Unsupervised Consonant-Vowel Prediction over Hundreds of Languages

Young-Bum Kim and Benjamin Snyder
Wisconsin-Madison

lost language

known languages

?

[ACL 2010] Computational Decipherment of *Ugaritic*

⇒ Leverages *related* languages

Remaining Challenges

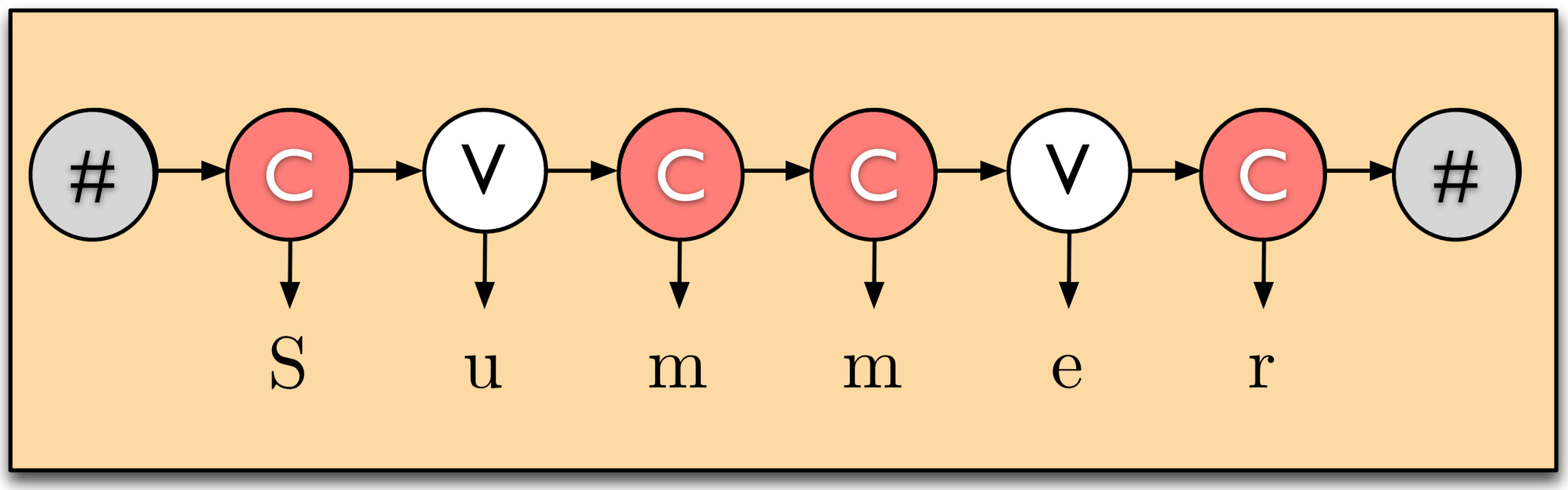
- Language family unknown
- No related living languages
  - ~ *Linear A*
  - ~ *Linear Elamite*
  - ~ *Byblos*
  - ~ *Cypro-Minoan*
  - ~ *Isthmian*

Undeciphered:
Linear A

- Syllabic writing systems

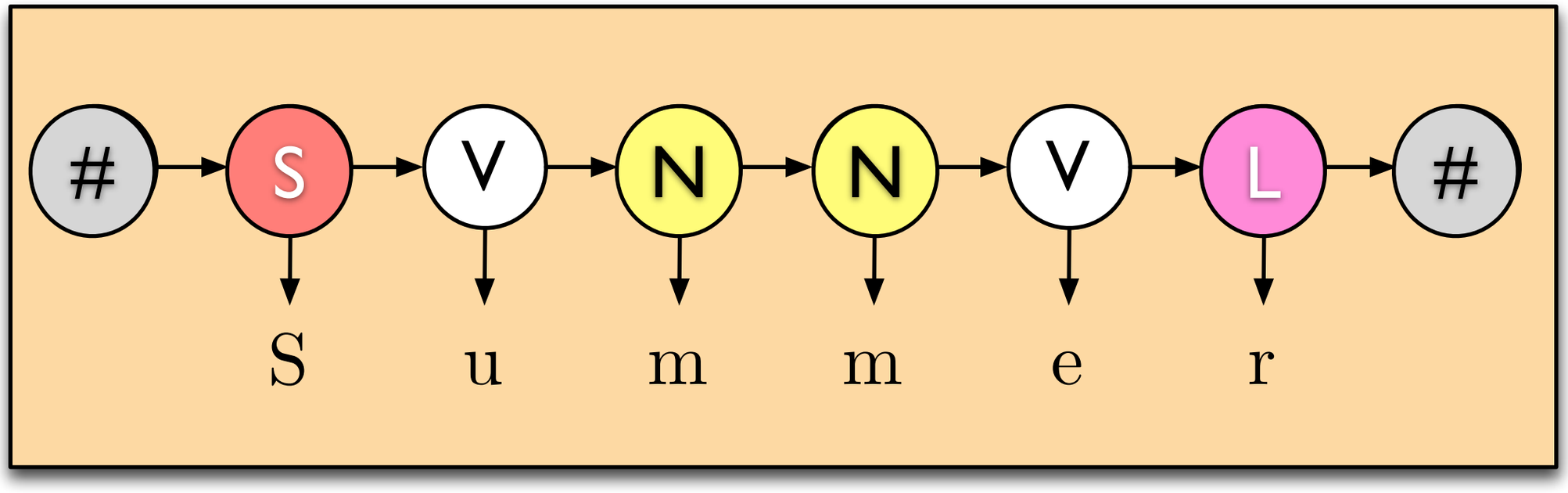
Knight et al 2006

- Trigram HMM over character sequences
  - ~ Two hidden states
  - ~ Estimated using EM
- Distinguishes Spanish consonants/vowels



Lingering Questions

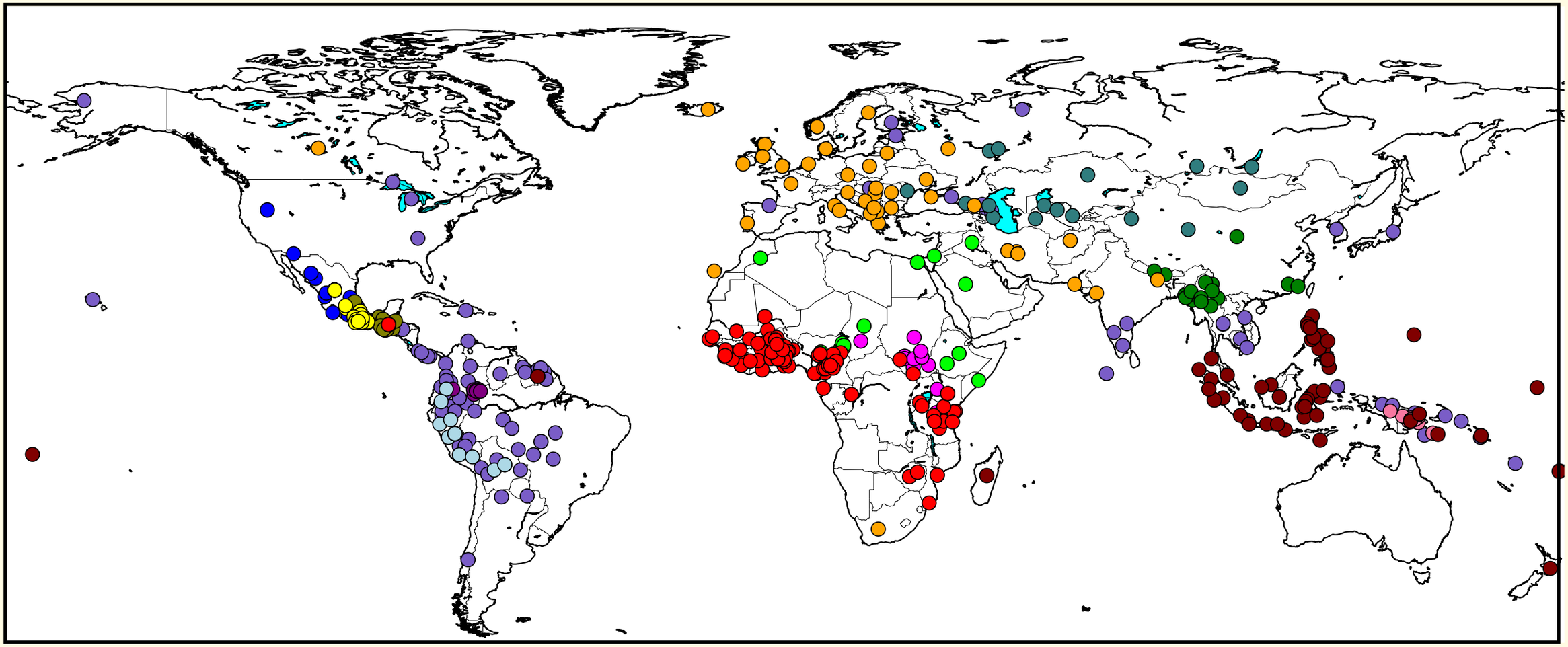
1. Just Spanish?
   - Georgian: prtskvna (*p*ee*l*i*n*g) : CCCCCCCCV
2. Leverage *known* languages?
3. sibilant, nasal, liquid, etc?



