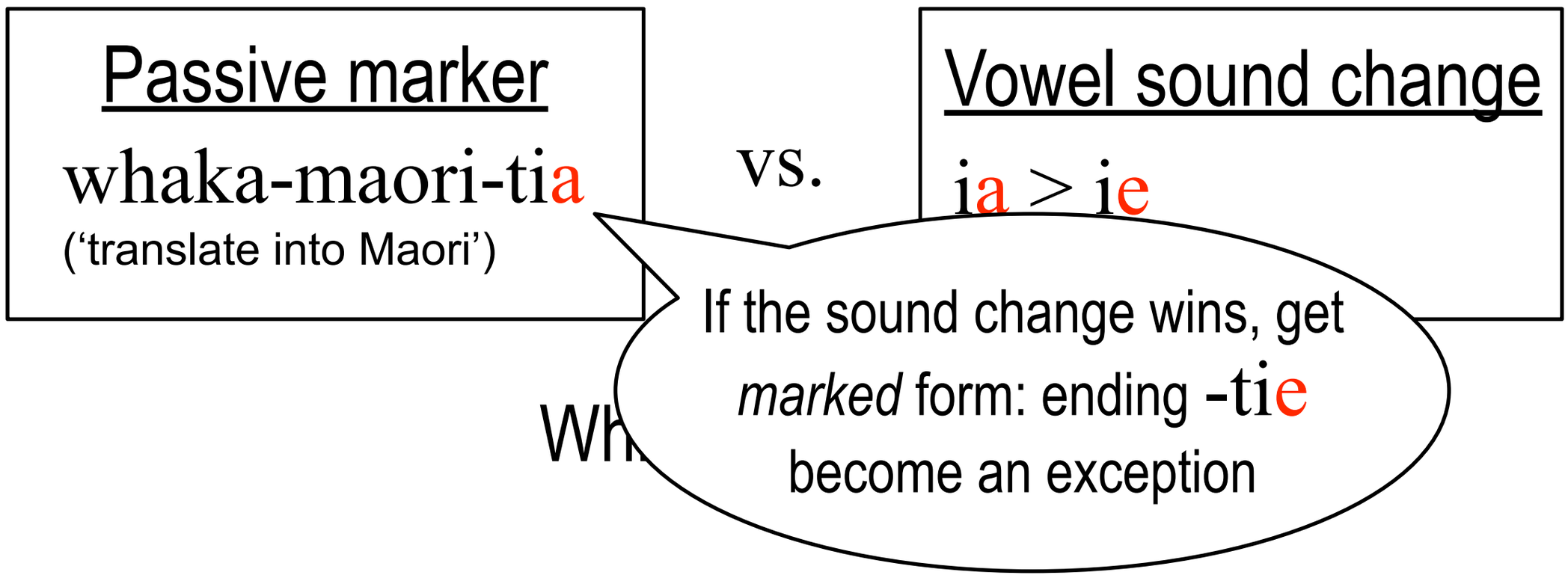
- Response to a concrete problem: sound changes are *not* exceptionless in real data
- Example: tension between a sound change and a morphological paradigm

Passive marker	vs.	Vowel sound change
whaka-maori-tia ('translate into Maori')		ia > ie

Which one wins?

Second improvement

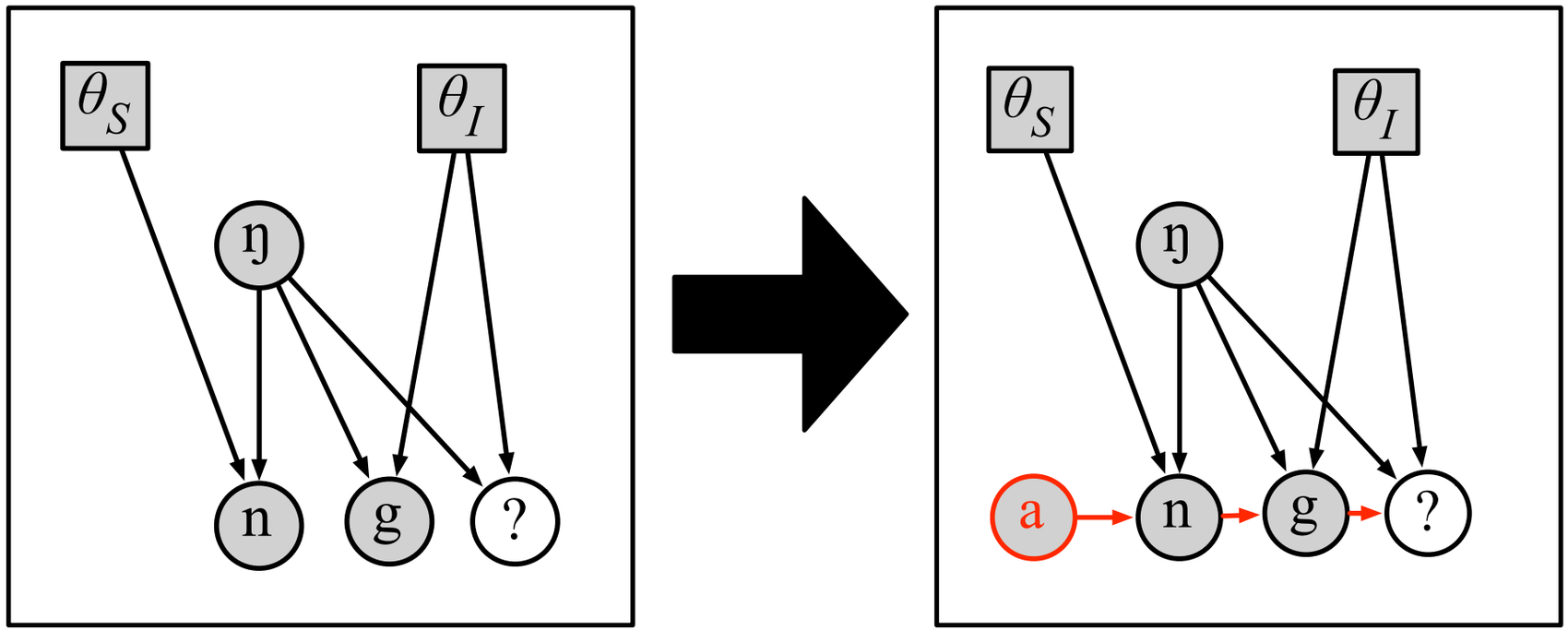
- Response to a concrete problem: sound changes are *not* exceptionless in real data
- Example: tension between a sound change and a morphological paradigm



University of California
C A L
N L P
Berkeley

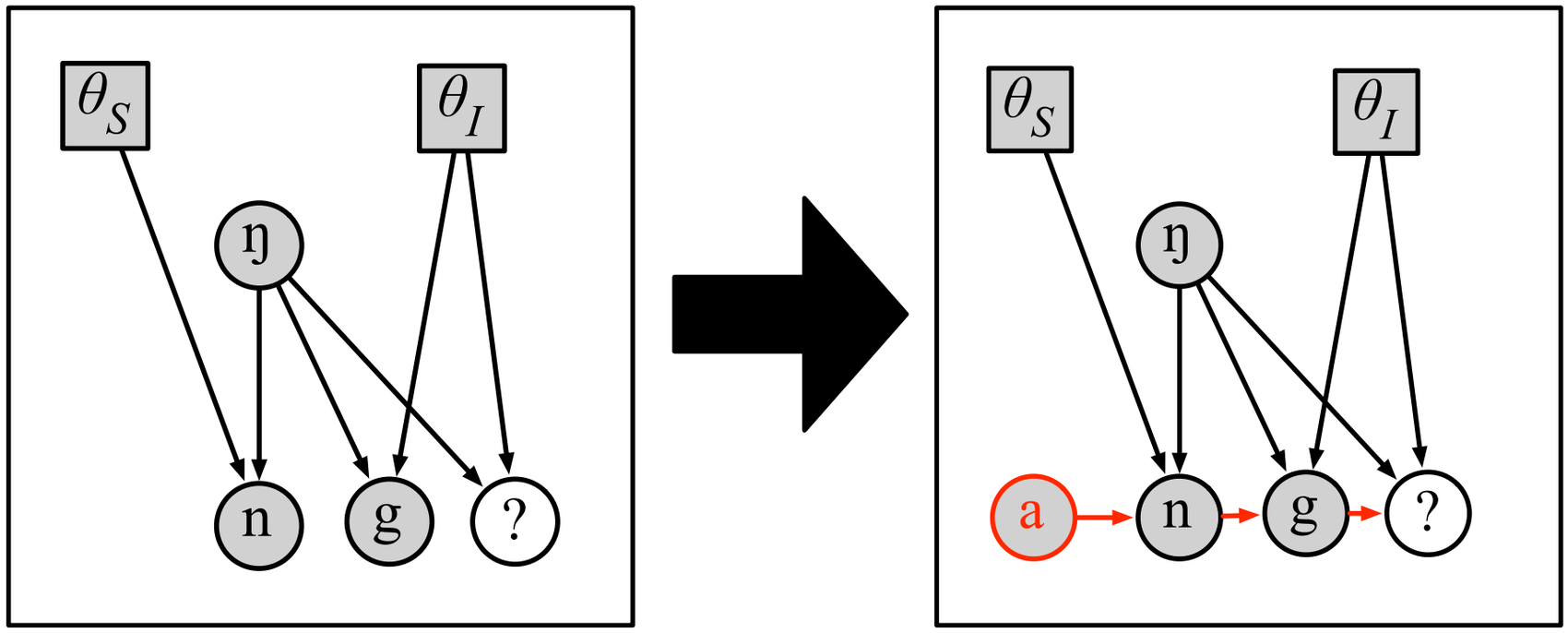
Adding markedness features

- Add dependencies in the string transducer model:



