Unsupervised Multilingual Grammar Induction

Benjamin Snyder
Tahira Naseem
Regina Barzilay
MIT

- Languages exhibit variations in patterns of ambiguity
- Variations as natural supervison

בראשית ברא אלהים את השמים ואת הארץ

في البداء خلق الله السموات والارض

Morphology:
acl 2008

POS tagging:
emnlp 2008
naacl 2009

Syntax:
acl 2009 (this talk)

Multilingual Cues

English: *I saw the student from MIT*

Multilingual Cues

$$\begin{array}{lcccccc} & \text{S} & \text{V} & & \text{O} & & \\ \text{English:} & \textit{I} & \textit{saw} & \textit{the} & \textit{student} & \textit{from} & \textit{MIT} \end{array}$$

Multilingual Cues

S V O

English: *I saw the student from MIT*

Urdu: *I MIT of student saw*

S O V

Multilingual Cues

English: ?
[*I saw*] *the student from MIT*

Urdu: *I MIT of student saw*

Multilingual Cues

X

English: [*I saw*] *the student from MIT*

Urdu: *I MIT of student saw*

Multilingual Cues

English: *I saw the student from MIT*

Urdu: [*I MIT*] *of student saw*

?

Multilingual Cues

English: *I saw the student from MIT*

Urdu: [*I MIT*] *of student saw*

X

Multilingual Cues

English: *I saw the student [from MIT]*

Urdu: *I MIT of student saw*

Multilingual Cues

?

English: *I [saw the student [from MIT]]*

Urdu: *I MIT of student saw*

Multilingual Cues

?

English: *I saw [the student [from MIT]]*

Urdu: *I MIT of student saw*

Multilingual Cues

English: *I saw the student [from MIT]*

Urdu: *I [MIT of] student saw*

Multilingual Cues

English: *I saw the student [from MIT]*

Urdu: *I [[MIT of] student] saw*

Multilingual Cues

English: *I saw* [*the student* [*from MIT*]]

Urdu: *I* [[*MIT of*] *student*] *saw*

Multilingual Cues

English: *I saw [the student [from MIT]]*

Urdu: *I [[MIT of] student] saw*

Main idea: learn from systematic variations in *phrase order* and *expression*

Key Technical Challenge

Represent shared cross-lingual syntactic structure

- *Linguistically plausible*
  - Allow full range of syntactic divergence and translational freedom
- *Computationally tractable*
  - Support probabilistic operations: argmax, marginalization, sampling

Prior Representations

Synchronous Grammars [Wu 1997; Melamed 2003; Chiang 2005; Smith&Smith 2004; Eisner 2005; Blunsom et al 2008]

- Employed for modeling phrase reordering in MT
- In basic form, isomorphic trees (up to sibling order)

Node Matching [Burkett&Klein 2008]

- Ignores tree structure
- Marginalization is #P-complete

Our Proposal

Probabilistic adaptation of *Unordered Tree Alignment* [Jiang et al 1995]

- Node alignments must respect tree structures
- Yet any number of nodes may remain *unaligned*
- Can marginalize and sample *all possible alignments* in linear time with dynamic program

For trees T_1 and T_2, an *alignment* A is obtained in the following way:

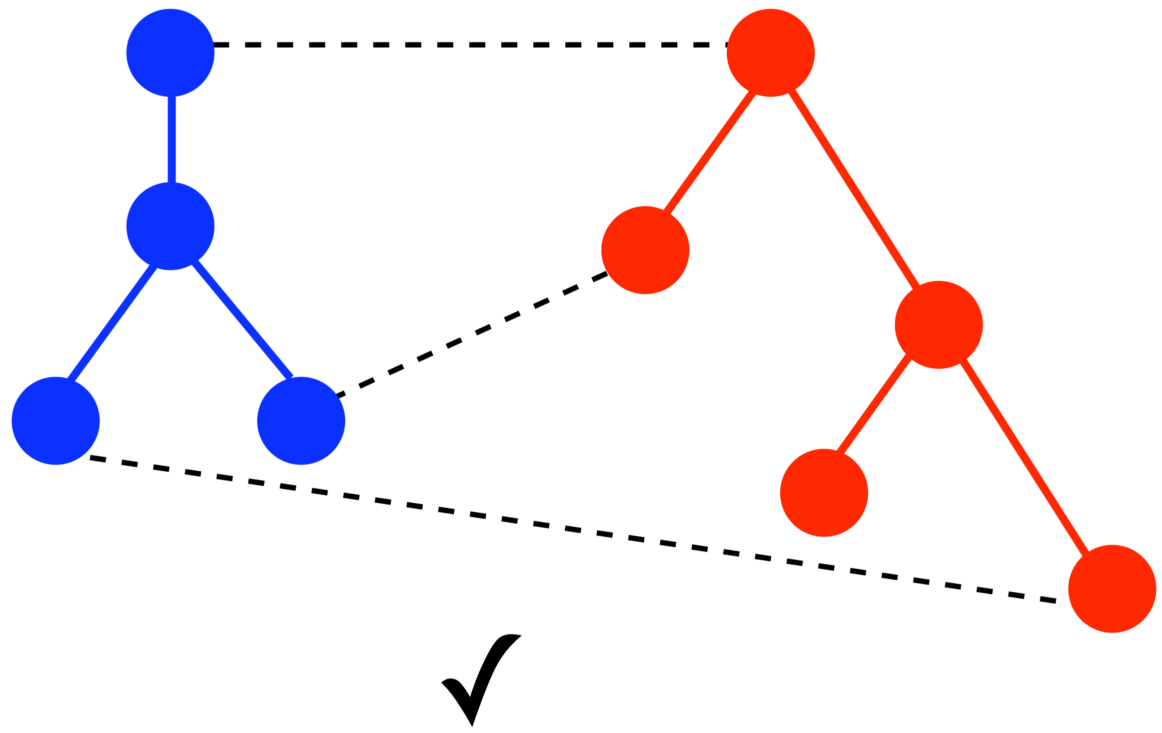
1. Insert empty nodes into T_1 and T_2 and swap sibling order, until they are isomorphic

2. Overlay the resulting trees T_1' and T_2' to obtain A

For trees T_1 and T_2, an *alignment* A is obtained in the following way:

1. Insert empty nodes into T_1 and T_2 and swap sibling order, until they are isomorphic

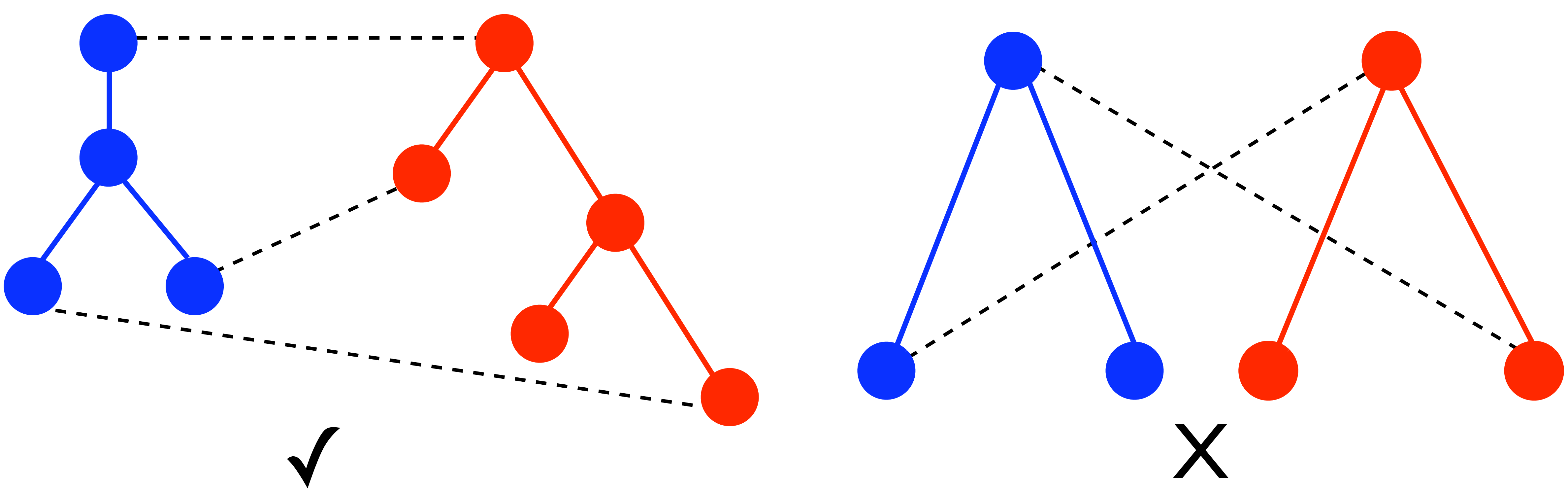
2. Overlay the resulting trees T_1' and T_2' to obtain A

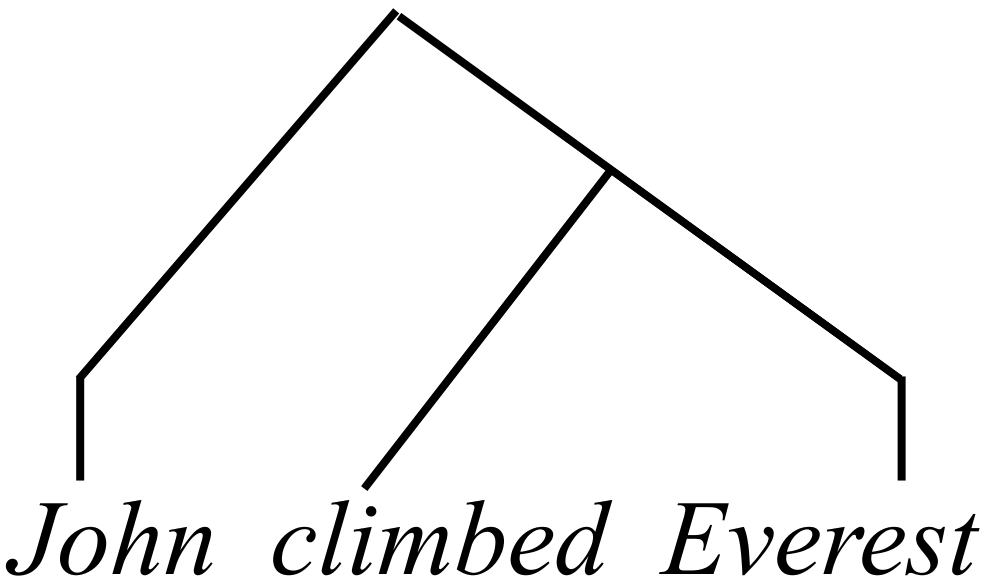


For trees T_1 and T_2, an *alignment* A is obtained in the following way:

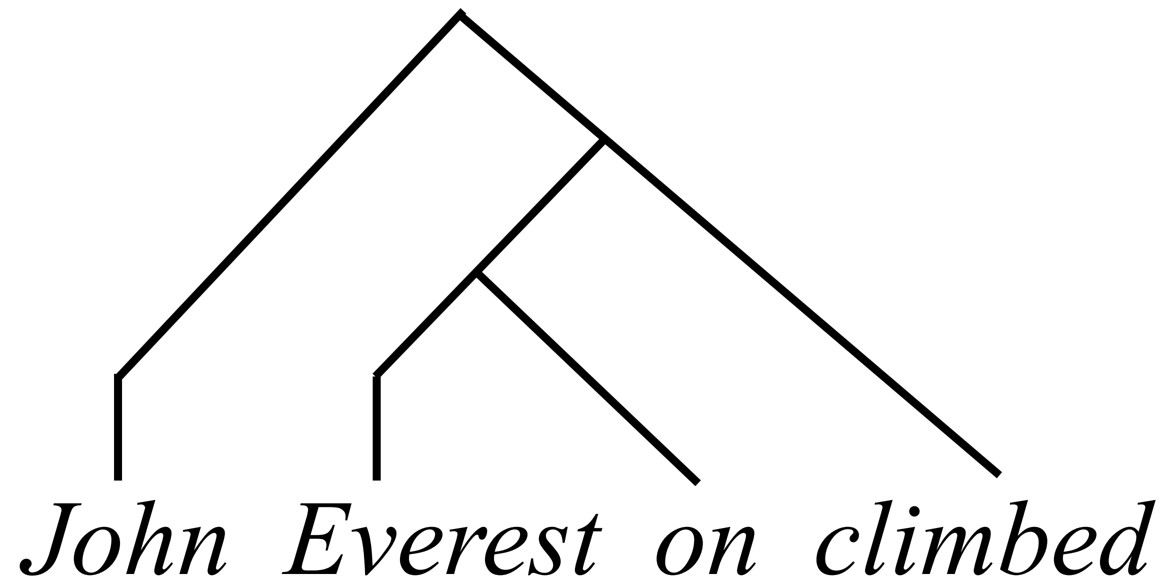
1. Insert empty nodes into T_1 and T_2 and swap sibling order, until they are isomorphic

