Why do languages have the sounds they do?

Insights from infant experiments and phylogenetic modeling

Muhammad Rehan (rehanmuh@g.ucla.edu)
Stacks Xchange: Field Notes, DataX Forum

University of California, Los Angeles

April 20, 2026

Overview

Joint work

(a) Megha Sundara, Department of Linguistics

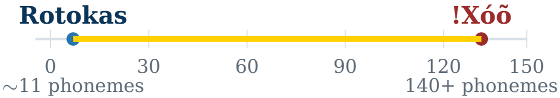
(b) David Goldstein, Department of Linguistics

Perception and phonological typology

DataX Forum

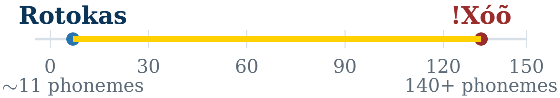
The scale of phonological diversity

Any model of phonological typology must explain both stability and extreme variation in the make-up of phoneme inventories



The scale of phonological diversity

Any model of phonological typology must explain both stability and extreme variation in the make-up of phoneme inventories

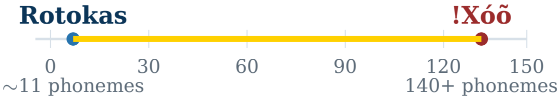


- inventories vary enormously in size

Sources: Maddieson, 1984; Moran and McCloy, 2019; Blasi et al., 2019

The scale of phonological diversity

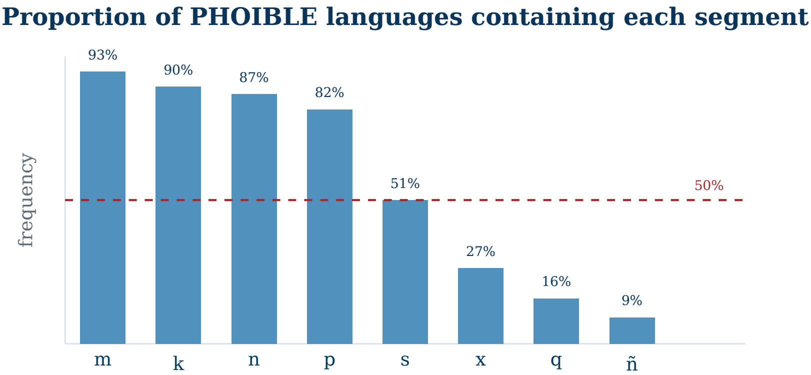
Any model of phonological typology must explain both stability and extreme variation in the make-up of phoneme inventories



- inventories vary enormously in size
- some segments are widespread; others rare and varying in stability

Sources: Maddieson, 1984; Moran and McCloy, 2019; Blasi et al., 2019

Cross-linguistic frequency



Sources: Moran and McCloy, 2019; Moran, Lester, and Eitan Grossman, 2021; Maddieson, 1984

Explaining phonological typology

Research Question (joint work w. Megha Sundara)

Why are some phonological contrasts more common than others?

Explaining phonological typology

Research Question (joint work w. Megha Sundara)

Why are some phonological contrasts more common than others?

- Some explanations:

Sources: Schwartz et al. (1997), Stevens (1989), Nikolaev and E. Grossman (2020), Flemming (2004), Lindblom (1986), and McGahay (2025)

Explaining phonological typology

Research Question (joint work w. Megha Sundara)

Why are some phonological contrasts more common than others?

- Some explanations:
  - Perceptual salience/Focalization

Sources: Schwartz et al. (1997), Stevens (1989), Nikolaev and E. Grossman (2020), Flemming (2004), Lindblom (1986), and McGahay (2025)

Explaining phonological typology

Research Question (joint work w. Megha Sundara)

Why are some phonological contrasts more common than others?

- Some explanations:
  - Perceptual salience/Focalization
  - Confusion minimization/Dispersion

Sources: Schwartz et al. (1997), Stevens (1989), Nikolaev and E. Grossman (2020), Flemming (2004), Lindblom (1986), and McGahay (2025)

Explaining phonological typology

Research Question (joint work w. Megha Sundara)

Why are some phonological contrasts more common than others?

- Some explanations:
  - Perceptual salience/Focalization
  - Confusion minimization/Dispersion
  - Feature economy

Sources: Schwartz et al. (1997), Stevens (1989), Nikolaev and E. Grossman (2020), Flemming (2004), Lindblom (1986), and McGahay (2025)

Explaining phonological typology

Research Question (joint work w. Megha Sundara)

Why are some phonological contrasts more common than others?

- Some explanations:
  - Perceptual salience/Focalization
  - Confusion minimization/Dispersion
  - Feature economy
  - Articulatory effort

Sources: Schwartz et al. (1997), Stevens (1989), Nikolaev and E. Grossman (2020), Flemming (2004), Lindblom (1986), and McGahay (2025)

Explaining phonological typology

Research Question (joint work w. Megha Sundara)

Why are some phonological contrasts more common than others?

- Some explanations:
  - Perceptual salience/Focalization
  - Confusion minimization/Dispersion
  - Feature economy
  - Articulatory effort

Claim

Perceptual salience predicts the frequency of phonological contrasts.

Sources: Schwartz et al. (1997), Stevens (1989), Nikolaev and E. Grossman (2020), Flemming (2004), Lindblom (1986), and McGahay (2025)

Why naïve salience measures are problematic

- Adult discrimination is shaped by:

Why naïve salience measures are problematic

- Adult discrimination is shaped by:
  - lexical frequency

Why naïve salience measures are problematic

- Adult discrimination is shaped by:
  - lexical frequency
  - category warping

Why naïve salience measures are problematic

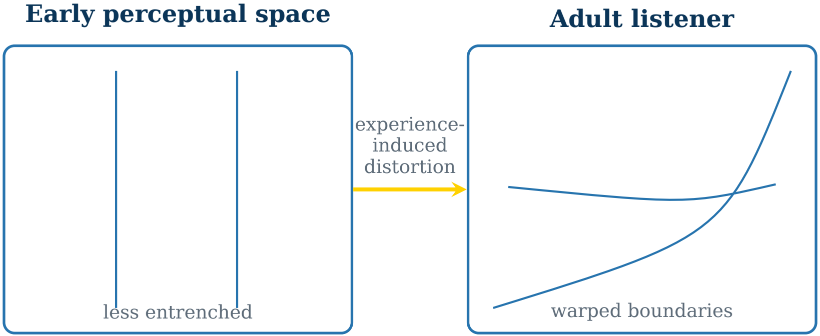
- Adult discrimination is shaped by:
  - lexical frequency
  - category warping
  - native phonological inventories

Why naïve salience measures are problematic

- Adult discrimination is shaped by:
  - lexical frequency
  - category warping
  - native phonological inventories
- These can bias estimates away from early perceptual structure.

Why naïve salience measures are problematic

- Adult discrimination is shaped by:
  - lexical frequency
  - category warping
  - native phonological inventories
- These can bias estimates away from early perceptual structure.



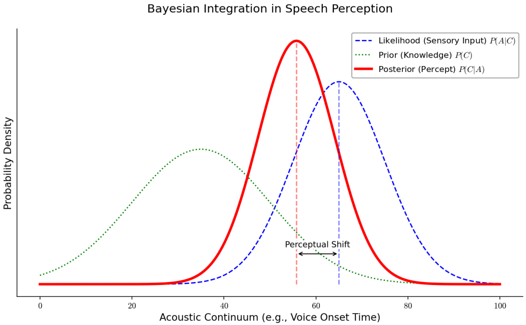
Bayesian speech perception

- Speech perception is **optimal Bayesian inference**.

Sources: Feldman, Griffiths, and Morgan, 2009; Kleinschmidt and Jaeger, 2015; Kleinschmidt, 2019; McGahay, 2026

Bayesian speech perception

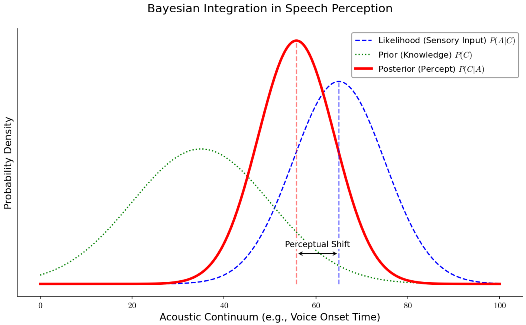
- Speech perception is **optimal Bayesian inference**.



Sources: Feldman, Griffiths, and Morgan, 2009; Kleinschmidt and Jaeger, 2015; Kleinschmidt, 2019; McGahay, 2026

Bayesian speech perception

- Speech perception is **optimal Bayesian inference**.



- In the face of noise and talker variability, listeners rely on **prior biases** from the lexicon and phonotactics.

Sources: Feldman, Griffiths, and Morgan, 2009; Kleinschmidt and Jaeger, 2015; Kleinschmidt, 2019; McGahay, 2026

How to measure innate perceptual salience?

Challenge

How to test the perceptual salience of contrasts without the confounds of language experience?

How to measure innate perceptual salience?

Challenge

How to test the perceptual salience of contrasts without the confounds of language experience?

- Who has no/minimal language experience?

How to measure innate perceptual salience?

Challenge

How to test the perceptual salience of contrasts without the confounds of language experience?

- Who has no/minimal language experience?

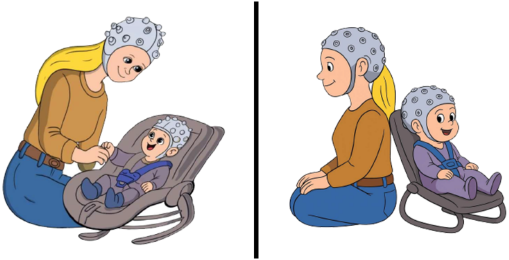


Figure: Infant in a neurophysiological response study (Endevelt-Shapira et al., 2025)

Why are infants special?

- Minimal lexical knowledge

Why are infants special?

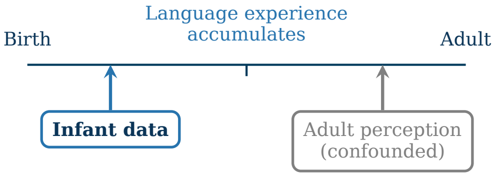
- Minimal lexical knowledge
- Weak or emerging phonological categories

Why are infants special?

- Minimal lexical knowledge
- Weak or emerging phonological categories
- Best available proxy for **early perceptual structure**

Why are infants special?

- Minimal lexical knowledge
- Weak or emerging phonological categories
- Best available proxy for **early perceptual structure**



Goal: estimate salience before the major confounds of lexical and phonological learning dominate discrimination.

