Adding More Languages Improves Unsupervised Multilingual Tagging

A Bayesian Non-Parametric Approach

Benjamin Snyder, Tahira Naseem
Jacob Eisenstein and Regina Barzilay
MIT

- Languages exhibit variations in patterns of ambiguity
- Multilingual cues as natural supervison

בראשית ברא אלהים את השמים ואת הארץ

في البداء خلق الله السموات والارض

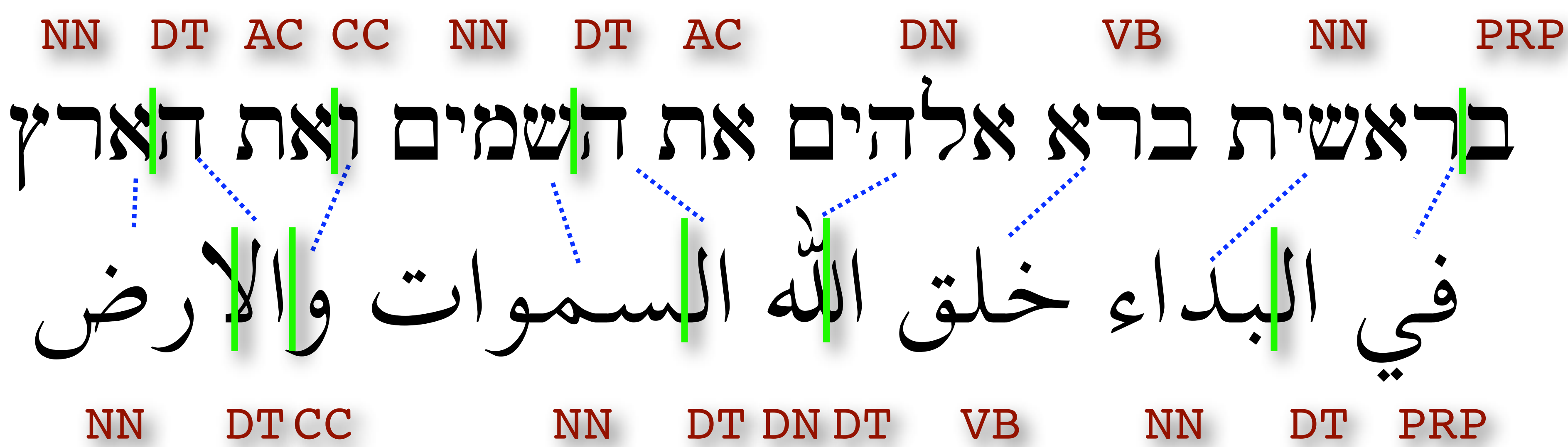
(acl 2008)

בראשית ברא אלהים את השמים ואת הארץ

في البداء خلق الله السموات والارض

(acl 2008)

(emnlp 2008)



(acl 2008)

(emnlp 2008)

(acl 2009)

This talk

NN DT AC CC NN DT AC DN VB NN PRP

בראשית ברא אלהים את השמים ואת הארץ

في البداء خلق الله السموات والارض

NN DT CC NN DT DN DT VB NN DT PRP

NN DT AC CC NN DT AC DN VB NN PRP

בראשית ברא אלהים את השמים ואת הארץ

ܒܪܫܝܬ ܒܪܐ ܐܠܗܐ ܝܬ ܫܡܝܐ ܘܝܬ ܐܪܥܐ

في البداء خلق الله السموات والارض

NN DT CC NN DT DN DT VB NN DT PRP

NN DT AC CC NN DT AC DN VB NN PRP

בראשית ברא אלהים את השמים ואת הארץ

ܒܪܫܝܬ ܒܪܐ ܐܠܗܐ ܝܬ ܫܡܝܐ ܘܝܬ ܐܪܥܐ

In the beginning God created the heaven and the earth

ܒܪܫܝܬ ܒܪܐ ܐܠܗܐ ܝܬ ܫܡܝܐ ܘܝܬ ܐܪܥܐ

في البداء خلق الله السموات والارض

NN DT CC NN DT DN DT VB NN DT PRP

NN DT AC CC NN DT AC DN VB NN PRP

בראשית ברא אלהים את השמים ואת הארץ

ܒܪܫܝܬ ܒܪܐ ܐܠܗܐ ܝܬ ܫܡܝܐ ܘܝܬ ܐܪܥܐ

આદિએ દેવે આકાશ તથા પૃથ્વી ઉત્પન્ન કર્યા

In the beginning God created the heaven and the earth

начале сотворил Бог небо и землю

ܒܪܫܝܬ ܒܪܐ ܐܠܗܐ ܝܬ ܫܡܝܐ ܘܝܬ ܐܪܥܐ

في البداء خلق الله السموات والارض

NN DT CC NN DT DN DT VB NN DT PRP

NN DT AC CC NN DT AC DN VB NN PRP

בראשית ברא אלהים את השמים ואת הארץ

ܒܪܫܝܬ ܒܪܐ ܐܠܗܐ ܝܬ ܫܡܝܐ ܘܝܬ ܐܪܥܐ

Au commencement, Dieu créa le ciel et la terre

આદિએ દેવે આકાશ તથા પૃથ્વી ઉત્પન્ન કર્યાં

In the beginning God created the heaven and the earth

начале сотворил Бог небо и землю

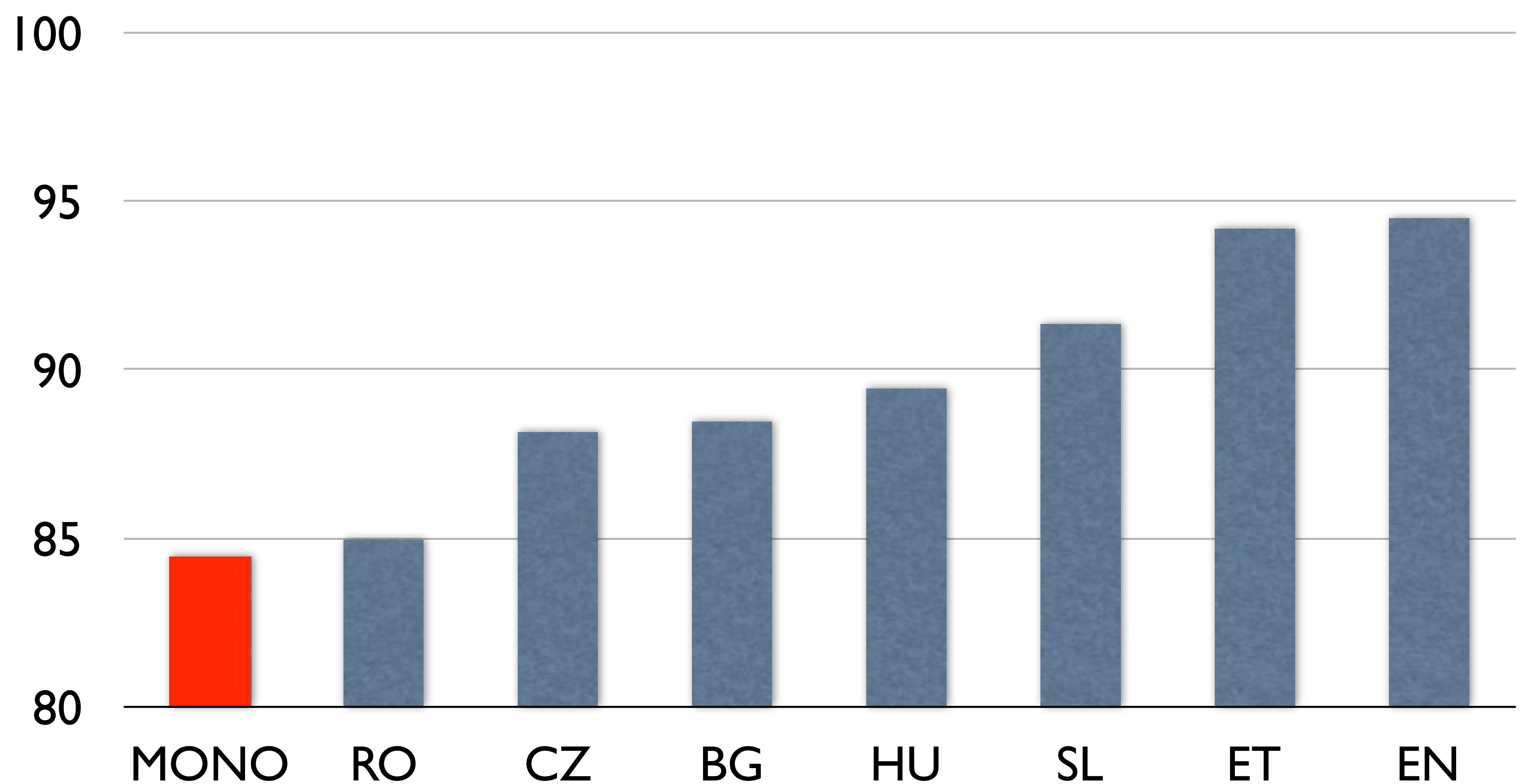
๑ในปฐมกาล พระเจาทรงเนรมิตสรางฟา

ܒܪܫܝܬ ܒܪܐ ܐܠܗܐ ܝܬ ܫܡܝܐ ܘܝܬ ܐܪܥܐ

في البداء خلق الله السموات والارض

NN DT CC NN DT DN DT VB NN DT PRP

Serbian, paired with...



Bilingual Model [Snyder et al 2008]

Multilingual Models

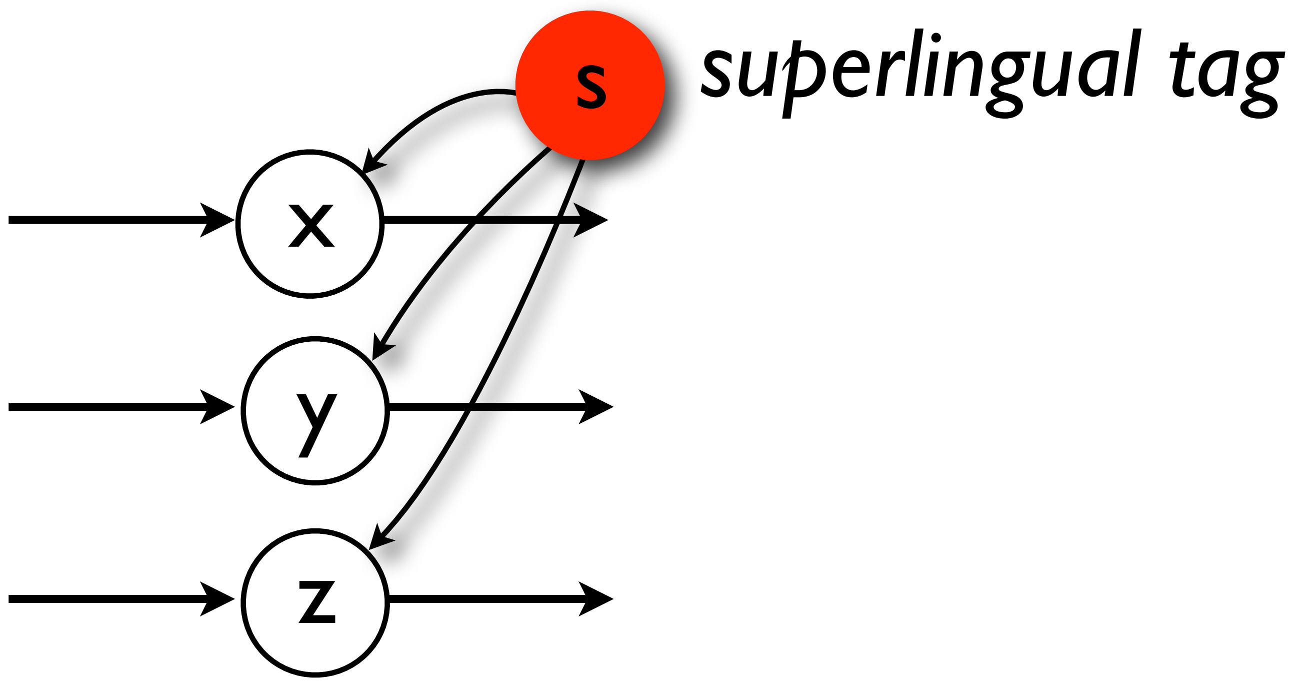
<u>Benefits:</u>

- Fully exploit large multiparallel corpora
- Benefit from richer set of multilingual cues

<u>Challenges</u>:

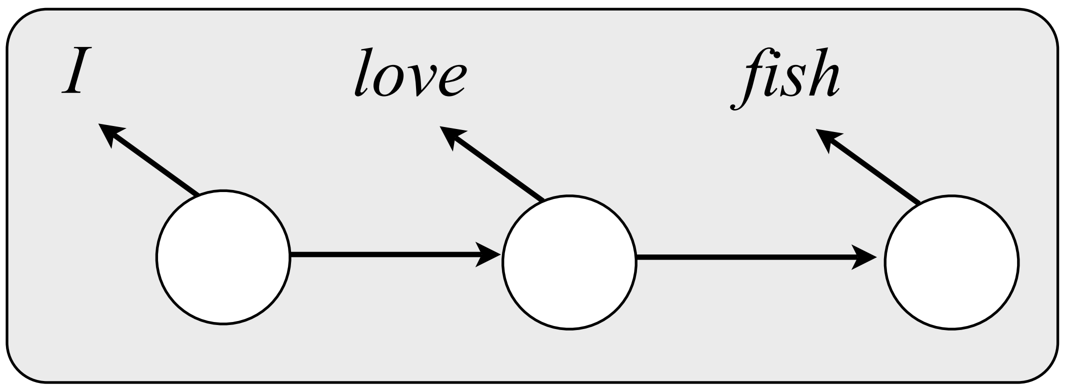
- Must allow for full diversity of language variation
- Must scale well with number of languages

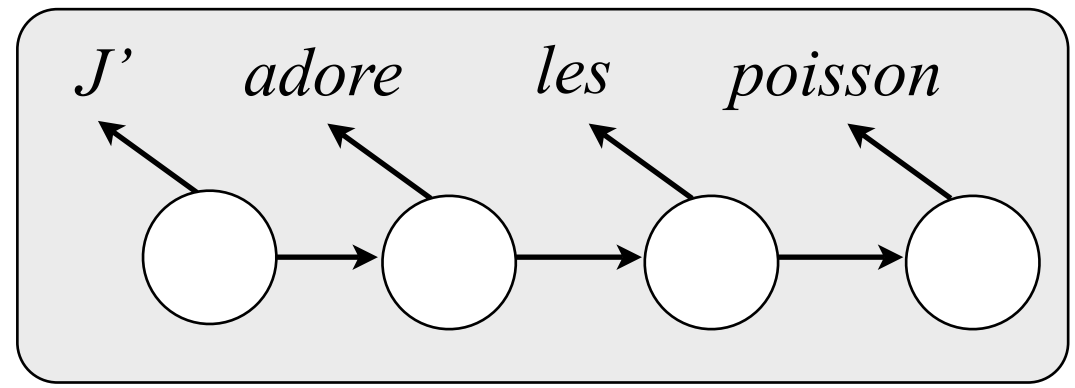
Latent Variable Parameterization

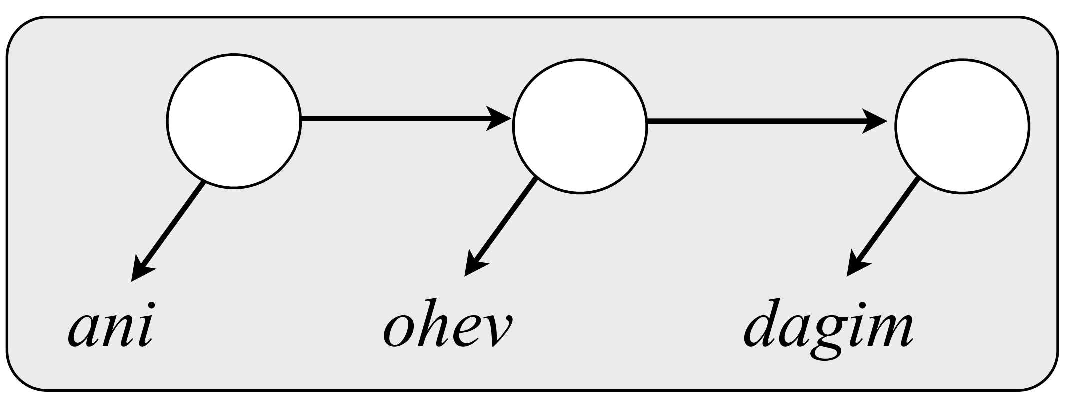


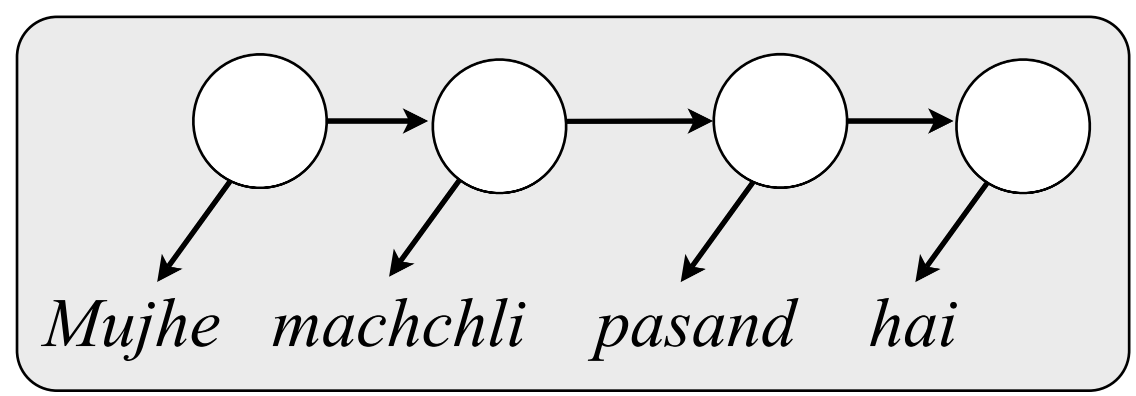
- *Scales linearly with number of languages*
- *Trade off between language-specific and multilingual cues*