Adding markedness features

- Add dependencies in the string transducer model:



- Also add new features:

word has /C V V/

word has /a #/ & mutation to Maori

Learning and Inference

University of California
C A L
N L P
Berkeley

Learning λ while reconstructing

- Monte Carlo EM
  - M step: not analytic but convex
  - E step: challenging; use MCMC

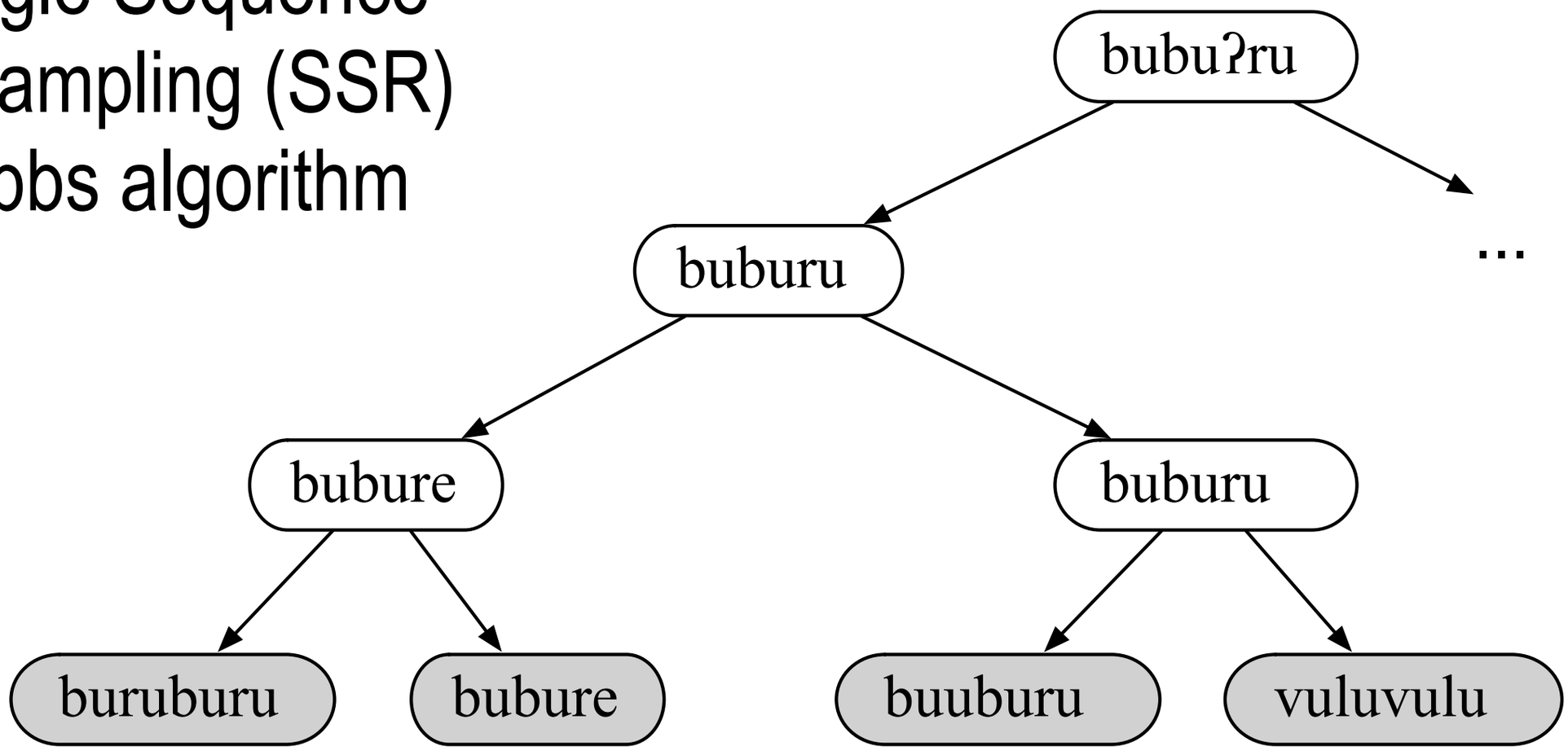
Learning λ while reconstructing

- Monte Carlo EM
  - M step: not analytic but convex
  - E step: challenging; use MCMC

- Hardness of inference (E step):
  - Horizontal links $\Longrightarrow$ (inference $\geq$ non-planar Ising inference)
  - Insertions, deletion $\Longrightarrow$ non-standard setup