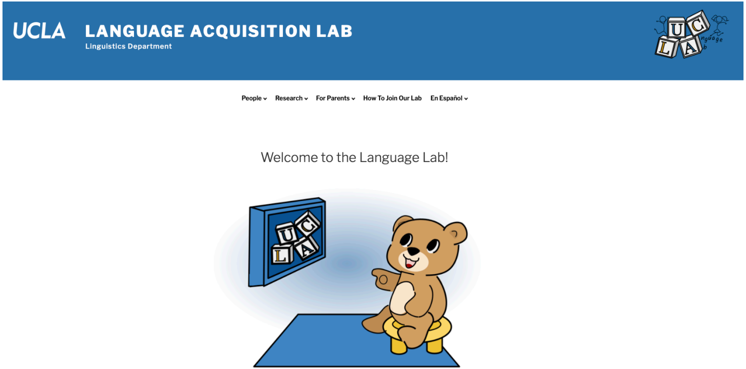
Phonological acquisition in infants

- Let's look at an example of an experimental paradigm called Conditioned Head Turn

Figure: Janet Werker testing an infant's discrimination abilities using the CHT paradigm

Language acquisition lab

Send your infants to the language acquisition lab!



Data

- Start by finding everything ever published on infants' discrimination of phonetic contrasts.

Data

- Start by finding everything ever published on infants’ discrimination of phonetic contrasts.
- 400 studies narrowed down to 175.

Data

- Start by finding everything ever published on infants’ discrimination of phonetic contrasts.
- 400 studies narrowed down to 175.
- 175 studies coded for 40 variables (with the help of 4 RAs: Lily Brawner, Mila Matos, Gabriel Scharff, and Kathryn Cunningham).

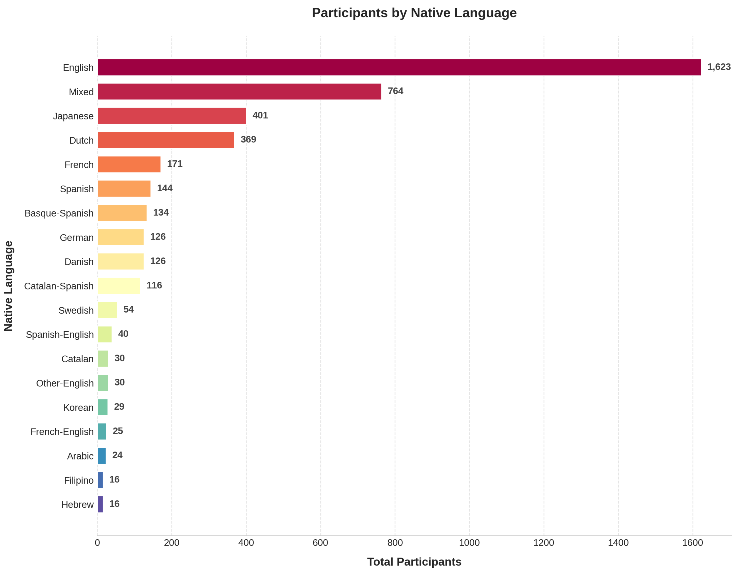
Data

- Start by finding everything ever published on infants' discrimination of phonetic contrasts.
- 400 studies narrowed down to 175.
- 175 studies coded for 40 variables (with the help of 4 RAs: Lily Brawner, Mila Matos, Gabriel Scharff, and Kathryn Cunningham).
- 141 experiments from 61 studies included where the effect of language experience could be ruled out.

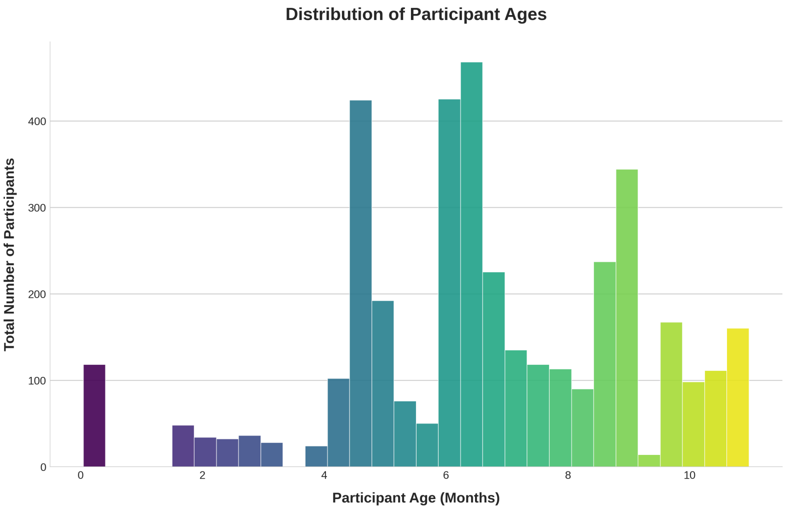
Data

- Start by finding everything ever published on infants' discrimination of phonetic contrasts.
- 400 studies narrowed down to 175.
- 175 studies coded for 40 variables (with the help of 4 RAs: Lily Brawner, Mila Matos, Gabriel Scharff, and Kathryn Cunningham).
- 141 experiments from 61 studies included where the effect of language experience could be ruled out.
- Phoneme frequency data from PHOIBLE.

Participant languages



Participant ages



Research Question and measurable data

Research Question

Are phonological contrasts that are **easier to discriminate** also more **typologically frequent**?

Research Question and measurable data

Research Question

Are phonological contrasts that are **easier to discriminate** also more **typologically frequent**?

- **Discrimination**: $|g|$ (Hedges' g)

Research Question and measurable data

Research Question

Are phonological contrasts that are **easier to discriminate** also more **typologically frequent**?

- **Discrimination**: $|g|$ (Hedges' g)
  - Hedges' g is a measure of effect size that takes sample size into account.

Research Question and measurable data

Research Question

Are phonological contrasts that are **easier to discriminate** also more **typologically frequent**?

- **Discrimination**: $|g|$ (Hedges' g)
  - Hedges' g is a measure of effect size that takes sample size into account.
  - Instead of asking whether the effect of a stimulus is statistically significant or not, we ask how big or small is it.

Research Question and measurable data

Research Question

Are phonological contrasts that are **easier to discriminate** also more **typologically frequent**?

- **Discrimination**: $|g|$ (Hedges' g)
  - Hedges' g is a measure of effect size that takes sample size into account.
  - Instead of asking whether the effect of a stimulus is statistically significant or not, we ask how big or small is it.
- **Stratified Frequency**: stratified proportion of languages

Research Question and measurable data

Research Question

Are phonological contrasts that are **easier to discriminate** also more **typologically frequent**?

- **Discrimination**: $|g|$ (Hedges' g)
  - Hedges' g is a measure of effect size that takes sample size into account.
  - Instead of asking whether the effect of a stimulus is statistically significant or not, we ask how big or small is it.

- **Stratified Frequency**: stratified proportion of languages
  - Persistent issue in typology: spatial and genetic autocorrelation.

Bayesian GLMM modeling

Data & Variables (for literature row $i = 1, \ldots, N$):

Bayesian GLMM modeling

Data & Variables (for literature row $i = 1, \ldots, N$):

- y_i: Number of stratified PHOIBLE languages with the contrast

Bayesian GLMM modeling

Data & Variables (for literature row $i = 1, \ldots, N$):

- y_i: Number of stratified PHOIBLE languages with the contrast
- n_i: Total PHOIBLE languages considered

Bayesian GLMM modeling

Data & Variables (for literature row $i = 1, \ldots, N$):

- y_i: Number of stratified PHOIBLE languages with the contrast
- n_i: Total PHOIBLE languages considered
- $g_i = |\text{Hedges' } g_i|$

Bayesian GLMM modeling

Data & Variables (for literature row $i = 1, \ldots, N$):

- y_i: Number of stratified PHOIBLE languages with the contrast
- n_i: Total PHOIBLE languages considered
- $g_i = |\text{Hedges' } g_i|$
- $s[i] \in \{1, \ldots, S\}$: Study index

Bayesian GLMM modeling

Data & Variables (for literature row $i = 1, \ldots, N$):

- y_i: Number of stratified PHOIBLE languages with the contrast
- n_i: Total PHOIBLE languages considered
- $g_i = |\text{Hedges' } g_i|$
- $s[i] \in \{1, \ldots, S\}$: Study index

Model Structure:

$$\begin{aligned} y_i &\sim \text{Binomial}(n_i, p_i) \\ \text{logit}(p_i) &= \beta_0 + \beta_1 g_i + u_{s[i]} \\ u_s &\sim \mathcal{N}(0, \sigma^2_{\text{study}}) \end{aligned}$$

Results

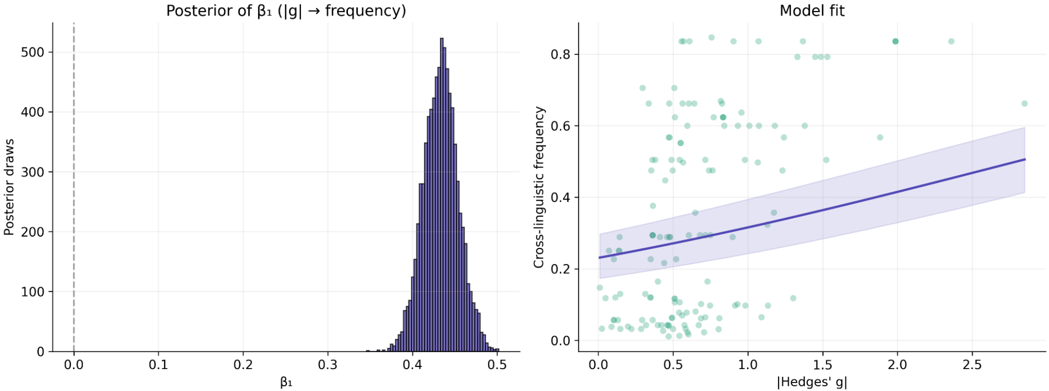


Figure: Infants' discrimination of a phonetic contrast has a statistically significant positive effect on the frequency of that contrast.

Interim Summary

- Languages show massive variation in the size and the composition of their phoneme inventories.

Interim Summary

- Languages show massive variation in the size and the composition of their phoneme inventories.
- Some scholars attribute this variation to perceptual salience.

Interim Summary

- Languages show massive variation in the size and the composition of their phoneme inventories.
- Some scholars attribute this variation to perceptual salience.
- Adults are not reliable measurement instruments for perceptual salience.

Interim Summary

- Languages show massive variation in the size and the composition of their phoneme inventories.
- Some scholars attribute this variation to perceptual salience.
- Adults are not reliable measurement instruments for perceptual salience.
- Young infants' discrimination abilities plausibly offer a reliable estimate of perceptual salience.

Interim Summary

- Languages show massive variation in the size and the composition of their phoneme inventories.
- Some scholars attribute this variation to perceptual salience.
- Adults are not reliable measurement instruments for perceptual salience.
- Young infants’ discrimination abilities plausibly offer a reliable estimate of perceptual salience.
- Infant discrimination of phonetic contrasts predicts their typological frequencies, linking perception with typology.

Discussion: From Perception to Evolution

- **Part 1: Perception & Typology**
  - Infant discrimination ≈ perceptual salience.

Discussion: From Perception to Evolution

- **Part 1: Perception & Typology**
  - Infant discrimination $\approx$ perceptual salience.
  - Salience predicts typological contrast frequency.

Discussion: From Perception to Evolution

- **Part 1: Perception & Typology**
  - Infant discrimination ≈ perceptual salience.
  - Salience predicts typological contrast frequency.
- **The Missing Piece: Historical Sound Change**
  - Current typological distributions are *not* stationary.