Constructing the Model

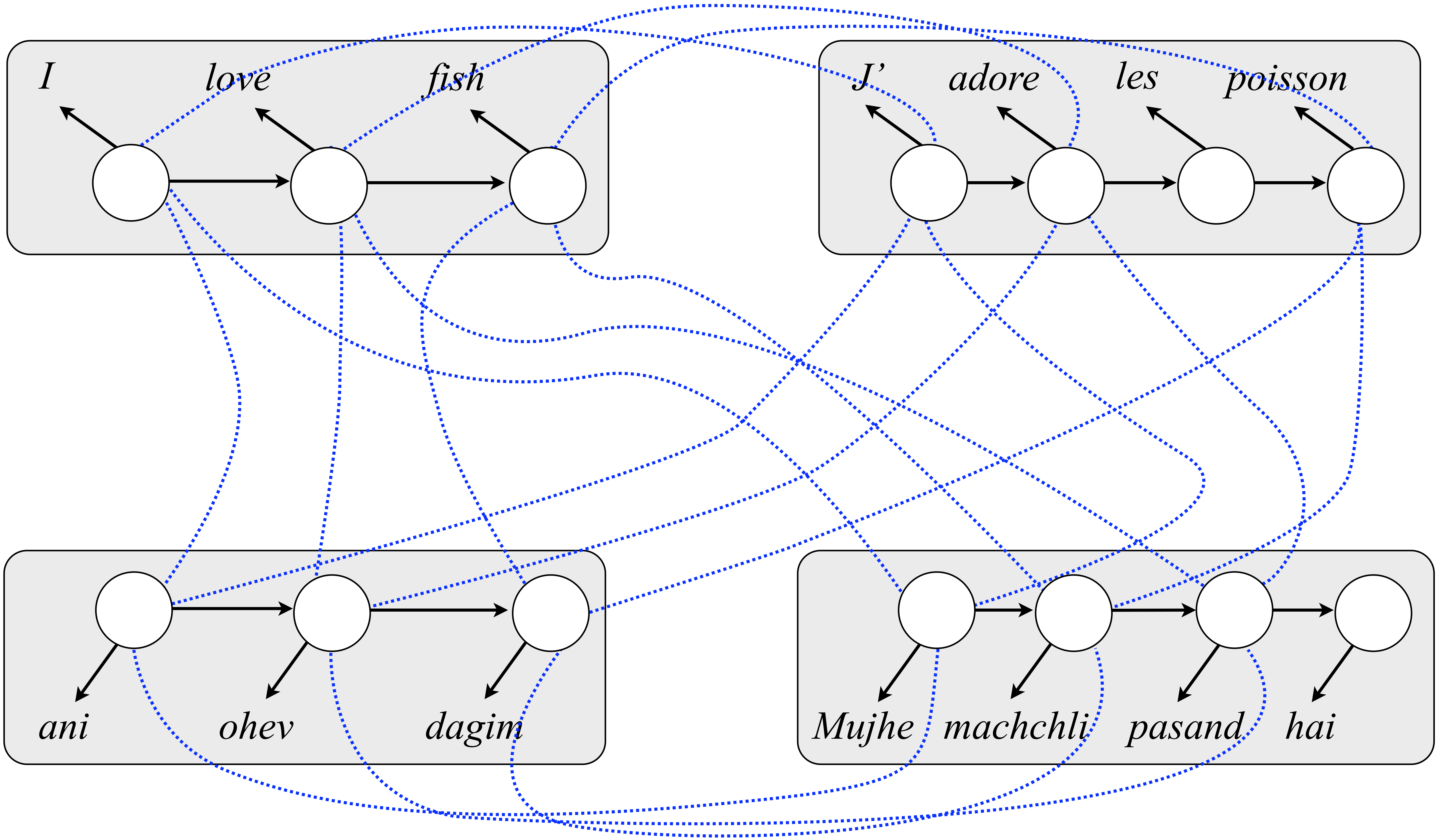

I
love
fish


J'
adore
les
poisson

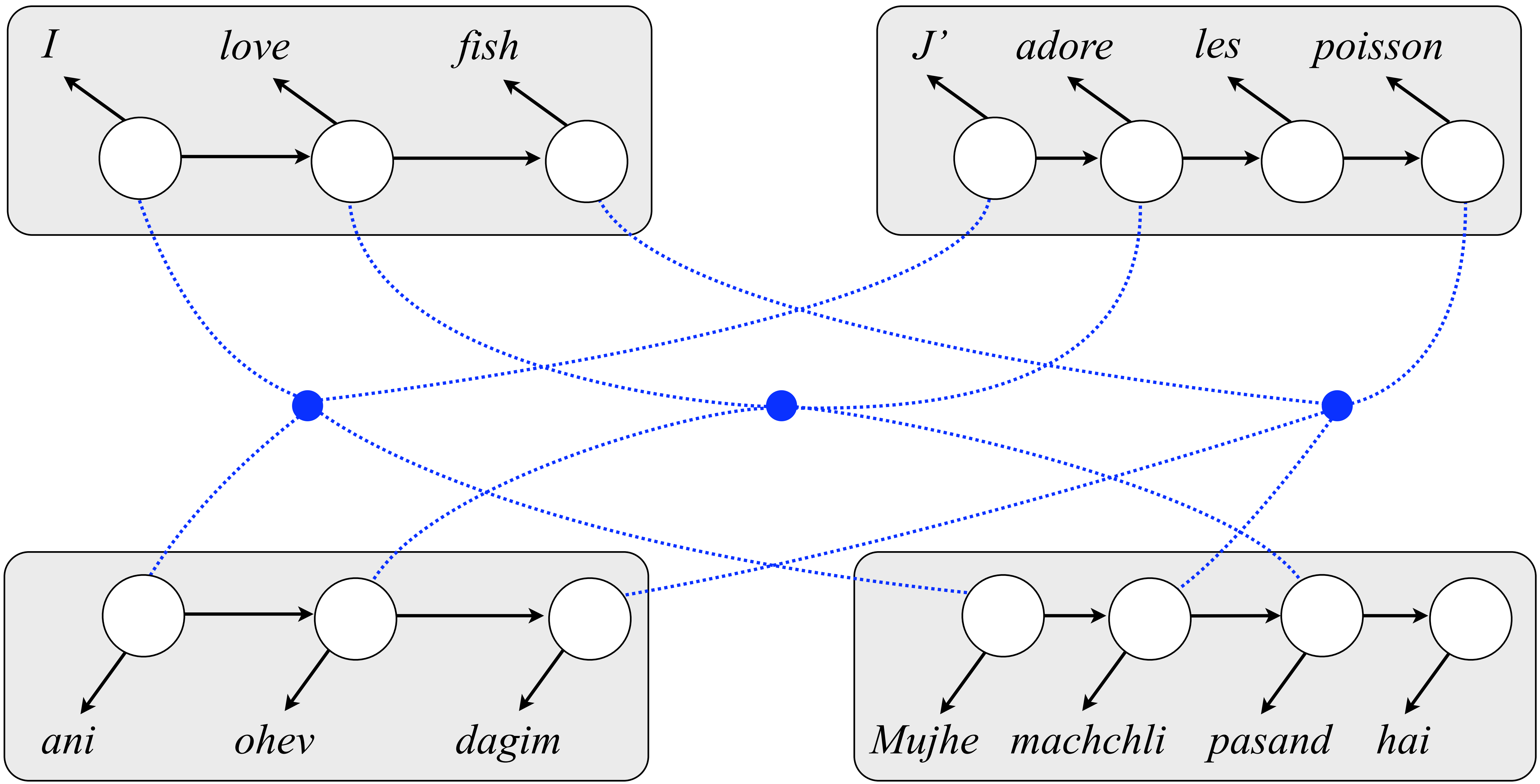

ani
ohev
dagim


Mujhe
machchli
pasand
hai

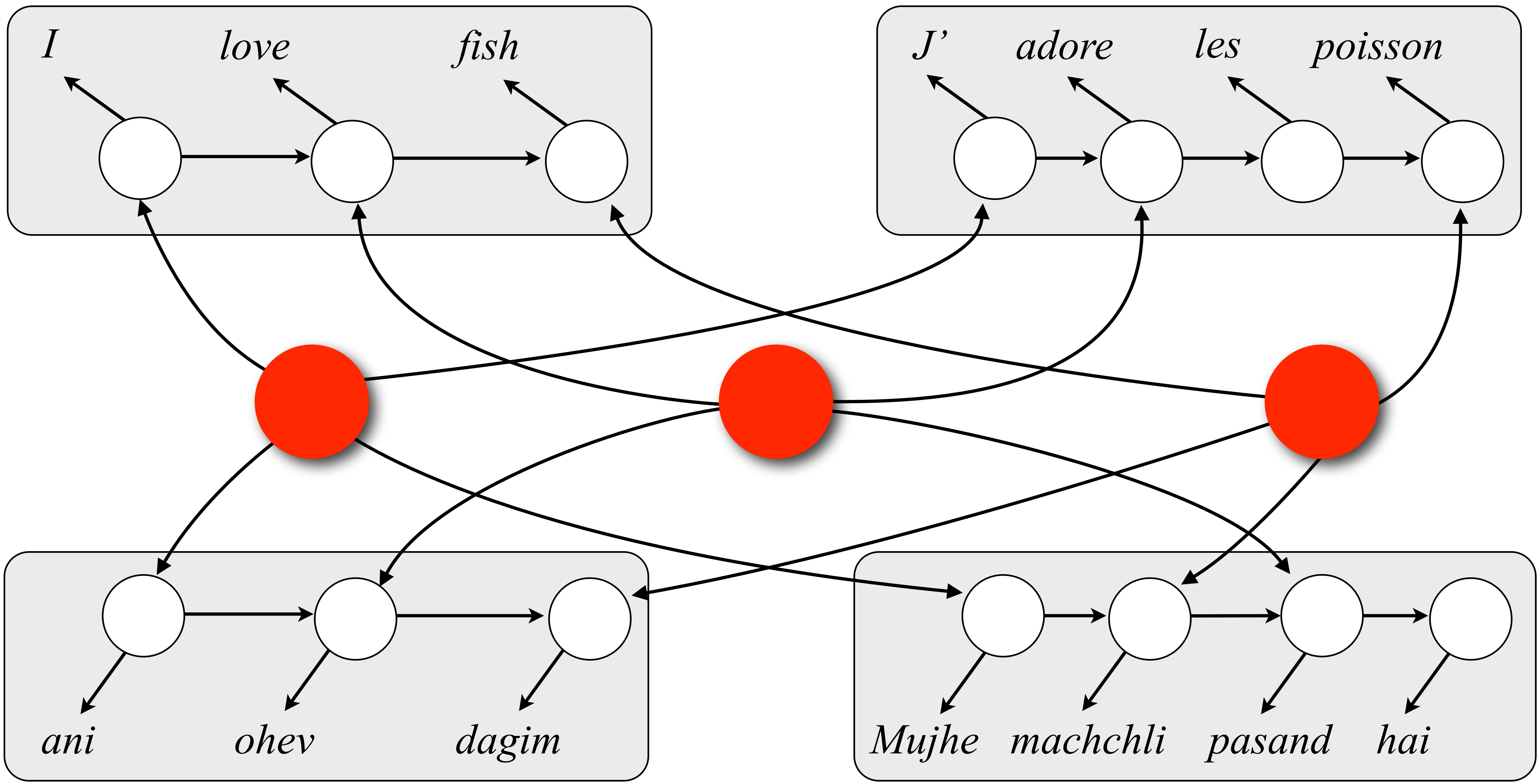
1. Gather lexical alignments (*giza++*)



2. Aggregate lexical alignments



3. Place *superlingual tag* on each clique



Superlingual Tags

- Each superlingual tag value s:
  - captures a multilingual context
  - indexes tag distribution for each language

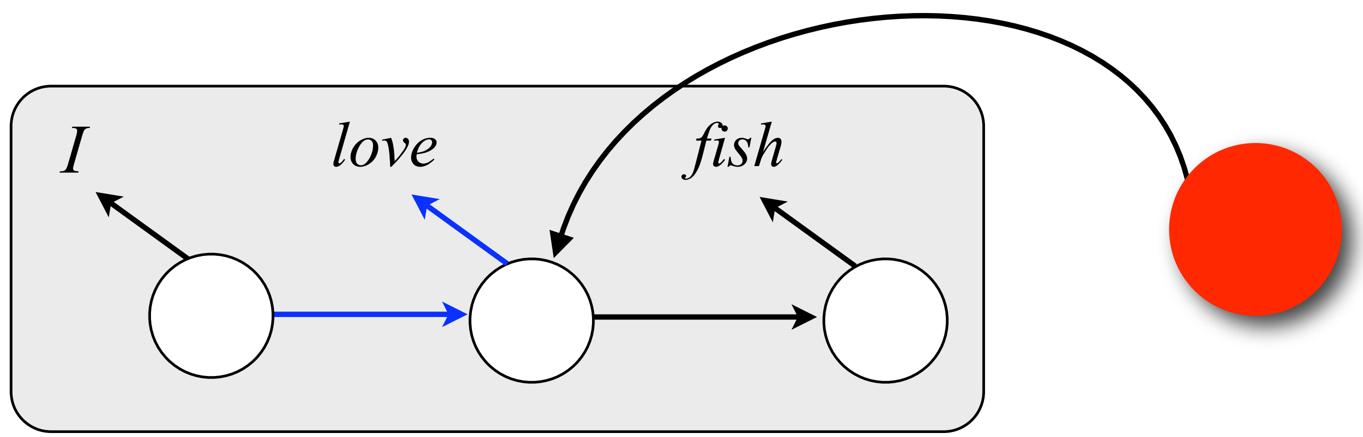
$$\Psi_s = \{\psi_s{}^1, \ldots, \psi_s{}^\ell\}$$

e.g. Superlingual tag value "2" may index distributions which prefers nouns across languages

- *Infinite* number of superlingual tag values
  - Dirichlet Process prior to find clusters of repeated multilingual patterns