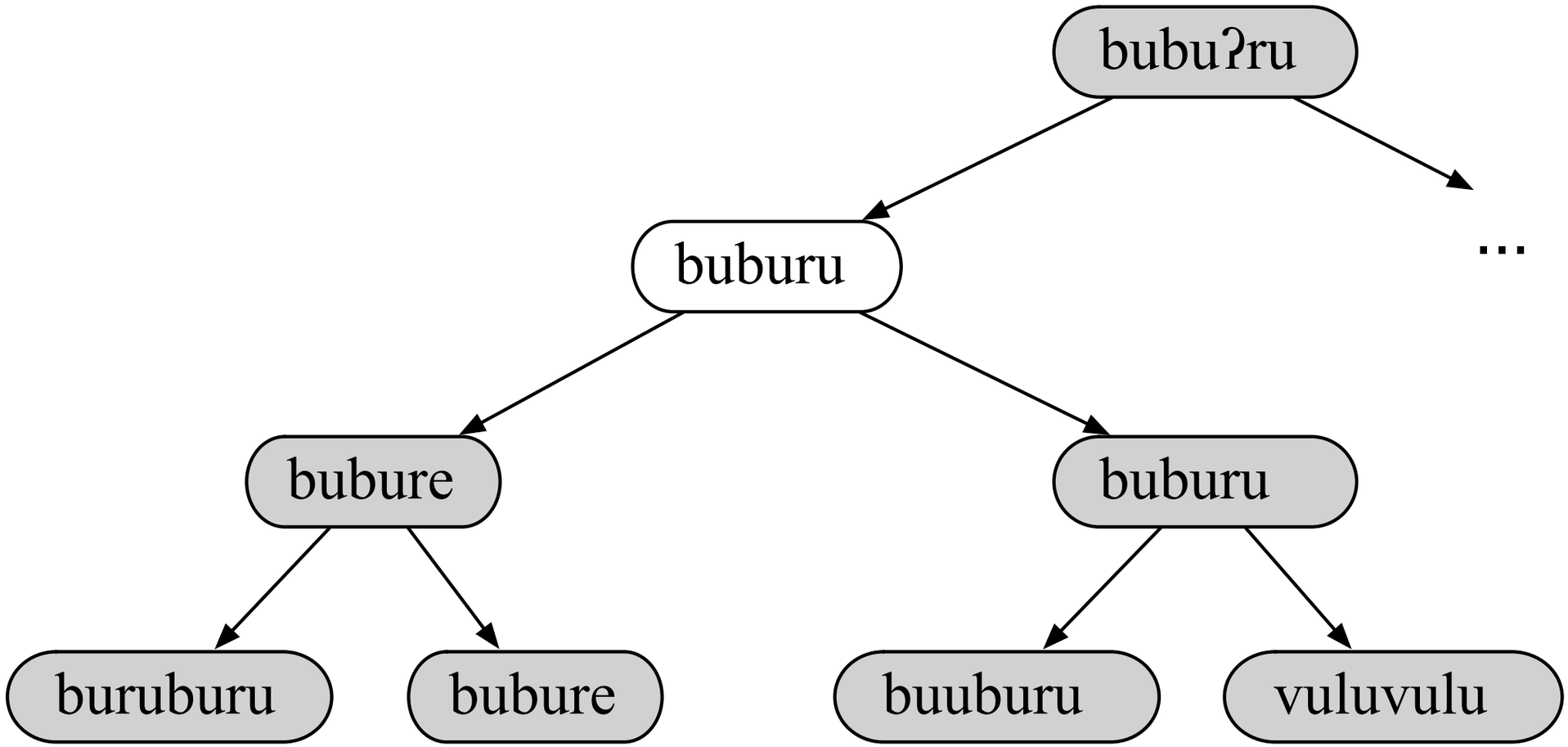
Our previous work

Single Sequence Resampling (SSR) Gibbs algorithm



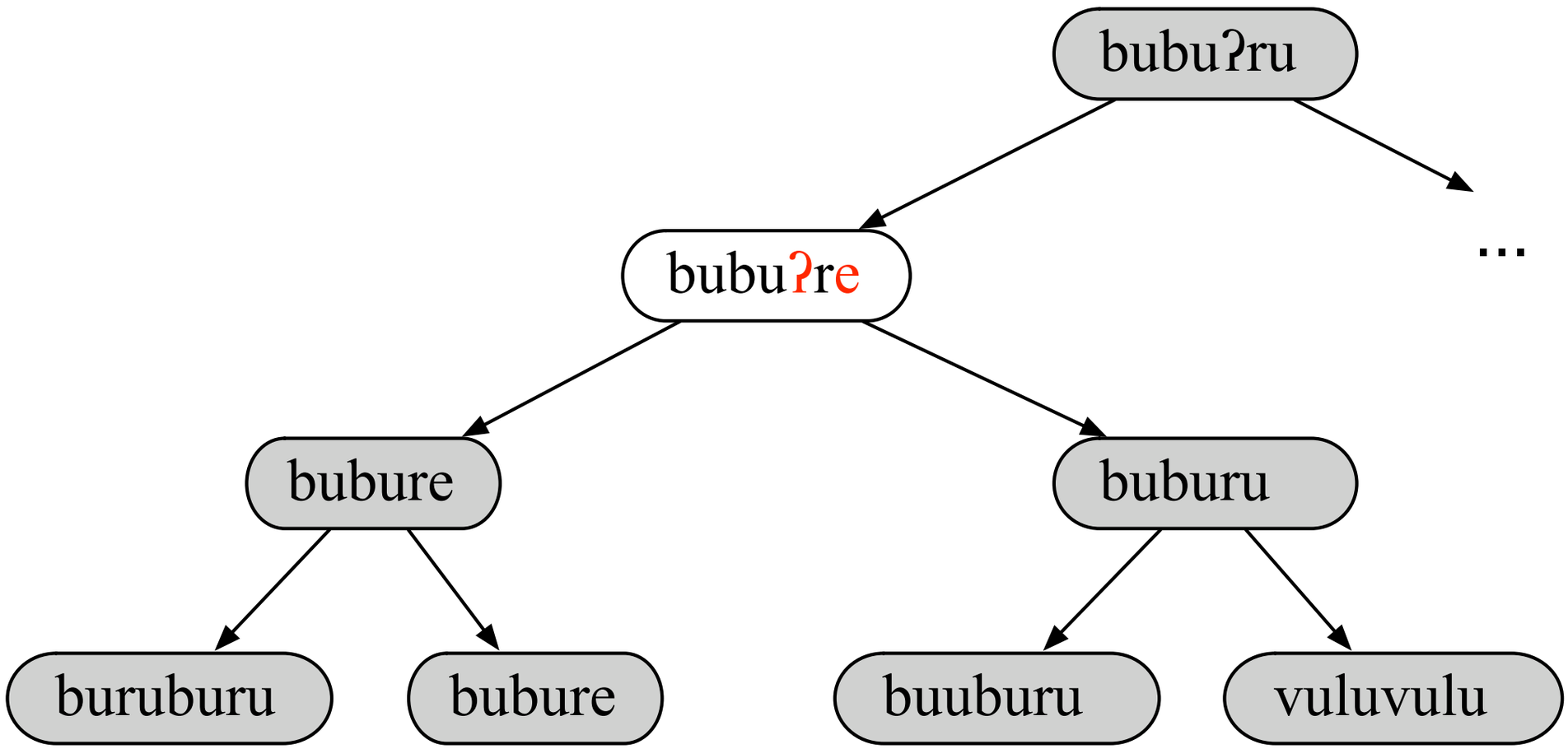
'grass'

Gibbs sampler



'grass'

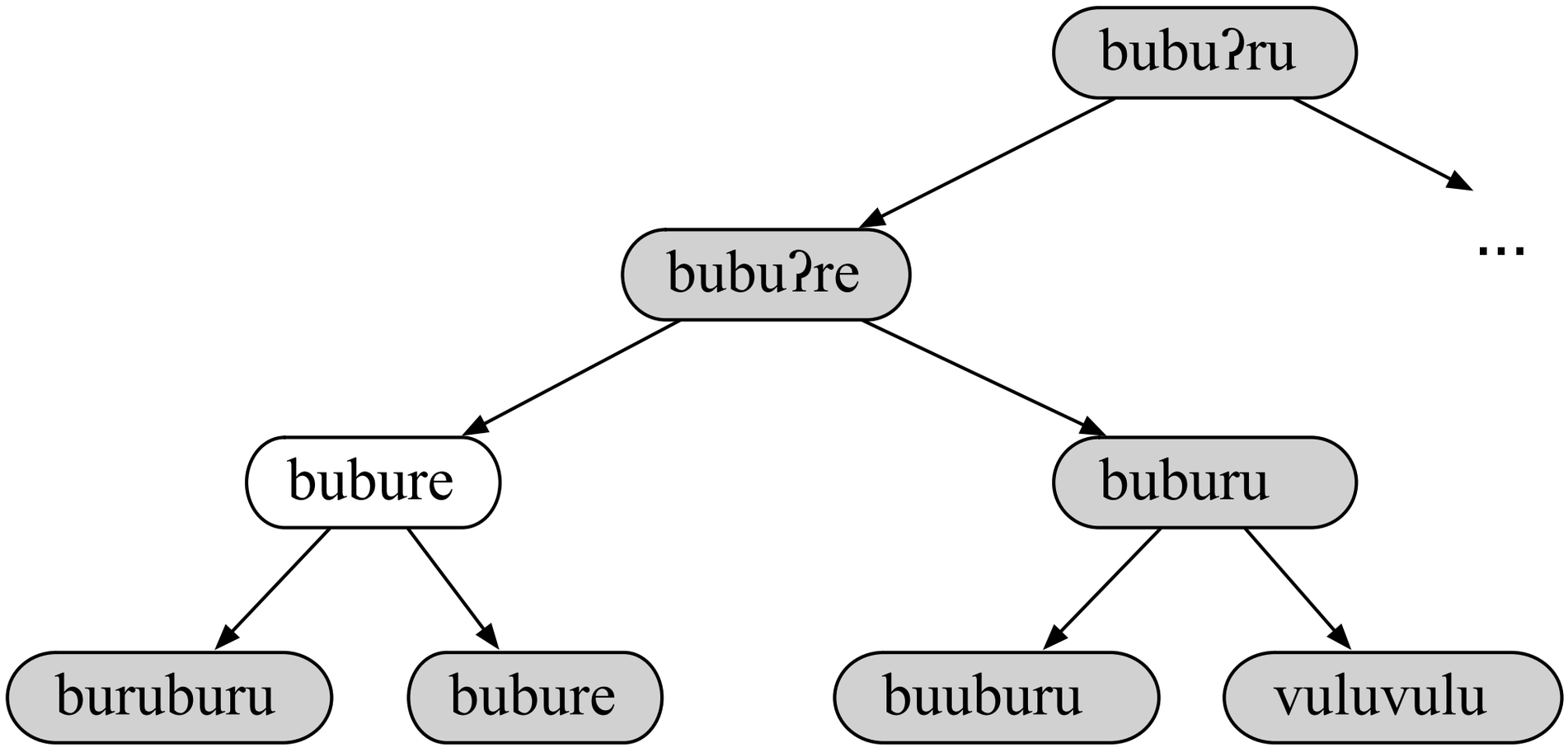
Gibbs sampler



‘grass’

University of California
Berkeley

Gibbs sampler

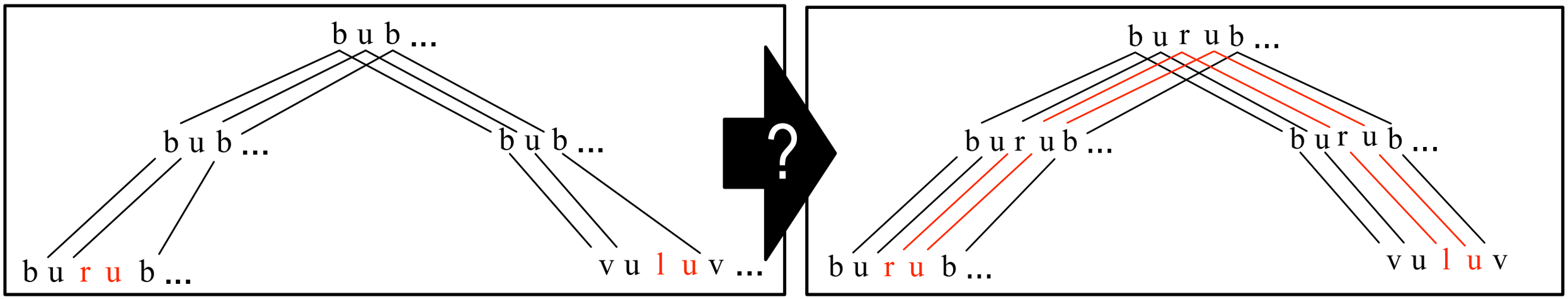


‘grass’

Gibbs sampler

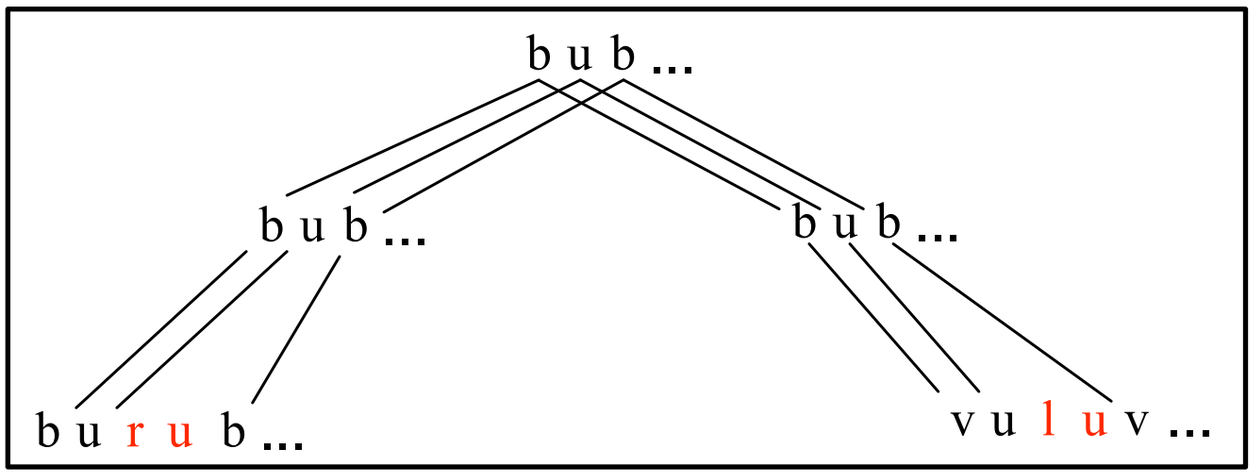
- Problems with the Single Gibbs sampler:
  - Extremely slow in phylogenetic trees with high branching (most linguistic trees)
  - Slow mixing in large trees