Discussion: From Perception to Evolution

- **Part 1: Perception & Typology**
  - Infant discrimination ≈ perceptual salience.
  - Salience predicts typological contrast frequency.
- **The Missing Piece: Historical Sound Change**
  - Current typological distributions are *not* stationary.
  - We must account for regular, directional segment evolution.

Discussion: From Perception to Evolution

- **Part 1: Perception & Typology**
  - Infant discrimination ≈ perceptual salience.
  - Salience predicts typological contrast frequency.
- **The Missing Piece: Historical Sound Change**
  - Current typological distributions are *not* stationary.
  - We must account for regular, directional segment evolution.

→ Up Next: Part 2

Dynamic modeling of phonological typology

Dyanmic modeling of phonological typology

DataX Forum

Sound systems are in constant flux

- Phoneme inventories are not static; they are the products of continuous diachronic change.

Sound systems are in constant flux

- Phoneme inventories are not static; they are the products of continuous diachronic change.

$$
\begin{tikzcd}[column sep=small]
& \text{Latin } p \ (\textit{pater}) & \\
\textbf{PIE *p} \arrow[ur] \arrow[dr] & & \\
& \textbf{PGmc *f} \arrow[r] & \text{English } f \ (\textit{father})
\end{tikzcd}
$$

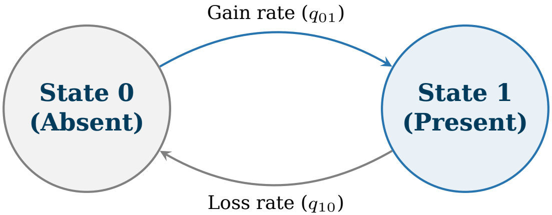
$$\text{OE } k \ (\textit{knīht}) \longrightarrow \text{ME } k \ (\textit{knight}) \xrightarrow{\text{loss}} \text{ModE } \emptyset \ (\textit{night})$$

Modeling the trajectories of change

- Can we quantitatively model the rates at which languages **acquire** and **discard** sounds over time?

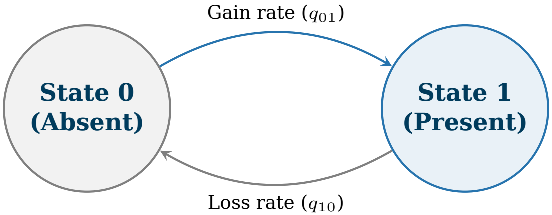
Modeling the trajectories of change

- Can we quantitatively model the rates at which languages **acquire** and **discard** sounds over time?



Modeling the trajectories of change

- Can we quantitatively model the rates at which languages **acquire** and **discard** sounds over time?



- We treat phonological evolution as a continuous-time Markov process.

The Optimization Hypothesis

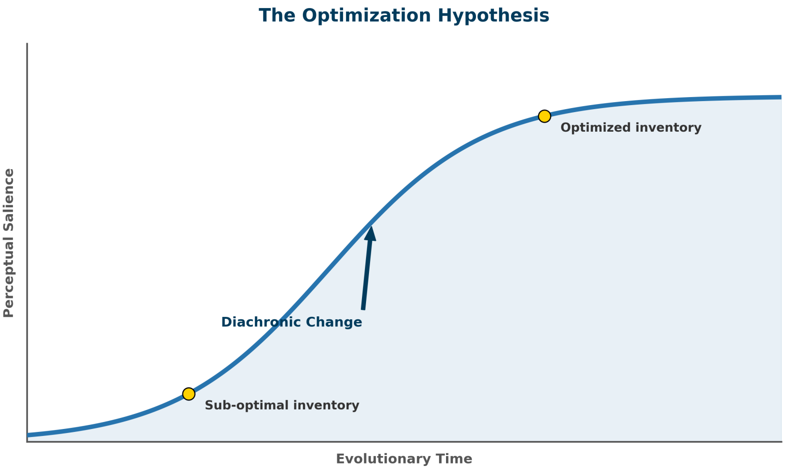


Figure: Hypothesis: Phoneme inventories change to increase perceptual salience

Sound change

What we know

- sound change is often regular
- conditioned by phonetic environments
- some changes recur across families

Sound change

What we know

- sound change is often regular
- conditioned by phonetic environments
- some changes recur across families

What we do not know

- how fast specific segments change
- how context changes those rates
- how much variation there is in change across language families

Sound change

What we know

- sound change is often regular
- conditioned by phonetic environments
- some changes recur across families

What we do not know

- how fast specific segments change
- how context changes those rates
- how much variation there is in change across language families

We lack a quantitative and principled model for modeling the evolution of segments over time on family trees.

Sources: Maddieson, 1984; Hock, 1991; Labov, 2020; Garrett, 2015

Phonetic grounding

Why some changes recur across lineages

Theoretical Consensus

Change is driven by misperception (Ohala, 1993), leading to recurrent patterns (Blevins, 2004) that are inherently phonetically biased and directional (Garrett and Johnson, 2013).

Sources: Ohala, 1981; Ohala, 1993; Blevins, 2004; Garrett and Johnson, 2013; Kirby, 2013

Phonetic grounding

Why some changes recur across lineages

Phonetic variation → **Listener (mis)parsing** → **Phonologization**

Theoretical Consensus

Change is driven by misperception (Ohala, 1993), leading to recurrent patterns (Blevins, 2004) that are inherently phonetically biased and directional (Garrett and Johnson, 2013).

Modeling implication

Rates of segmental change must be structured and not arbitrary.

Sources: Ohala, 1981; Ohala, 1993; Blevins, 2004; Garrett and Johnson, 2013; Kirby, 2013

What a probabilistic model should look like

Regularity & pooling

Same local context ⇒ same change parameters across lexical items.

Sources: Hayes & Wilson 2008; Hayes & Steriade 2004; Hayes & White 2015; Zuraw 2010; Moore-Cantwell 2016

What a probabilistic model should look like

Regularity & pooling

Same local context $\Rightarrow$ same change parameters across lexical items.

Gradient exceptionality

Skew, partial productivity, and near-categorical trends.

Sources: Hayes & Wilson 2008; Hayes & Steriade 2004; Hayes & White 2015; Zuraw 2010; Moore-Cantwell 2016

What a probabilistic model should look like

Regularity & pooling

Same local context ⇒ same change parameters across lexical items.

Gradient exceptionality

Skew, partial productivity, and near-categorical trends.

Naturalness

*/p/ > [b] should look more plausible than */p/ > [r]. Asymmetries should enter the prior.

Sources: Hayes & Wilson 2008; Hayes & Steriade 2004; Hayes & White 2015; Zuraw 2010; Moore-Cantwell 2016

What a probabilistic model should look like

Regularity & pooling

Same local context ⇒ same change parameters across lexical items.

Gradient exceptionality

Skew, partial productivity, and near-categorical trends.

Naturalness

*/p/ > [b] should look more plausible than */p/ > [r]. Asymmetries should enter the prior.

Parameter Learnability

Restricted enough that finite cognate data can identify something interpretable.

Sources: Hayes & Wilson 2008; Hayes & Steriade 2004; Hayes & White 2015; Zuraw 2010; Moore-Cantwell 2016

Previous computational work in this domain

Model / Study	Context	Rate Var.	Naturalness	Insertions	Deletions
Bouchard-Côté et al. (2013)	✓	✓	✗	✓	✓
Hruschka et al. (2015)	✓	✗	✗	✗	✗
Moran et al. (2021)	✗	✓	✗	✗	✗
He et al. (2024)	✓	✗	✗	✓	✓
Cathcart & Wandl (2020)	✓	✓	✓	✗	✗
Auderset (2025)	✗	✓	✗	✗	✗
Goldstein et al. (2026)	✗	✓	✓*	✓	✓
Desired	✓	✓	✓	✓	✓

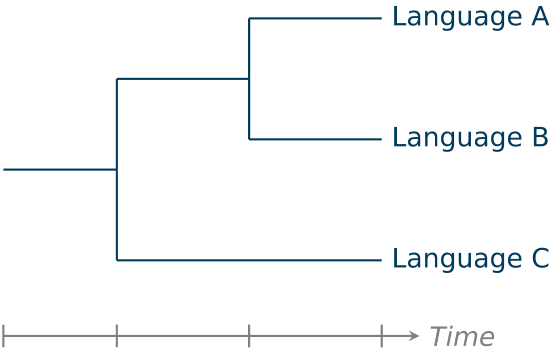
None of the former computational models of sound change model all aspects of local segmental change.

Sources: Greenhill et al., 2017; Bouchard-Côté et al., 2013; Hruschka et al., 2015; Moran, Lester, and Eitan Grossman, 2021; He and Zhao, 2024; Auderset, 2025; Cathcart and Wandl, 2020, Goldstein et al., 2026

This section

Goal: Build a probabilistic model that infers *rates of segmental change* while encoding *phonetic similarity, directional biases, local context*, and *regularity*.

Input Data 1: Phylogeny

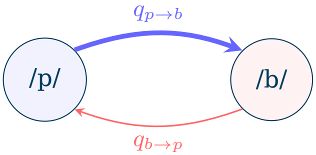


Input Data 2: Aligned Cognates

	Site 1	Site 2	Site 3	Site 4
L1	p	a	t	e
L2	b	a	d	e
L3	f	a	--	e

Homologous Column

Input Data 3: Segment Features



[+cons]	[+cons]
[-son]	[-son]
[-cont]	[-cont]
[+lab]	[+lab]
[-voice]	**[+voice]**
[-nas]	[-nas]

State space and edit types

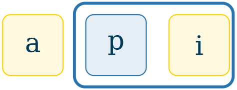
Illustrative inventory

$\{p, t, k, b, d, g, f, s, h, m, n, l, r, a, i, u, \varnothing\}$ 16 segments
1 gap state __
1 boundary marker #

Three transition types

$x > y$	substitution
$\varnothing > y$	insertion
$x > \varnothing$	deletion

Local context as part of the rate structure



target + neighbors = context

- edge vs. medial position
- neighboring features from parent states

Feature engineering

PanPhon-style feature vectors as covariates

Representation

Segment $x \rightarrow$ feature vector $\mathbf{f}(x)$

- **Consonantal:** place, manner, voicing
- **Vocalic:** height, backness, rounding

Feature engineering

PanPhon-style feature vectors as covariates

Representation

Segment $x \rightarrow$ feature vector $\mathbf{f}(x)$

- **Consonantal:** place, manner, voicing
- **Vocalic:** height, backness, rounding

Features as Predictors

Features act as covariates to estimate weight vector $\mathbf{w}$:

- **Symmetric:** weights on shared/differing features
- **Directional:** weights on specific transitions (e.g., $+ \rightarrow -$)

Sources: Mortensen et al., 2016; Ohala, 1993; Blevins, 2004; Garrett and Johnson, 2013