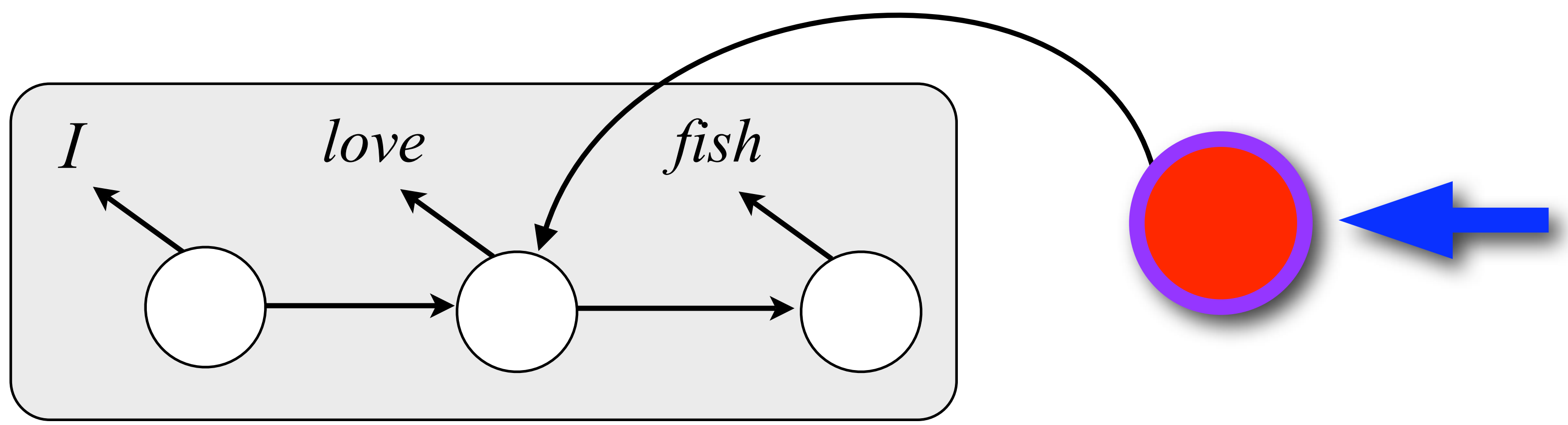
The Generative Story

Generative Story: *parameters*



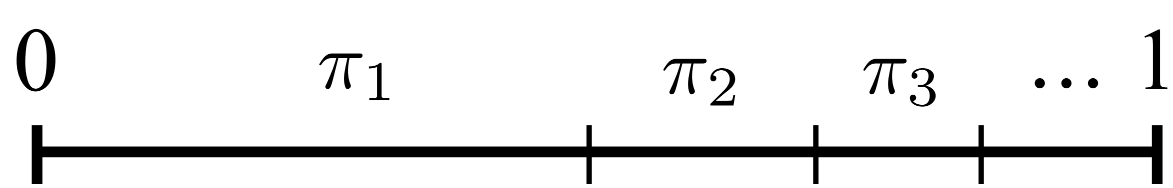
- HMM *transitions* and *emissions* from Dirichlet priors

Generative Story: *parameters*

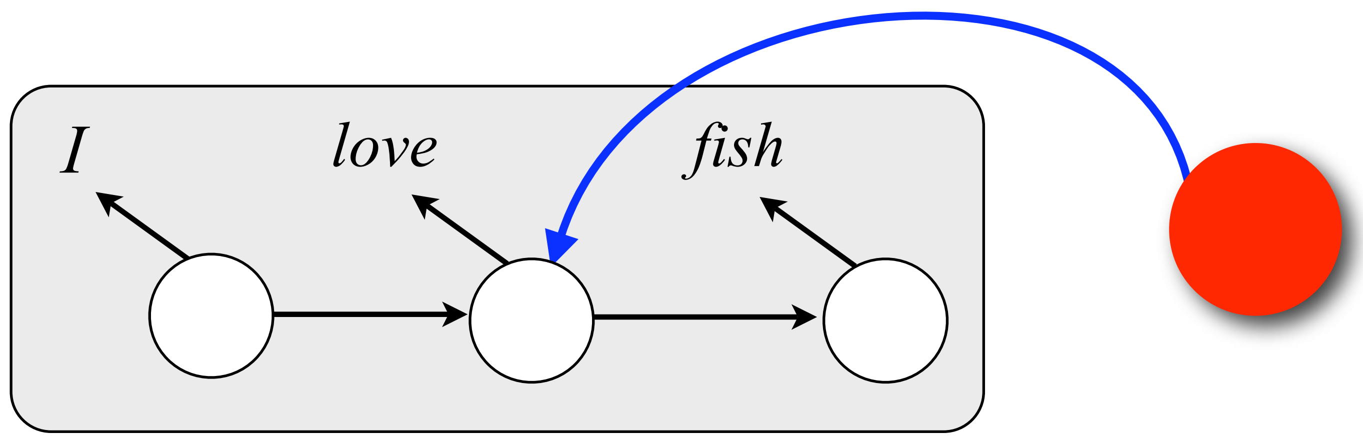


- HMM *transitions* and *emissions* from Dirichlet priors
- *Superlingual tag probabilities*: infinite sequence of mixing parameters $\pi_1, \pi_2, \ldots$ from stick breaking process

$$\pi_{s} : \text{ prob of superlingual tag } s$$



Generative Story: *parameters*



- HMM *transitions* and *emissions* from Dirichlet priors
- *Superlingual tag probabilities*: infinite sequence of mixing parameters $\pi_1, \pi_2, \ldots$ from stick breaking process
- Infinite sequence of *sets of tag distributions* from Dirichlet priors: $\Psi_1, \Psi_2, \ldots$

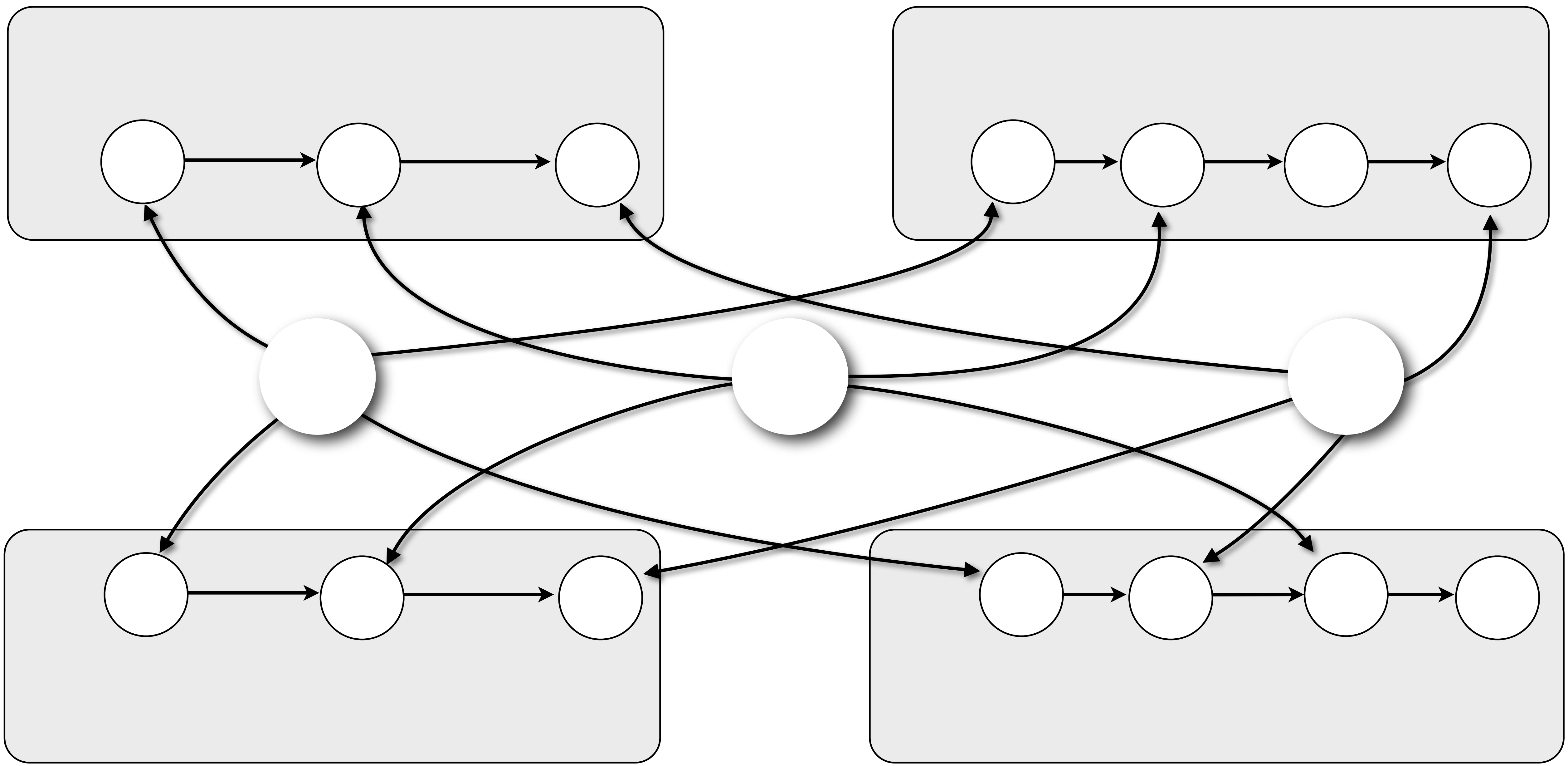
for superlingual tag $\boldsymbol{s}$: $\Psi_s = \{\psi_s^1, \ldots, \psi_s^\ell\}$

tag dist. for each lang.

Generative Story: *sentences*

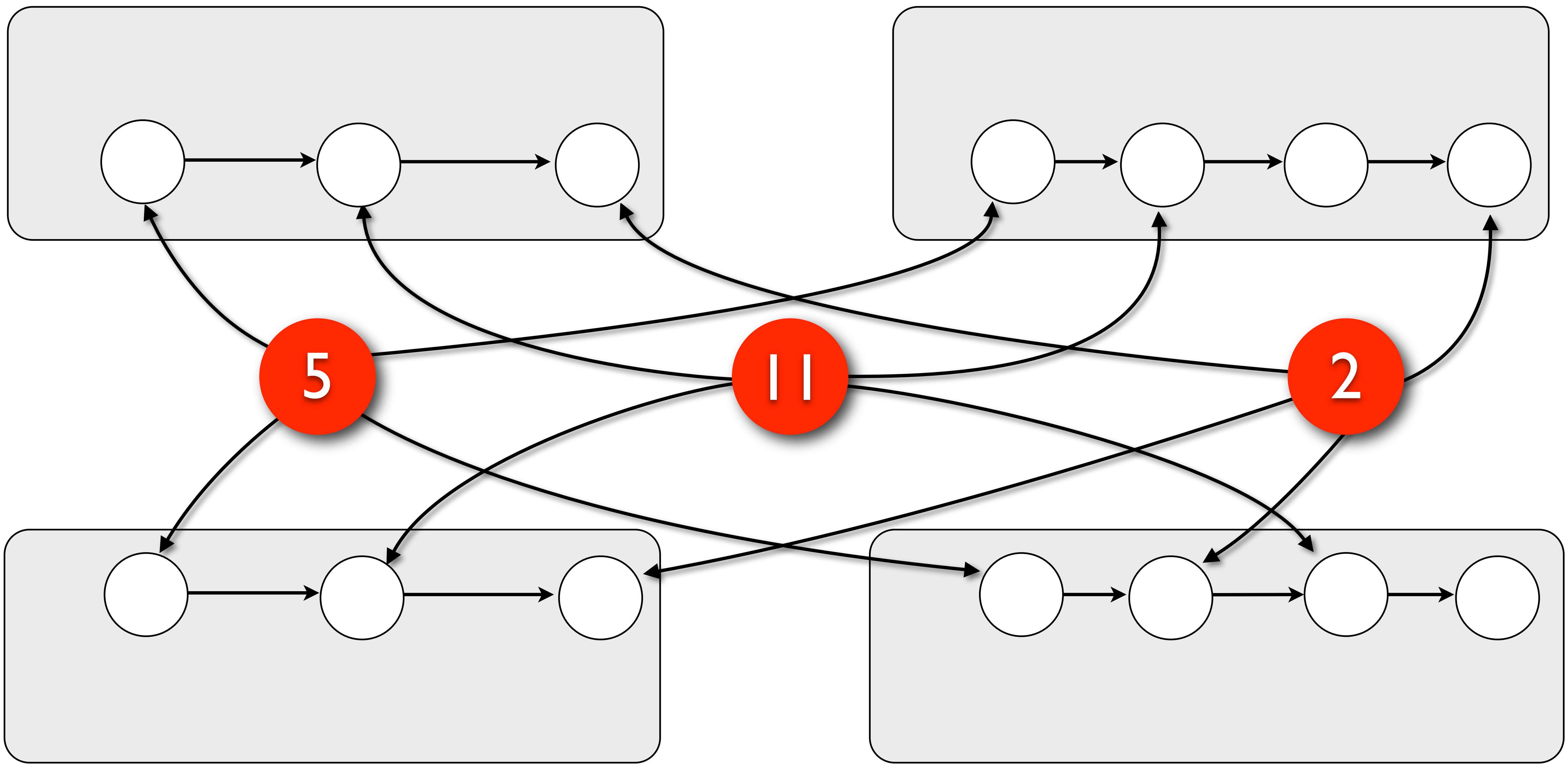
1. Draw *alignment template:*

[1,1,1,1]
[3,3,2,4]
[2,2,3,_]