2. Overlay the resulting trees T_1' and T_2' to obtain A

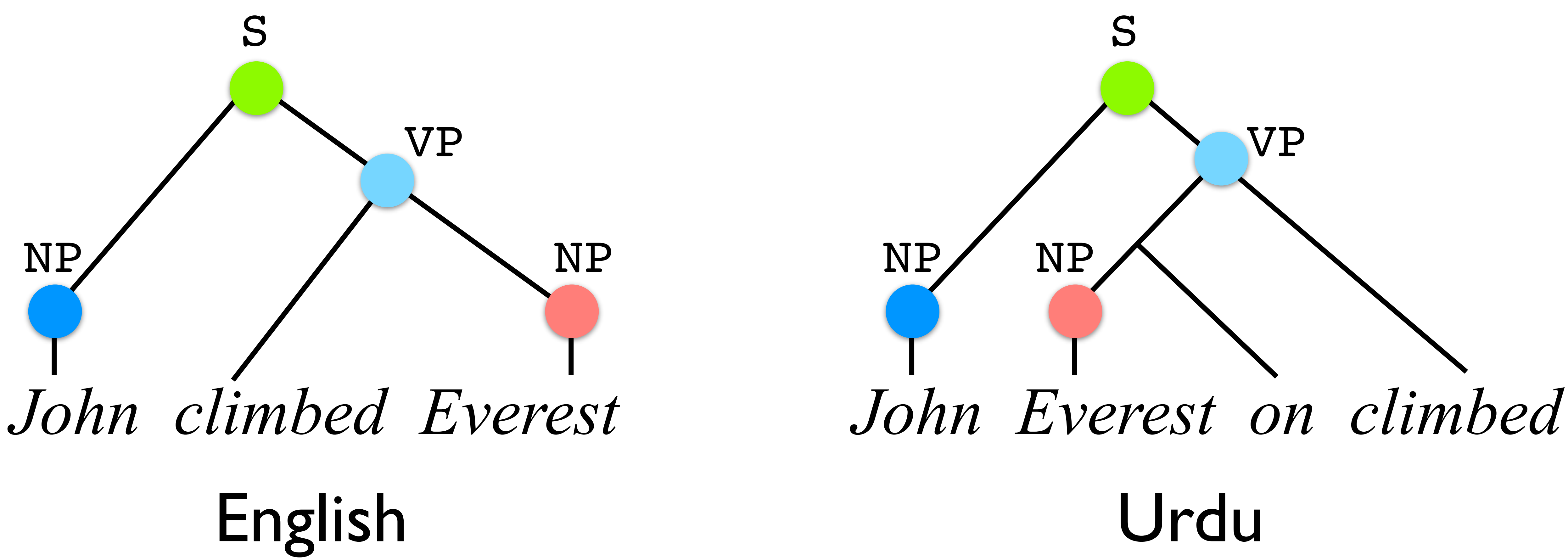


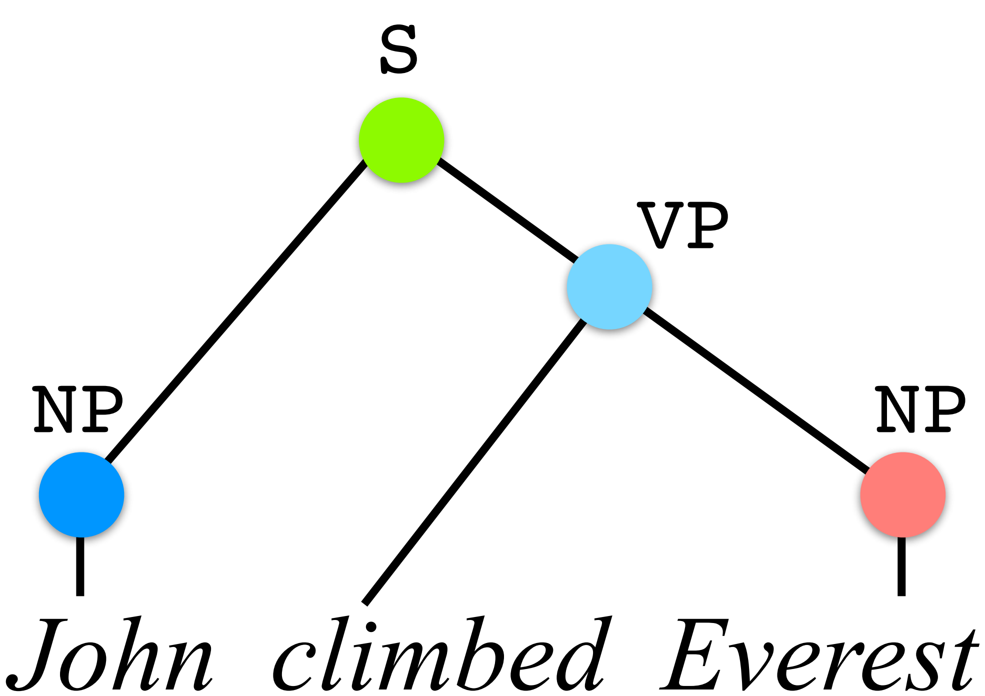


English

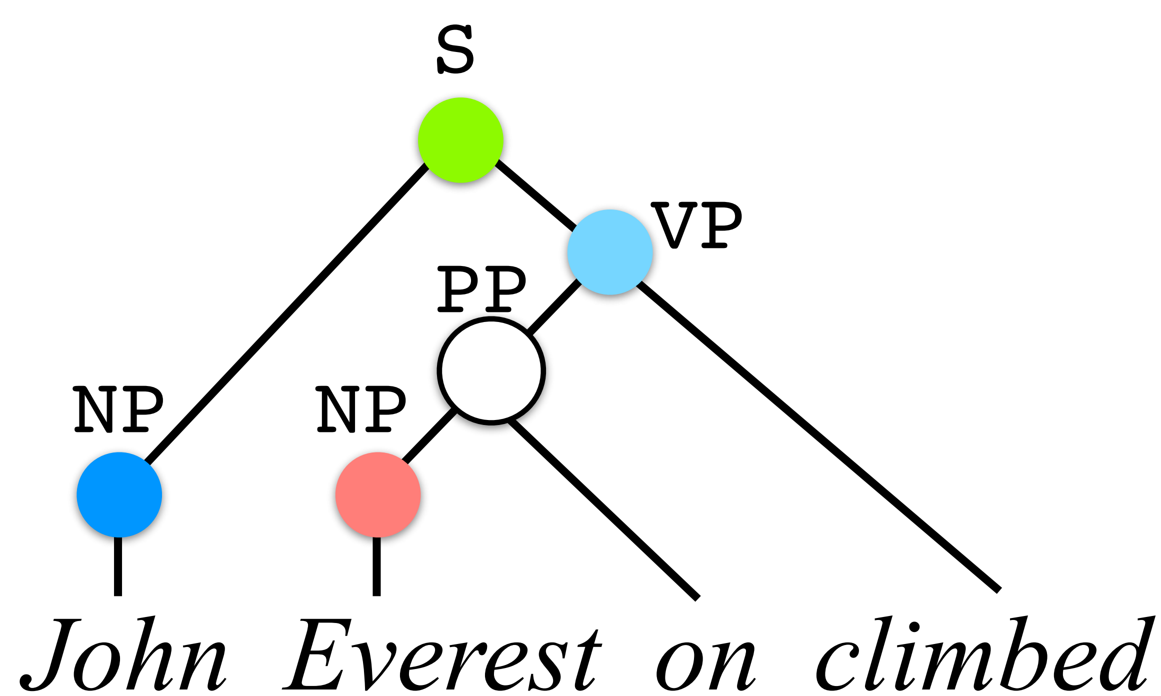


Urdu


S
VP
NP
NP
John climbed Everest
English
S
VP
NP
NP
John Everest on climbed
Urdu



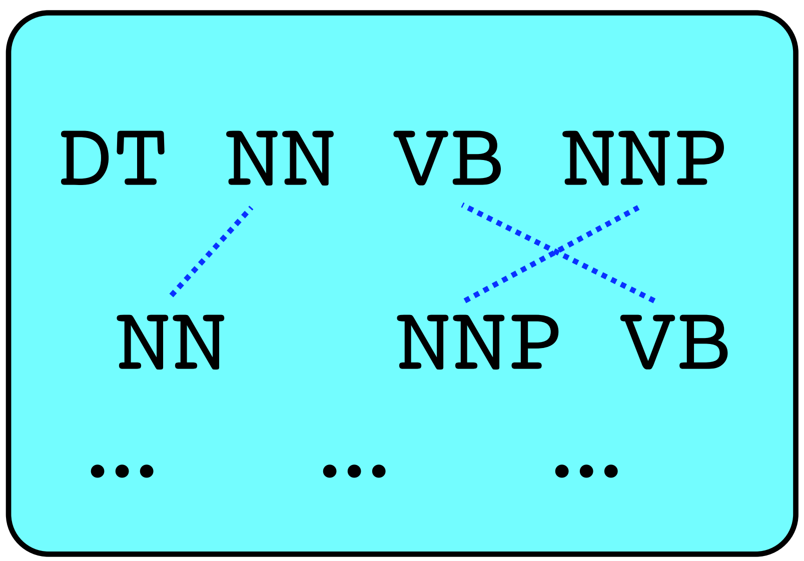
English



Urdu

A Generative Model

We observe:



A Generative Model

We observe:

DT NN (VB NNP)

NN NNP VB

... ...