Slow mixing

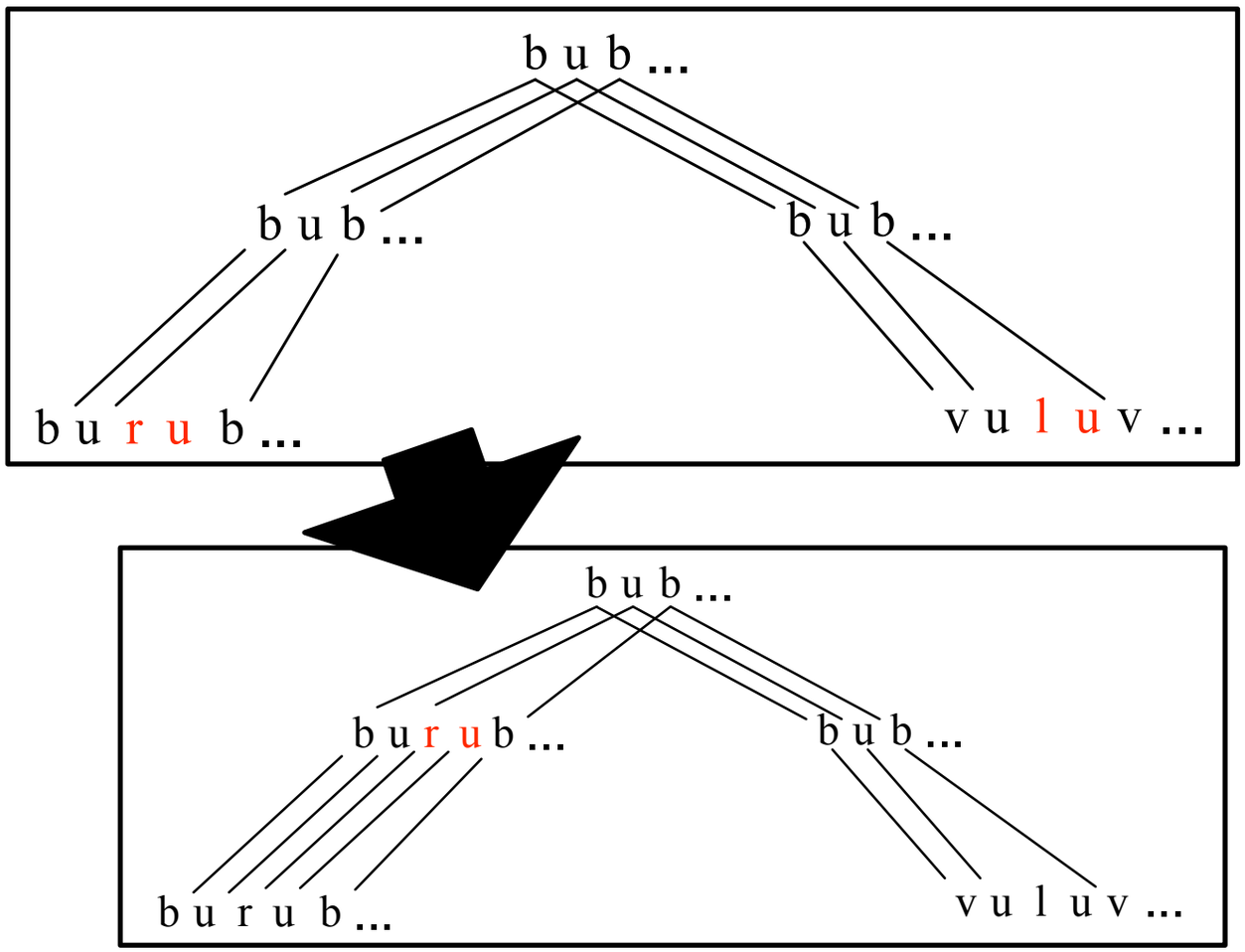


How to jump to a state where the liquids /r/ and /l/ have a common ancestor?