Feature Vectors in Practice

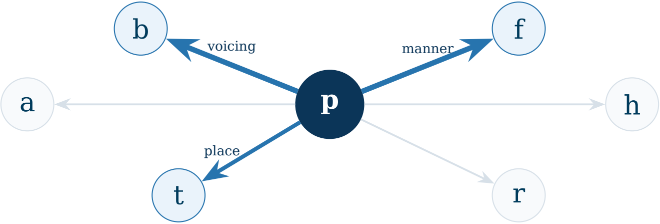
/m/

$$\mathbf{f}(\text{/m/}) \in \{1, 0, -1\}^n$$

Covariates and Weights

Feature	Value	Weight
Syllabic (syl)	-1	w_{syl}
Consonantal (cons)	1	w_{cons}
Sonorant (son)	1	w_{son}
Coronal (cor)	-1	w_{cor}
Labial (lab)	1	w_{lab}
Nasal (nas)	1	w_{nas}
Voiced (voi)	1	w_{voi}

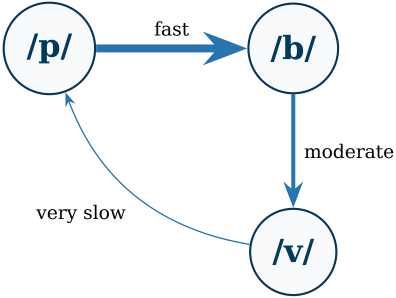
Feature distance shapes transition probability



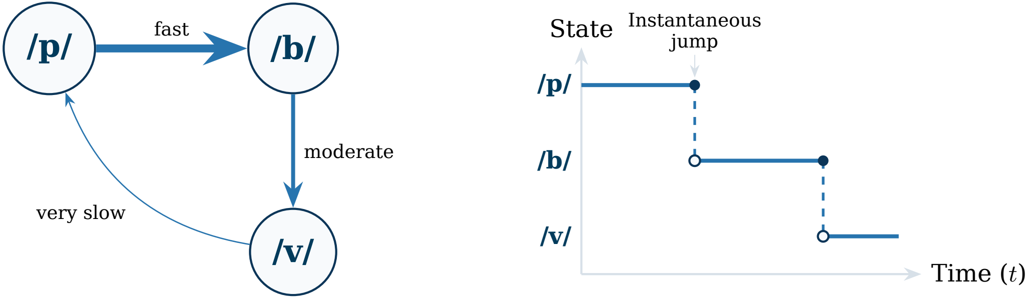
thick = small distance → higher rate

thin = large distance → lower rate

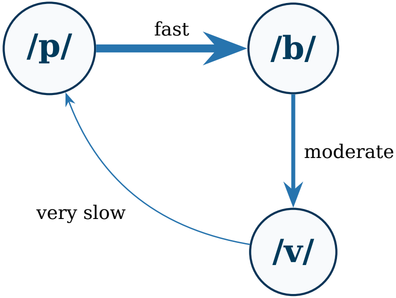
Continuous-Time Markov Chains

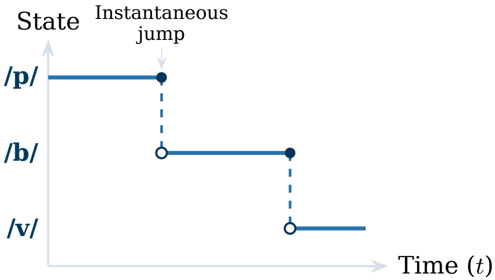


Continuous-Time Markov Chains



Continuous-Time Markov Chains



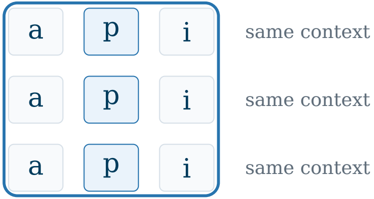


Instantaneous Propensity

- **Rates (Q):** Innate phonetic biases.
- **Evolutionary Process:** Continuous time, discrete categories.
- **Segmental Change:** Instantaneous jumps (e.g., /p/ > /b/).

Regularity encoded in contextual modeling

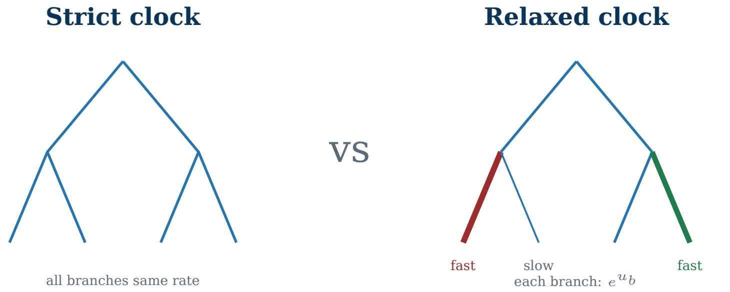
Same context ⇒ same generator



All sites with /p/ between /a/ and /i/ share $Q(a\_\_i)$.

This models Neo-grammarian regularity in probabilistic form.

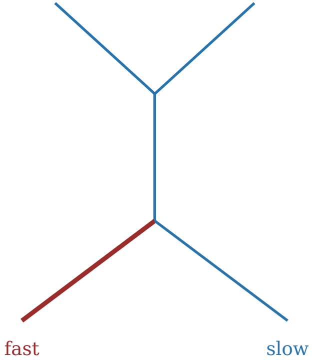
Strict vs. relaxed clock



Language families routinely violate the strict clock.

Sources: Bromham et al., 2015; Greenhill et al., 2017; Bowern and Atkinson, 2012

Branch-specific rate heterogeneity



- multiplicative rate effect e^{u_b} per branch
- hierarchical shrinkage priors over branch-rate heterogeneity

Sources: cf. Bromham et al., 2015; Greenhill et al., 2017; Chang et al., 2015; Yang, 2000

What determines the rate of change ($x > y$)?

$$Q_{xy}(\mathbf{c}) \propto \pi_y \exp\{\eta_{xy} + \kappa(\mathbf{c}) + u_b\}$$

Phonetic & Structural Drivers

π_y **Stationary frequency**

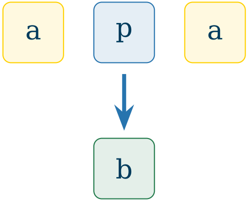
η_{xy} **Similarity & Directionality**
Phonetic similarity and directional biases (e.g., lenition) encoded via covariates.

Context & Phylogeny

$\kappa(\mathbf{c})$ **Local context**
Effect of surrounding environment.

u_b **Branch multiplier**
Lineage-specific rate variation.

Example: intervocalic lenition



$\eta_{p,b}$: small feature distance (voicing)

β_{dir}: lenition bias

$\kappa(V\_V)$: intervocalic boost

All add up $\Rightarrow$ high $Q_{p,b}(V\_V)$

$Q_{p,r}(V\_V)$ stays low: no similarity or bias support

Continuous-Time Evolution: Probabilities

The Outcome Over Time ($P(t)$)

Ancestor

/p/

Time duration (t)

Descendant

/p/ : 75%
/b/ : 20%
/v/ : 5%

Accumulation of Change

- **Process:** Integrate rates (Q) over time (t).
- **Result:** A probability distribution over jumps in discrete states.
- **Impact:** Longer branches = more accumulated change.

Simulation analysis

1. simulate tree and branch lengths

Simulation analysis

1. simulate tree and branch lengths
2. set ground-truth parameters

Simulation analysis

1. simulate tree and branch lengths
2. set ground-truth parameters
3. generate cognate-site data forward

Simulation analysis

1. simulate tree and branch lengths
2. set ground-truth parameters
3. generate cognate-site data forward
4. hide internal states; keep tips only

Simulation analysis

1. simulate tree and branch lengths
2. set ground-truth parameters
3. generate cognate-site data forward
4. hide internal states; keep tips only
5. run full inference pipeline

Simulation analysis

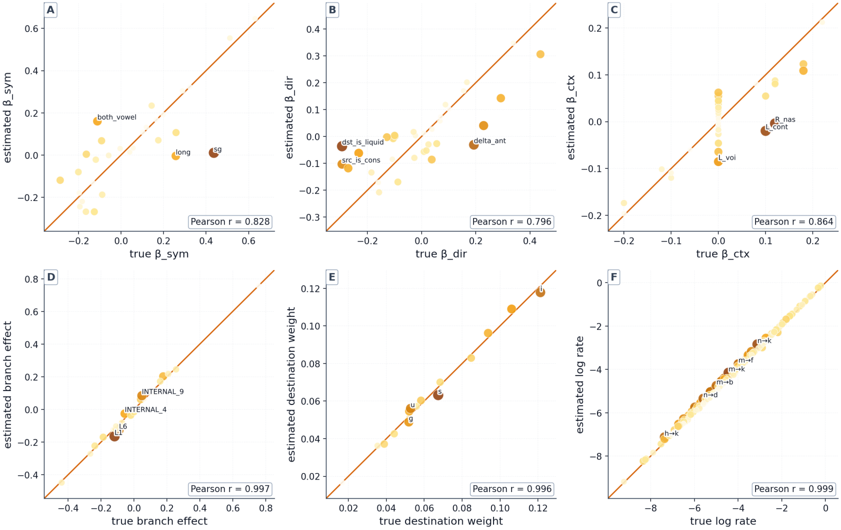
1. simulate tree and branch lengths
2. set ground-truth parameters
3. generate cognate-site data forward
4. hide internal states; keep tips only
5. run full inference pipeline
6. compare recovered vs. true values

Simulation analysis

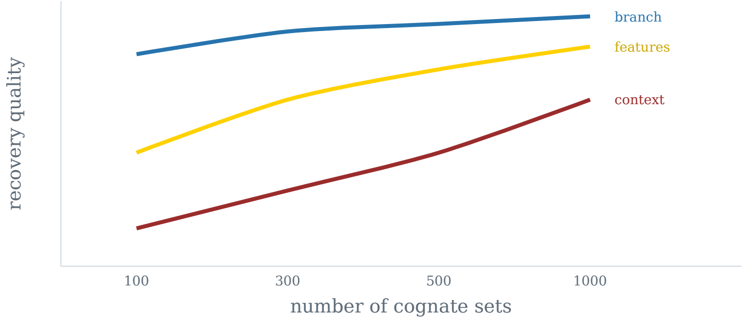
1. simulate tree and branch lengths
2. set ground-truth parameters
3. generate cognate-site data forward
4. hide internal states; keep tips only
5. run full inference pipeline
6. compare recovered vs. true values

Inference goal: show that the model developed here can recover parameters defining the probabalistic sound change process *in principle*.

Parameter recovery



How identifiability scales with data



Context coefficients benefit most from larger datasets.