Bible Data for 503 alphabetic languages

1. Cons vs Vowel
2. Nasal vs Cons vs Vowel

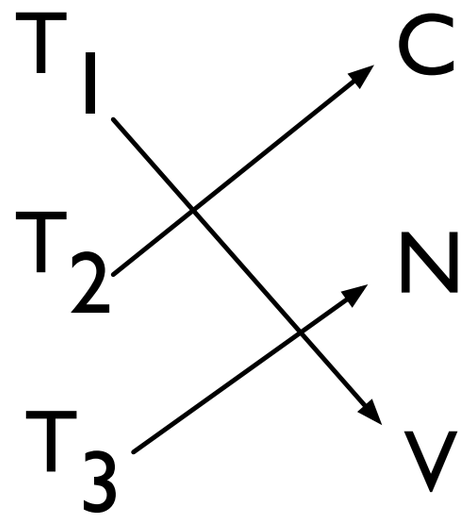
- Niger-Congo (113)
- Austronesian (68)
- Indo-European (43)
- Oto-Manguean (40)
- Mayan (33)
- Afro-Asiatic (21)
- Sino-Tibetan (18)
- Altaic (17)
- Quechuan (17)
- Uto-Aztecan (15)
- Trans-New Guinea (15)
- Nilo-Saharan (14)
- Tucanoan (9)
- Others (99)



HMM Results

C / V	N / C / V
93.4%	74.6%

- 503 alphabetic languages
- 10 EM restarts
- Oracle tag mappings:



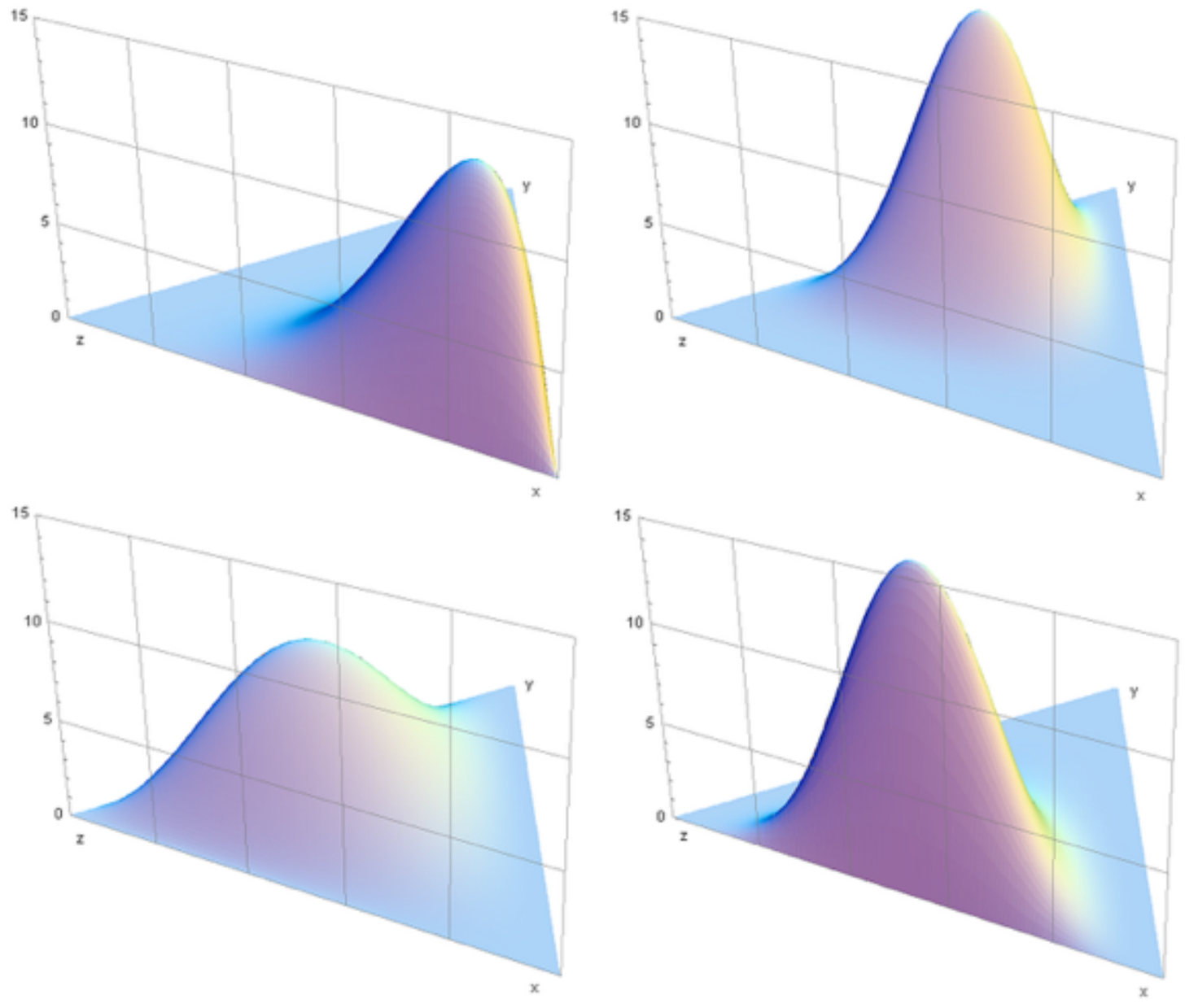
⇒ [C / V] dominant natural pattern

⇒ [N / C / V] less so...

Model Design Goals

1. Cluster letters:
   - HMM does *soft* clustering
   - letters typically in single phonetic class
2. Learn from hundreds of languages
   - Avoid oracle tag mappings
3. Model typologically coherent groupings

Bayesian HMM: Dirichlet Priors



$$\alpha_1,\ \alpha_2,\ \alpha_3 > 0$$

Hard Bayesian HMM

- One tag per character type

  (stochastic *deterministic* FSA)

- *Dimensionality* of tag emissions unknown:

Emission distribution

$$\phi|N \sim \text{Dirichlet}(\beta_1 \ldots \beta_N)$$

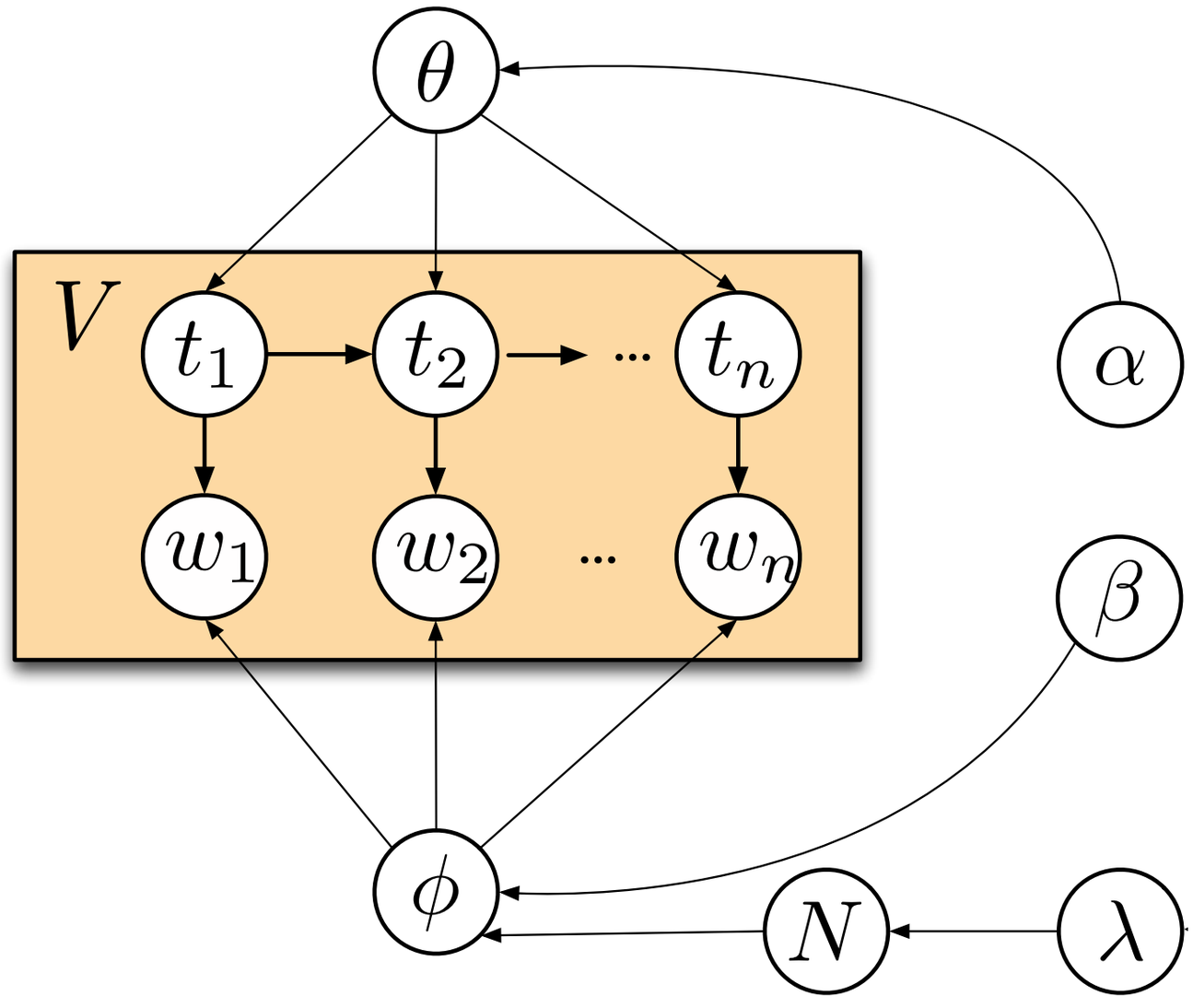
symbols

$$N|\lambda \sim \text{Poisson}(\lambda)$$

Model 1: SYMM

Assume fixed, symmetric priors

transition dist



Dirichlet priors

Poisson prior

Emission dist

symbols

Inference: Gibbs Sampling

- Integrate out parameters
- Iteratively sample tag assignment t_w of symbol types W :

$$f(t_w|\mathbf{t}_{-w}, w_1 \ldots w_n) \propto$$

Inference: Gibbs Sampling

- Integrate out parameters
- Iteratively sample tag assignment t_w of symbol types W :

$$f(t_w|\mathbf{t}_{-w}, w_1 \dots w_n) \propto$$

$$\int f(N|\lambda) f(\lambda)\ d\lambda \qquad \text{(type counts)} \qquad (1)$$

$$\int f(t_1 \dots t_n|\theta) f(\theta)\ d\theta \qquad \text{(tags)} \qquad (2)$$

$$\int f(w_1 \dots w_n|t_1 \dots t_n, \phi) f(\phi|N)\ d\phi \qquad \text{(symbols)} \qquad (3)$$