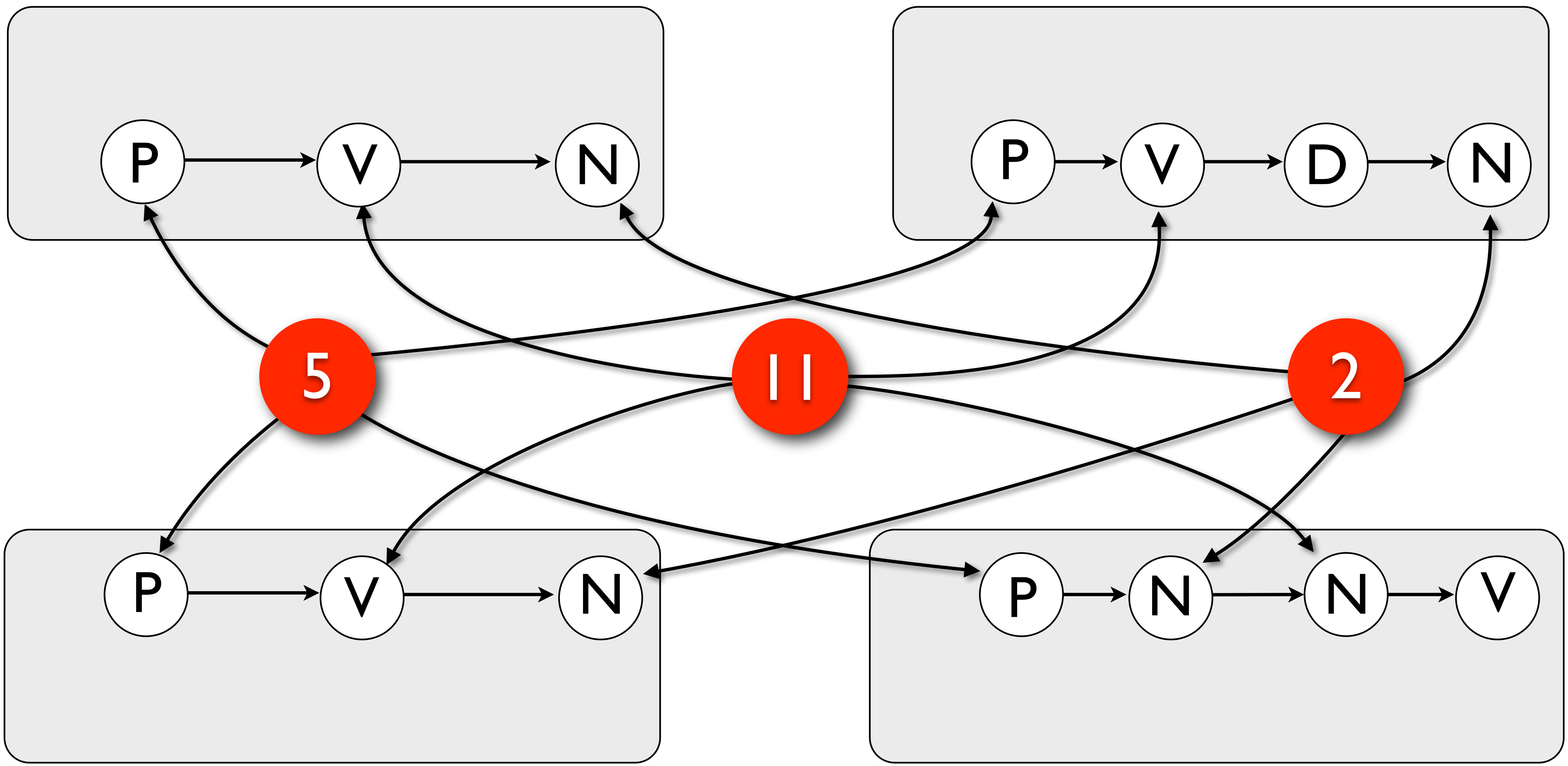
Generative Story: *sentences*

2. Draw *superlingual tags:* $s_i \sim \boldsymbol{\pi}$



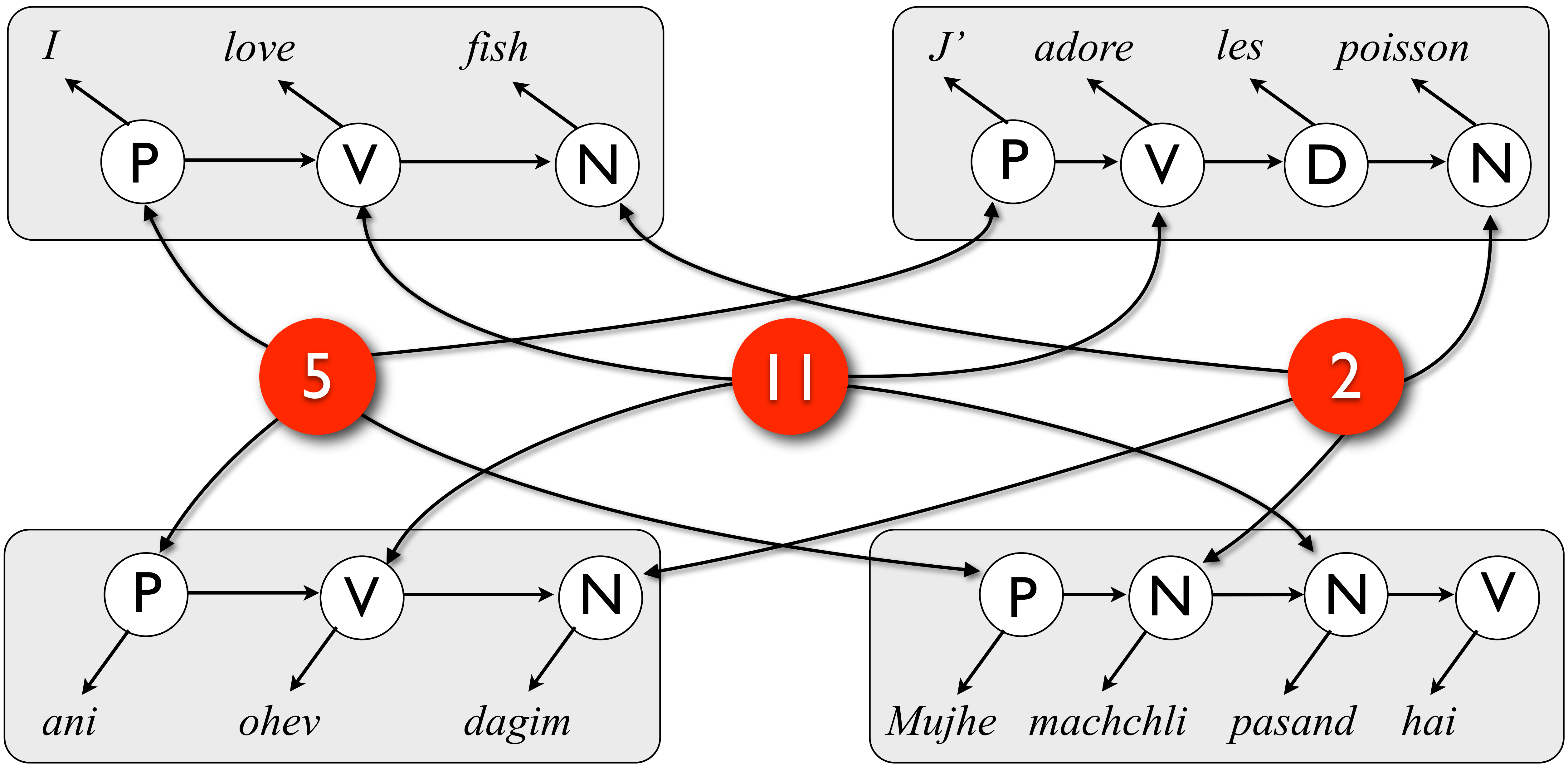
Generative Story: *sentences*

3. Draw *POS tags:* $y_i \sim \frac{trans(y_i|y_{i-1}) \cdot \ \psi_s^\ell(y_i)}{Z}$



Generative Story: *sentences*

4. Emit words: $x_i \sim emit(x_i|y_i)$



Inference: Gibbs Sampling

- Marginalize over emission, transition, and superlingual tag distributions using standard closed forms.
- Explicitly sample each *POS tag* and *superlingual tag*, conditioned on others

Sampling POS Tags

$$P(y_i^\ell | \mathbf{y}_{-(\ell,i)}, \mathbf{x}, \mathbf{a}, \mathbf{s}) \propto$$
$$P(x_i^\ell | \mathbf{x}_{-i}^\ell, \mathbf{y}^\ell) P(y_{i+1}^\ell | y_i^\ell, \mathbf{y}_{-(\ell,i)}, \mathbf{a}, \mathbf{s}) P(y_i^\ell | \mathbf{y}_{-(\ell,i)}, \mathbf{a}, \mathbf{s})$$

Posteriors proportional to:

1. Emission probability of word
2. Probability of next tag

   (given superlingual tags and current tag)
3. Probability of current tag

   (given superlingual tags and previous tag)

Sampling Superlingual Tags

$$P\big(s_i \big| \mathbf{s}_{-i}, \mathbf{y}\big) \propto$$

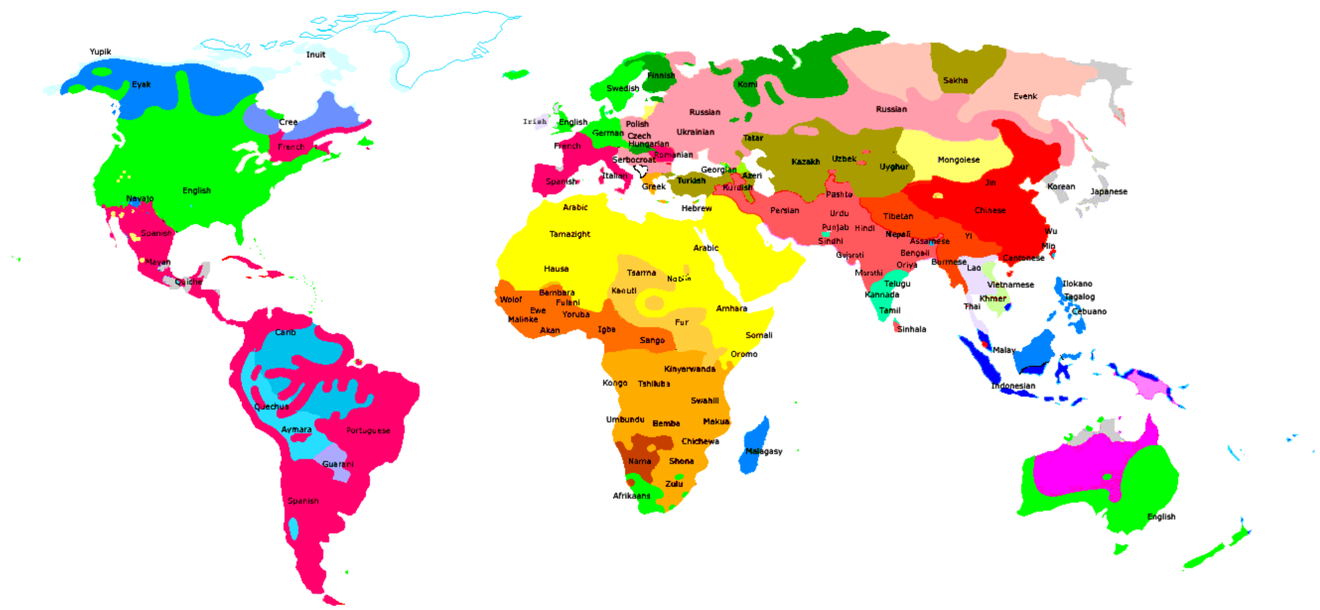
$$\prod_{\ell} P\big(y_i^{\ell} \big| s_i, \mathbf{s}_{-i}, \mathbf{y}_{-(\ell,i)}\big) \cdot \begin{cases} \frac{1}{k+\alpha} count(s_i, \mathbf{s}_{-i}) & \text{if } s_i \in \mathbf{s}_{-i} \\ \frac{\alpha}{k+\alpha} & \text{otherwise} \end{cases}$$

Posteriors proportional to:

1. Probabilities of aligned POS tags

2. Chinese Restaurant Process [Antoniak '74]

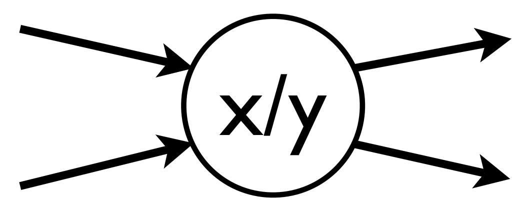
Corpus



- Orwell's <u>Nineteen Eighty Four</u> (~100k words)
  - Bulgarian, Czech, Serbian, Slovene
  - Hungarian, Estonian
  - Romanian
  - English
- 14 coarse POS tags (Multext v3)
- Train on parallel data, evaluate on *monolingual*

Experiments

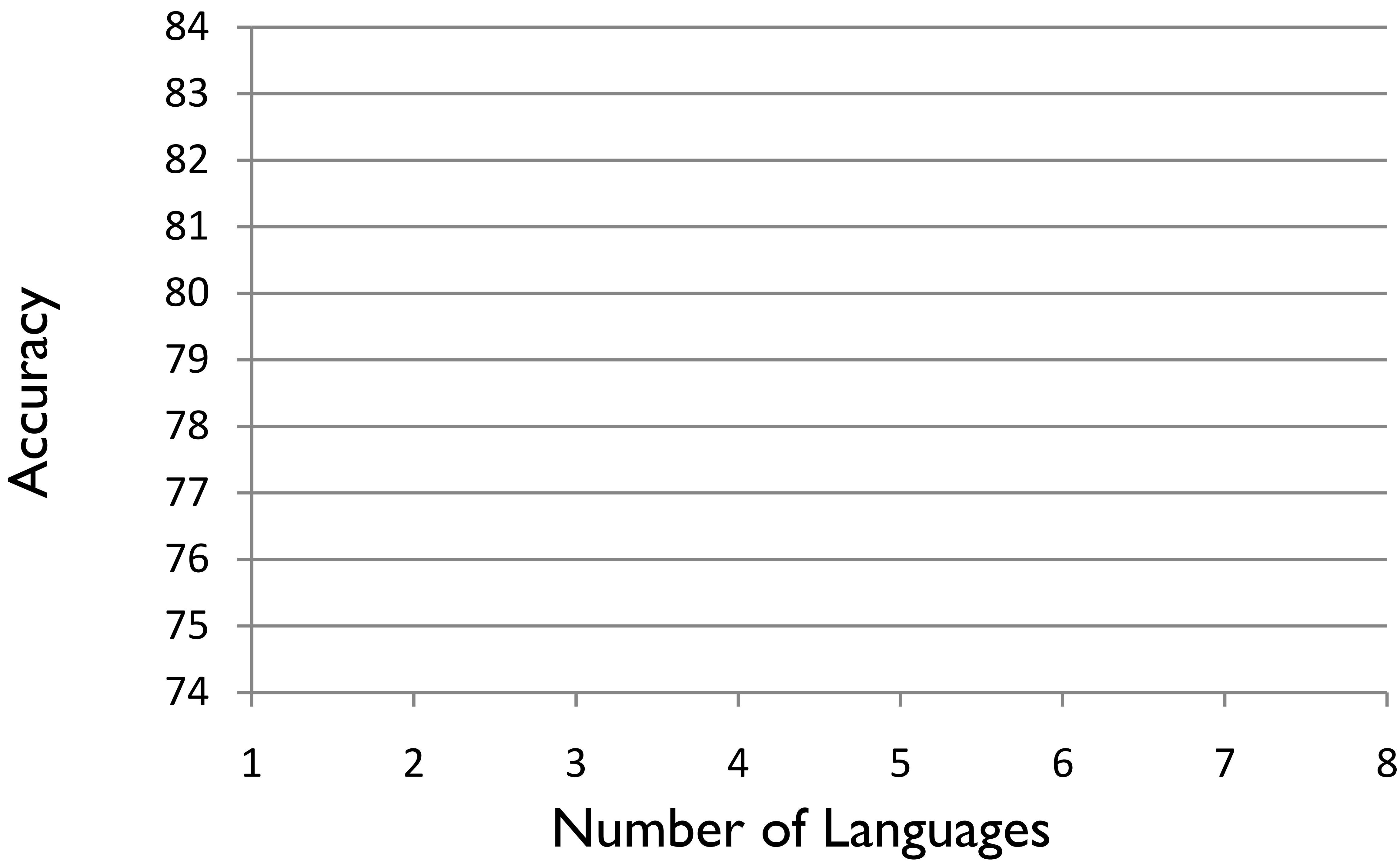
- Baselines:
  1. Monolingual BHMM [Goldwater & Griffiths 2007]
  2. Bilingual model [Snyder et al 2008]
     a. Avg
     b. Oracle



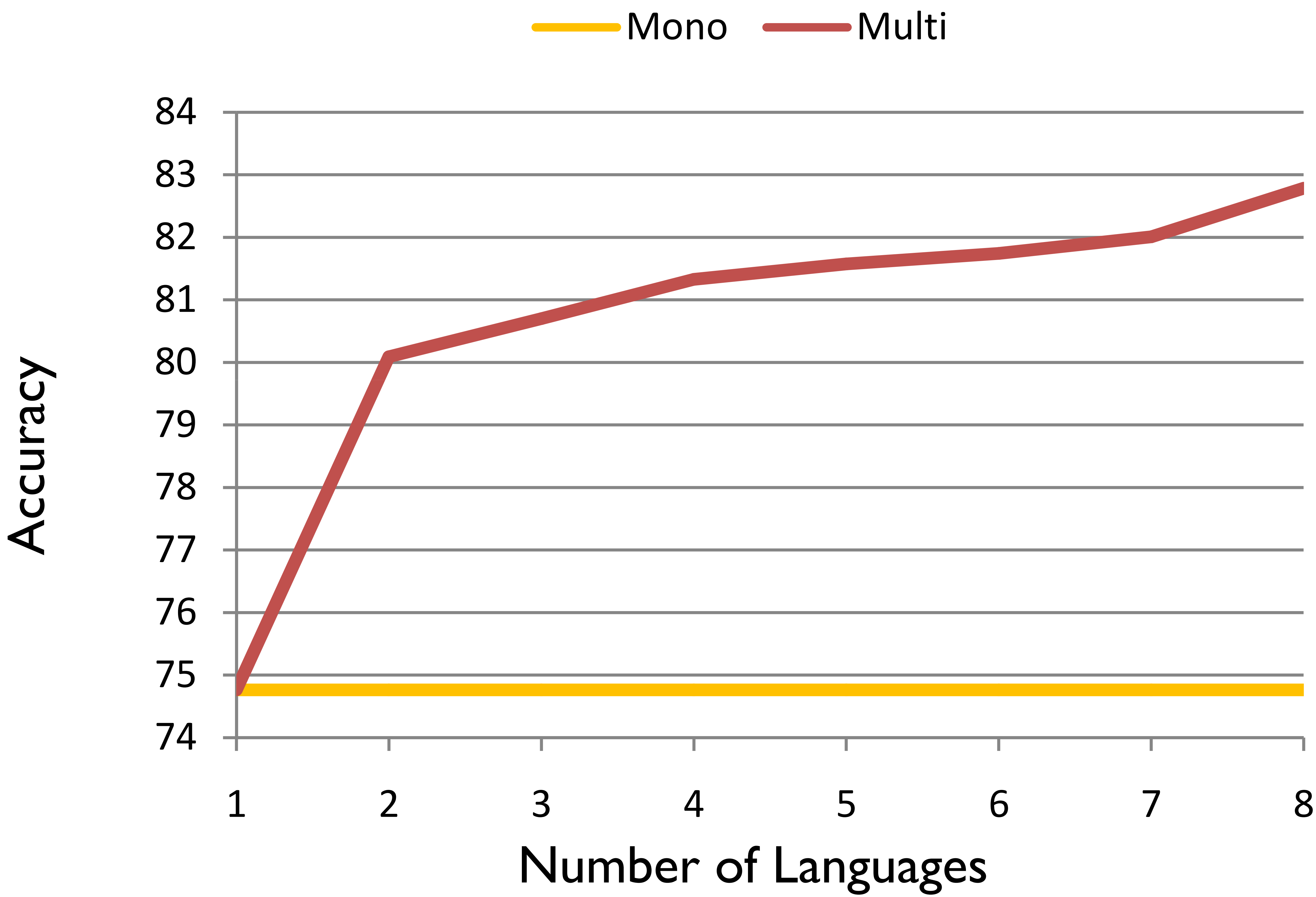
- Three scenarios:

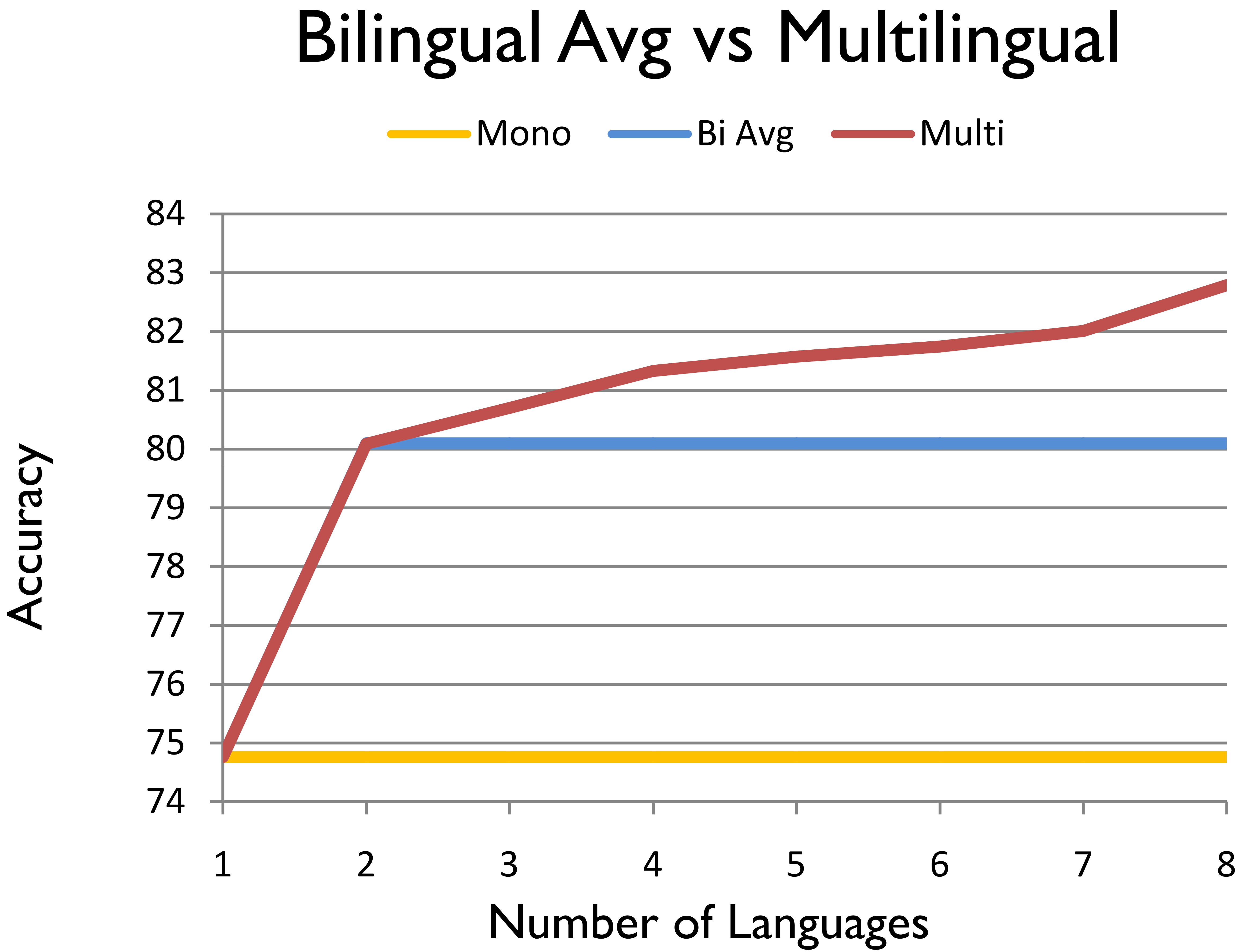
  Full lexicon
  Reduced lexicon: count > 5
  Reduced lexicon: count > 10

Reduced Lexicon: *n > 5*



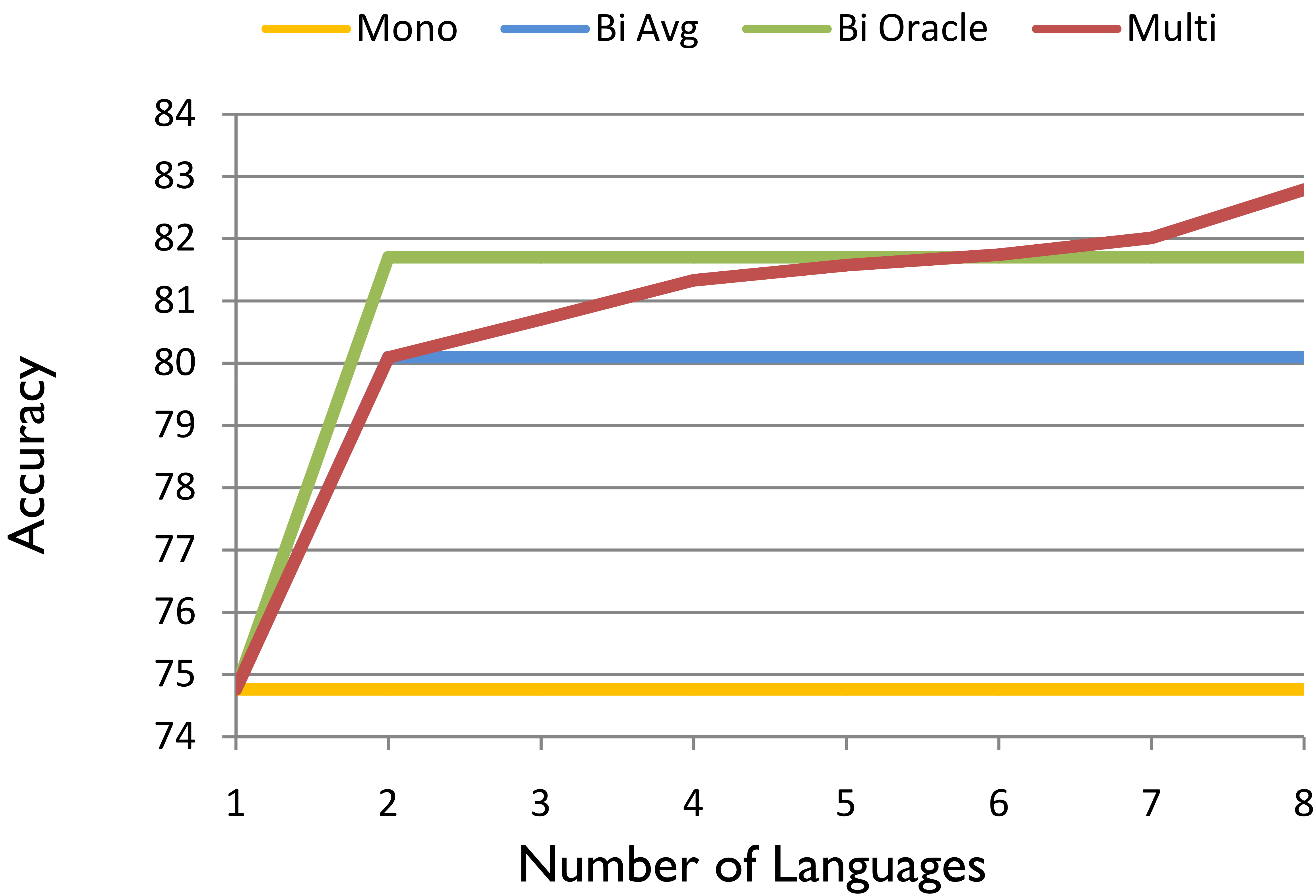
Monolingual vs Multilingual



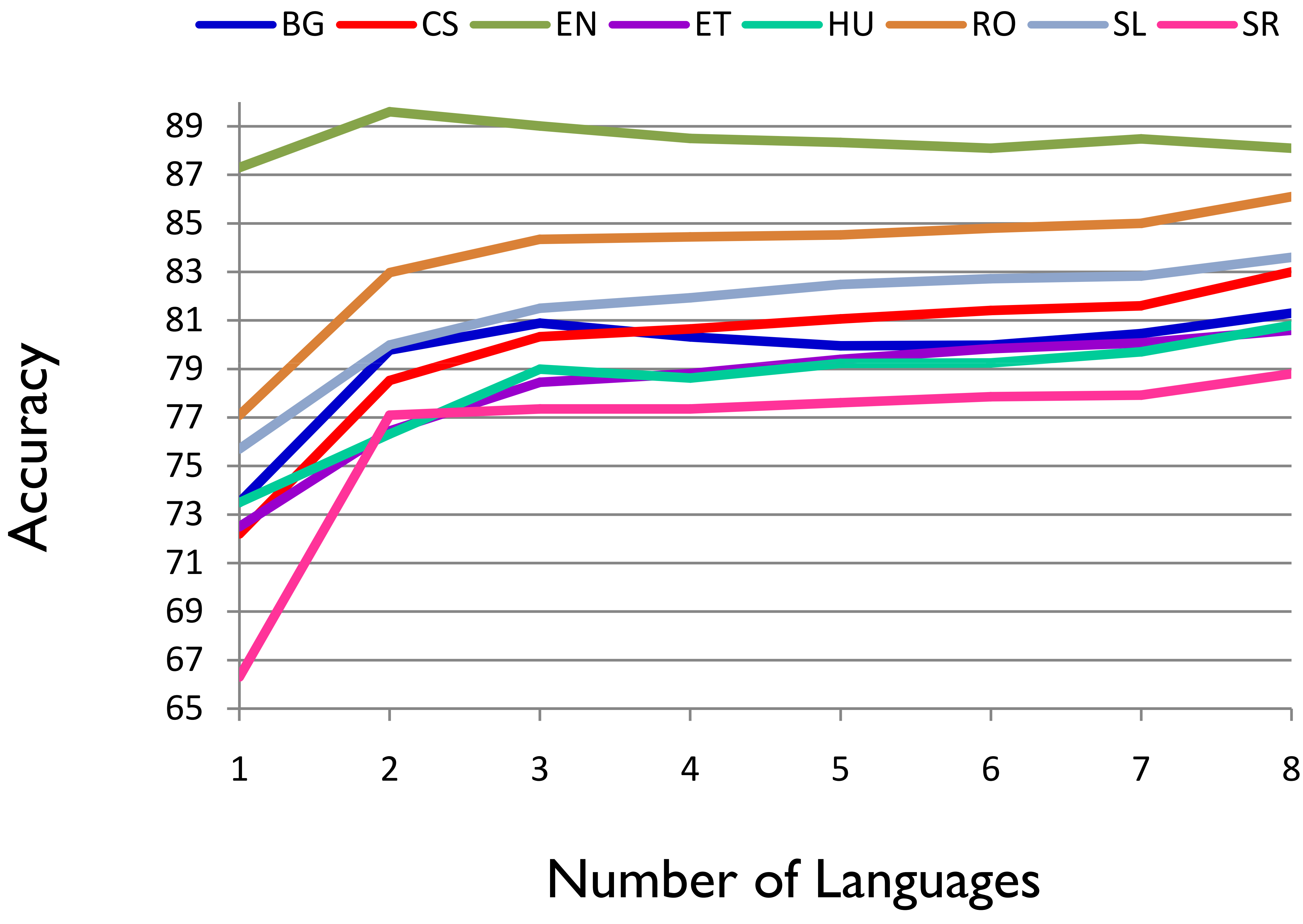


Bilingual Avg vs Multilingual
Mono
Bi Avg
Multi
84
83
82
81
80
79
78
77
76
75
74
Accuracy
1
2
3
4
5
6
7
8
Number of Languages

Bilingual Oracle vs Multilingual



Breakdown by Language...



Related Work

- Multi-source MT

  [Och & Ney 2001; Utiyama & Isahara 2006; Cohn & Lapata 2007; Chen et al 2008; Bertoldi et al 2008]

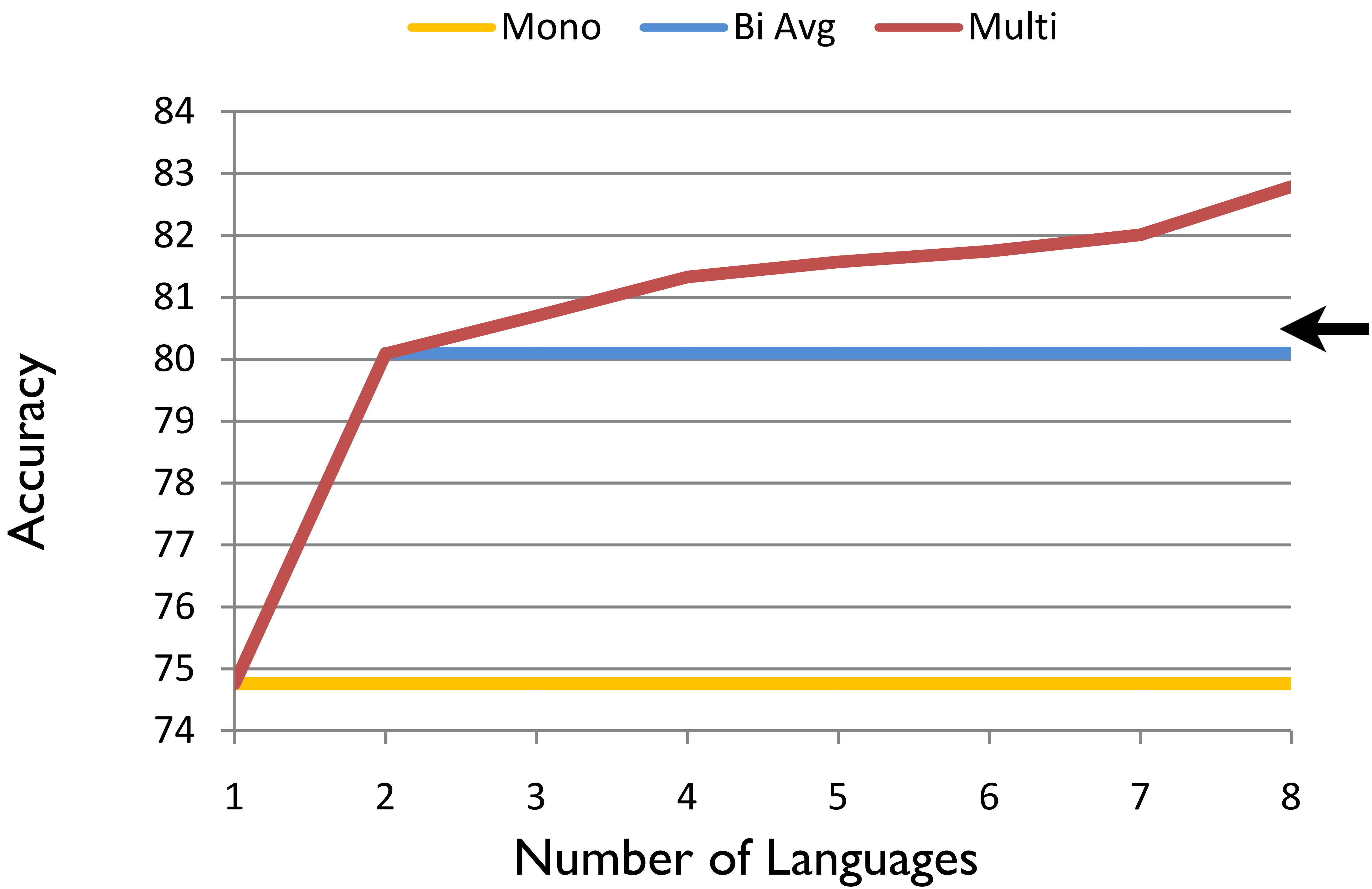
- Multilingual lexicon induction

  [Genzel 2005]

General Trend: *Combining bilingual models*

Our Approach: *Joint multilingual model*

Bilingual: Voting



Analysis: Superlingual Tags

- As languages added, number of superlingual tags increases: 11 *(pairs)* ➔ 20 *(8 languages)*
- Most superlingual tags model a single dominant POS:

$s = 6$

bg	N=.91	A=.04	...
en	N=.98	V=.01	...
hu	N=.85	A=.07	...
sl	N=.94	A=.04	...