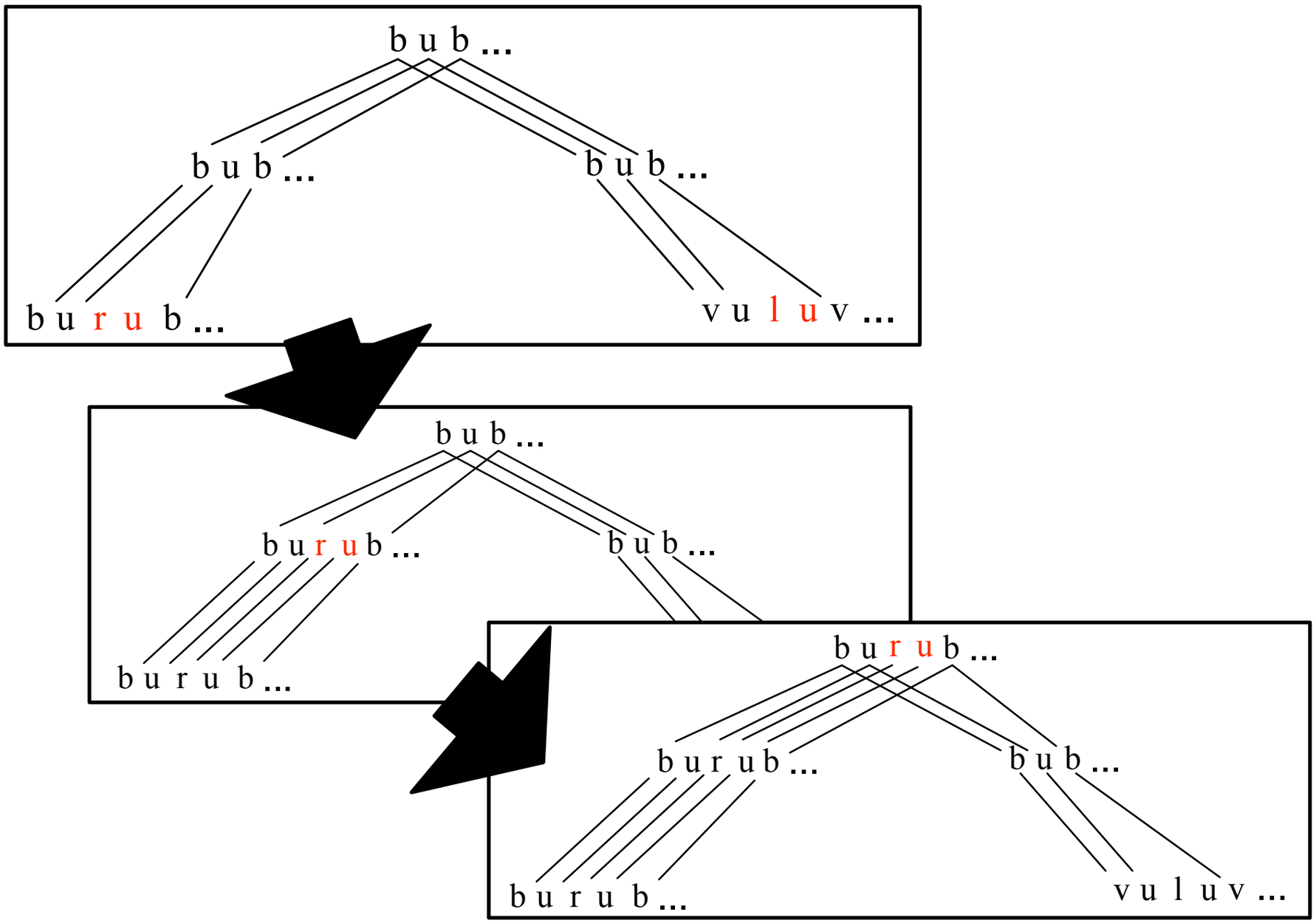
Slow mixing



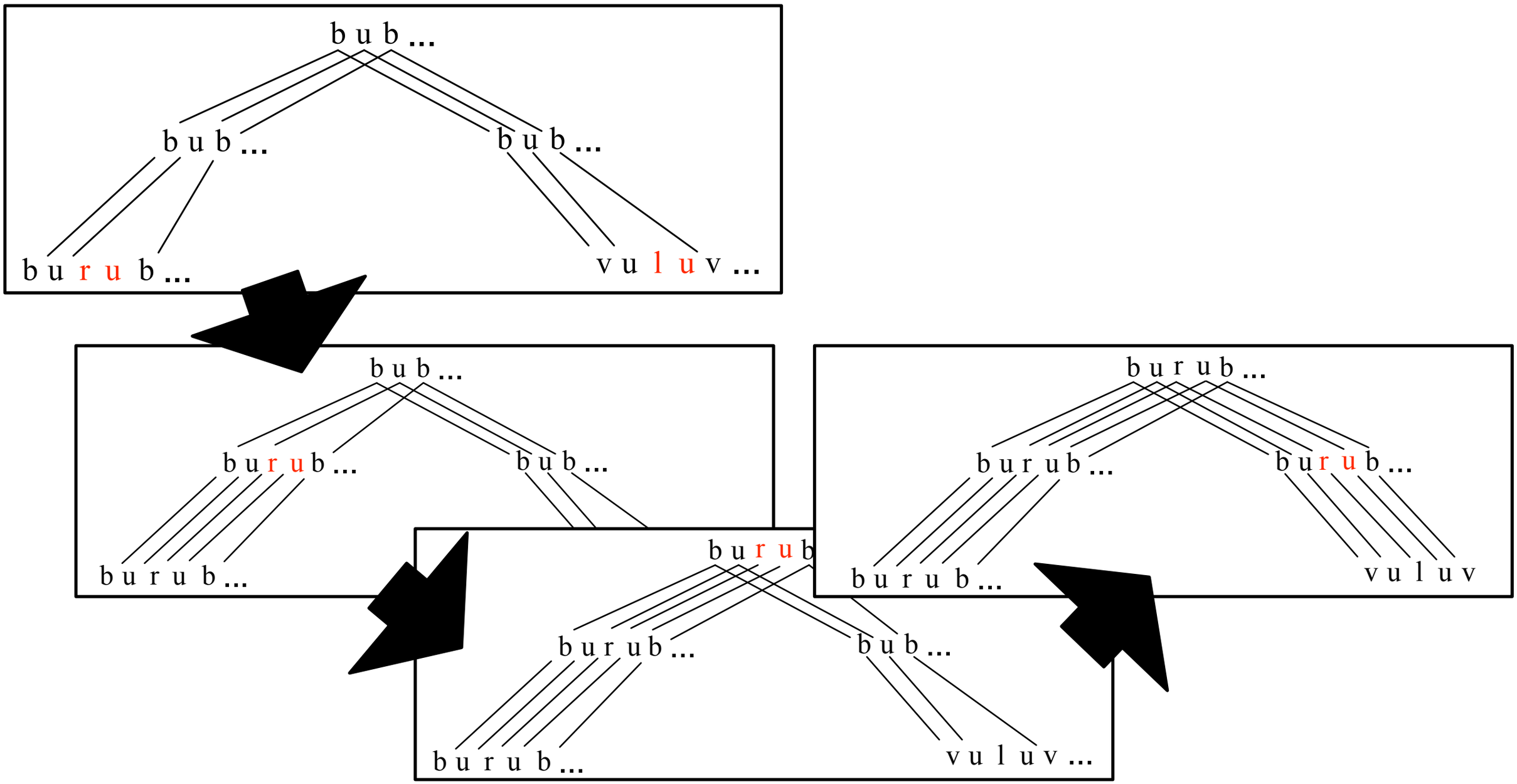
Slow mixing



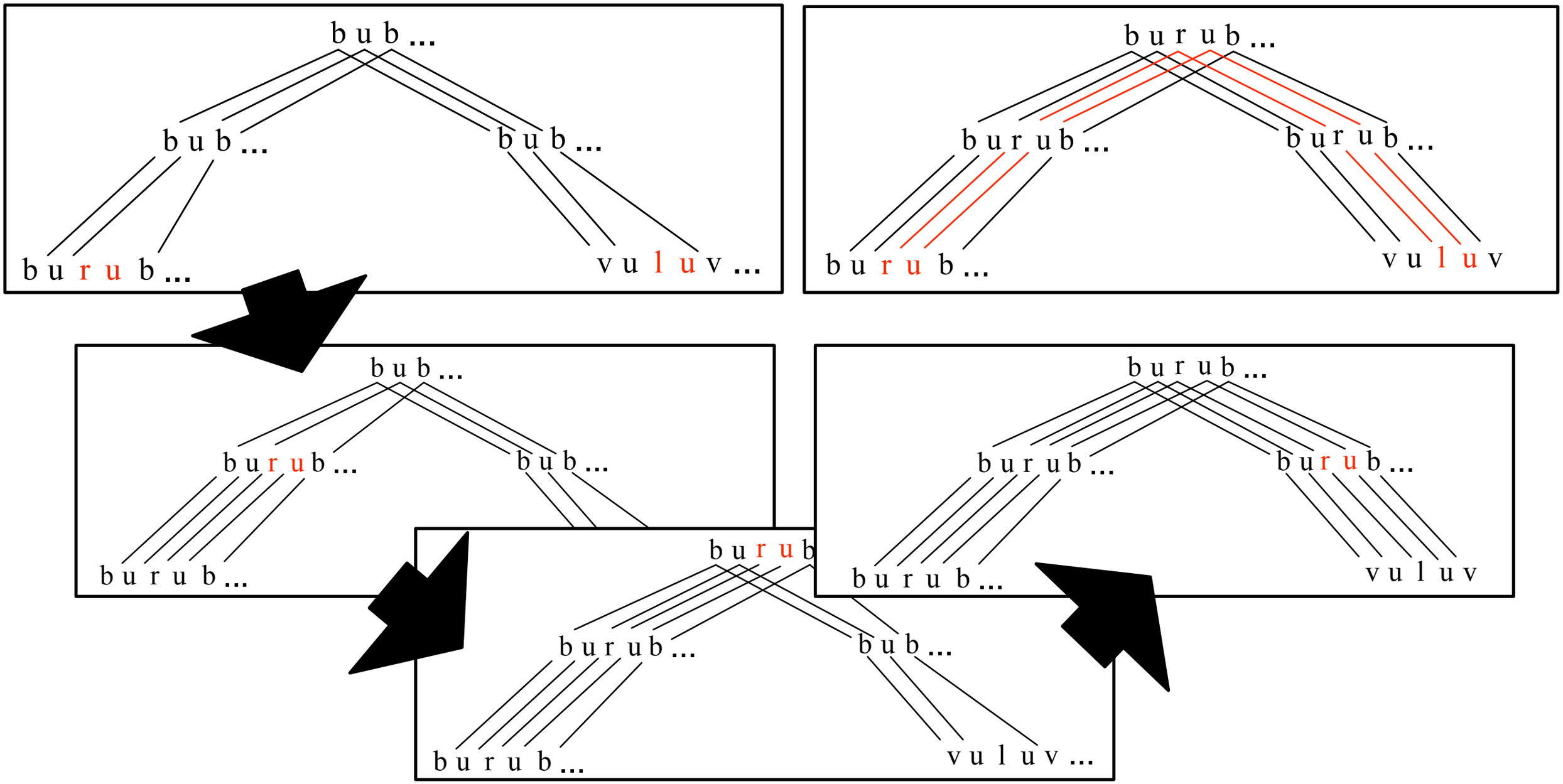
Slow mixing


b u b …
b u b …
b u b …
b u r u b …
v u l u v …
b u b …
b u r u b …
b u b …
b u r u b …
b u r u b …
b u b …
b u r u b …
b u r u b …
b u b …
b u r u b …
v u l u v …

Slow mixing

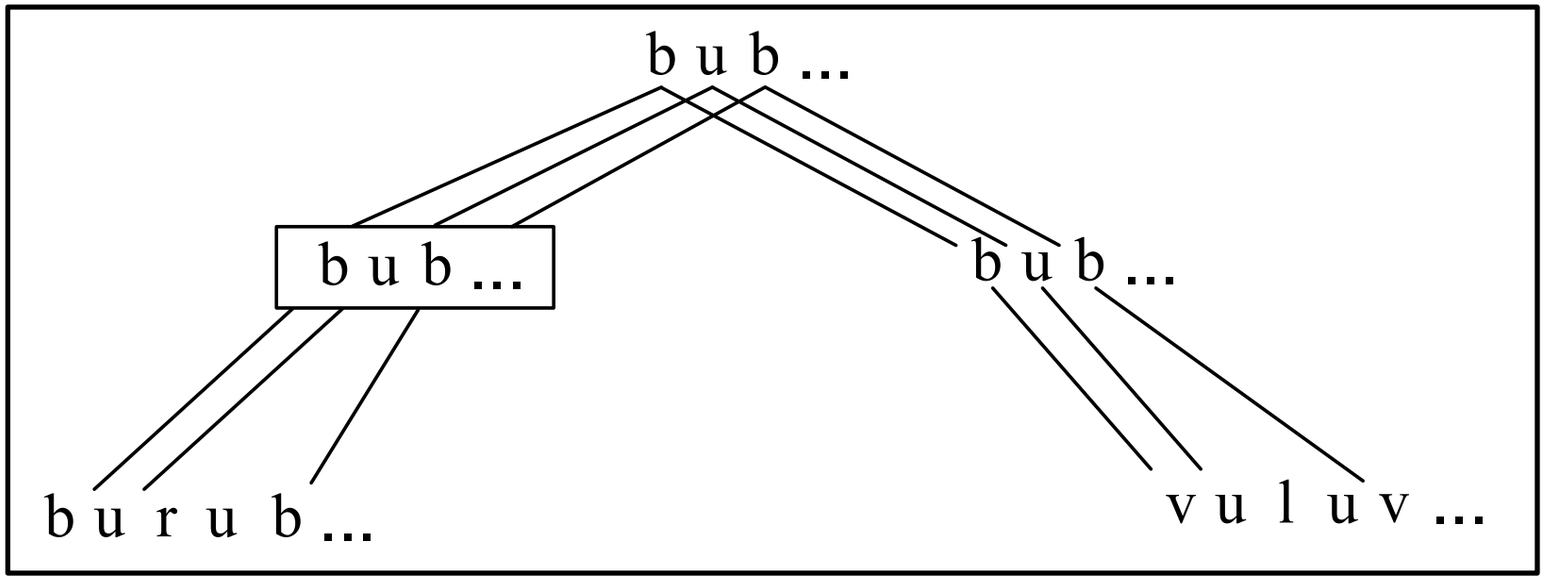

b u b ...
b u b ...
b u b ...
b u r u b ...
v u l u v ...
b u b ...
b u r u b ...
b u b ...
b u r u b ...
b u r u b ...
b u r u b ...
b u b ...
b u r u b ...
v u l u v ...
b u r u b ...
b u r u b ...
b u r u b ...
b u r u b ...
v u l u v