Tag Predictive Distribution

Dirichlet hyperparameter

tag bigram counts *excluding* symbol w

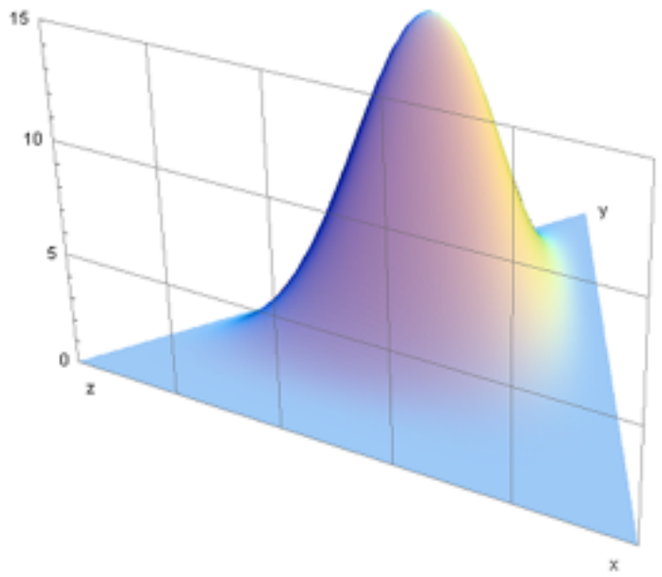
tag bigram counts *only* with *symbol* w

$$\frac{\prod_{t,t'} (\alpha_{t,t'} + n(t,t'))^{[\delta(t,t')]}}{\prod_t \quad \cdots \quad \cdots}$$

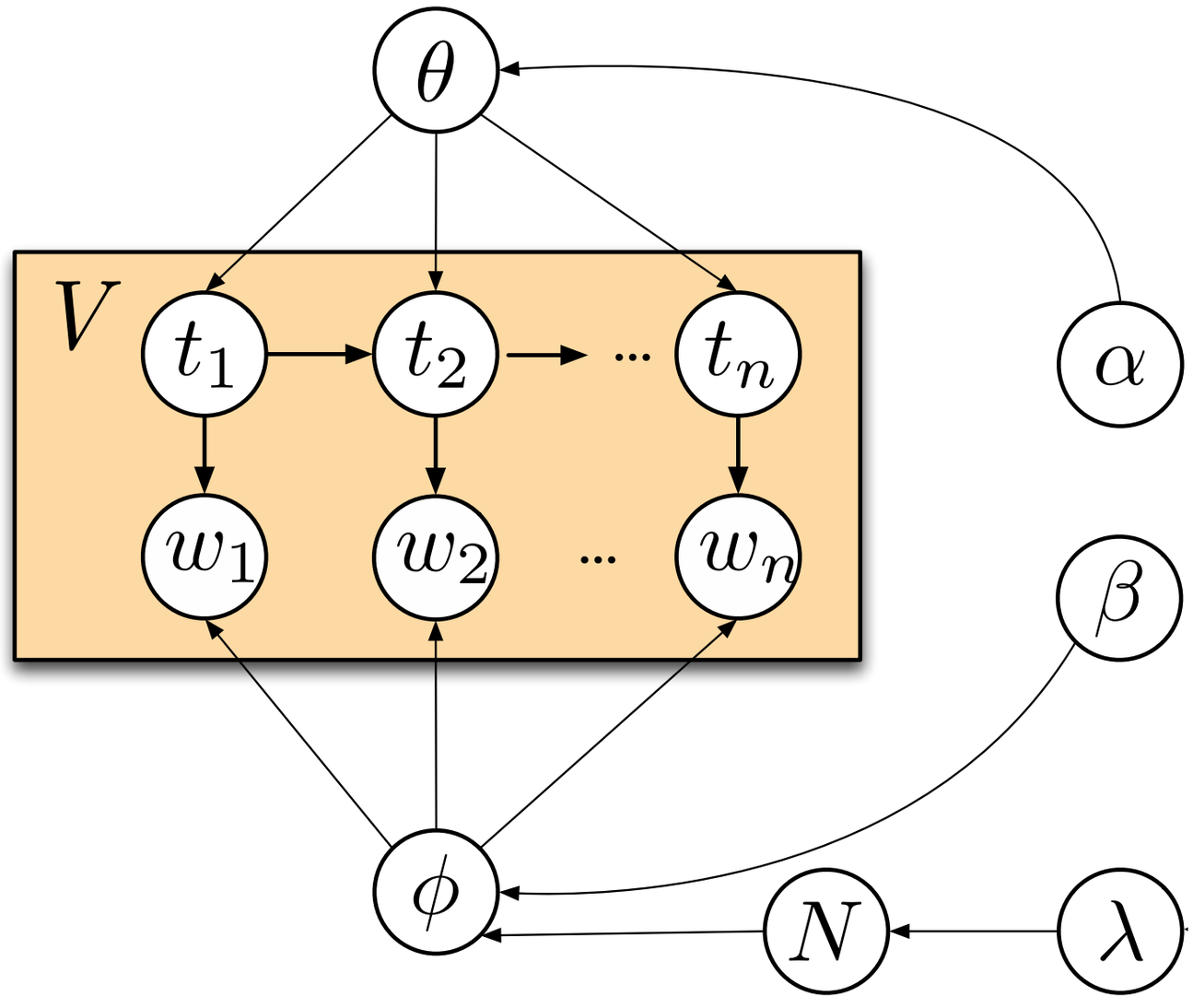
Next Idea

- Trying to decipher an unknown language
- Learn from our 502 other languages where the tags are *known*