Inference on real data: Manual alignments from Romance

```
namespace Mountain {
  namespace Primary {
    Latin      = "montem";
    Romanian   = "munte-";
    Portuguese = "mõ-tɨ-";
    Spanish    = "monte-";
    Catalan    = "mun---";
    French     = "mɔ˜----";
    Walloon    = "mɔ˜----";
    Friulian   = "mɔnt--";
    Italian    = "monte-";
  }
}
```

Figure: Sample alignment for ‘mountain’ from Goldstein et al. (2026)

Inference

- Fit the model on aligned word forms from Romance languages

Inference

- Fit the model on aligned word forms from Romance languages
- Implenent an EM-based reconstruction of latent context

Inference

- Fit the model on aligned word forms from Romance languages
- Implenent an EM-based reconstruction of latent context
- Approximate posterior inference using Automatic Differential Variational Inference (ADVI)

Inference

- Fit the model on aligned word forms from Romance languages
- Implenent an EM-based reconstruction of latent context
- Approximate posterior inference using Automatic Differential Variational Inference (ADVI)
- 6 parallel VI draws with random initializations to traverse the multimodal posterior better

Rate of change in the Romance languages

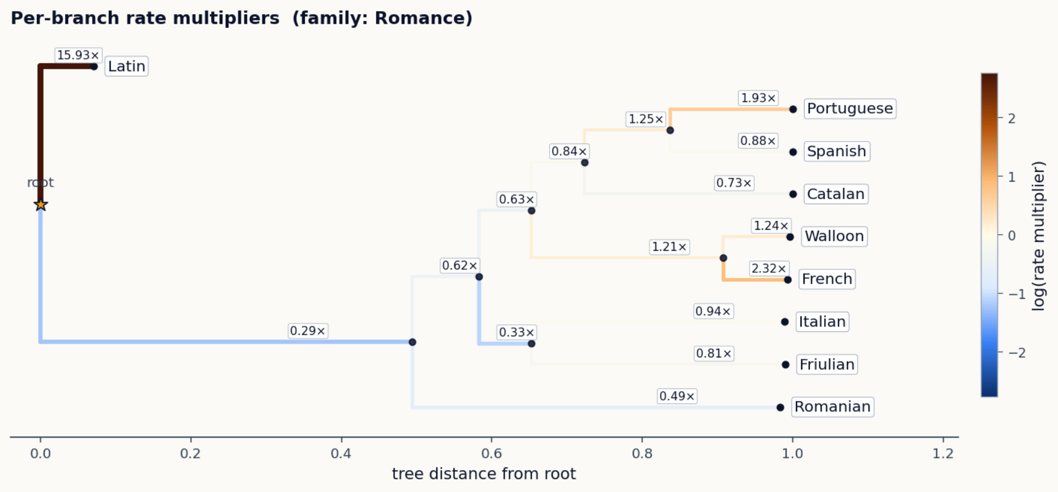


Figure: Inferred rates of phonological change in the histories of Romance languages

Feature and context coefficients

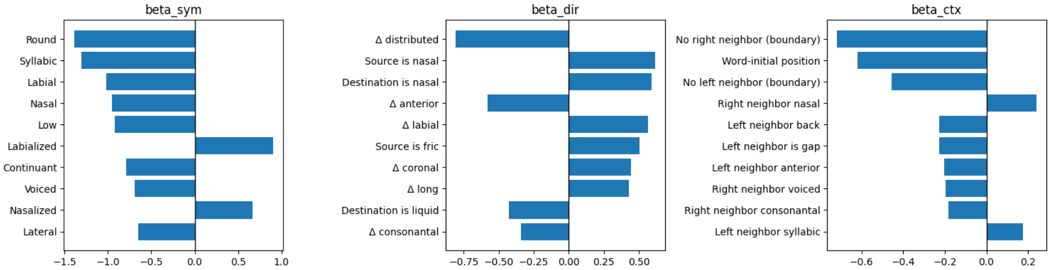


Figure: Symmetric, directional, and contextual contributions to sound (in)stability

Learned perceptual metric from phonological features

Phonological features define each contrast; the model learns which feature dimensions matter most for infant discrimination.

Learned perceptual metric from phonological features

Phonological features define each contrast; the model learns which feature dimensions matter most for infant discrimination.

$$d_{\mathbf{w}}(a,b) = \left(\sum_k w_k (x_{ak} - x_{bk})^2 \right)^{1/2}$$

$$\bar{g}_c = \alpha + \gamma\, d_{\mathbf{w}}(c) + \varepsilon_c$$

Learned perceptual metric from phonological features

Phonological features define each contrast; the model learns which feature dimensions matter most for infant discrimination.

$$d_{\mathbf{w}}(a,b) = \left(\sum_k w_k (x_{ak} - x_{bk})^2\right)^{1/2}$$
$$\bar{g}_c = \alpha + \gamma\, d_{\mathbf{w}}(c) + \varepsilon_c$$

- separate fits for consonants and vowels

Learned perceptual metric from phonological features

Phonological features define each contrast; the model learns which feature dimensions matter most for infant discrimination.

$$d_{\mathbf{w}}(a,b) = \left(\sum_k w_k (x_{ak} - x_{bk})^2\right)^{1/2}$$

$$\bar{g}_c = \alpha + \gamma\, d_{\mathbf{w}}(c) + \varepsilon_c$$

- separate fits for consonants and vowels
- larger weights indicate more perceptually important dimensions

Correlating instability with infant discrimination

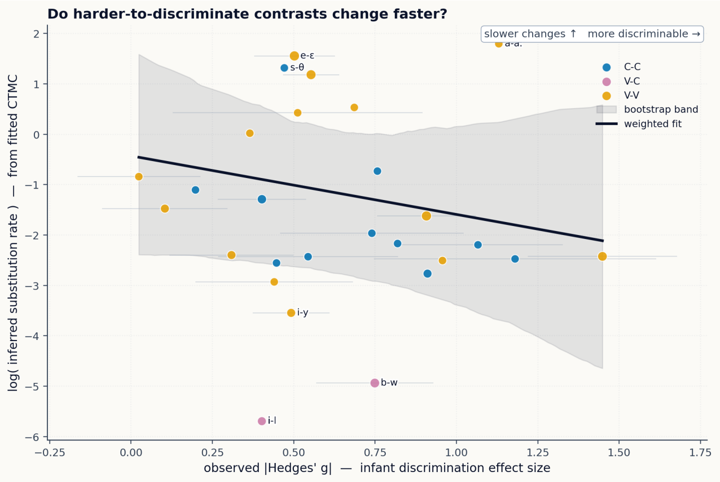


Figure: Sounds tend to change the slowest along the dimensions of contrasts most salient for discrimination.

Conclusion

- Phonological typology is shaped by perceptual salience of contrasts.

Conclusion

- Phonological typology is shaped by perceptual salience of contrasts.
- Infant discrimination provides a minimally confounded estimate of this perceptual salience.

Conclusion

- Phonological typology is shaped by perceptual salience of contrasts.
- Infant discrimination provides a minimally confounded estimate of this perceptual salience.
- Perceptual salience predicts which contrasts are typologically frequent.

Conclusion

- Phonological typology is shaped by perceptual salience of contrasts.
- Infant discrimination provides a minimally confounded estimate of this perceptual salience.
- Perceptual salience predicts which contrasts are typologically frequent.
- But typology is the outcome of **diachronic processes**, not a static equilibrium.

Conclusion

- Phonological typology is shaped by perceptual salience of contrasts.
- Infant discrimination provides a minimally confounded estimate of this perceptual salience.
- Perceptual salience predicts which contrasts are typologically frequent.
- But typology is the outcome of **diachronic processes**, not a static equilibrium.
- A robust account of phonological typology requires **dynamic models of sound change** with structured, context-sensitive rates to provide.

Conclusion

- Phonological typology is shaped by perceptual salience of contrasts.
- Infant discrimination provides a minimally confounded estimate of this perceptual salience.
- Perceptual salience predicts which contrasts are typologically frequent.
- But typology is the outcome of **diachronic processes**, not a static equilibrium.
- A robust account of phonological typology requires **dynamic models of sound change** with structured, context-sensitive rates to provide.

Phonological systems emerge from the interaction of **early perceptual biases** and **evolutionary dynamics**.

Thank you![1]

Questions?

[1] Thank you to members of the PIES Graduate Seminar, the UCLA Phonology seminar, and the audience at SCAMP 2026 for helpful comments and discussion. Special thanks to Marc Suchard and Tandy Warnow for their guidance in the development of the phylogenetic models.

References I

- Auderset, Sandra (2025). “Mixtec Sound Change Database 2.0: Integrating Tone Change”. In: *Journal of Open Humanities Data* 11, p. 37. doi: 10.5334/johd.322.
- Baković, Eric (2007). “Local Assimilation and Constraint Interaction”. In: *The Cambridge Handbook of Phonology*. Ed. by Paul de Lacy. Cambridge: Cambridge University Press, pp. 335–352.
- Blasi, Damián E. et al. (2019). “Human Sound Systems Are Shaped by Post-Neolithic Changes in Bite Configuration”. In: *Science* 363.6432, eaav3218. doi: 10.1126/science.aav3218.
- Blevins, Juliette (2004). *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge: Cambridge University Press.
- Bouchard-Côté, Alexandre et al. (2013). “Automated Reconstruction of Ancient Languages Using Probabilistic Models of Sound Change”. In: *Proceedings of the National Academy of Sciences* 110.11, pp. 4224–4229. doi: 10.1073/pnas.1204678110.

References II

- Bowern, Claire and Quentin Atkinson (2012). “Computational Phylogenetics and the Internal Structure of Pama-Nyungan”. In: *Language* 88.4, pp. 817–845.
- Bromham, Lindell et al. (2015). “Rate of Language Evolution Is Affected by Population Size”. In: *Proceedings of the National Academy of Sciences* 112.7, pp. 2097–2102. doi: 10.1073/pnas.1419704112.
- Bybee, Joan (2001). *Phonology and Language Use*. Cambridge: Cambridge University Press.
- — (2002). “Word Frequency and Context of Use in the Lexical Diffusion of Phonetically Conditioned Sound Change”. In: *Language Variation and Change* 14.3, pp. 261–290. doi: 10.1017/S0954394502143018.
- — (2017). “Grammatical and Lexical Factors in Sound Change: A Usage-Based Approach”. In: *Language Variation and Change* 29.3, pp. 273–300. doi: 10.1017/S0954394517000199.

References III

- Cathcart, Chundra and Florian Wandl (July 2020). "In search of isoglosses: continuous and discrete language embeddings in Slavic historical phonology". In: *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Ed. by Garrett Nicolai, Kyle Gorman, and Ryan Cotterell. Online: Association for Computational Linguistics, pp. 233–244. doi: 10.18653/v1/2020.sigmorphon-1.28. url: https://aclanthology.org/2020.sigmorphon-1.28/.
- Chang, Will et al. (2015). "Ancestry-constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis". In: *Language* 91.1, pp. 194–244. doi: 10.1353/lan.2015.0005.
- Chen, Matthew Y. and William S.-Y. Wang (1975). "Sound Change: Actuation and Implementation". In: *Language* 51.2, pp. 255–281. doi: 10.2307/412854.
- Dediu, Dan and Scott R. Moisik (2019). "Pushes and Pulls from Below: Anatomical Variation, Articulation and Sound Change". In: *Glossa* 4.1, p. 7.

References IV

- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society: Series B* 39.1, pp. 1–38.
- Endevelt-Shapira, Yaara et al. (May 2025). "How mother-infant inter-brain synchrony is affected by the context of the interaction and speech input?" In: *Proceedings of the Society for Research in Child Development (SRCD) Biennial Meeting*. Poster presentation. Minneapolis, MN, USA.
- Feldman, N. H., T. L. Griffiths, and J. L. Morgan (2009). "The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference". In: *Psychological Review* 116.4, p. 752.
- Felsenstein, Joseph (1981). "Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach". In: *Journal of Molecular Evolution* 17.6, pp. 368–376. doi: 10.1007/BF01734359.