Slow mixing

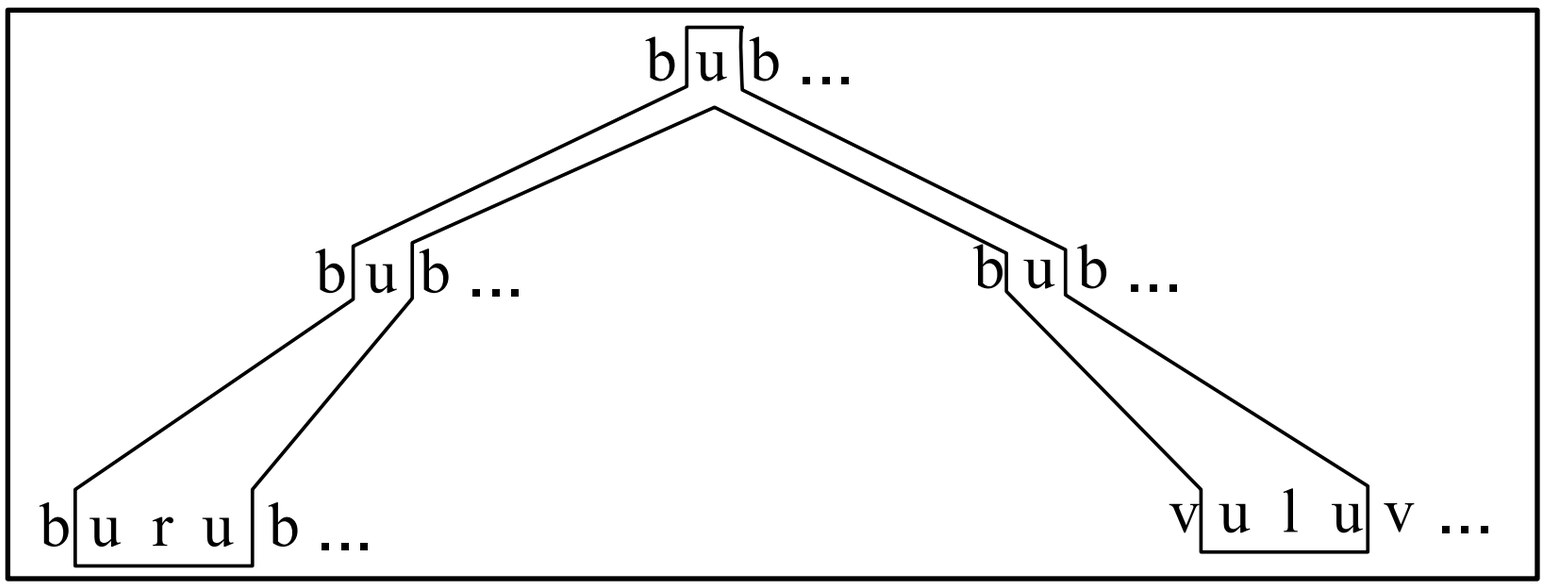

b u b ...
b u b ...
b u b ...
b u r u b ...
v u l u v ...
b u r u b ...
b u r u b ...
b u r u b ...
v u l u v
b u b ...
b u r u b ...
b u b ...
b u r u b ...
b u r u b ...
b u r u b ...
b u r u b ...
b u b ...
b u r u b ...
b u r u b ...
v u l u v ...
b u r u b ...
b u r u b ...
b u r u b ...
v u l u v

Solution: taking vertical slices

SSR

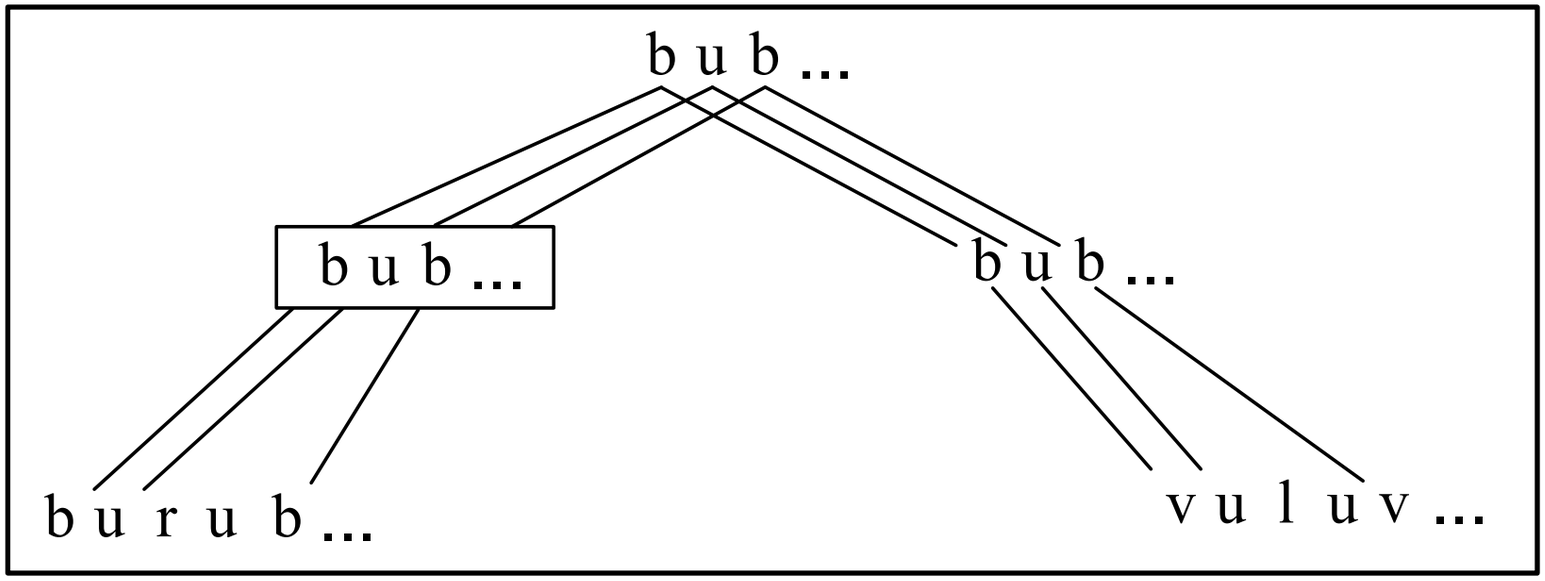


Ancestry resampling

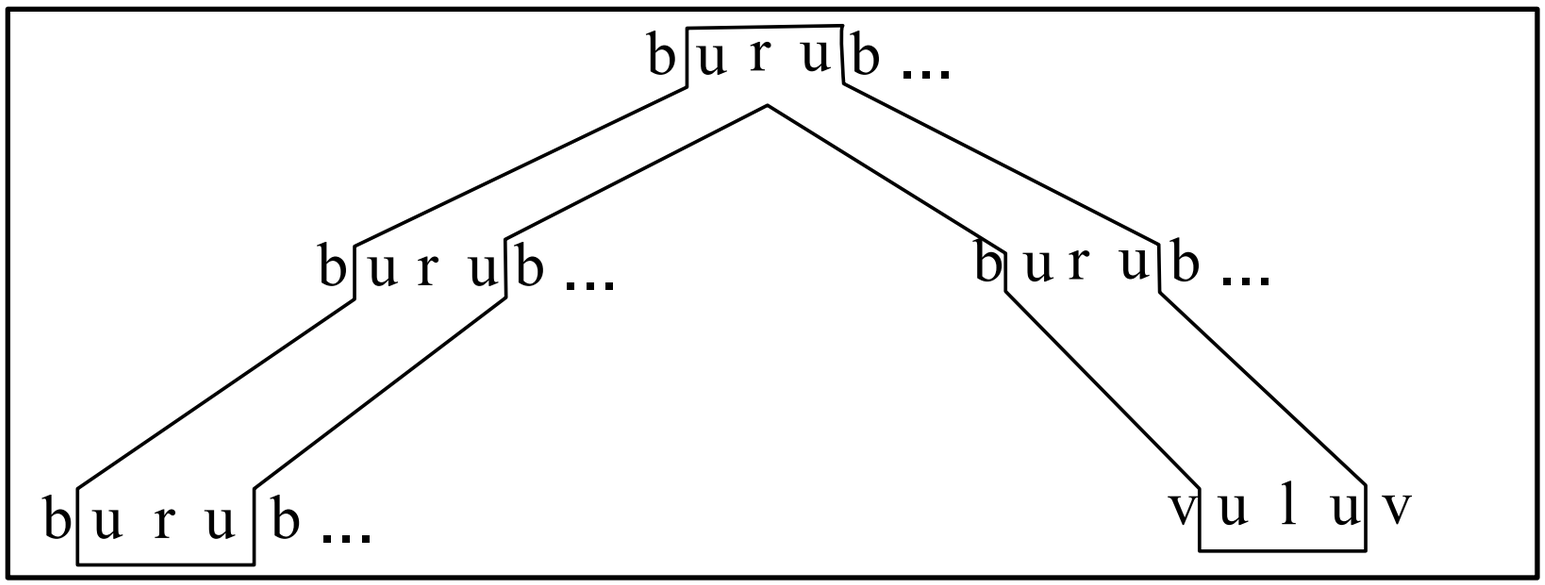


Solution: taking vertical slices

SSR



Ancestry resampling



Experiments

Comparison to other methods

- Evaluation: edit distance from a reconstruction made by a linguist (lower is better)