⇒ Assume shared, unknown priors across languages



transition dist



Dirichlet priors

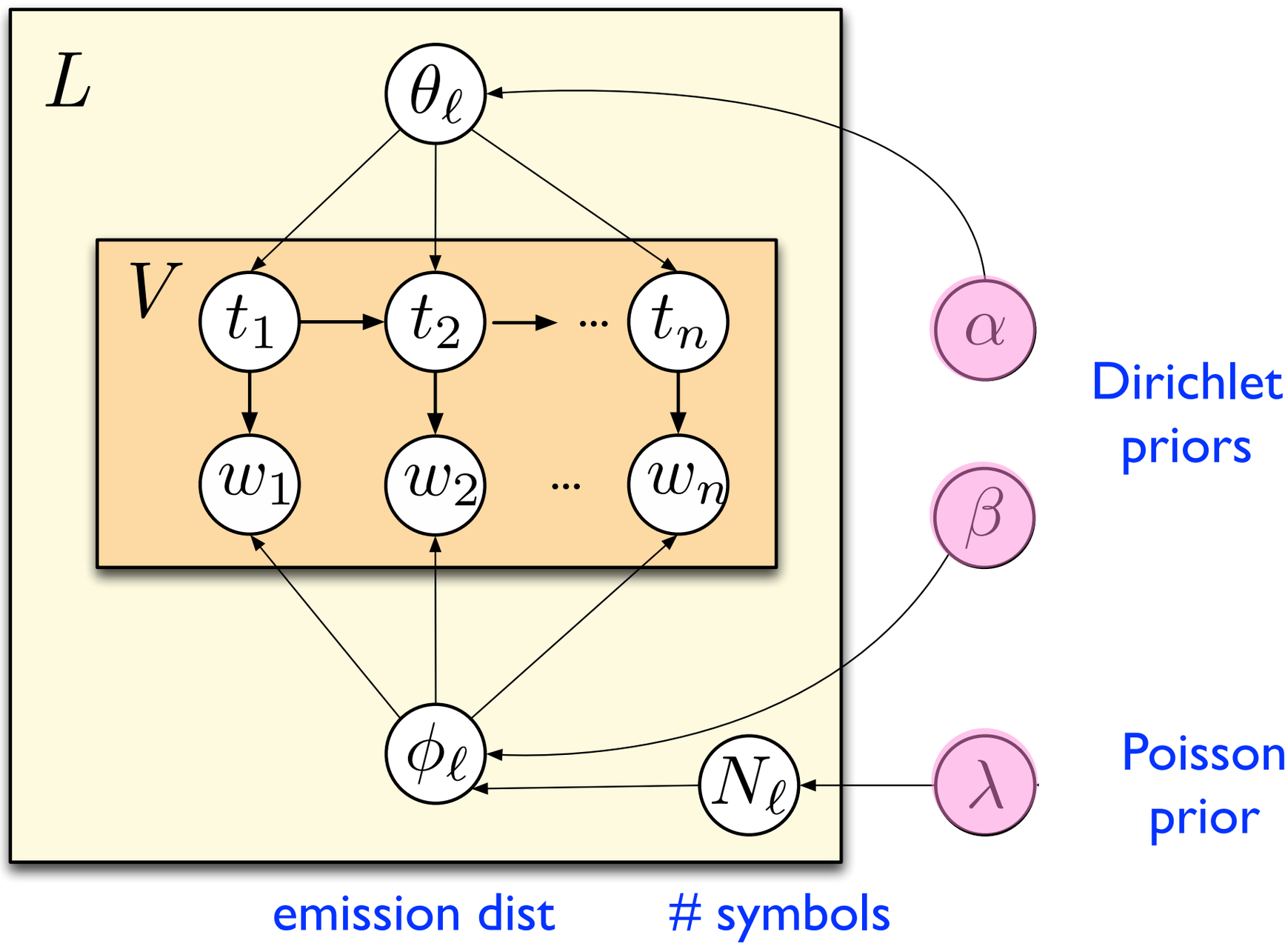
Poisson prior

emission dist

symbols

Model 2: MERGED

Infer universal language priors



Inference: Gibbs Sampling

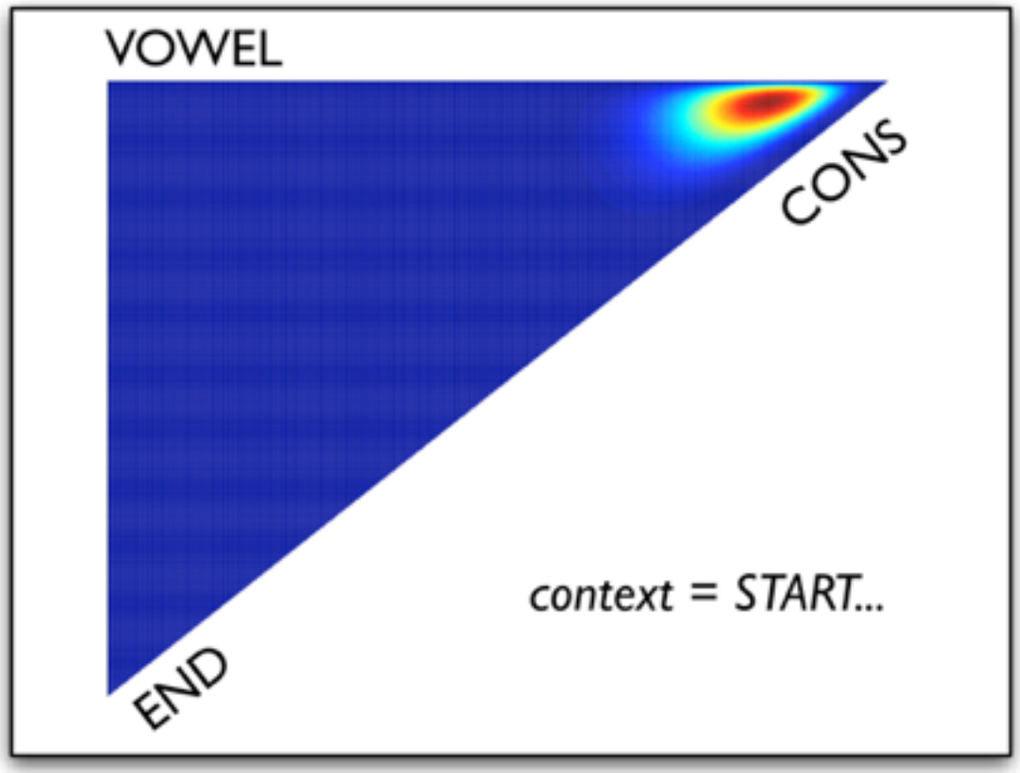
- Slice sample hyperparameters
  - ~ *product of* predictive distributions with common prior
  - ~ combine terms across languages for efficiency

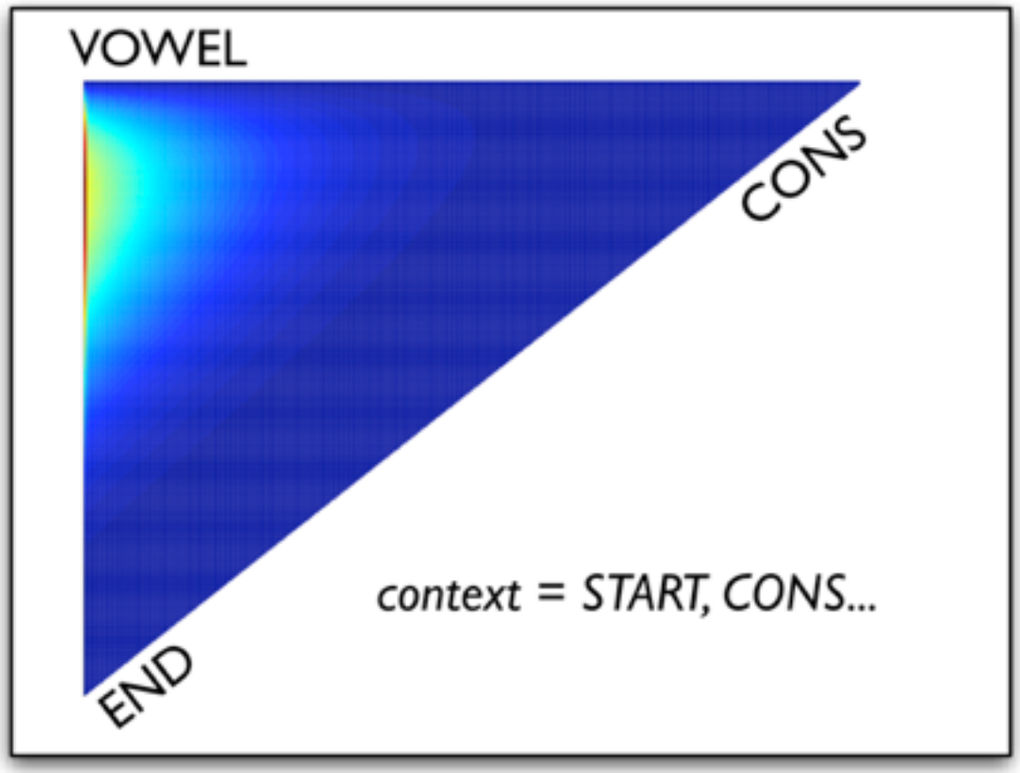
number of languages which have ≥ k occurrences of bigram (t, t')

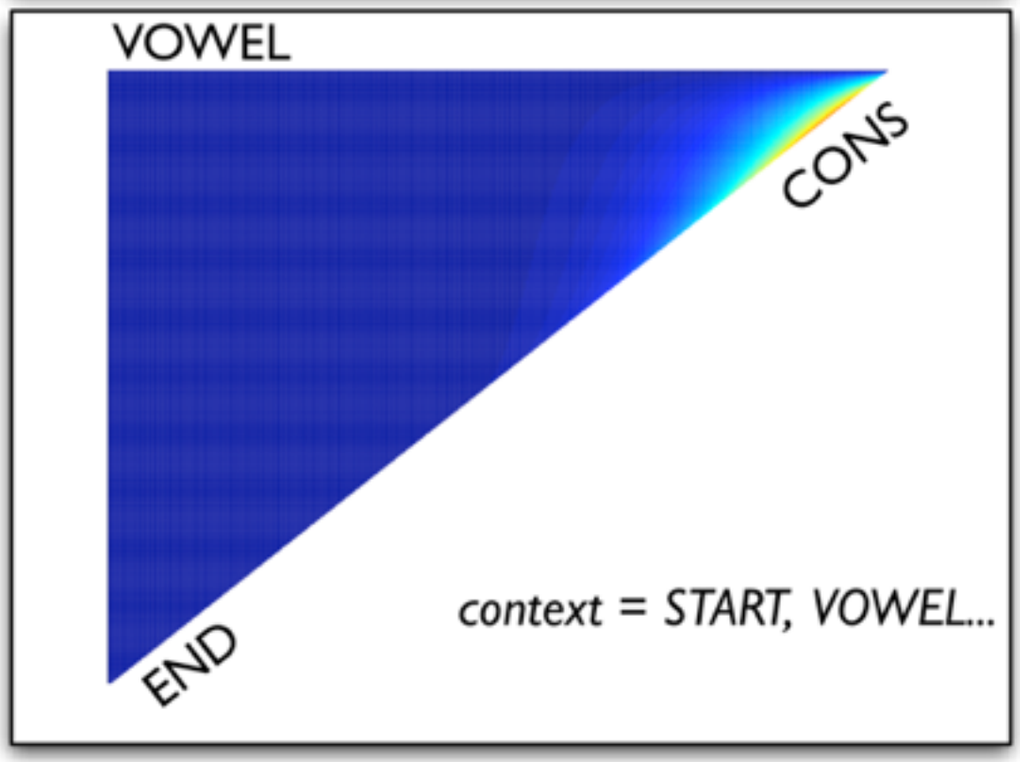
$$f(\alpha_{t,t'} | \ldots) \propto$$

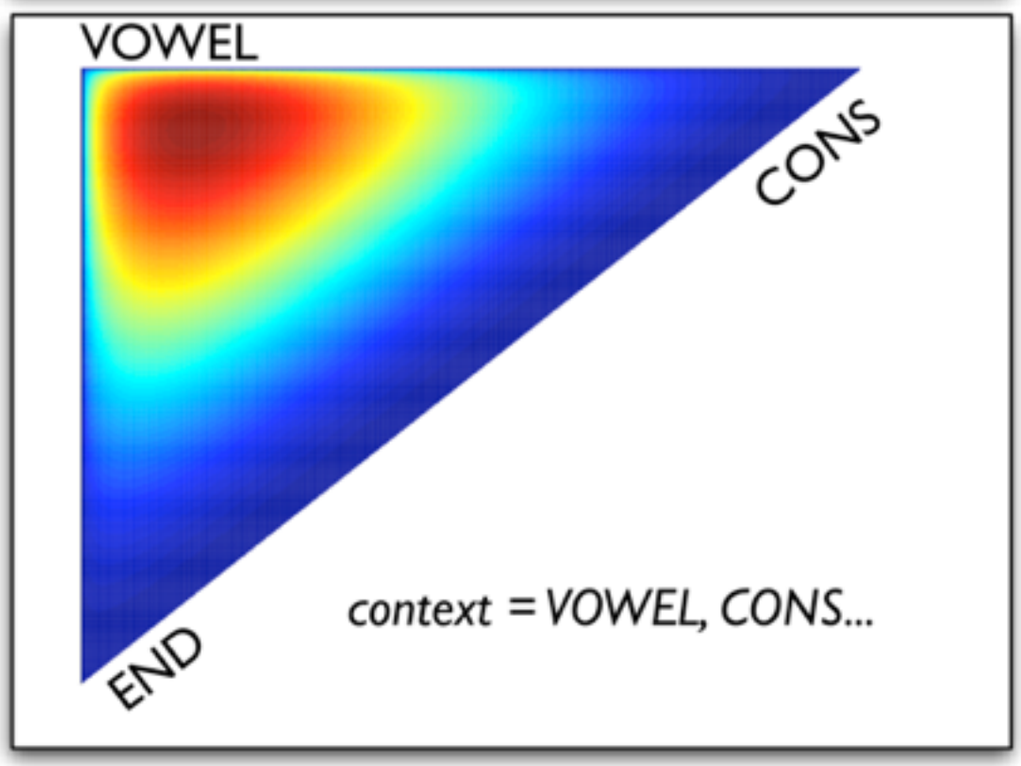
$$\prod_k \frac{(\alpha_{t,t'} + k)^{\#(k,t,t')}}{\ldots \ldots}$$