Analysis: Superlingual Tags

- But some superlingual tags model more complex multilingual patterns

$s = 14$

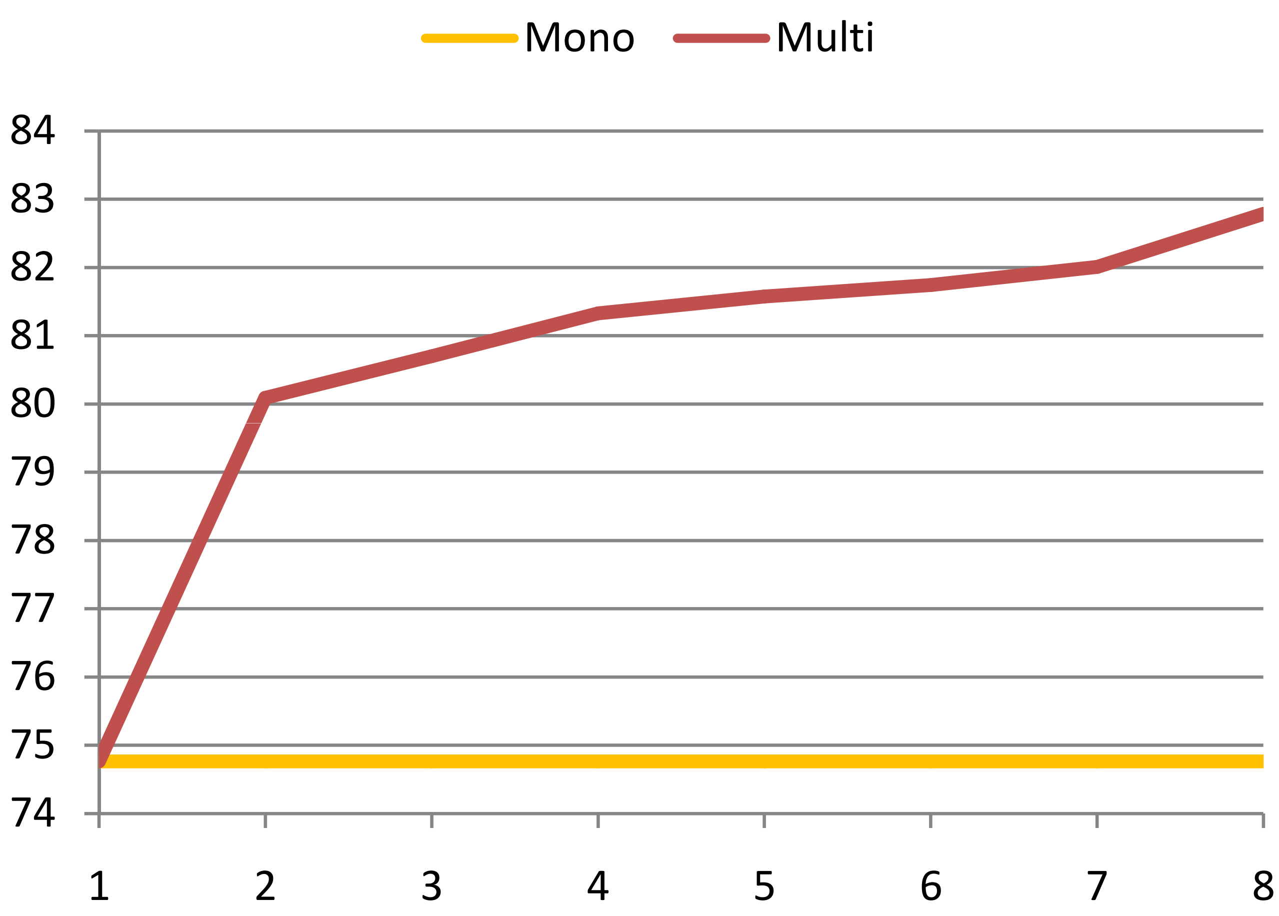
bg	V=.66	N=.21	...
en	V=.55	N=.25	...
et	N=.52	V=.30	...
hu	N=.44	V=.34	...

$s = 15$

cs	PRN=.61	...
en	DT=.99	...
sl	V=.96	...
sr	V=.89	...

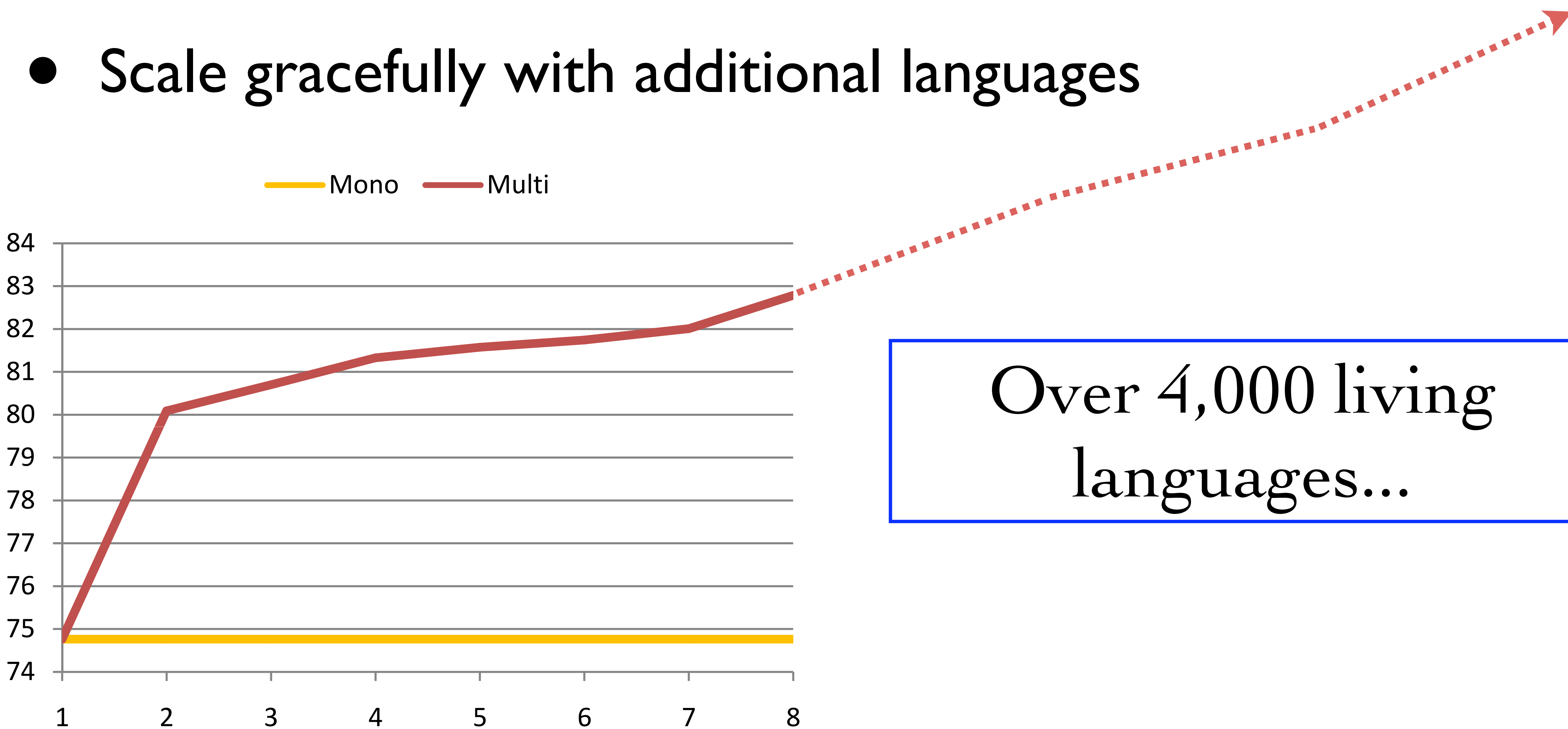
Conclusions

- Capture multilingual patterns using non-parametric latent variables
- Scale gracefully with additional languages

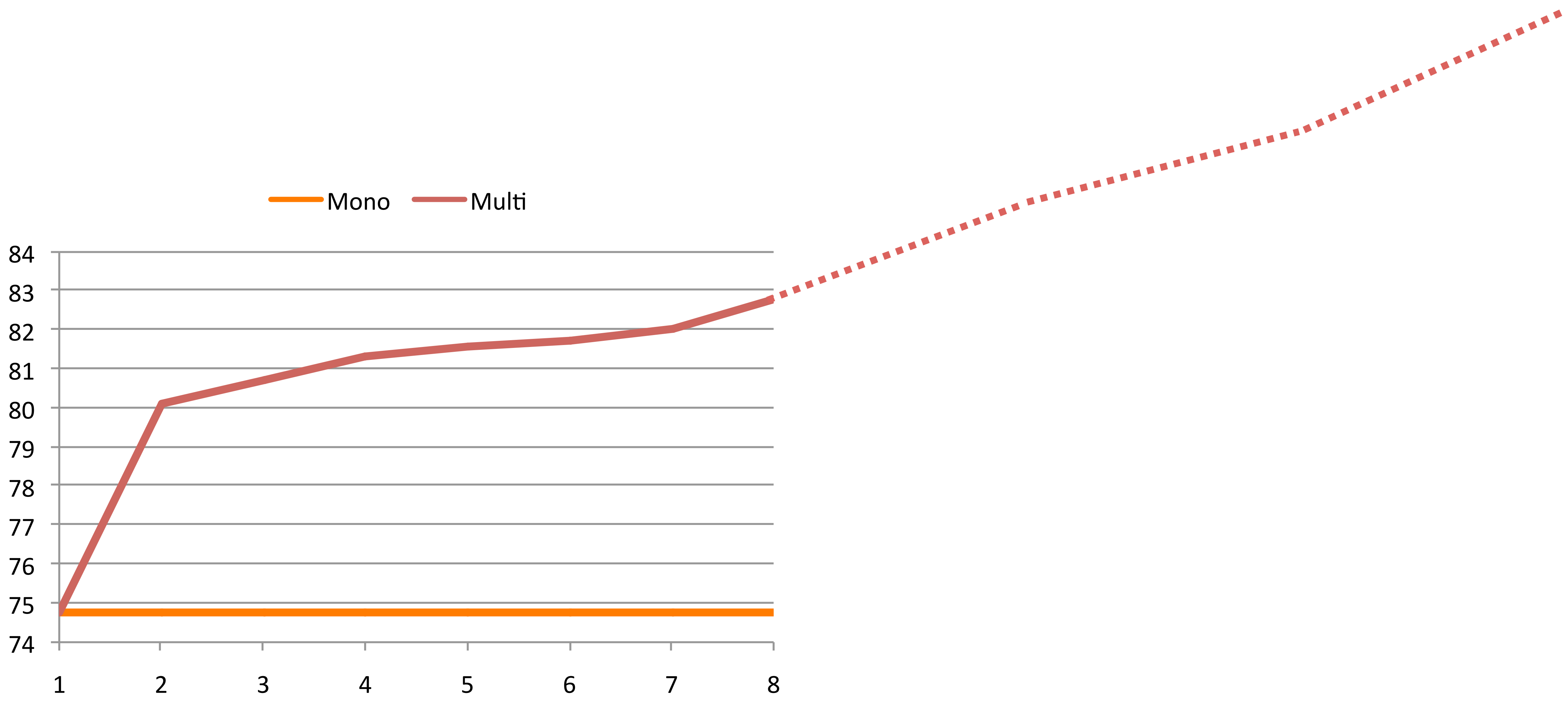


Conclusions

- Capture multilingual patterns using non-parametric latent variables
- Scale gracefully with additional languages

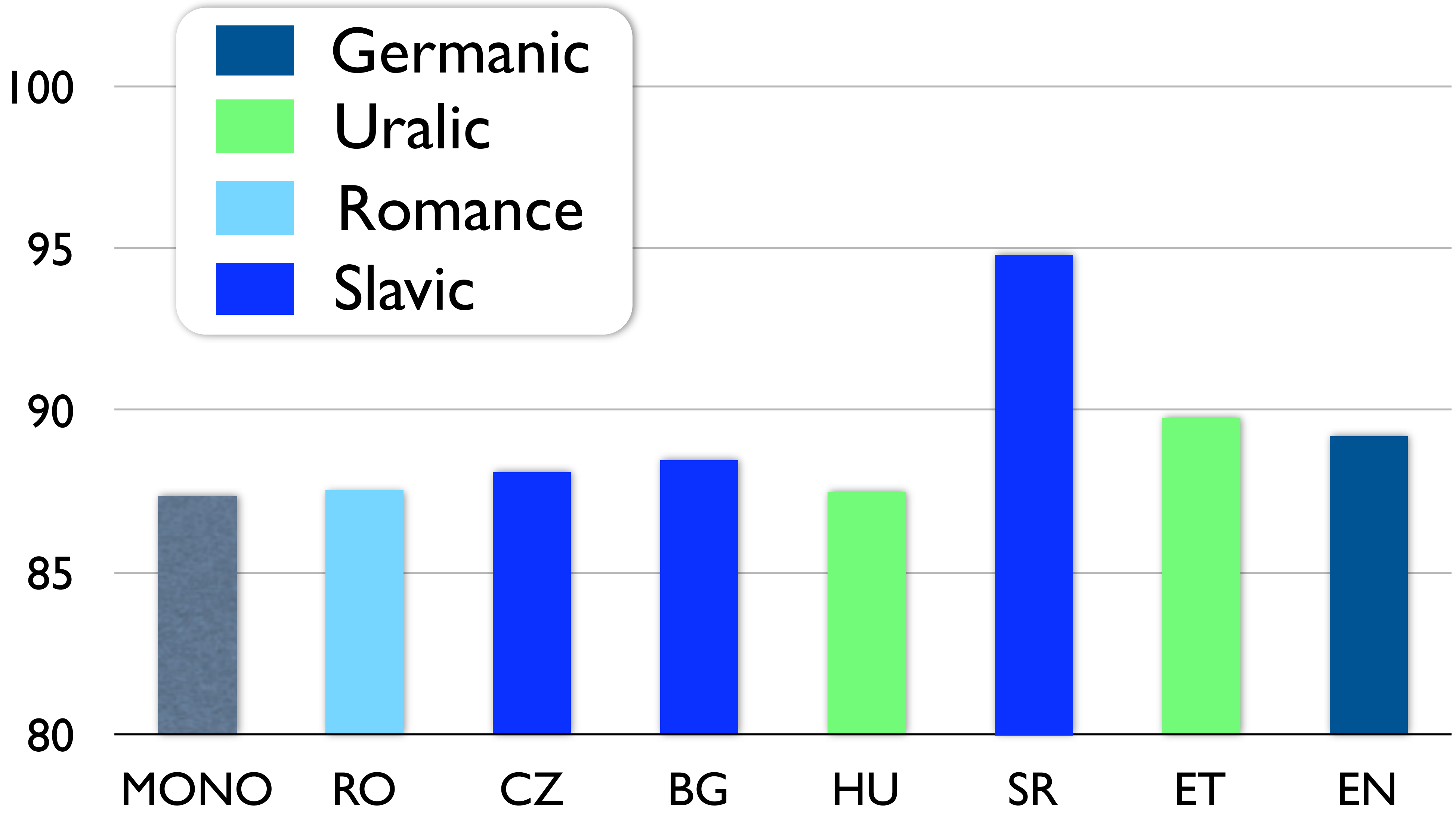


Over 4,000 living languages…

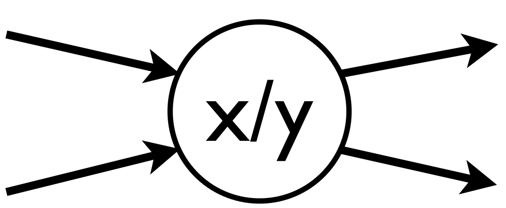


Mono
Multi
84
83
82
81
80
79
78
77
76
75
74
1
2
3
4
5
6
7
8

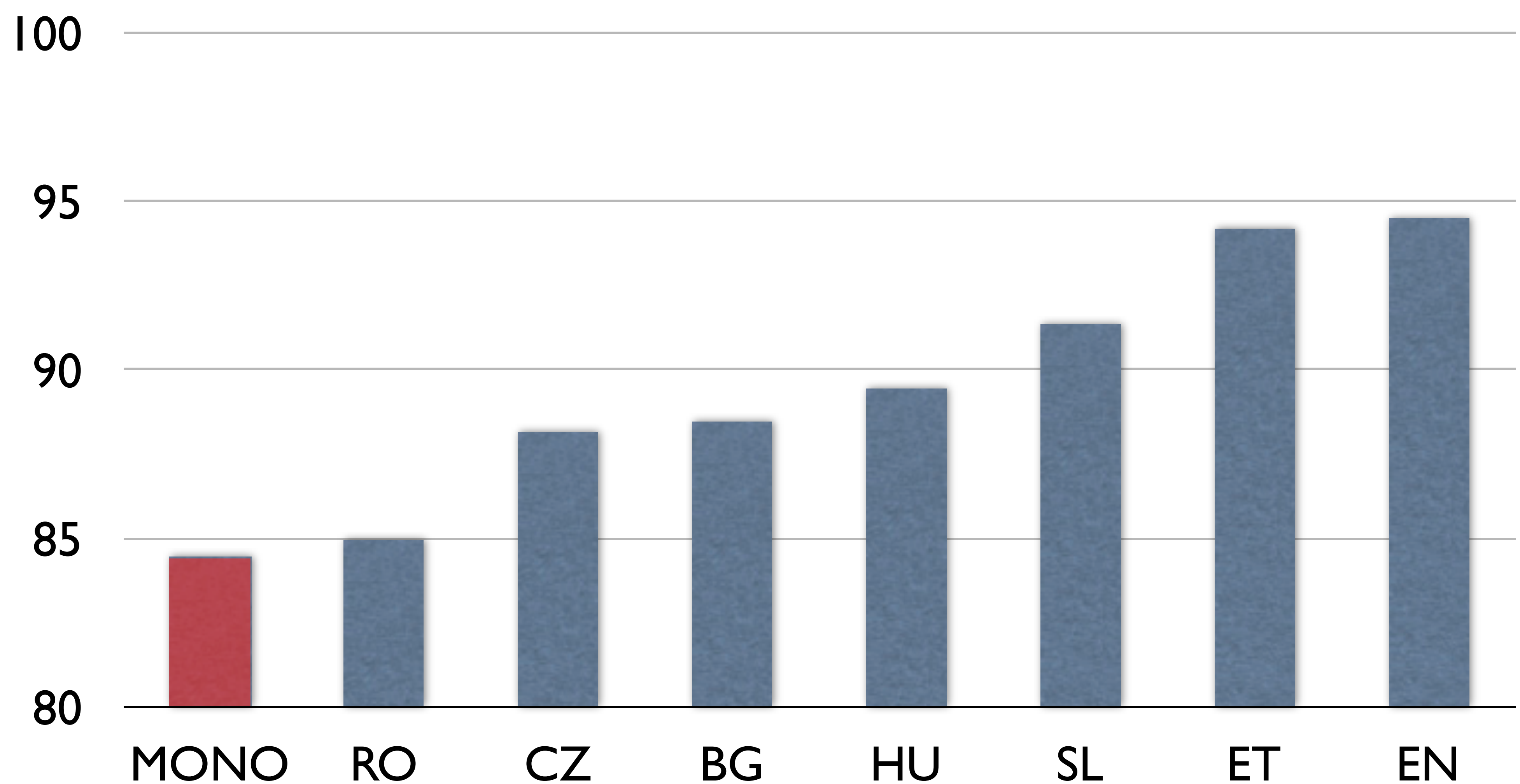
Slovene, paired with...