References V

- Flemming, E. (2004). "Contrast and perceptual distinctiveness". In: *Phonetically based phonology*. Ed. by B. Hayes, R. Kirchner, and D. Steriade. Cambridge: Cambridge University Press, pp. 232–276. doi: 10.1017/CBO9780511486401.008.
- Garrett, Andrew (2015). "Sound Change". In: *The Routledge Handbook of Historical Linguistics*. Routledge, pp. 227–248.
- Garrett, Andrew and Keith Johnson (2013). "Phonetic Bias in Sound Change". In: *Origins of Sound Change: Approaches to Phonologization*. Ed. by Alan C. L. Yu. Oxford: Oxford University Press, pp. 51–97.
- Geyer, Charles J. (1991). "Markov Chain Monte Carlo Maximum Likelihood". In: *Computing Science and Statistics*, pp. 156–163.
- Goldstein, David et al. (2026). "Phylogenetic inference from cognate word forms". In: *Journal of Language Evolution* 11.1. doi: 10.1093/jole/lzag001.

References VI

- Greenhill, Simon J. et al. (2017). “Evolutionary Dynamics of Language Systems”. In: *Proceedings of the National Academy of Sciences* 114.42, E8822–E8829. doi: 10.1073/pnas.1700388114.
- Hay, Jennifer and Paul Foulkes (2016). “The Evolution of Medial /t/ over Real and Remembered Time”. In: *Language* 92.2, pp. 298–330.
- He, Siqi and Wei Zhao (2024). *Exploring Sound Change Over Time: A Review of Computational and Human Perception*. arXiv: 2407.05092 [cs.CL]. url: https://arxiv.org/abs/2407.05092.
- Hock, Hans Henrich (1991). *Principles of Historical Linguistics*. 2nd ed. Berlin: Mouton de Gruyter.
- Hruschka, Daniel J. et al. (2015). “Detecting Regular Sound Changes in Linguistics as Events of Concerted Evolution”. In: *Current Biology* 25.1, pp. 1–9. doi: 10.1016/j.cub.2014.10.064.
- Huelsenbeck, John P., Rasmus Nielsen, and Jonathan P. Bollback (2003). “Stochastic Mapping of Morphological Characters”. In: *Systematic Biology* 52.2, pp. 131–158. doi: 10.1080/10635150390192780.

References VII

- Kiparsky, Paul (2016). "Labov, Sound Change, and Phonological Theory". In: *Journal of Sociolinguistics* 20.4, pp. 464–488. doi: 10.1111/josl.12196.
- Kirby, James (2013). "The role of probabilistic enhancement in phonologization". In: *Origins of Sound Change: Approaches to Phonologization*. Oxford University Press. doi: 10.1093/acprof:oso/9780199573745.003.0011.
- Kleinschmidt, D. F. (2019). "Structure in talker variability: How much is there and how much can it help?" In: *Language, Cognition and Neuroscience* 34.1, pp. 43–68.
- Kleinschmidt, D. F. and T. F. Jaeger (Apr. 2015). "Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel". In: *Psychological Review* 122.2, pp. 148–203. doi: 10.1037/a0038695.
- Labov, William (2020). "The Regularity of Regular Sound Change". In: *Language* 96.1, pp. 42–59. doi: 10.1353/lan.2020.0001.

References VIII

- Lindblom, B. (1986). “Phonetic universals in vowel systems”. In: *Experimental Phonology*. Ed. by J. Ohala and J. Jaeger. Orlando: Academic Press, pp. 13–44.
- Maddieson, Ian (1984). *Patterns of Sounds*. Cambridge: Cambridge University Press.
- Martinet, André (1952). “Function, Structure, and Sound Change”. In: *Word* 8.1, pp. 1–32.
- Mayer, Connor (2021). “Capturing Gradience in Long-distance Phonology Using Probabilistic Tier-based Strictly Local Grammars”. In: *Proceedings of the Society for Computation in Linguistics 2021*. Online: Association for Computational Linguistics, pp. 39–50.
- McGahay, John (2025). “Modeling Vowel System Typology Using Iterated Confusion Minimization”. In: *Proc. Interspeech 2025*, pp. 2955–2959. doi: 10.21437/Interspeech.2025-2192.

References IX

McGahay, John (2026). “A computational model of allophony as Dispersion over phonetic sequences”. Presentation at the Linguistic Society of America Annual Meeting, New Orleans, LA, January 8–11.

Milroy, Lesley (1987). *Language and Social Networks*. 2nd ed. Oxford: Blackwell.

Moran, Steven, Nicholas A. Lester, and Eitan Grossman (2021). “Inferring Recent Evolutionary Changes in Speech Sounds”. In: *Philosophical Transactions of the Royal Society B* 376.1824, p. 20200198. doi: 10.1098/rstb.2020.0198.

Moran, Steven and Daniel McCloy (2019). *PHOIBLE 2.0*. Jena. doi: 10.5281/zenodo.2593234.

Moreton, Elliott (2008). “Analytic Bias and Phonological Typology”. In: *Phonology* 25.1, pp. 83–127. doi: 10.1017/S0952675708001413.

Mortensen, David R. et al. (2016). “PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors”. In: *Proceedings of COLING 2016*, pp. 3475–3484.

References X

- Nielsen, Rasmus (2002). “Mapping Mutations on Phylogenies”. In: *Systematic Biology* 51.5, pp. 729–739.
- Nikolaev, D. and E. Grossman (2020). “Consonant co-occurrence classes and the feature-economy principle”. In: *Phonology* 37.3, pp. 419–451. doi: 10.1017/S0952675720000226.
- Ohala, John J. (1981). “The Listener as a Source of Sound Change”. In: *Papers from the Parasession on Language and Behavior*. Ed. by Carrie S. Masek, Roberta A. Hendrick, and Mary Frances Miller. Chicago: Chicago Linguistic Society, pp. 178–203.
- — (1993). “The Phonetics of Sound Change”. In: *Speech Communication* 13.1–2, pp. 153–161.
- Paster, Mary Elizabeth (2006). “Phonological Conditions on Affixation”. PhD thesis. University of California, Berkeley.
- Phillips, Betty S. (1984). “Word Frequency and the Actuation of Sound Change”. In: *Language* 60.2, pp. 320–342.

References XI

- Ramsammy, Michael (2015). “The Life Cycle of Phonological Processes: Accounting for Dialectal Microtypologies”. In: *Language and Linguistics Compass* 9.1, pp. 33–54. doi: 10.1111/lnc3.12102.
- Rose, Sharon (2011). “Long-distance Assimilation of Consonants”. In: *The Blackwell Companion to Phonology*. Ed. by Marc van Oostendorp et al. Malden, MA: Wiley-Blackwell, pp. 1–27.
- Rose, Sharon and Rachel Walker (2004). “A Typology of Consonant Agreement as Correspondence”. In: *Language* 80.3, pp. 475–531.
- Round, Erich R. (2022). “Evolution and Trade-Off Dynamics of Functional Load”. In: *Entropy* 24.4, p. 507. doi: 10.3390/e24040507.
- Schwartz, Jean-Luc et al. (1997). “The Dispersion-Focalization Theory of vowel systems”. In: *Journal of Phonetics* 25.3, pp. 255–286. issn: 0095-4470. doi: https://doi.org/10.1006/jpho.1997.0043. url: https://www.sciencedirect.com/science/article/pii/S0095447097900437.

References XII

- Sóskuthy, Márton (2015). "Understanding Change through Stability: A Computational Study of Sound Change Actuation". In: *Lingua* 163, pp. 40–60.
- Stevens, Kenneth N. (1989). "On the quantal nature of speech". In: *Journal of Phonetics* 17.1, pp. 3–45. issn: 0095-4470. doi: https://doi.org/10.1016/S0095-4470(19)31520-7. url: https://www.sciencedirect.com/science/article/pii/S0095447019315207.
- Tavaré, Simon (1986). "Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences". In: *Lectures on Mathematics in the Life Sciences* 17, pp. 57–86.
- Todd, Simon J., Janet B. Pierrehumbert, and Jennifer B. Hay (2019). "Word Frequency Effects in Sound Change as a Consequence of Perceptual Asymmetries: An Exemplar-based Model". In: *Cognition* 185, pp. 1–20. doi: 10.1016/j.cognition.2019.01.004.

References XIII

- Wedel, Andrew, Abby Kaplan, and Scott Jackson (2013). "High Functional Load Inhibits Phonological Contrast Loss: A Corpus Study". In: *Cognition* 128.2, pp. 179–186. doi: 10.1016/j.cognition.2013.03.002.
- Weinreich, Uriel, William Labov, and Marvin I. Herzog (1968). "Empirical Foundations for a Theory of Language Change". In: *Directions for Historical Linguistics*. Ed. by Winfred P. Lehmann and Yakov Malkiel. Austin: University of Texas Press, pp. 95–195.
- Yang, Ziheng (2000). "Maximum Likelihood Estimation on Large Phylogenies and Analysis of Adaptive Evolution in Human Influenza Virus A". In: *Journal of Molecular Evolution* 51.5, pp. 423–432.
- — (2006). *Computational Molecular Evolution*. Oxford: Oxford University Press.

Mapping out the perceptual space

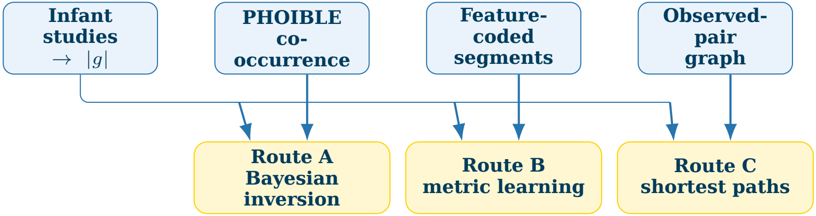
Infant studies $\rightarrow \ |g|$

PHOIBLE co-occurrence

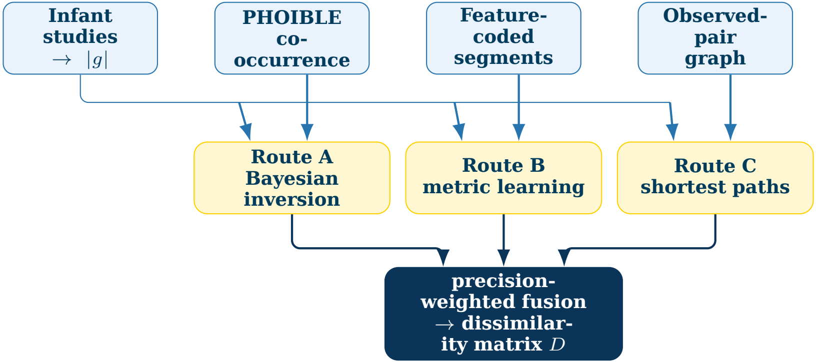
Feature-coded segments

Observed-pair graph

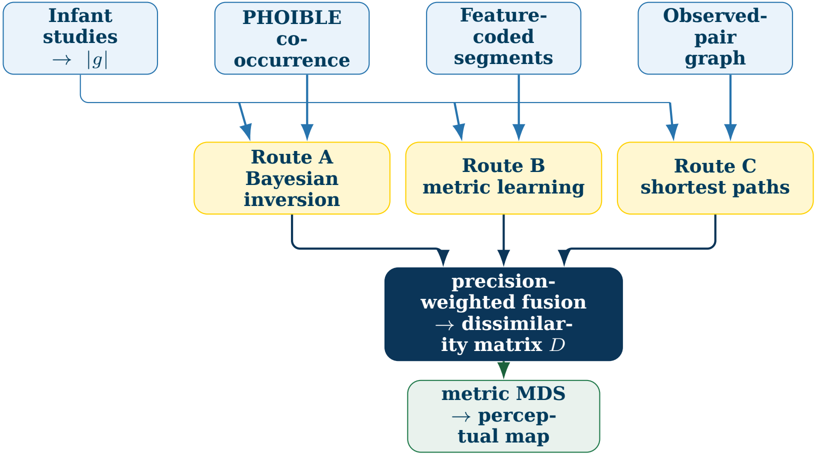
Mapping out the perceptual space



Mapping out the perceptual space



Mapping out the perceptual space



What Route A does

Use the fitted meta-regression to convert an observed PHOIBLE pair frequency f_{ab} into an implied discrimination magnitude $|g|$.