MAP Transition Priors: 503 languages

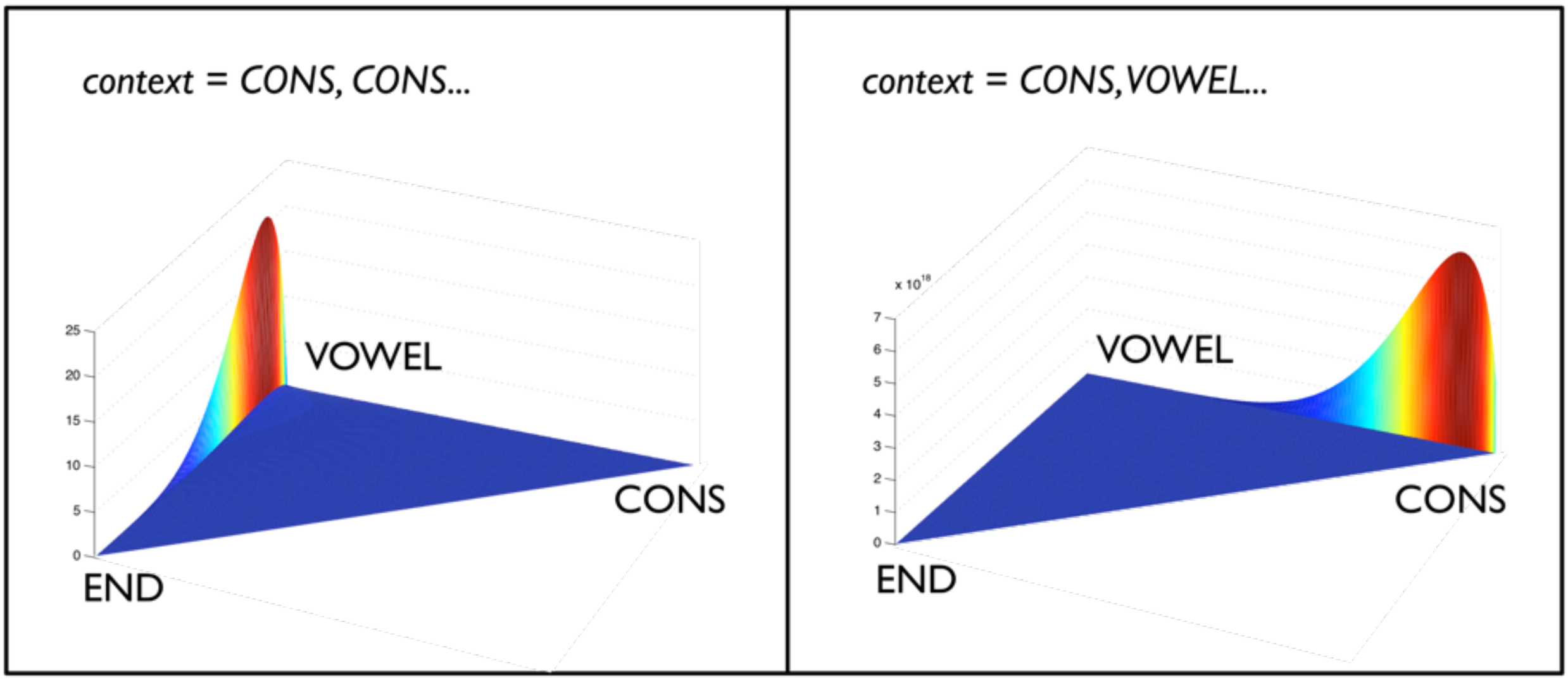








MAP Transition Priors: 503 languages



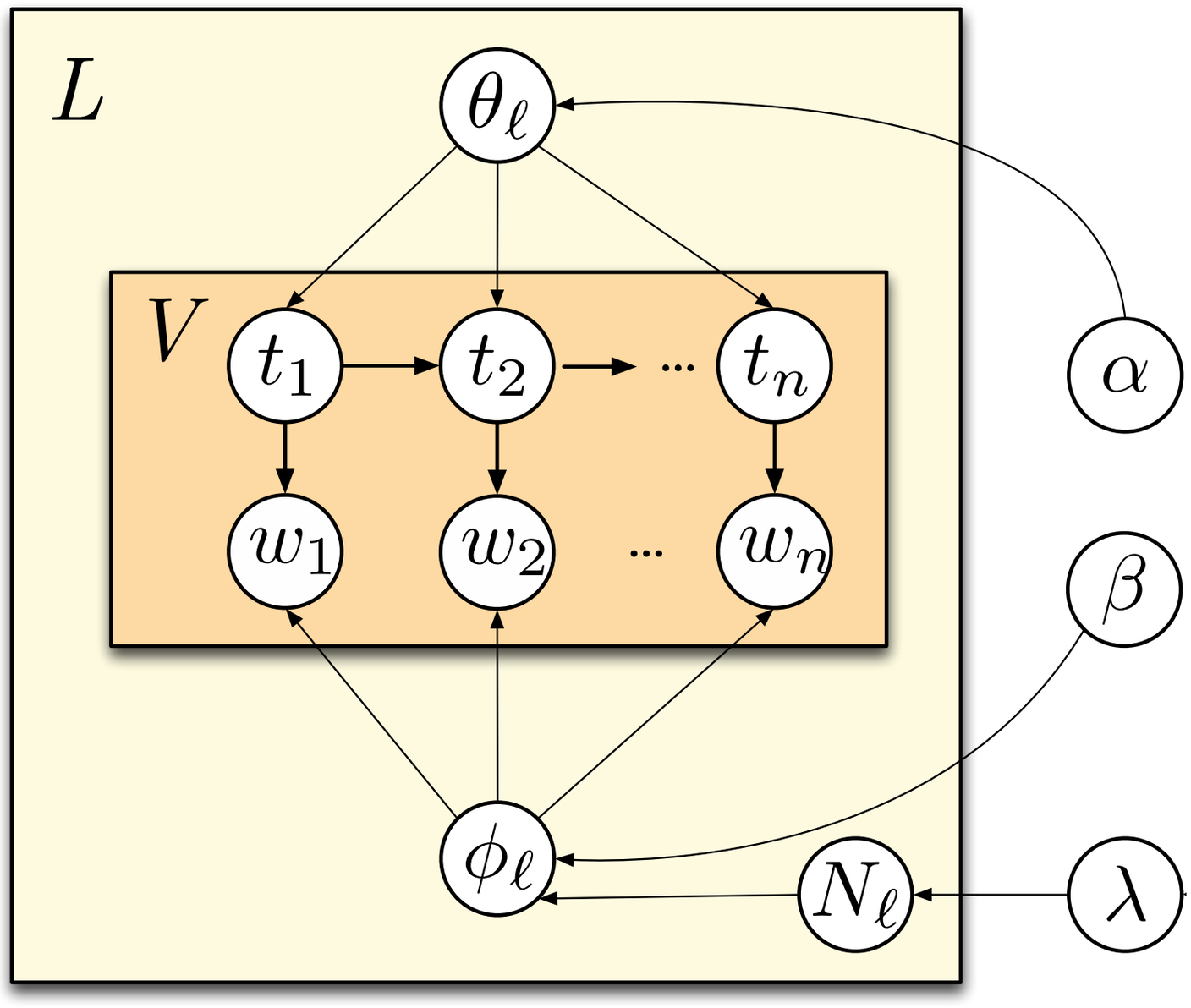
Final Idea

- Languages vary typologically

⇒ Assume latent clustering of languages

⇒ Hyperparameters shared at the cluster level

Transition dist



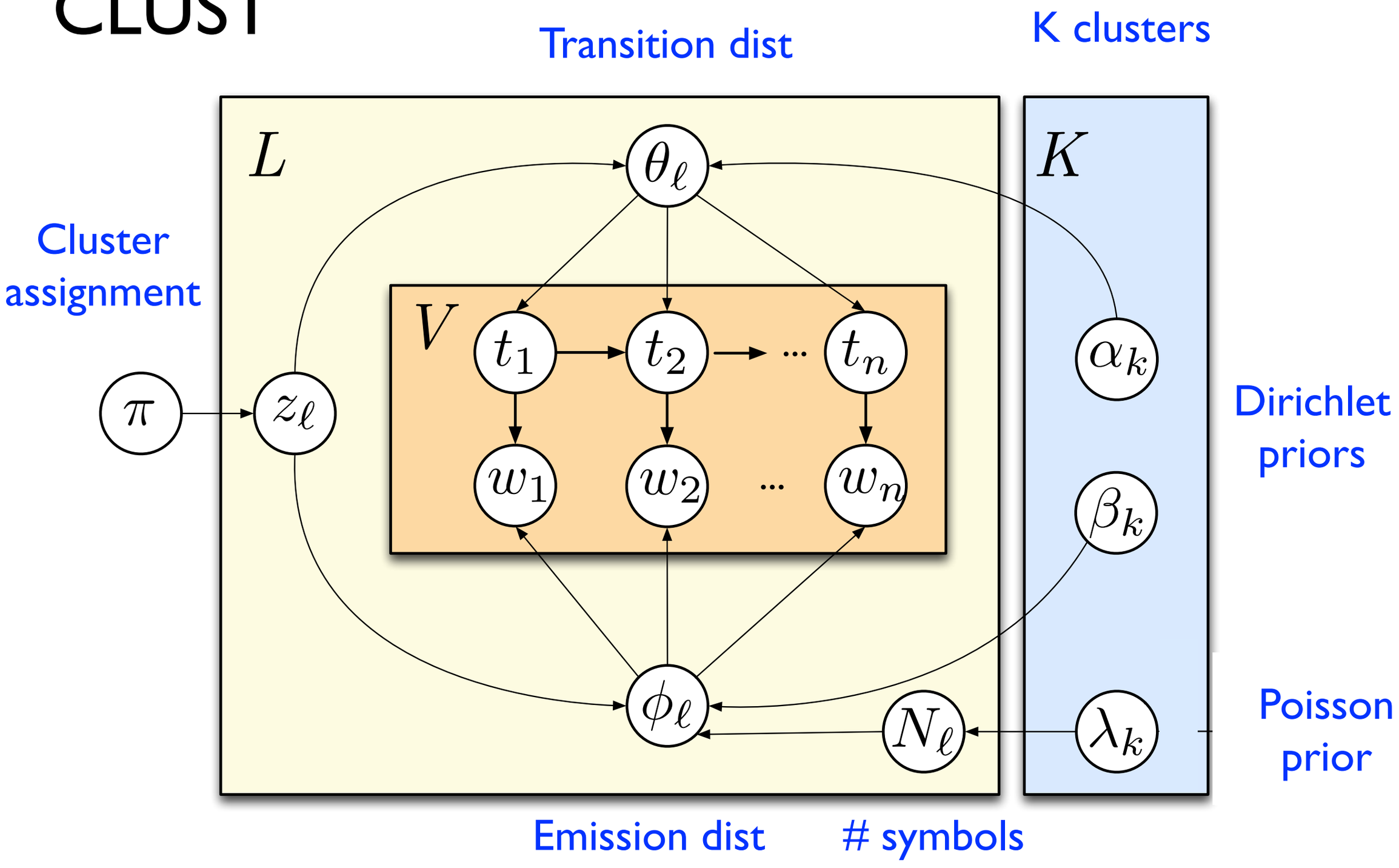
Dirichlet priors

Poisson prior

Emission dist

symbols

Model 3: CLUST