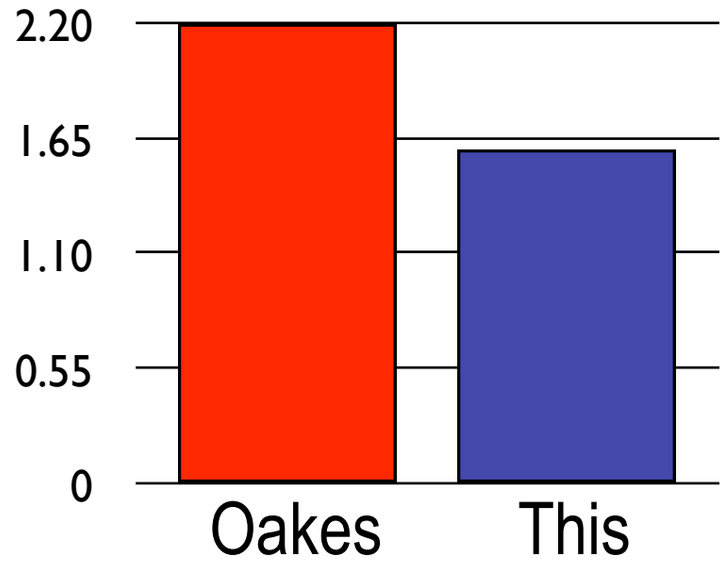
Comparison to other methods

- Evaluation: edit distance from a reconstruction made by a linguist (lower is better)

- Oakes 2000

- Uses exact inference and deterministic rules

- Reconstruction of Proto-Malayo-Javanic (recontructed in Nothefer 1975)

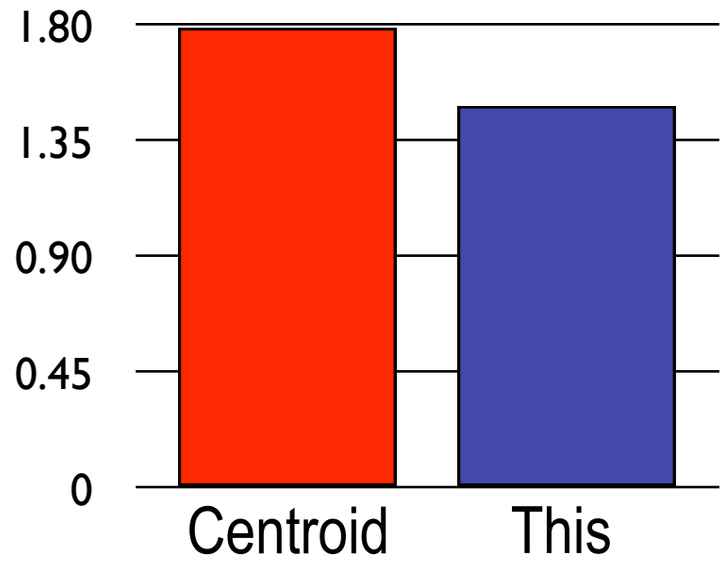


Comparison in large phylogenies

- Centroid: a novel heuristic based on an approximation to the Minimum Bayes risk

  - Reconstruction of Proto-Oceanic (reconstructed in Blust 1993)

  - Both algorithms use 64 modern languages



Back to the initial puzzle

- Can we harness more modern languages to improve reconstructions?

Back to the initial puzzle

- Can we harness more modern languages to improve reconstructions?

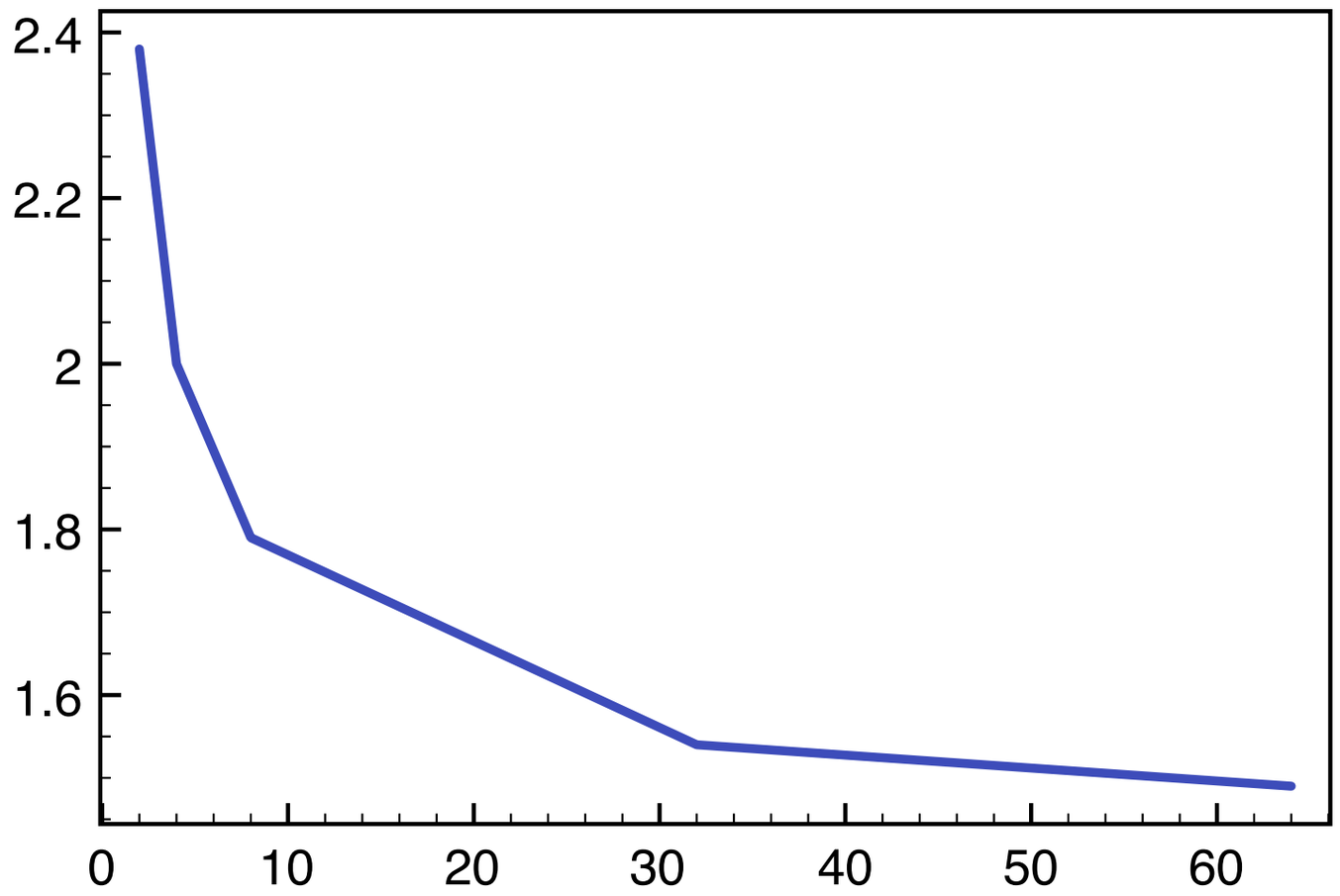
- Using previous model (NIPS 2008): NO (!)