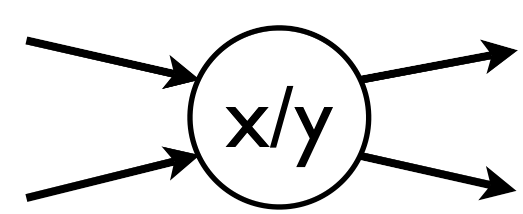
Bilingual merged-node Model
[Snyder et al 2008]



Serbian, paired with...



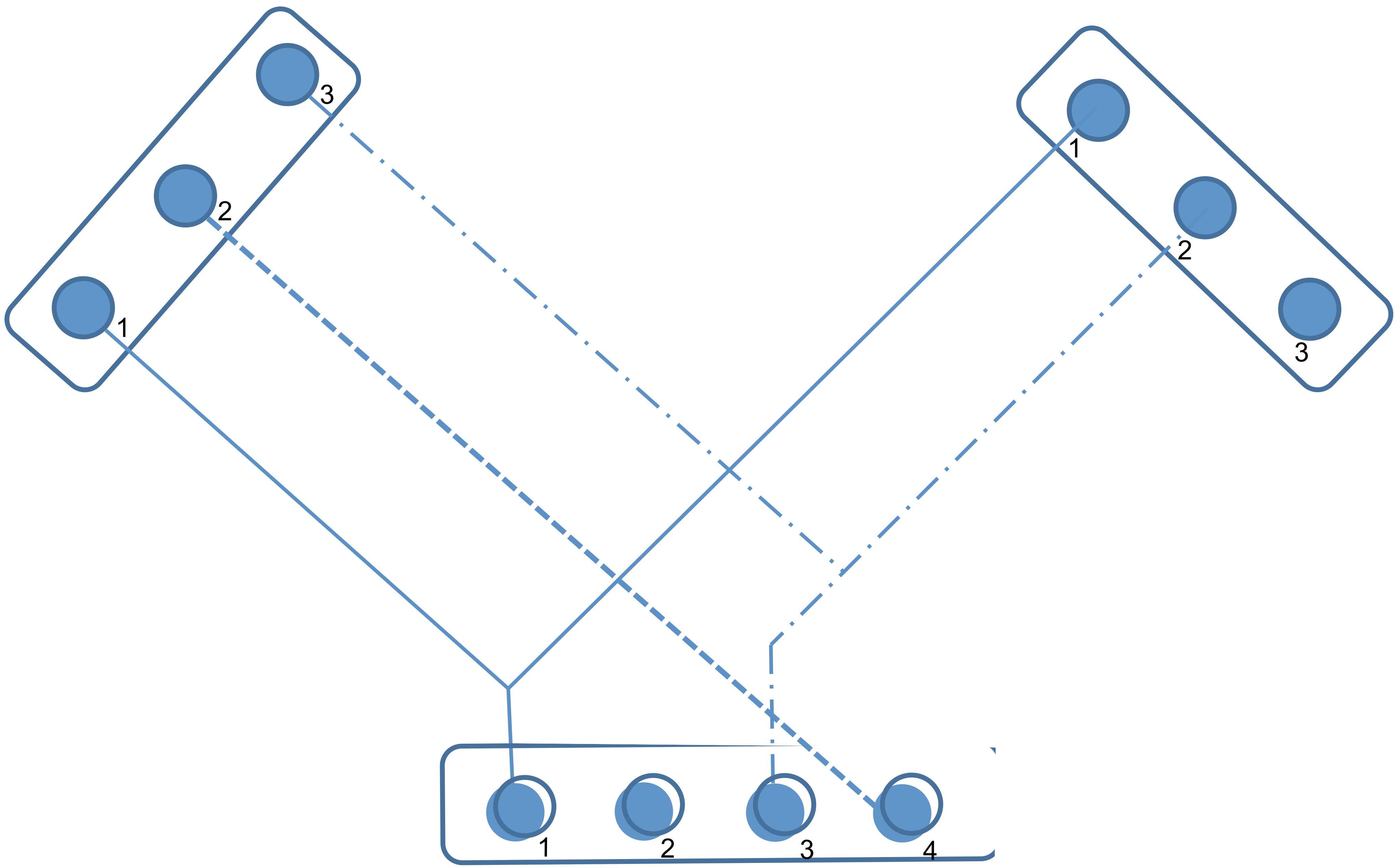
Bilingual merged-node Model
[Snyder et al 2008]



Generative Story: *sentences*

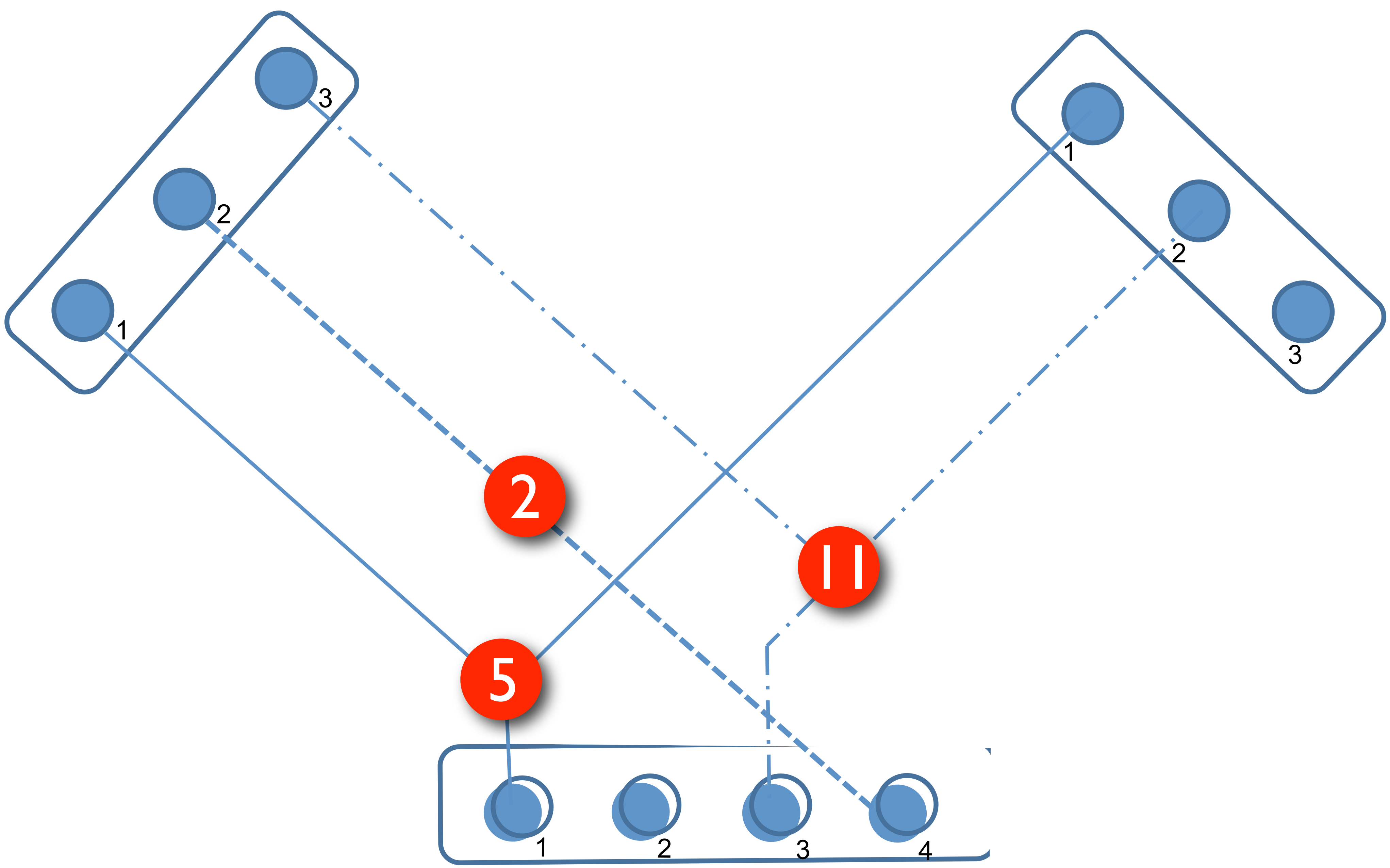
1. Draw *alignment template:*

[1, 1, 1]
[2, 4, _]
[3, 3, 2]



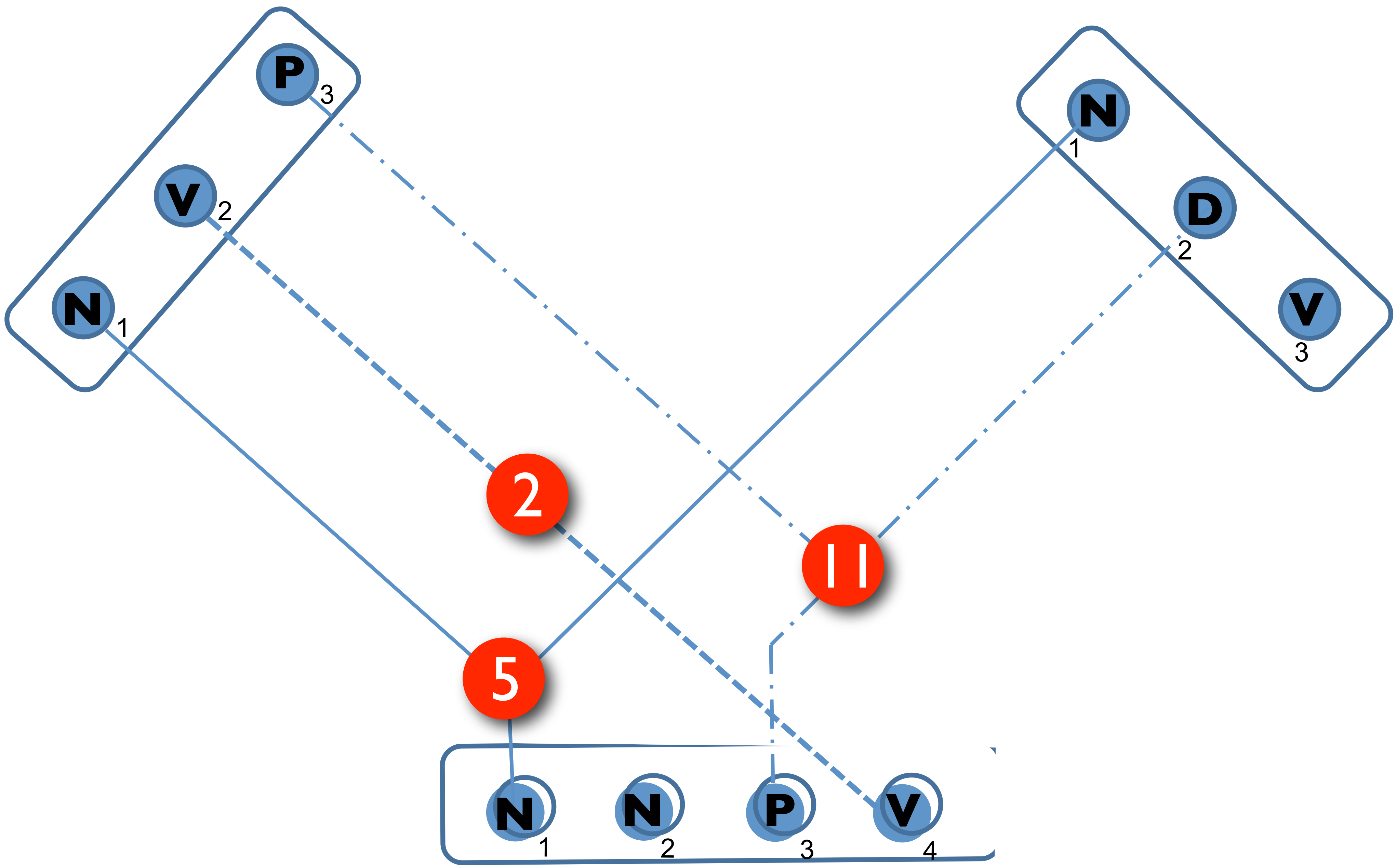
Generative Story: *sentences*

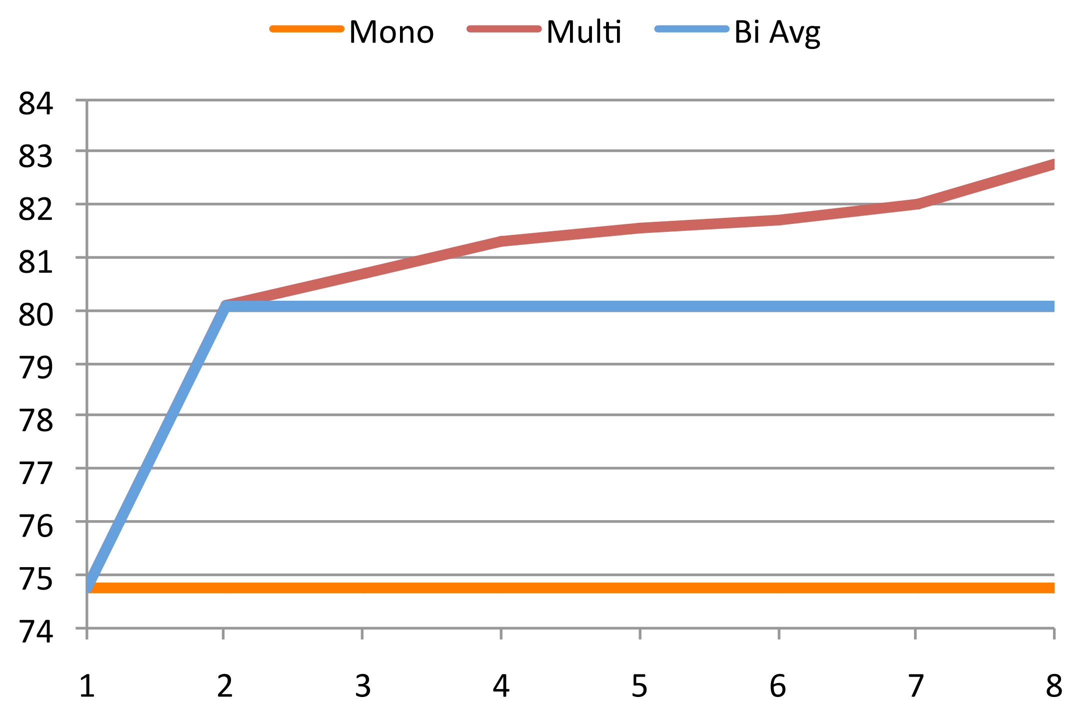
2. Draw *superlingual tags:* $s_i \sim \pi$