Evaluation

- Token-level accuracy
  - ~ 503 languages
  - ~ leave-one-out average

Two Tasks

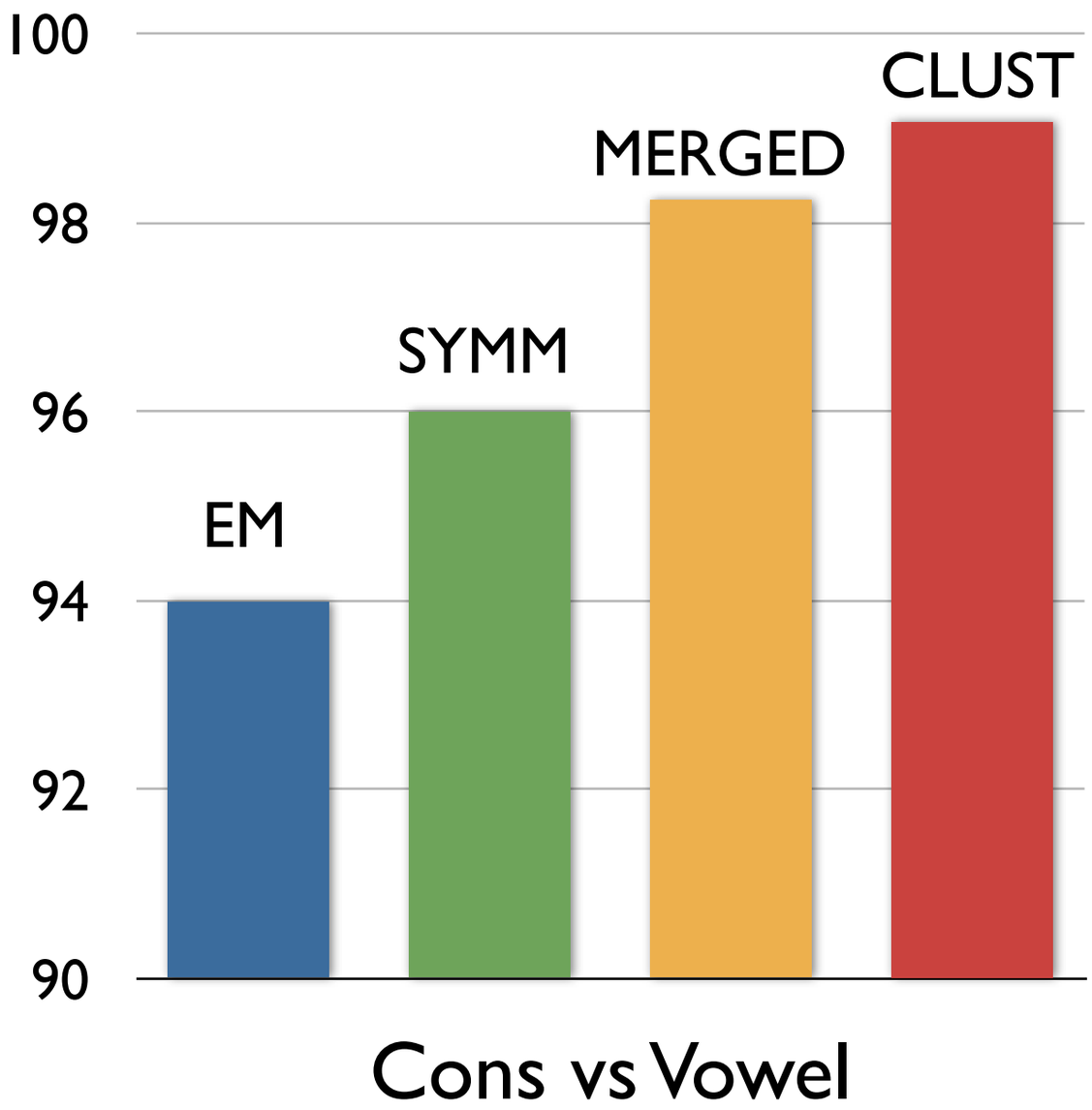
Cons / Vowel

Nasal / Cons / Vowel

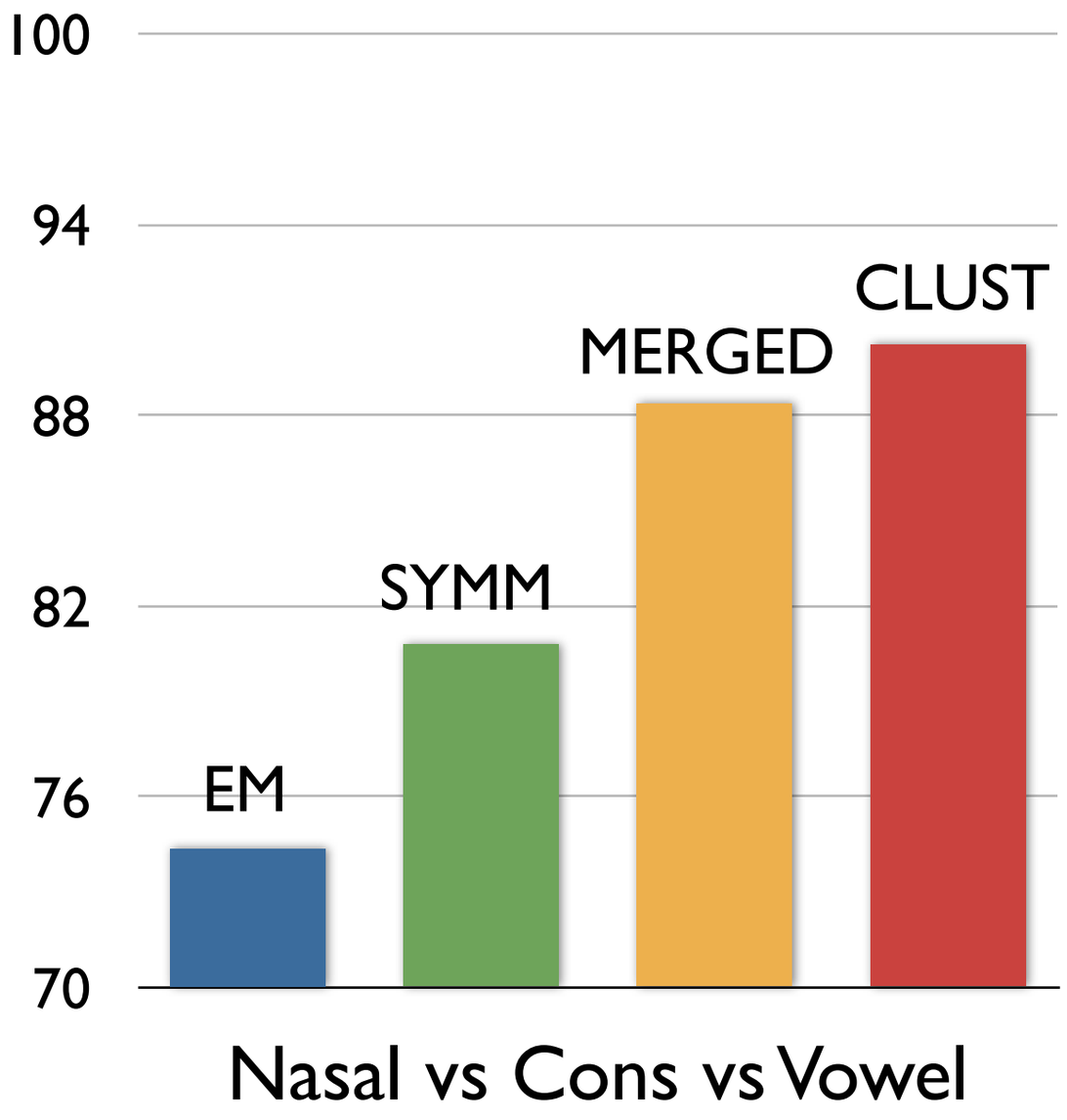
- Baseline:
  - ~ HMM (EM) [Knight 2006] ← oracle mappings
- Hard Bayesian HMM
  - ~ SYMM: symmetric fixed priors ← oracle mappings
  - ~ MERGE: uses 502 observed languages
  - ~ CLUST: 20 latent language clusters