  - No sharing across branches

Back to the initial puzzle

Performance of our model:

Mean edit distance to Blust's reconstruction



Number of modern languages:
close to POc ➔ far (less useful)

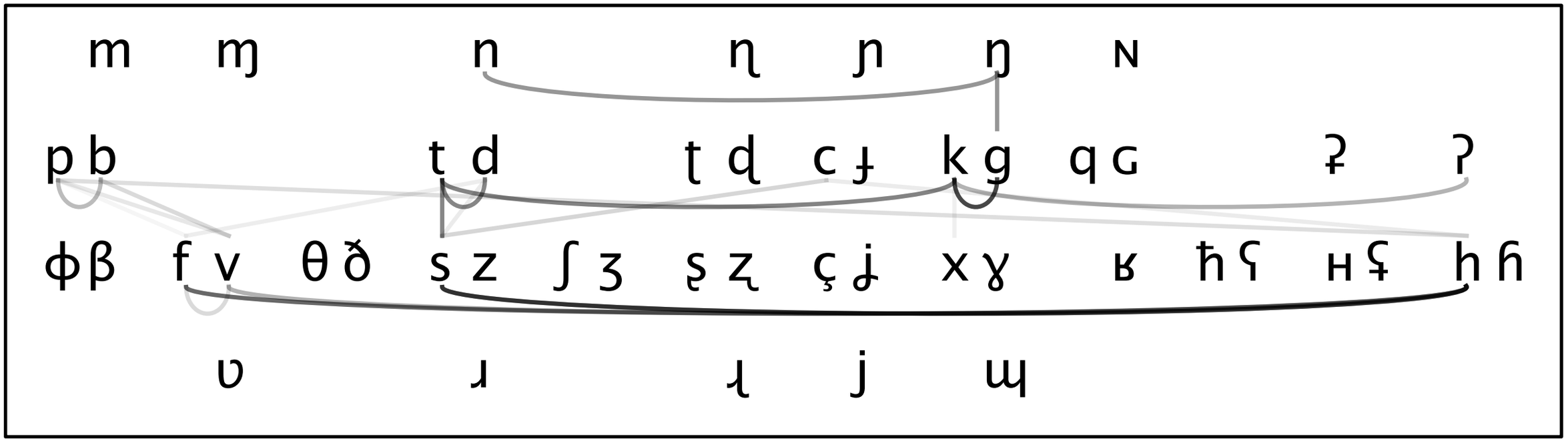
Visualization of learned universals

- For each pair of phonemes, there is a link with grayscale value proportional to the weight of that transition

- Organized in the shape of a IPA chart for convenience

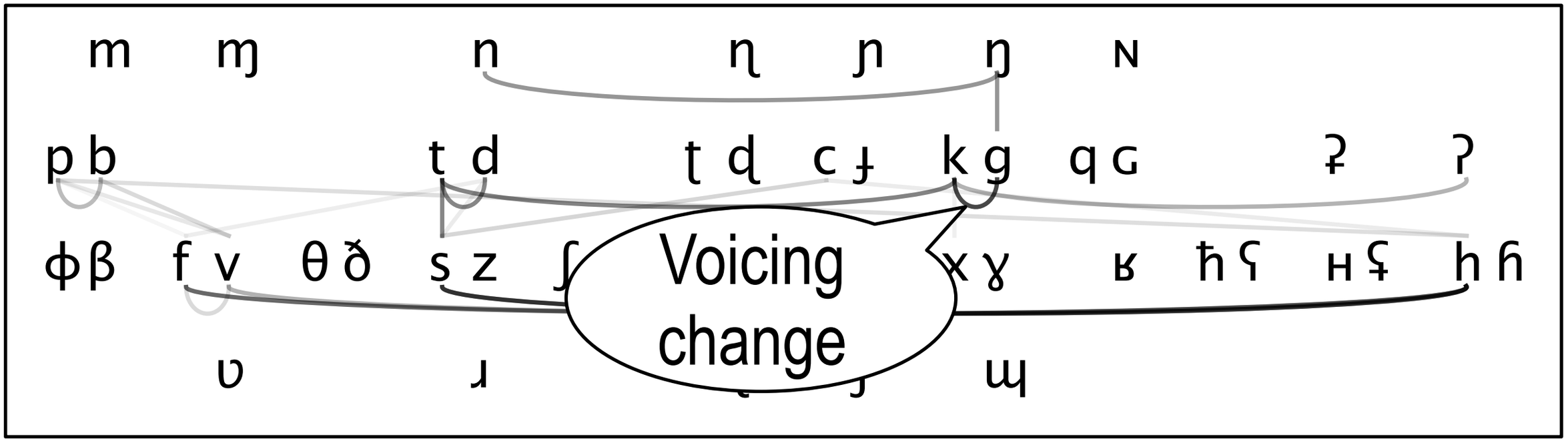
Manner \ Place of articulation																				
	m		ɱ				n				ɳ		ɲ		ŋ		ɴ			
	p	b					t	d			ʈ	ɖ	c	ɟ	k	g	q	ɢ		ʡ		ʔ
	ɸ	β	f	v	θ	ð	s	z	ʃ	ʒ	ʂ	ʐ	ç	ʝ	x	ɣ		ʁ	ħ	ʕ	ʜ	ʢ	h	ɦ
				ʋ				ɹ				ɻ		j		ɰ								

Visualization of learned universals



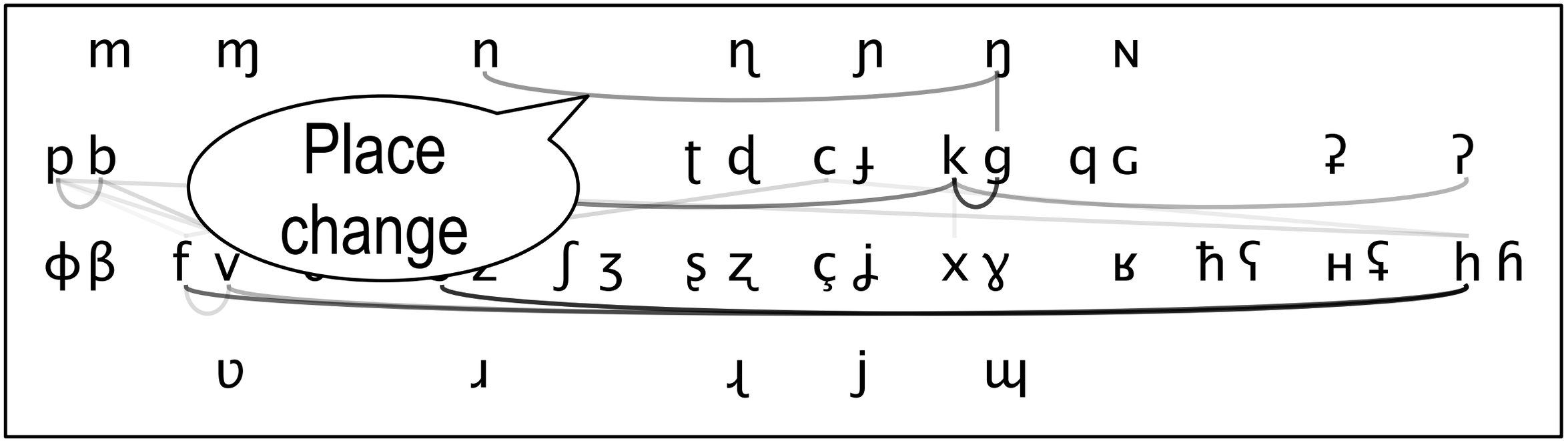
*The model did *not* have features encoding natural classes

Visualization of learned universals



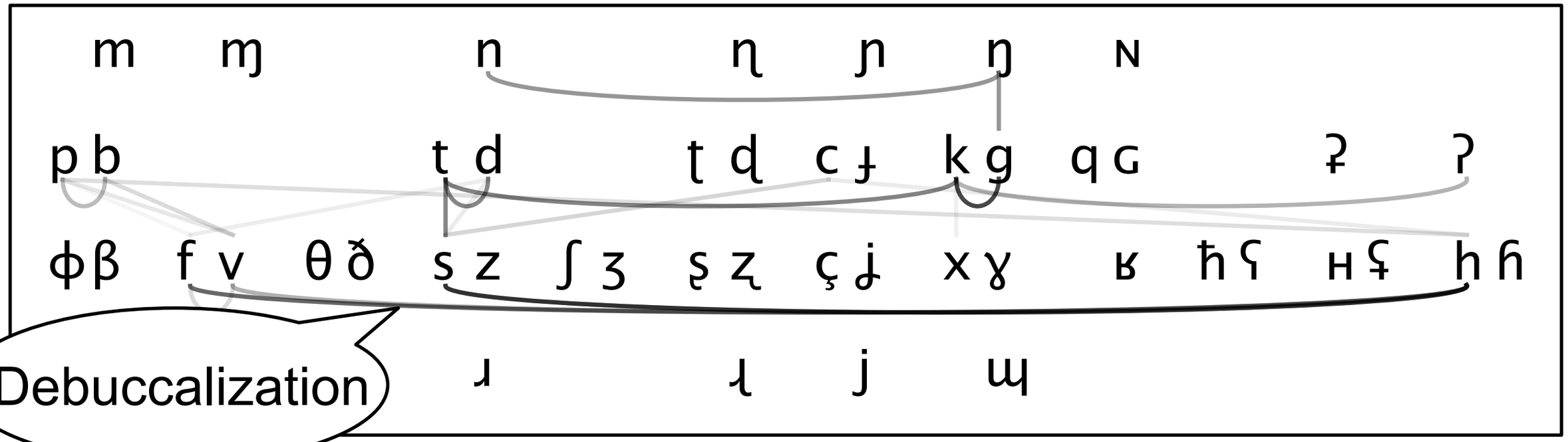
University of California
CALNLP
Berkeley

Visualization of learned universals



University of California
C A L
N L P
Berkeley

Visualization of learned universals



Conclusion

- We proposed three improvements
  - Markedness of internal reconstructions
  - Cross-linguistic universals
  - Using a new inference method to scale up

- Results:
  - We outperform previous approaches
  - We show that using more languages improves reconstructions

- Current work: using the model to attack open questions in historical linguistics

Thank you!

Acknowledgments

- Simon Greenhill, Robert Blust and Russell Gray for sharing their Austronesian dataset

- Michael Oakes for sharing his dataset and results

- Anna Rafferty and our anonymous reviewers for their comments

- Research funded by NSERC and NSF BCS-0631518