What Route A does

Use the fitted meta-regression to convert an observed PHOIBLE pair frequency f_{ab} into an implied discrimination magnitude $|g|$.

$$\mathrm{logit}(f_{ab}) = \beta_0^{(m)} + \beta_1^{(m)} g_{ab}^{(m)} + u^{(m)}$$

$$u^{(m)} \sim \mathcal{N}\left(0, (\sigma_{\text{study}}^{(m)})^2\right)$$

What Route A does

Use the fitted meta-regression to convert an observed PHOIBLE pair frequency f_{ab} into an implied discrimination magnitude $|g|$.

$$\mathrm{logit}(f_{ab}) = \beta_0^{(m)} + \beta_1^{(m)} g_{ab}^{(m)} + u^{(m)}$$
$$u^{(m)} \sim \mathcal{N}\Big(0, (\sigma_{\mathrm{study}}^{(m)})^2\Big)$$

Solving for $g_{ab}^{(m)}$ gives

$$g_{ab}^{(m)} = \frac{\mathrm{logit}(f_{ab}) - \beta_0^{(m)} - u^{(m)}}{\beta_1^{(m)}}.$$

What Route A does

Use the fitted meta-regression to convert an observed PHOIBLE pair frequency f_{ab} into an implied discrimination magnitude $|g|$.

$$\mathrm{logit}(f_{ab}) = \beta_0^{(m)} + \beta_1^{(m)} g_{ab}^{(m)} + u^{(m)}$$
$$u^{(m)} \sim \mathcal{N}\Big(0, (\sigma_{\text{study}}^{(m)})^2\Big)$$

Solving for $g_{ab}^{(m)}$ gives

$$g_{ab}^{(m)} = \frac{\mathrm{logit}(f_{ab}) - \beta_0^{(m)} - u^{(m)}}{\beta_1^{(m)}}.$$

We turn cross-linguistic co-occurrence into an implied perceptual distance.

Route B: learned perceptual metric from phonological features

Phonological features define each contrast; the model learns which feature dimensions matter most for infant discrimination.

Route B: learned perceptual metric from phonological features

Phonological features define each contrast; the model learns which feature dimensions matter most for infant discrimination.

$$d_{\mathbf{w}}(a,b) = \left(\sum_k w_k (x_{ak} - x_{bk})^2\right)^{1/2}$$

$$\bar{g}_c = \alpha + \gamma\, d_{\mathbf{w}}(c) + \varepsilon_c$$

Route B: learned perceptual metric from phonological features

Phonological features define each contrast; the model learns which feature dimensions matter most for infant discrimination.

$$d_{\mathbf{w}}(a,b) = \left(\sum_k w_k (x_{ak} - x_{bk})^2\right)^{1/2}$$

$$\bar{g}_c = \alpha + \gamma\, d_{\mathbf{w}}(c) + \varepsilon_c$$

- separate fits for consonants and vowels

Route B: learned perceptual metric from phonological features

Phonological features define each contrast; the model learns which feature dimensions matter most for infant discrimination.

$$d_{\mathbf{w}}(a,b) = \left(\sum_k w_k (x_{ak} - x_{bk})^2\right)^{1/2}$$

$$\bar{g}_c = \alpha + \gamma\, d_{\mathbf{w}}(c) + \varepsilon_c$$

- separate fits for consonants and vowels
- larger weights indicate more perceptually important dimensions

From dissimilarities to a perceptual map

Input: Scores from different routes

Precision-weighted fusion gives a dissimilarity matrix D over sound pairs.

From dissimilarities to a perceptual map

Input: Scores from different routes

Precision-weighted fusion gives a dissimilarity matrix D over sound pairs.

- MDS finds a low-dimensional configuration whose Euclidean distances approximate D

From dissimilarities to a perceptual map

Input: Scores from different routes

Precision-weighted fusion gives a dissimilarity matrix D over sound pairs.

- MDS finds a low-dimensional configuration whose Euclidean distances approximate D
- far apart = easier to discriminate

From dissimilarities to a perceptual map

Input: Scores from different routes

Precision-weighted fusion gives a dissimilarity matrix D over sound pairs.

- MDS finds a low-dimensional configuration whose Euclidean distances approximate D
- far apart = easier to discriminate
- close together = more perceptually similar

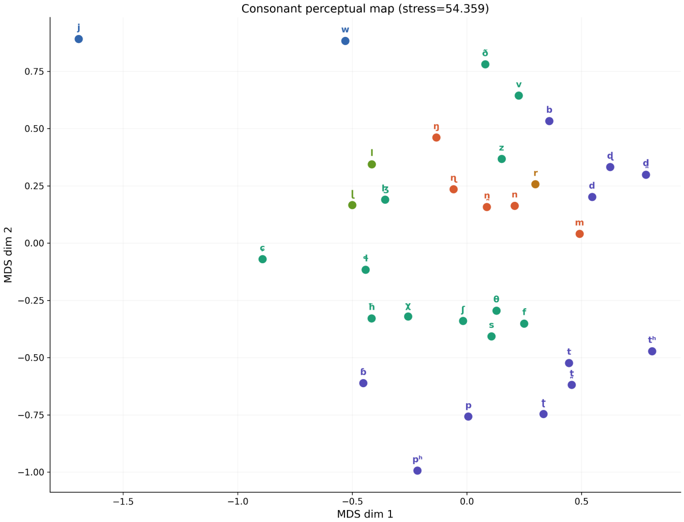
From dissimilarities to a perceptual map

Input: Scores from different routes

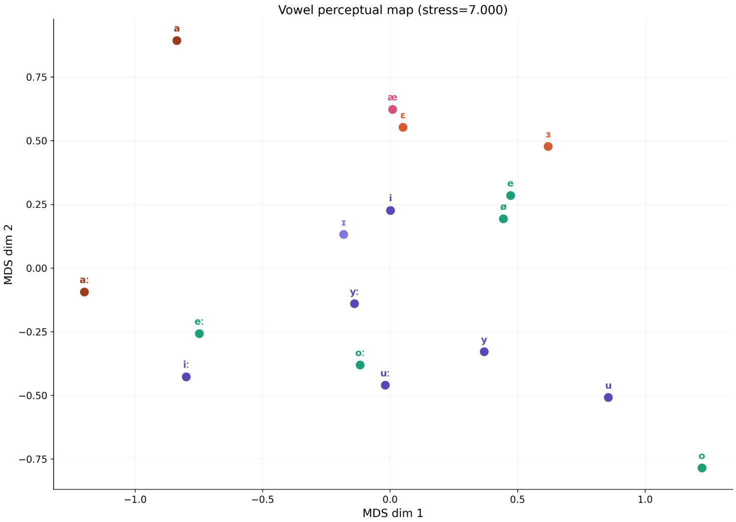
Precision-weighted fusion gives a dissimilarity matrix D over sound pairs.

- MDS finds a low-dimensional configuration whose Euclidean distances approximate D
- far apart = easier to discriminate
- close together = more perceptually similar
- stress = 2D distortion

Perceptual space of selected consonants



Perceptual space of selected vowels



Differences in perceptual mapping of consonants and vowels

Consonants

- held-out correlation: $r = 0.652$
- Shepard correlation: 0.769
- 2D stress: 54.36
- harder to compress into a clean 2D map

Vowels

- held-out correlation: $r = 0.442$
- Shepard correlation: 0.903
- 2D stress: 7.00
- much more compatible with a low-dimensional geometry

Lexical diffusion

Lexical diffusion

Some changes propagate unevenly through the lexicon; frequency and neighborhood density affect which words change first.

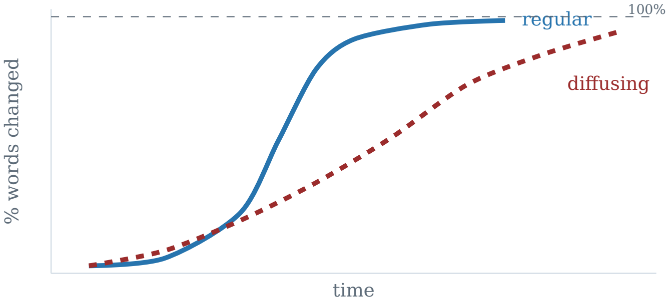
- **wang1969**: competing changes create residue
- Bybee (2002): sound change shows usage and frequency effects

A quantitative model should treat regularity as a *testable baseline*, but not enforce it.

Sources: **wang1969**; Chen and Wang, 1975; Phillips, 1984; Bybee, 2001; Bybee, 2002; Todd, Pierrehumbert, and J. B. Hay, 2019

Diffusion dynamics: the S-curve

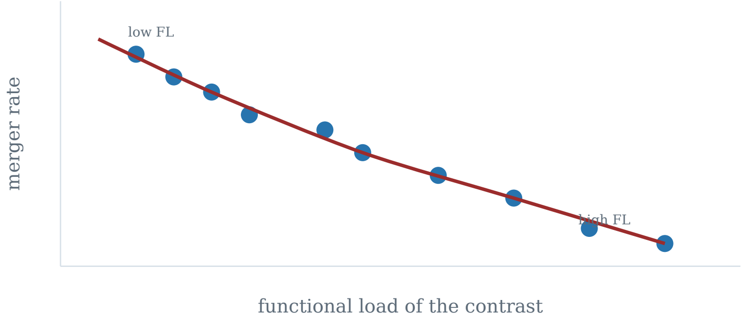
Both regularity and diffusion predict S-shaped adoption



The empirical question: do lexical items differ in their intrinsic rate of adopting a change?

Sources: **labov1994**; Bybee, 2002; Sóskuthy, 2015; J. Hay and Foulkes, 2016

Functional load as a rate inhibitor



High-FL contrasts resist merger. This is absent from version 1 but is a natural extension.

Sources: Martinet, 1952; Wedel, Kaplan, and Jackson, 2013; Round, 2022

External and demographic pressures

- network density shapes diffusion
- contact and isolation influence the tempo of change
- anatomical and climatic factors may bias inventories

Actuation problem
Why this change, here, now? A rate model should represent *heterogeneity across branches*.