Results

- 503 alphabetic languages



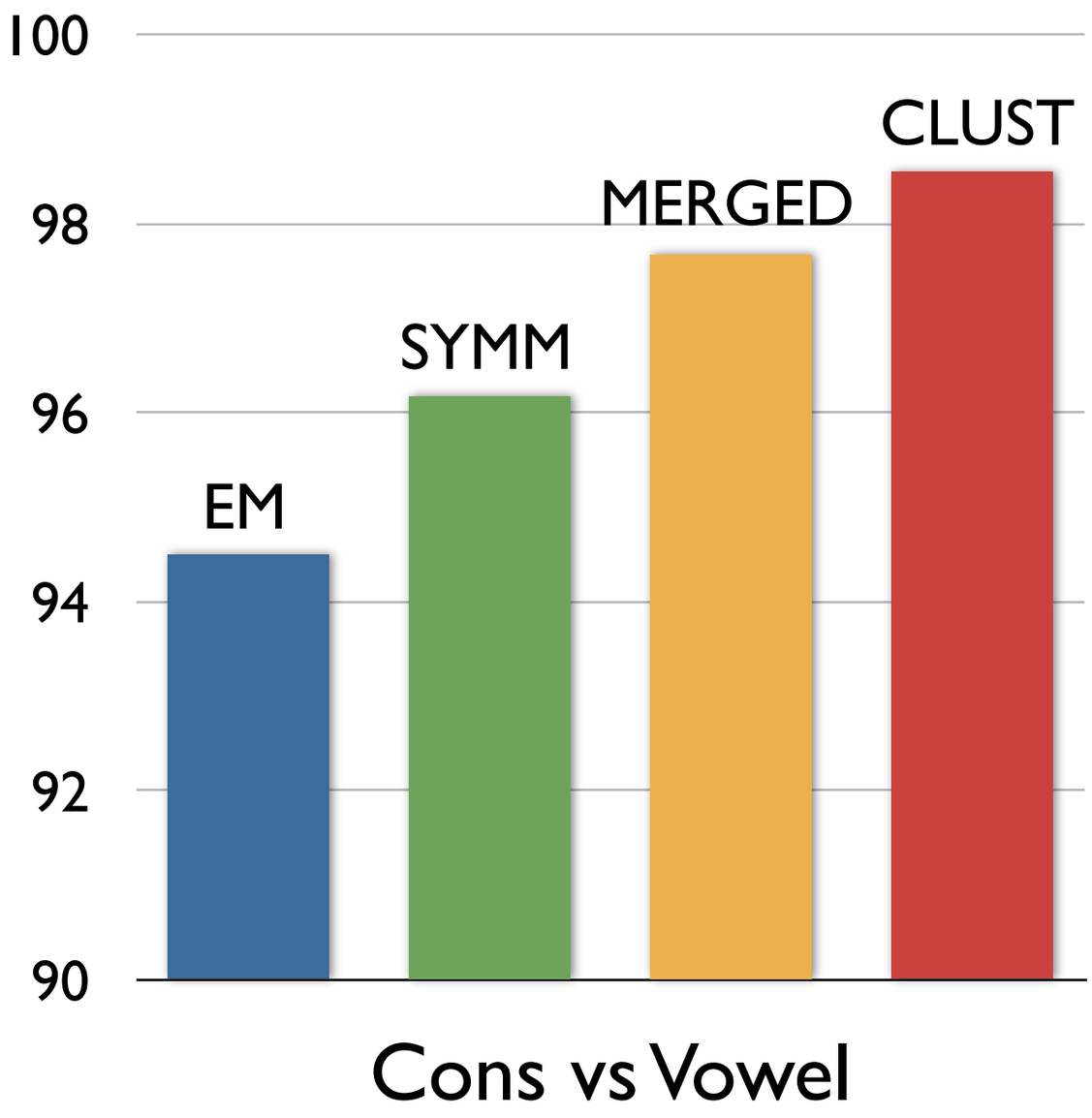
Cons vs Vowel



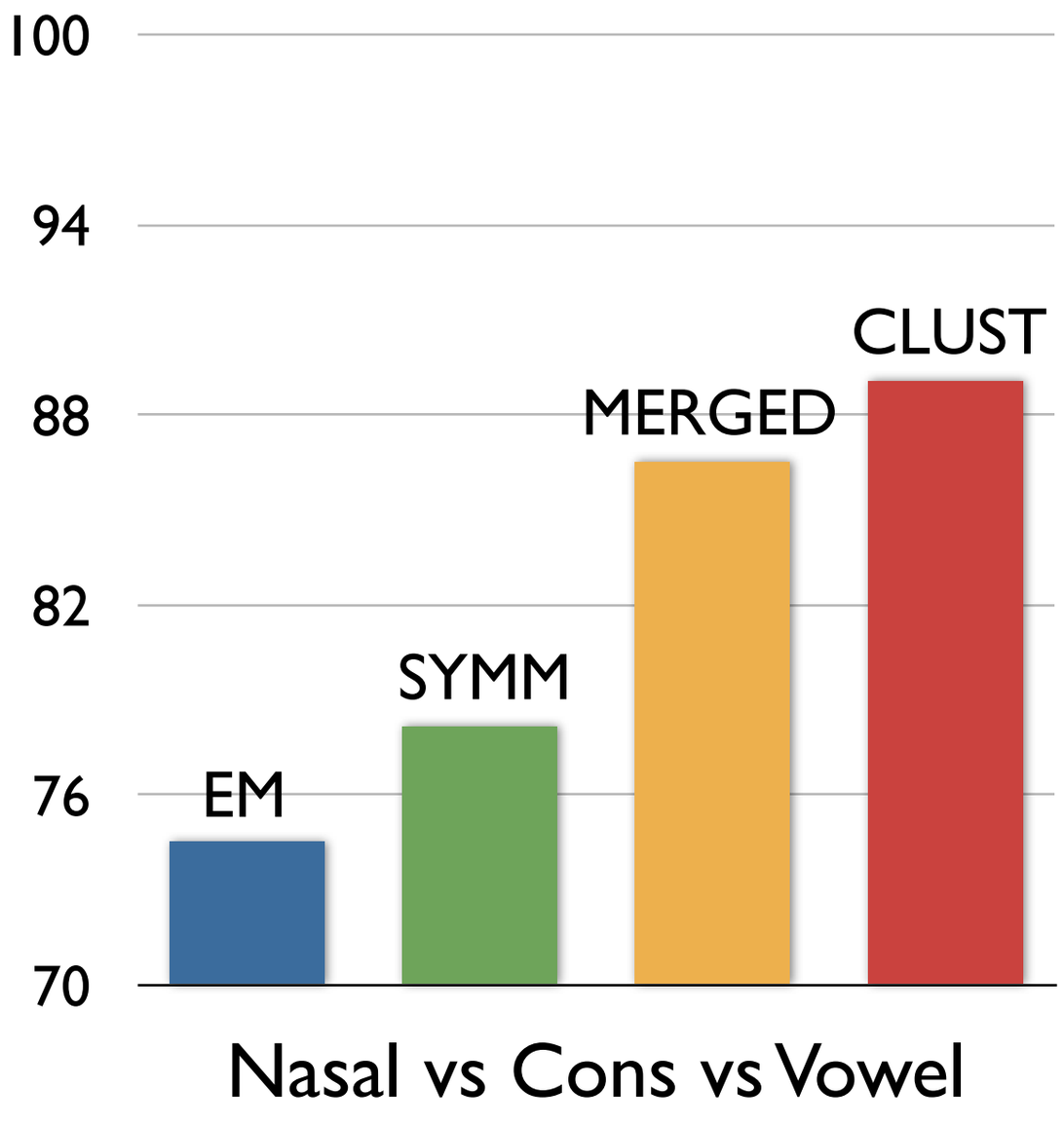
Nasal vs Cons vs Vowel

Results

- 30 language isolates

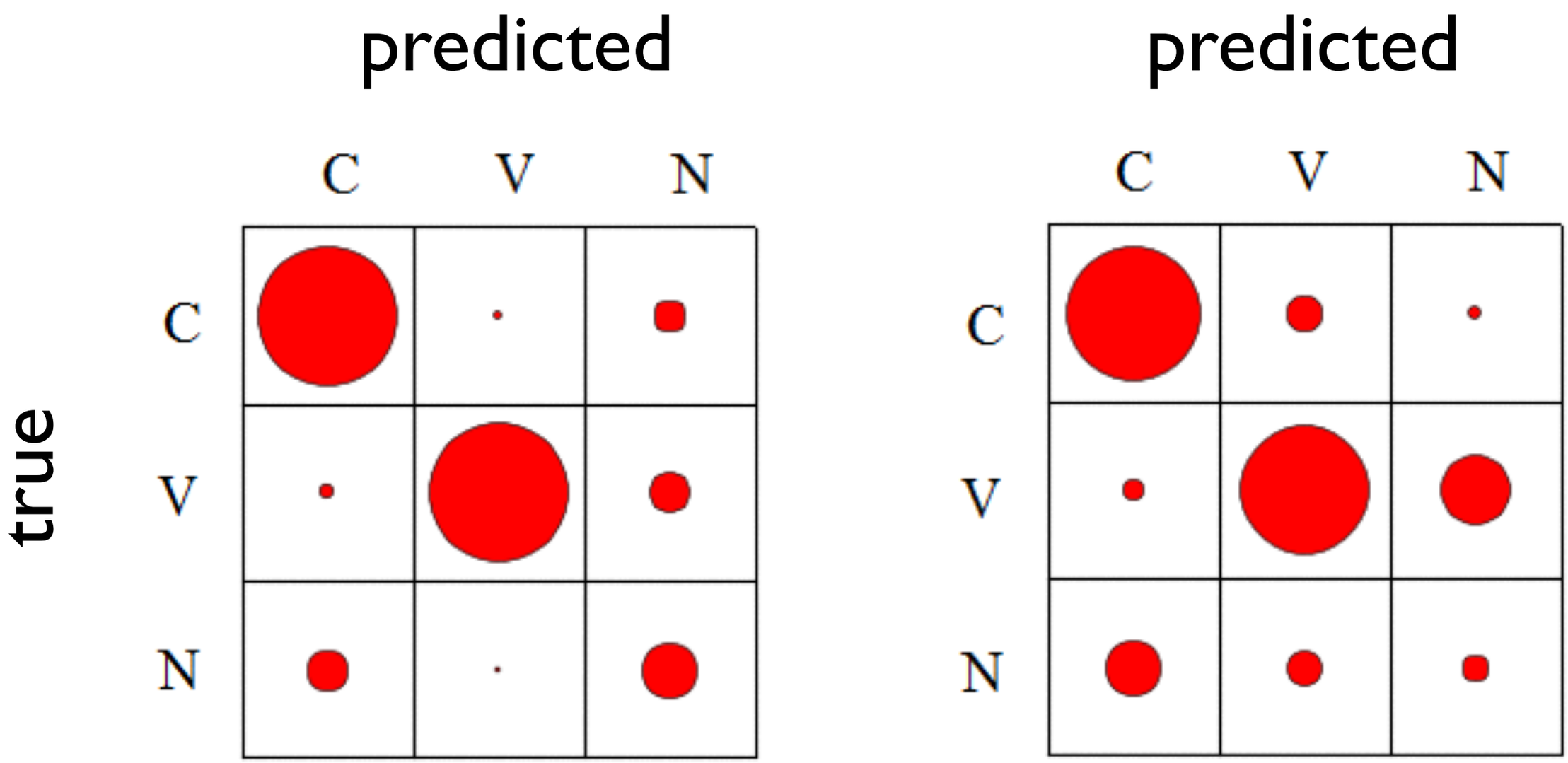


Cons vs Vowel



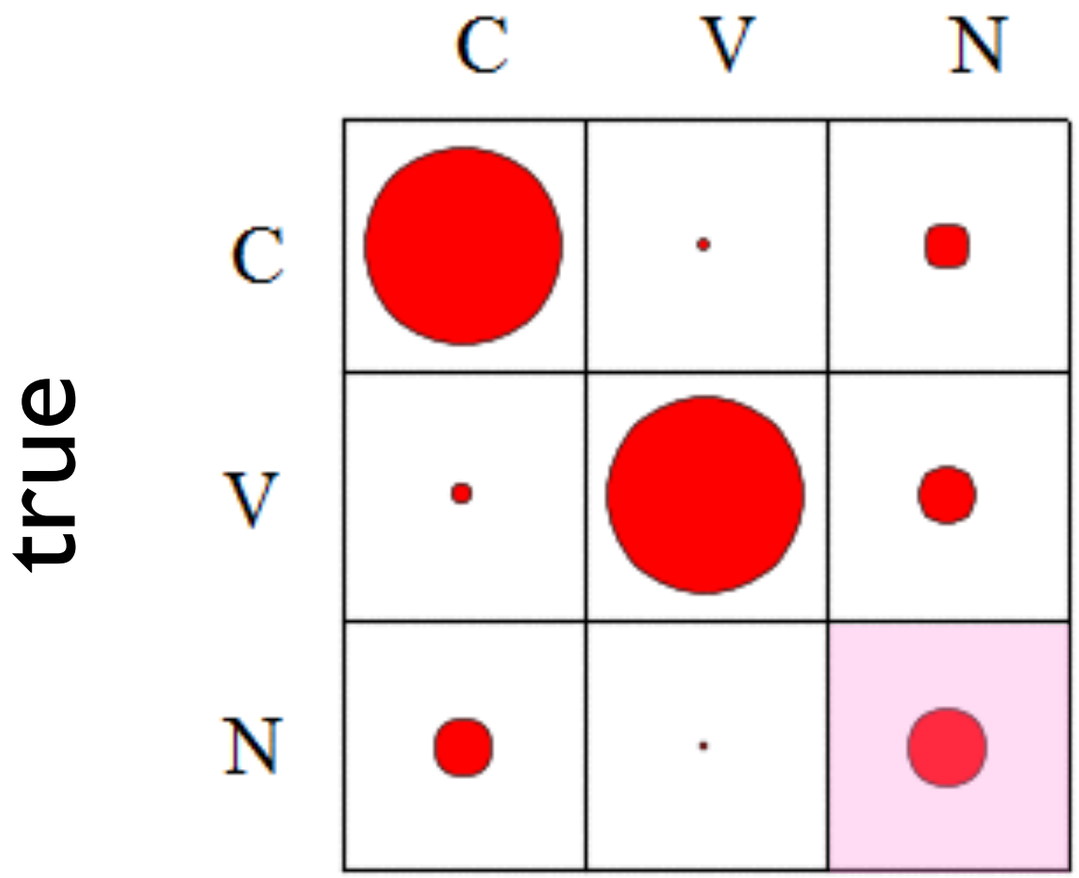
Nasal vs Cons vs Vowel

Confusion matrix



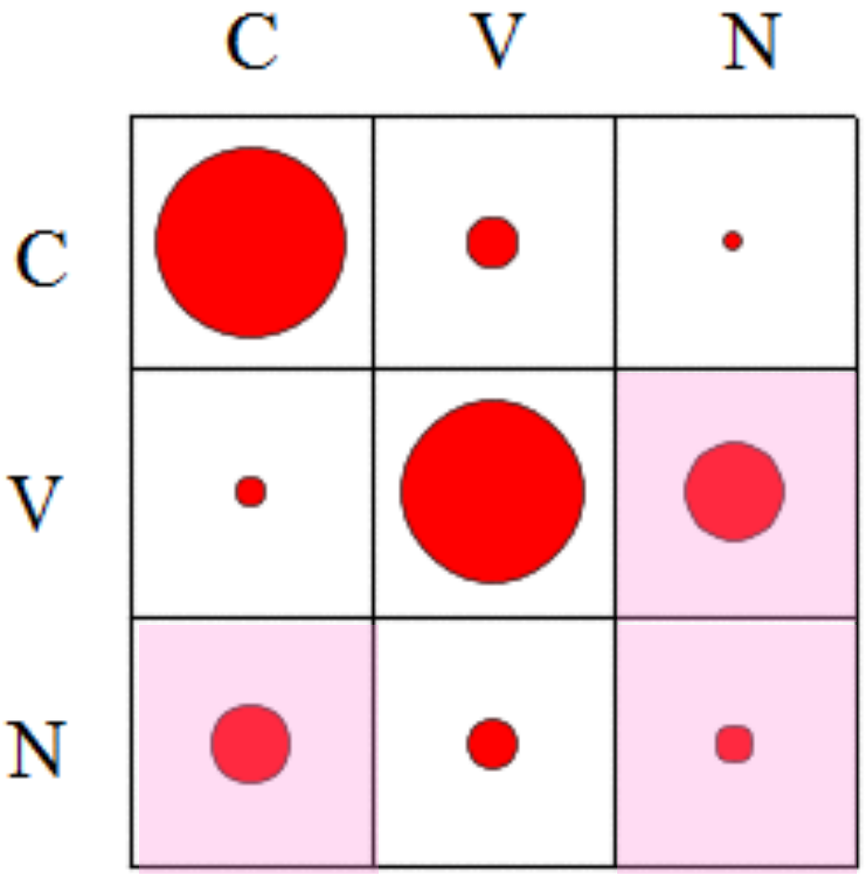
Confusion matrix

predicted



CLUST

predicted



baseline

Cluster Analysis

Plurality Family	*Proportion (%)*
Indo-European	38 24 21
Quechuan	**89**
Mayan	64
Oto-Manguean	55
Maipurean	25
Tucanoan	20
Uto-Aztecan	40
Altaic	44

Plurality Family	*Proportion (%)*
Austronesian	**91** 71 24
Niger-Congo	**100** 78 74 68 67 50 24

Future Work

- Finer-grained phonetic categories
- Syllable writing systems
  - ~ Syllables as hidden variables in our data
- Apply to Linear A

Thank you!