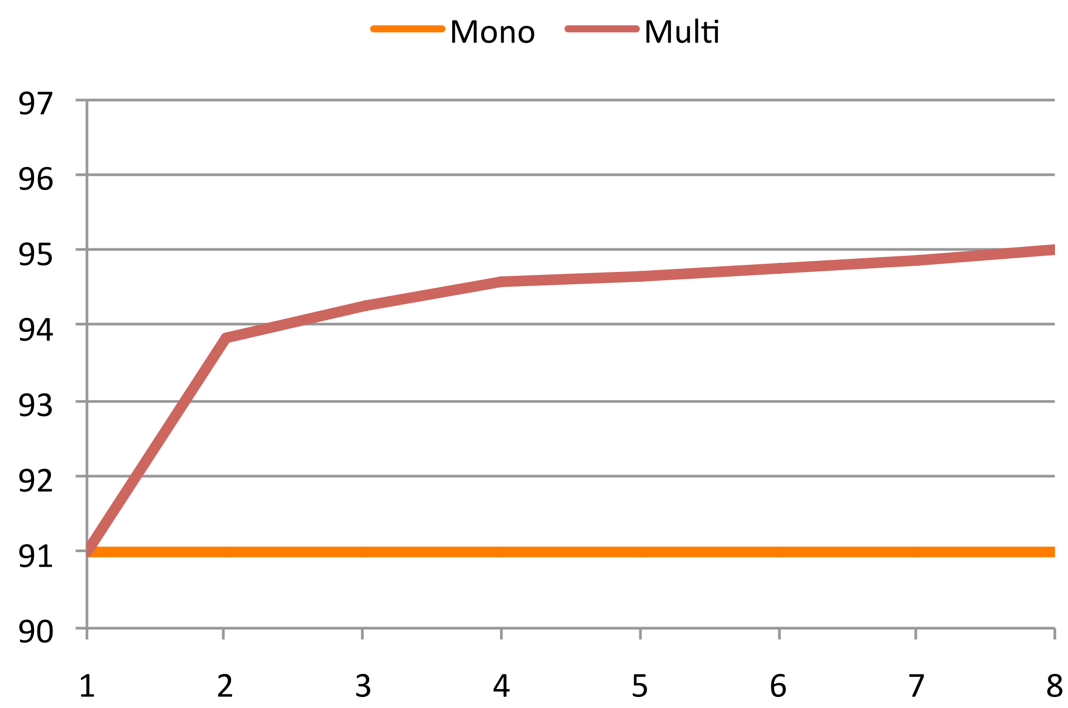


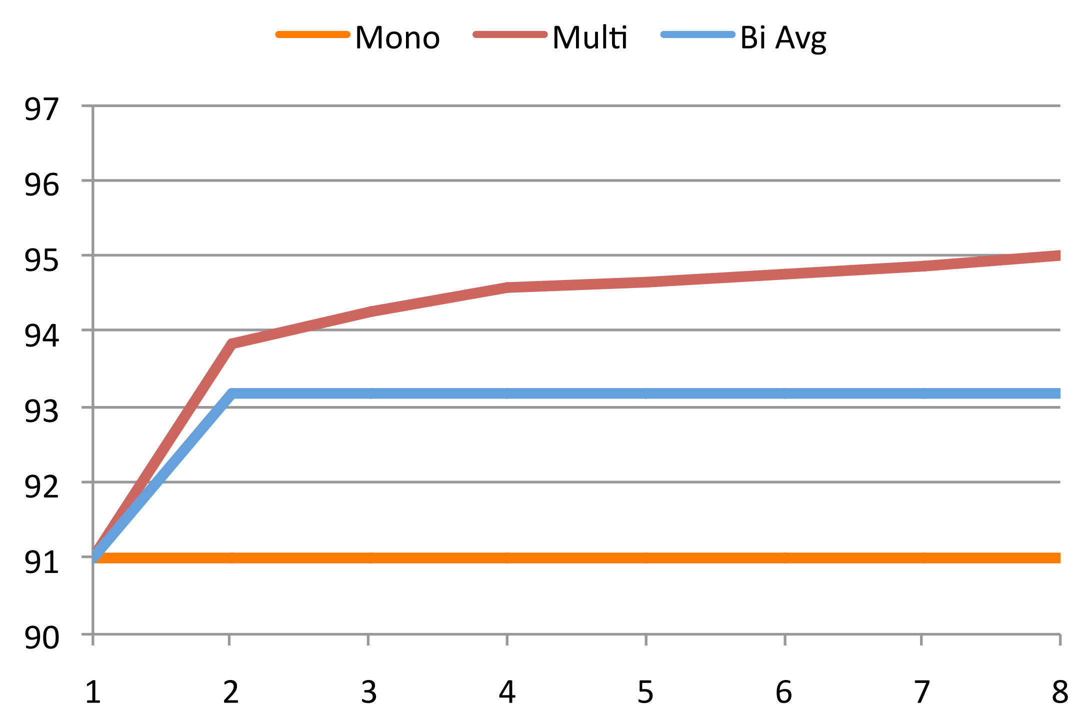
Generative Story: *sentences*

3. Draw *POS tags*: $y_i \sim \frac{trans(y_i|y_{i-1}) \cdot \ \psi_s^{\ell}(y_i)}{Z}$



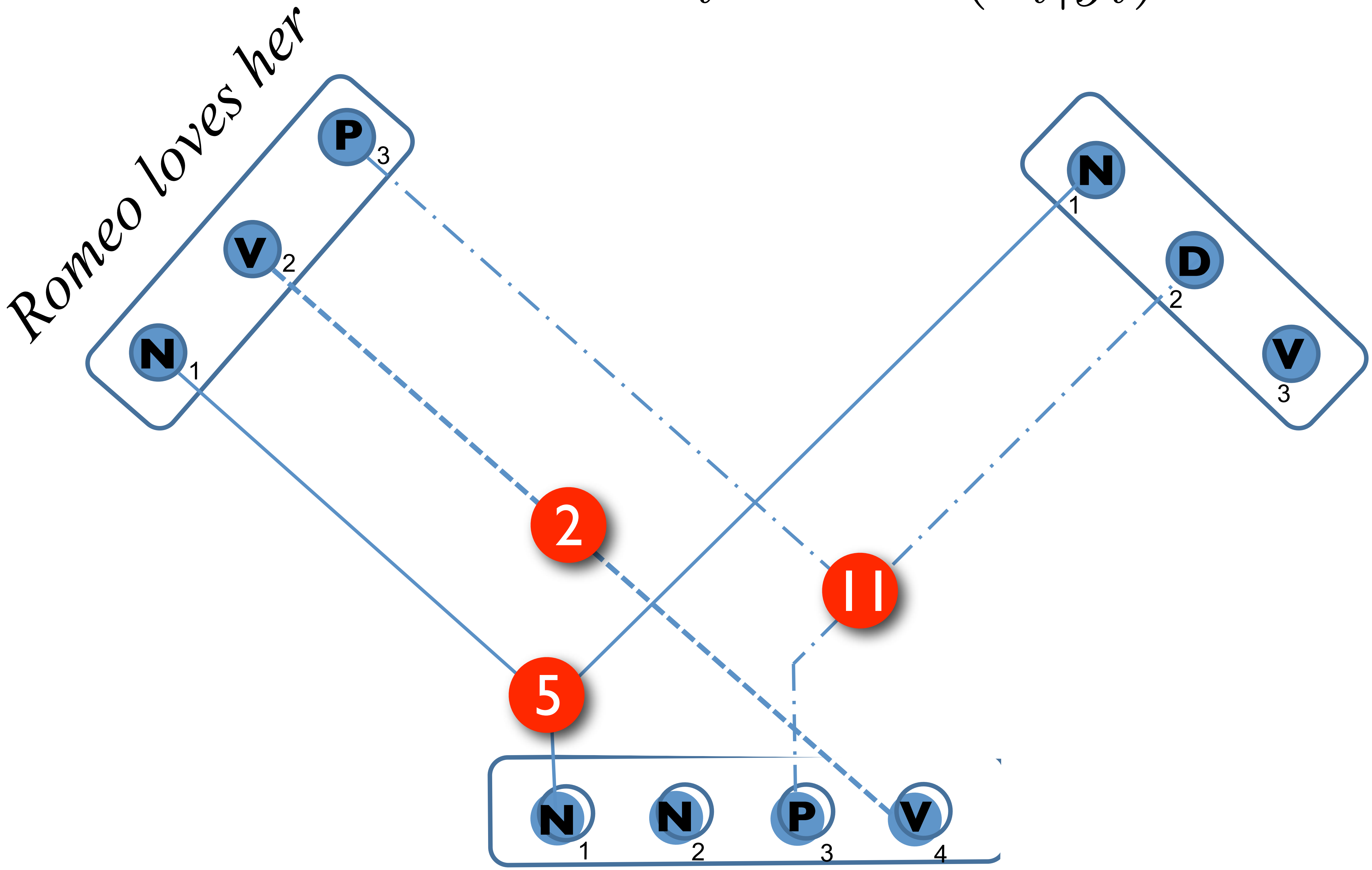

Mono
Multi
Bi Avg
84
83
82
81
80
79
78
77
76
75
74
1
2
3
4
5
6
7
8


Mono
Multi
97
96
95
94
93
92
91
90
1
2
3
4
5
6
7
8

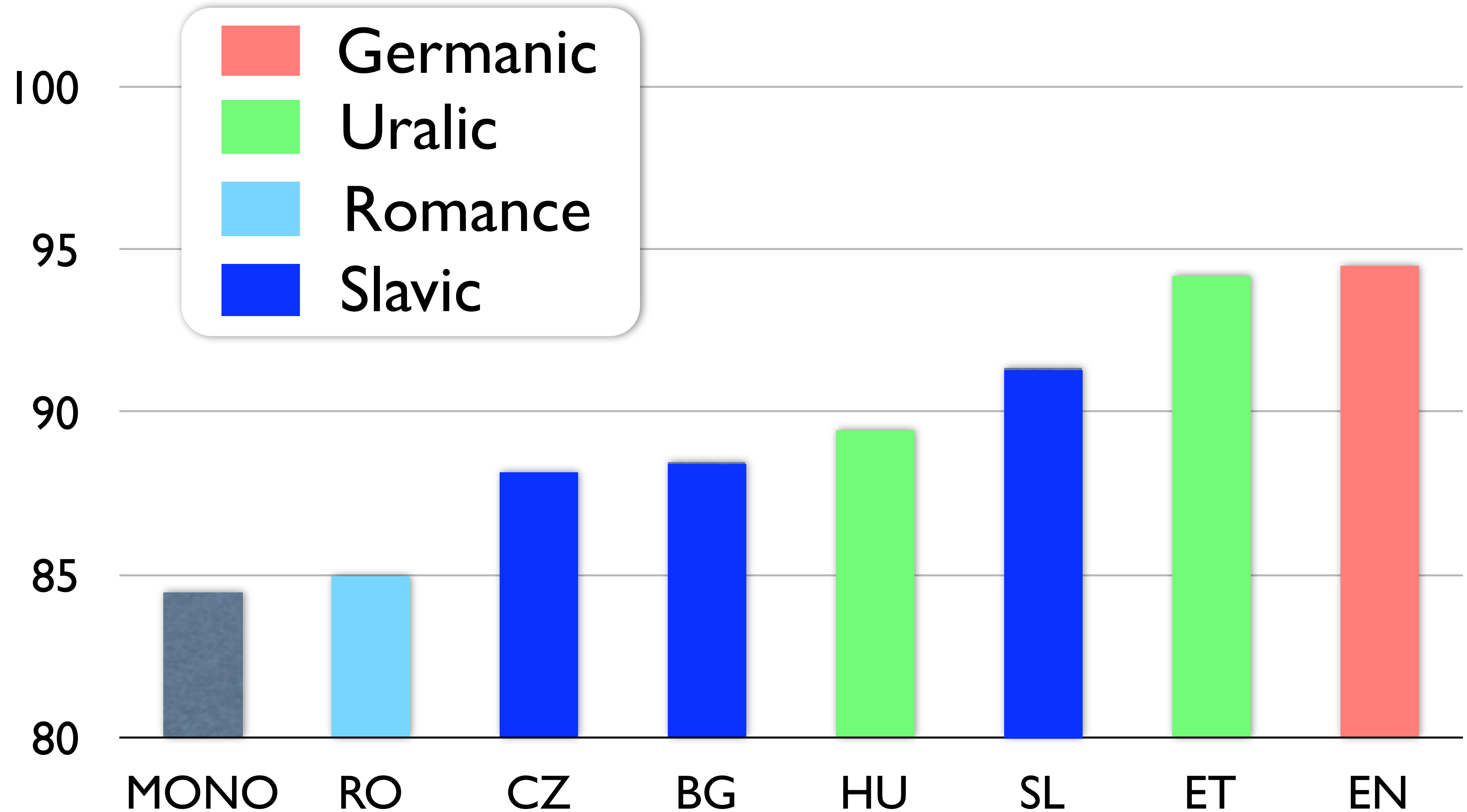

Mono
Multi
Bi Avg
97
96
95
94
93
92
91
90
1
2
3
4
5
6
7
8

Generative Story: *sentences*

4. Draw *words*: $x_i \sim emit(x_i | y_i)$

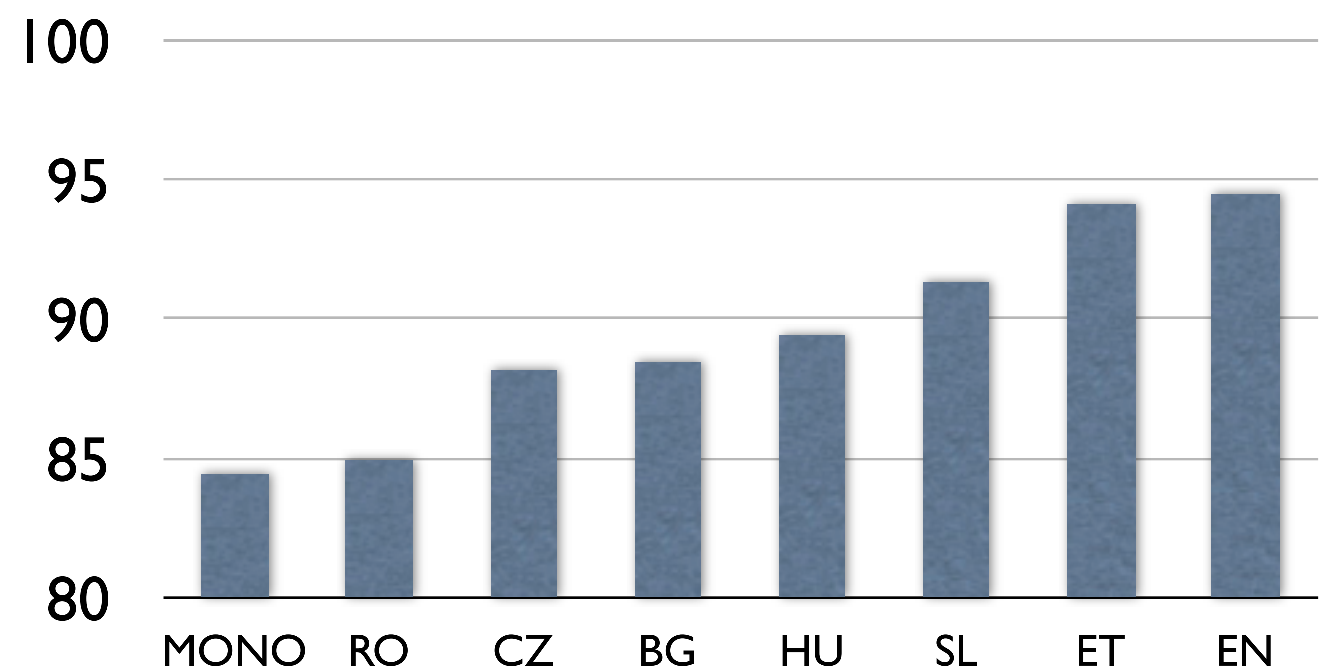


Serbian, paired with...



Bilingual Model [Snyder et al 2008]

Multilingual Performance Goals



<u>Minimum</u>: Beat avg bilingual performance

<u>Ideally</u>: Beat *best* bilingual performance