Sources: Weinreich, Labov, and Herzog, 1968; Milroy, 1987; Bromham et al., 2015; Greenhill et al., 2017; Dediu and Moisik, 2019

Defining the Context

Locality, similarity, and phonological domains

1. Strictly Local

Assimilation, lenition, and boundary effects driven by **adjacent** material.

2. Long-Distance

Harmony and correspondence sensitive to **non-local** similarity.

3. Domain-Sensitive

Rules behaving differently inside roots vs. across affixes (**morphological** domains).

Modeling Assumption

The local-context approach taken here is a principled **first approximation**, not a theoretical claim that all phonological structure is strictly local, or should be modeled as such.

Sources: Baković, 2007; Rose and Walker, 2004; Rose, 2011; Mayer, 2021; Paster, 2006

Continuous-Time Markov Chains

At a single site

A segment evolves as a continuous-time Markov chain, a probabilistic model for discrete jumps between states.

Sources: Felsenstein, 1981; Bouchard-Côté et al., 2013; Yang, 2006

Continuous-Time Markov Chains

The Transition Rate Matrix Q

$$Q = \begin{pmatrix} -\sum_{j\neq 1} q_{1j} & q_{12} & \cdots & q_{1n} \\ q_{21} & -\sum_{j\neq 2} q_{2j} & \cdots & q_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ q_{n1} & q_{n2} & \cdots & -\sum_{j\neq n} q_{nj} \end{pmatrix}$$

- **Off-diagonal entries ($q_{ij} \geq 0$):** Represent the instantaneous rate of transitioning from state i to state j.
- **Diagonal entries ($q_{ii} \leq 0$):** Defined such that each row sums to zero.
- **Transition Probabilities:** Calculated via the matrix exponential $P(t) = e^{Qt}$.

Sources: Felsenstein, 1981; Bouchard-Côté et al., 2013; Yang, 2006

Instantaneous rates of change > probabilities of change

Generator Q

instantaneous rates Q_{xy}: propensity to move $x > y$

$\xrightarrow{e^{Qt}}$

Transition $P(t)$

probability of state y after time t, given started in x

Longer branches $\Rightarrow$ more change

Sources: Felsenstein, 1981; Yang, 2006; Tavaré, 1986

APPENDIX

Mathematical details

A1. Site expansion and root priors

Lexical root prior

$P(X_r{=}x \mid \text{lex}) = (1{-}\varepsilon_{\text{lex}})\pi_x$ for $x \in \mathcal{S}_{\text{real}}$; $\quad \varepsilon_{\text{lex}}$ for $x{=}\varnothing$.

Ghost root prior

$P(X_r{=}x \mid \text{ghost}) = \varepsilon_{\text{ghost}}\pi_x$ for $x \in \mathcal{S}_{\text{real}}$; $\quad 1{-}\varepsilon_{\text{ghost}}$ for $x{=}\varnothing$.

A2. Generator matrix

$$Q_{xy} = \exp(\log \mu + \kappa + \eta_{xy})\, \pi_y, \qquad Q_{xx} = -\sum_{y \neq x} Q_{xy}$$

$$\eta_{xy} = \langle \mathbf{X}_{xy}^{\mathrm{sym}}, \boldsymbol{\beta}_{\mathrm{sym}} \rangle + \langle \mathbf{X}_{xy}^{\mathrm{dir}}, \boldsymbol{\beta}_{\mathrm{dir}} \rangle, \qquad \kappa = \langle \mathbf{c}, \boldsymbol{\beta}_{\mathrm{ctx}} \rangle$$

A3. Branch-rate heterogeneity

$$u_b^{\mathrm{raw}} = \tau \lambda_b z_b, \quad z_b \sim \mathcal{N}(0,1); \qquad \nu_b = u_b^{\mathrm{raw}} - \overline{u}; \qquad \tilde{\ell}_b = \ell_b e^{\nu_b}$$

Centering makes branch effects relative.

A4. Likelihood and pruning

Pattern-compressed likelihood

$\log p(\mathbf{y} \mid \theta) = \sum_u n_u \log P(\mathbf{y}^{(u)} \mid \theta)$

Felsenstein recursion

$L_v(k) = \prod_{c \in \text{children}(v)} \sum_j P_{b_c}(k, j)\, L_c(j);$ root: $\sum_k \rho(k)\, L_{\text{root}}(k)$.

A5. Parallel tempering

Tempered targets $p(\theta \mid \mathbf{y})^{1/T_m}$; swap acceptance:

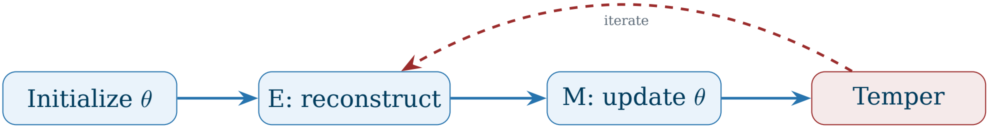
$$\log a_{\text{swap}} = \left(\frac{1}{T_m} - \frac{1}{T_{m+1}}\right)\left[\log p(\theta_{m+1}) - \log p(\theta_m)\right]$$

Sources: Geyer, 1991

Inference

Core difficulty

Rates depend on *local context*, but context is determined by *latent ancestral states*.



Sources: Felsenstein, 1981; Bouchard-Côté et al., 2013; Dempster, Laird, and Rubin, 1977; Geyer, 1991

A6. Stochastic mapping

Set $\omega \geq \max_k(-Q_{kk})$, define $B = I + Q/\omega$. Conditional uniformization samples Poisson event counts and skeleton paths to estimate posterior event histories.

Sources: Huelsenbeck, Nielsen, and Bollback, 2003; Nielsen, 2002

A7. Ancestral reconstruction

$$P(X_r{=}i \mid \mathbf{y}) \propto \rho(i)\, L_r(i); \qquad P(X_c{=}j \mid X_p{=}i, \mathbf{y}) \propto P_{ij}^{(g,b)}\, L_c(j)$$

Simulation design

Tree

20-tip balanced tree; log-normal branch lengths; one fast clade, one slow.

Data

500 cognate sets, 3–8 sites each; 17-state inventory; ground truth mimics known asymmetries.

A8. Regularization

- Gaussian priors on coefficients to cause shrinkage
- Half-normal priors on branch scales
- Structured generator prevents combinatorial explosion

A9. Relation to MaxEnt / Harmonic Grammar

Shared	Different
Weighted structures preserving gradience and restricting hypothesis space.	MaxEnt/HG: output probabilities. CTMC: transition intensities over time.

Sources: Hayes & Wilson 2008; Moore-Cantwell 2016; Shih 2017; Pater 2016

A10. Why rates matter for theory

Different theories make different predictions about *where* change should be fast or slow:

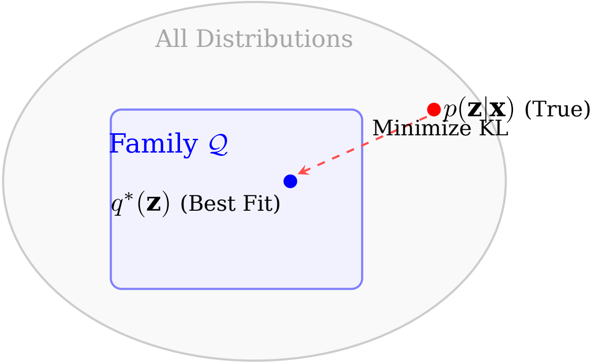
- **Phonetic-bias:** rates highest in articulatorily/perceptually pressured contexts
- **Analytic-bias:** rates relate to learnability, even with comparable phonetic grounding
- **Usage-based:** rates relate to frequency and predictability
- **Grammar-internal:** rates cluster by domain and life-cycle stage

Sources: Blevins, 2004; Moreton, 2008; Garrett and Johnson, 2013; Bybee, 2017; Kiparsky, 2016; Ramsammy, 2015

The Intuition: Finding the Closest Match

The Roadblock
True posterior $p(\mathbf{z}|\mathbf{x})$ is intractable. We can't calculate the evidence $p(\mathbf{x})$.

The VI Workaround
1. Define a simple, tractable family of distributions $\mathcal{Q}$ (e.g., Gaussian).
2. Search $\mathcal{Q}$ for the specific distribution $q^*(\mathbf{z})$ that is **closest** to the true posterior.



The Math: Maximizing the ELBO

We measure ”closeness” using KL Divergence, but we can’t compute it directly. Instead, we use this identity:

$$\underbrace{\log p(\mathbf{x})}_{\text{Fixed Data}} = \underbrace{\mathcal{L}(\phi)}_{\textbf{ELBO (Computable)}} + \underbrace{\mathrm{KL}(q_{\phi}(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x}))}_{\text{Divergence } \geq 0}$$

The Insight: Maximizing the ELBO perfectly minimizes the KL Divergence.

The ELBO Trade-off

$$\mathcal{L}(\phi) = \underbrace{\mathbb{E}_{q_{\phi}}[\log p(\mathbf{x}|\mathbf{z})]}_{\textbf{1. Reconstruction (Data Fit)}} - \underbrace{\mathrm{KL}(q_{\phi}(\mathbf{z}) \parallel p(\mathbf{z}))}_{\textbf{2. Penalty (Keep it simple)}}$$

The Execution: Reparameterization Trick

How do we get gradients through a random sampling process $\mathbf{z} \sim q_\phi$ to update our neural network? **We reroute the randomness.**

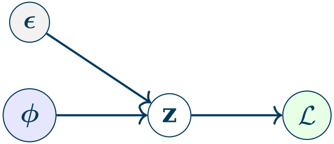
Standard Sampling

(Blocks gradients)

$$\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$$

Reparameterized

(Uninterrupted gradients)



$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$$

Deep Dive: The Mechanics of VI

1. **The Core Objective**
   - We seek parameters $\phi^* = \arg\min_\phi \mathrm{KL}(q_\phi(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x}))$.
   - Because $p(\mathbf{z}|\mathbf{x}) = p(\mathbf{x}, \mathbf{z})/p(\mathbf{x})$ and $p(\mathbf{x})$ is intractable, we optimize the ELBO instead.
2. **The ELBO Expansions**

$$\begin{aligned}
\mathcal{L}(\phi) &= \mathbb{E}_{q_\phi}[\log p(\mathbf{x}, \mathbf{z})] - \mathbb{E}_{q_\phi}[\log q_\phi(\mathbf{z})] && \textit{(Joint/Entropy Form)} \\
&= \mathbb{E}_{q_\phi}[\log p(\mathbf{x}|\mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z}) \parallel p(\mathbf{z})) && \textit{(Reconstruction/Prior Form)}
\end{aligned}$$

3. Primary Optimization Pathways

- **Coordinate Ascent (CAVI):** Assumes q fully factorizes ($q(\mathbf{z}) = \prod q_j(z_j)$). Updates one coordinate at a time using exact expectations. Best for conditionally conjugate models.
- **Stochastic VI (SVI):** Uses mini-batches and noisy gradients ($\nabla_\phi \mathcal{L}$) to scale to massive datasets. Relies heavily on the reparameterization trick to keep variance low.