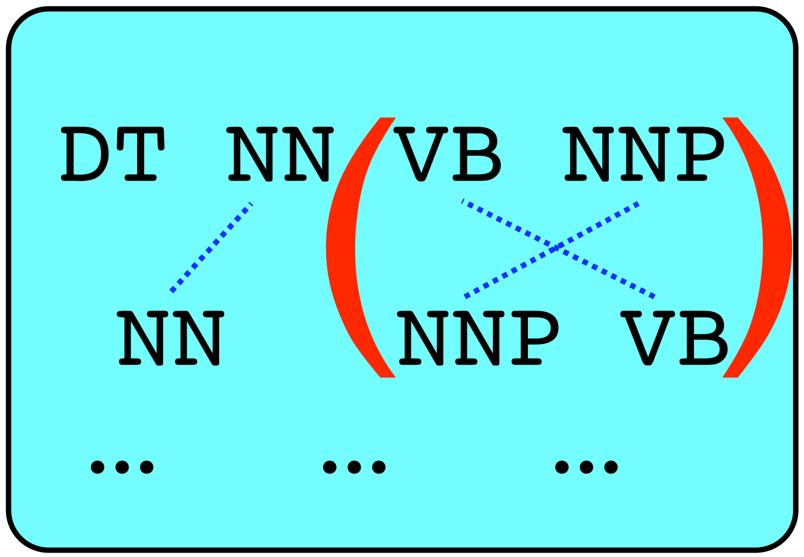
A Generative Model

We observe:

```
DT  NN (VB  NNP)
  NN   (NNP  VB)
 ...    ...    ...
```

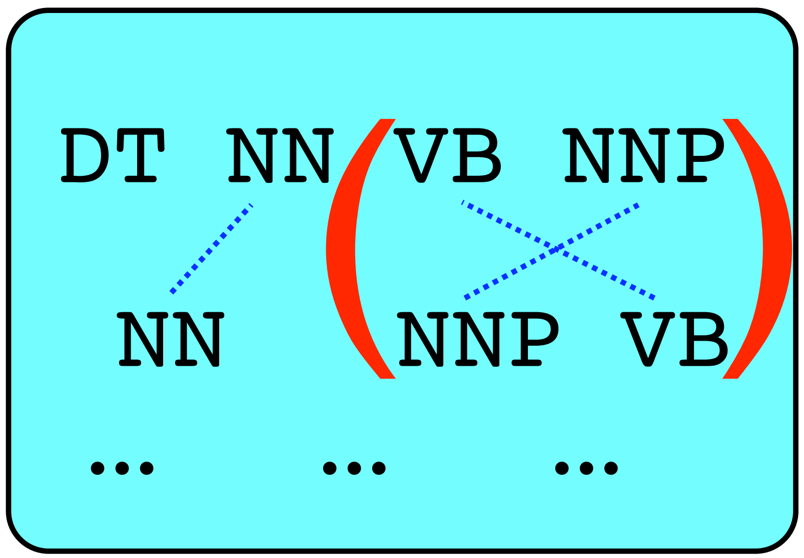
A Generative Model

We observe:



A Generative Model

We observe:

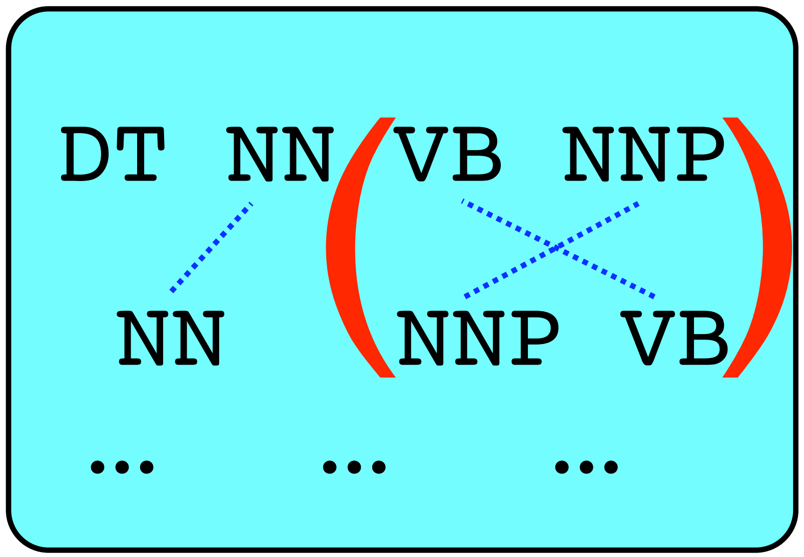


Hypothesize aligned trees that best explain:

- frequent POS sequence pairs
- lexical alignments

A Generative Model

We observe:



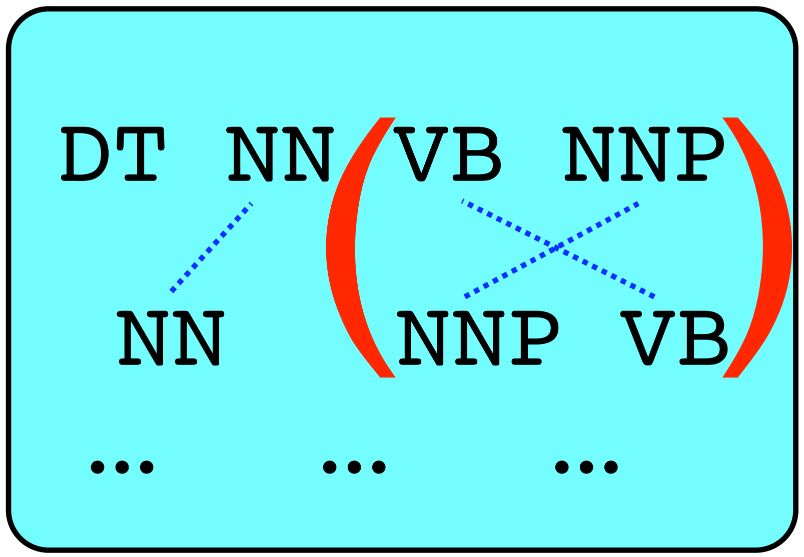
Parameters to learn

Hypothesize aligned trees that best explain:

- frequent POS sequence pairs
- lexical alignments

A Generative Model

We observe:



Hypothesize aligned trees that best explain:

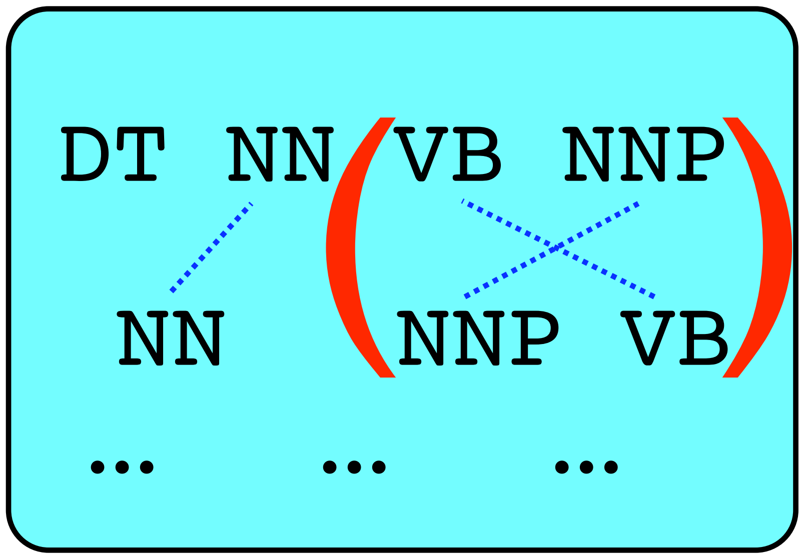
- frequent POS sequence pairs
- lexical alignments

__*Parameters to learn*__

ω Probability of constituent pairs of *aligned* nodes

A Generative Model

We observe:



Hypothesize aligned trees that best explain:

- frequent POS sequence pairs
- lexical alignments

<u>*Parameters to learn*</u>

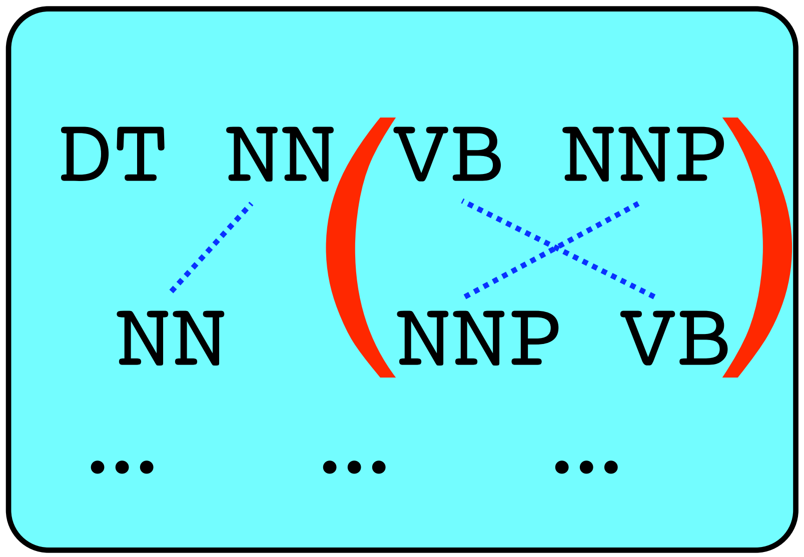
ω Probability of constituent pairs of *aligned* nodes

ϕ^+ Distribution on num. of word alignments between *aligned* nodes

ϕ^- Distribution on num. of word alignments between *unaligned* nodes

A Generative Model

We observe:



Hypothesize aligned trees that best explain:

- frequent POS sequence pairs
- lexical alignments

<u>*Parameters to learn*</u>

ω Probability of constituent pairs of *aligned* nodes

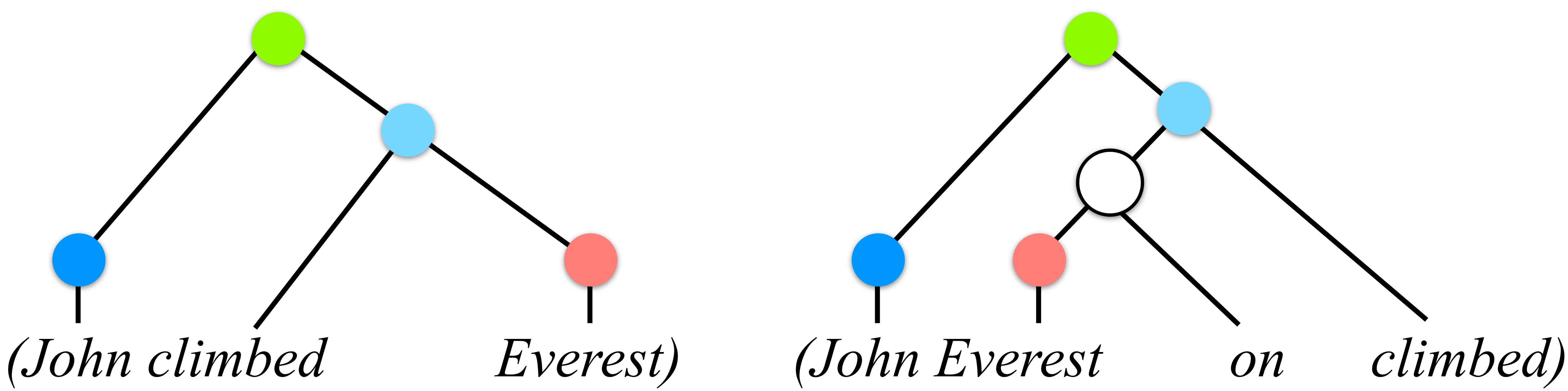
ϕ^{+} Distribution on num. of word alignments between *aligned* nodes

ϕ^{-} Distribution on num. of word alignments between *unaligned* nodes

(language-specific parameters for unaligned nodes [Klein&Manning 2002])

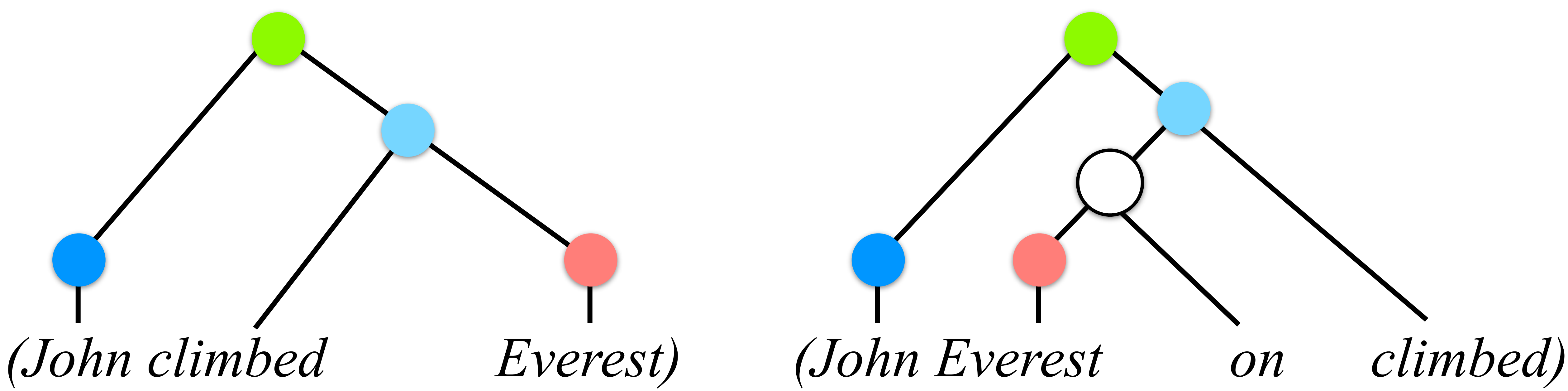
Generative Story

Draw alignment tree *template* (T_1, T_2, A)
from uniform distribution:



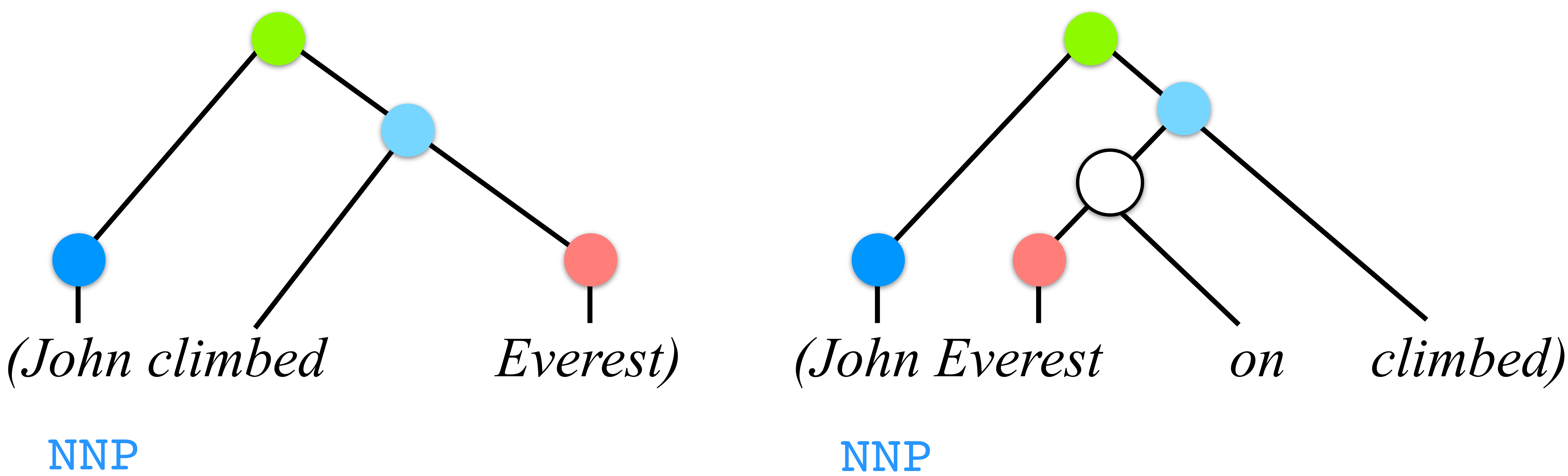
Generative Story

For each *aligned* node pair, draw a *constituent pair* jointly from ω:



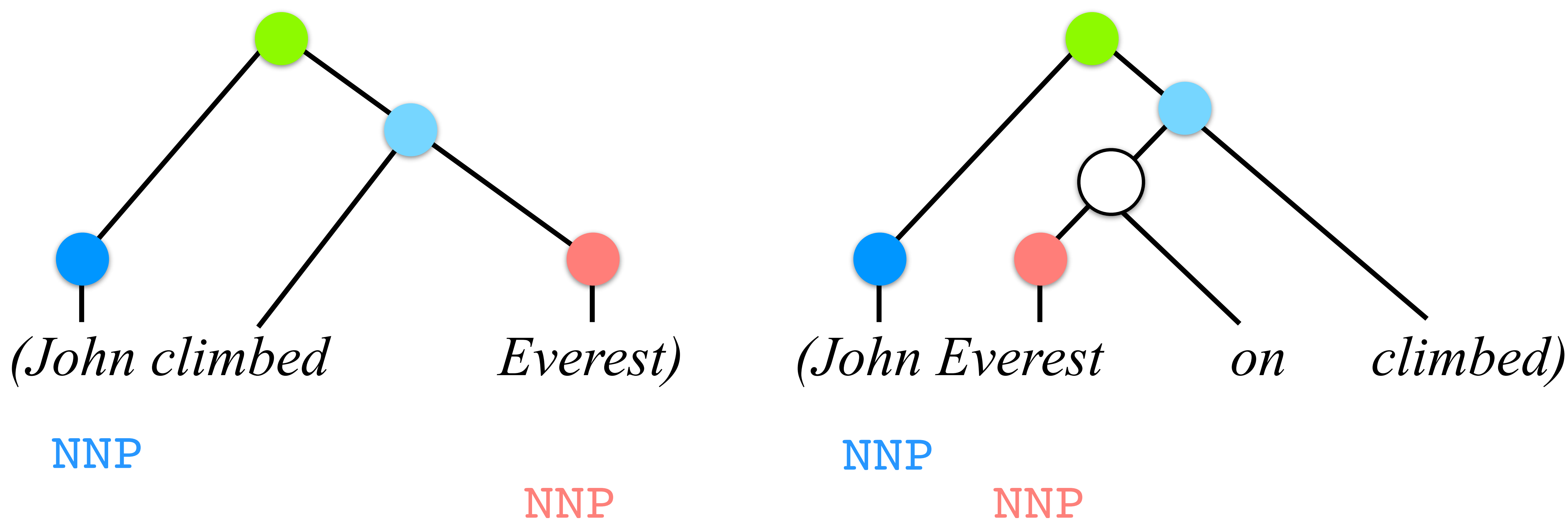
Generative Story

For each *aligned* node pair, draw a *constituent pair* jointly from ω:



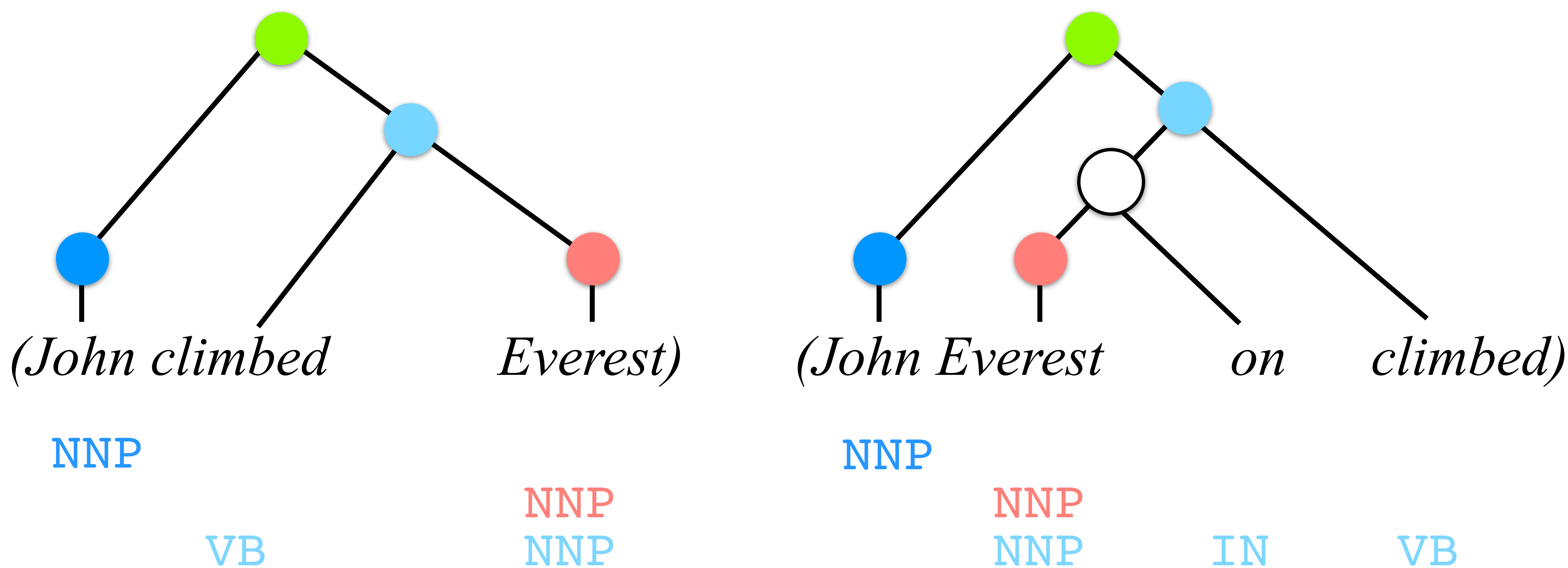
Generative Story

For each *aligned* node pair, draw a *constituent pair* jointly from ω:



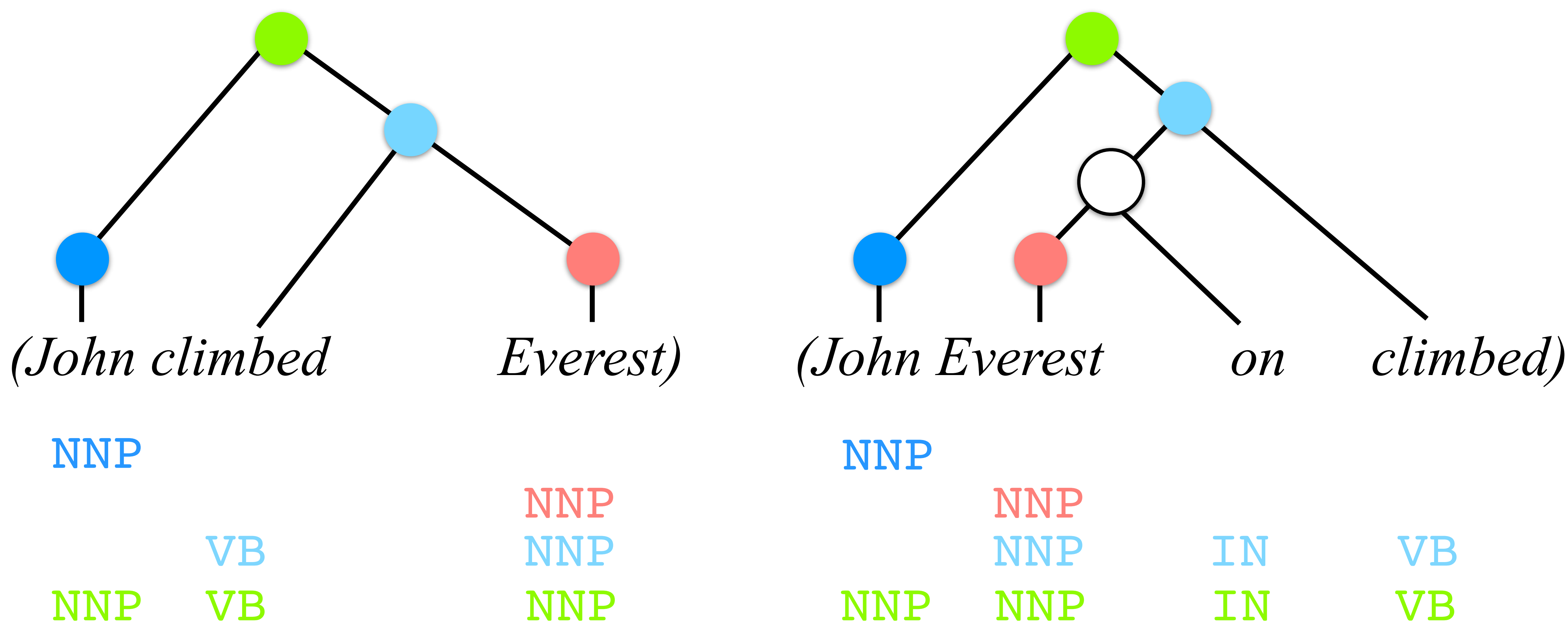
Generative Story

For each *aligned* node pair, draw a *constituent pair* jointly from ω:



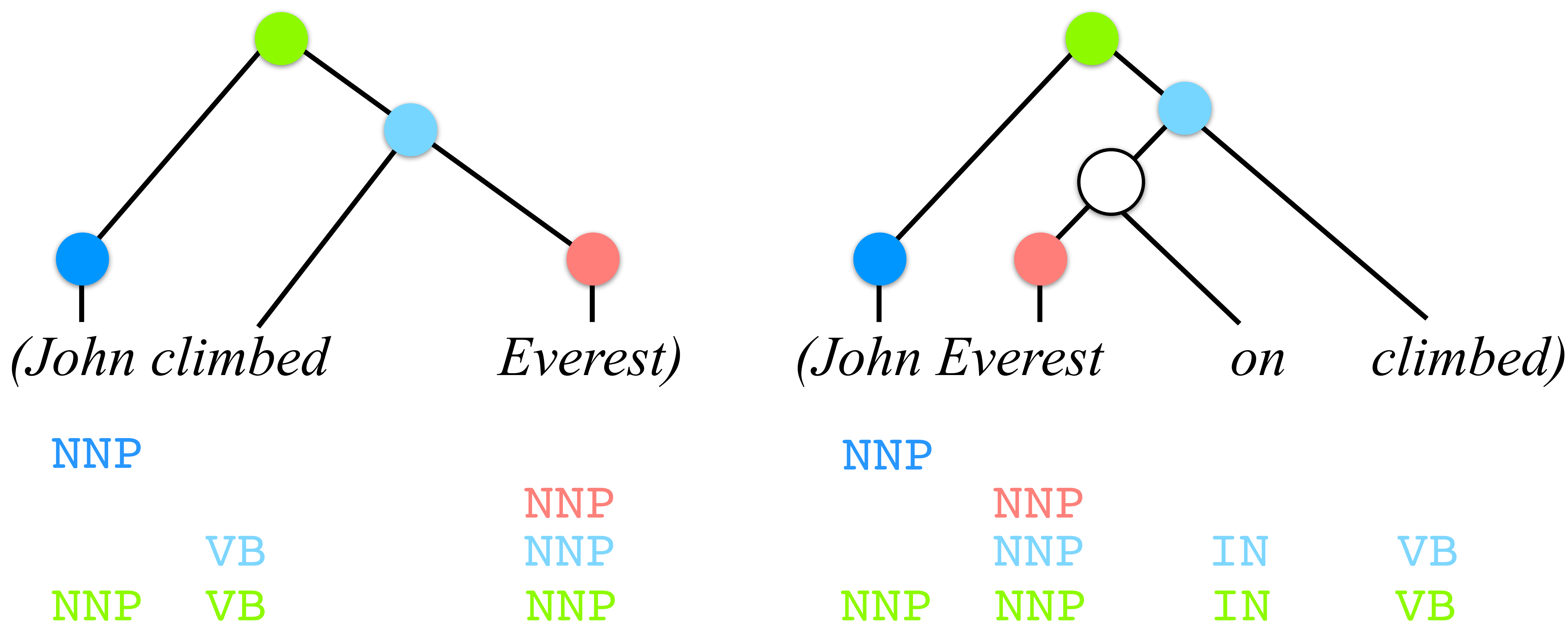
Generative Story

For each *aligned* node pair, draw a *constituent pair* jointly from ω:



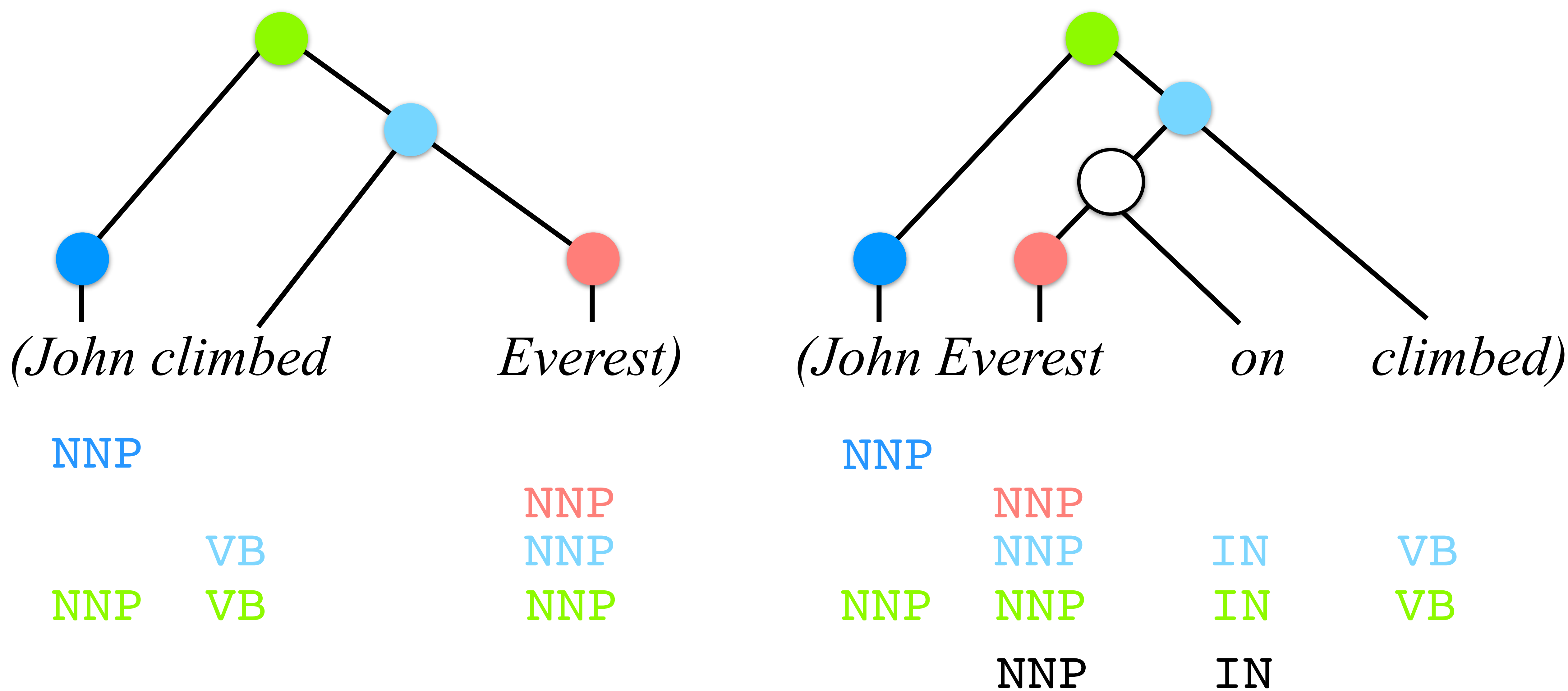
Generative Story

For each *unaligned* node, draw a *constituent* from language-specific parameters:



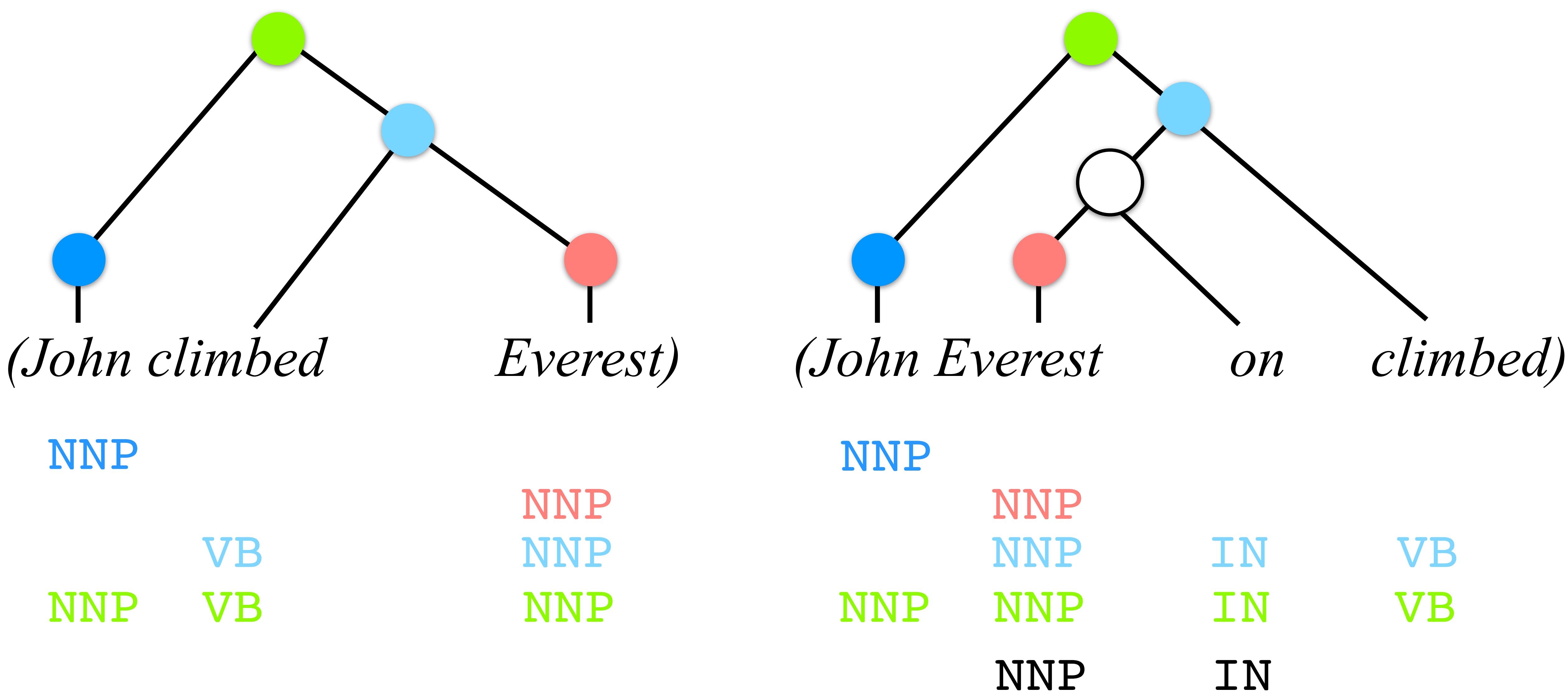
Generative Story

For each *unaligned* node, draw a *constituent* from language-specific parameters:



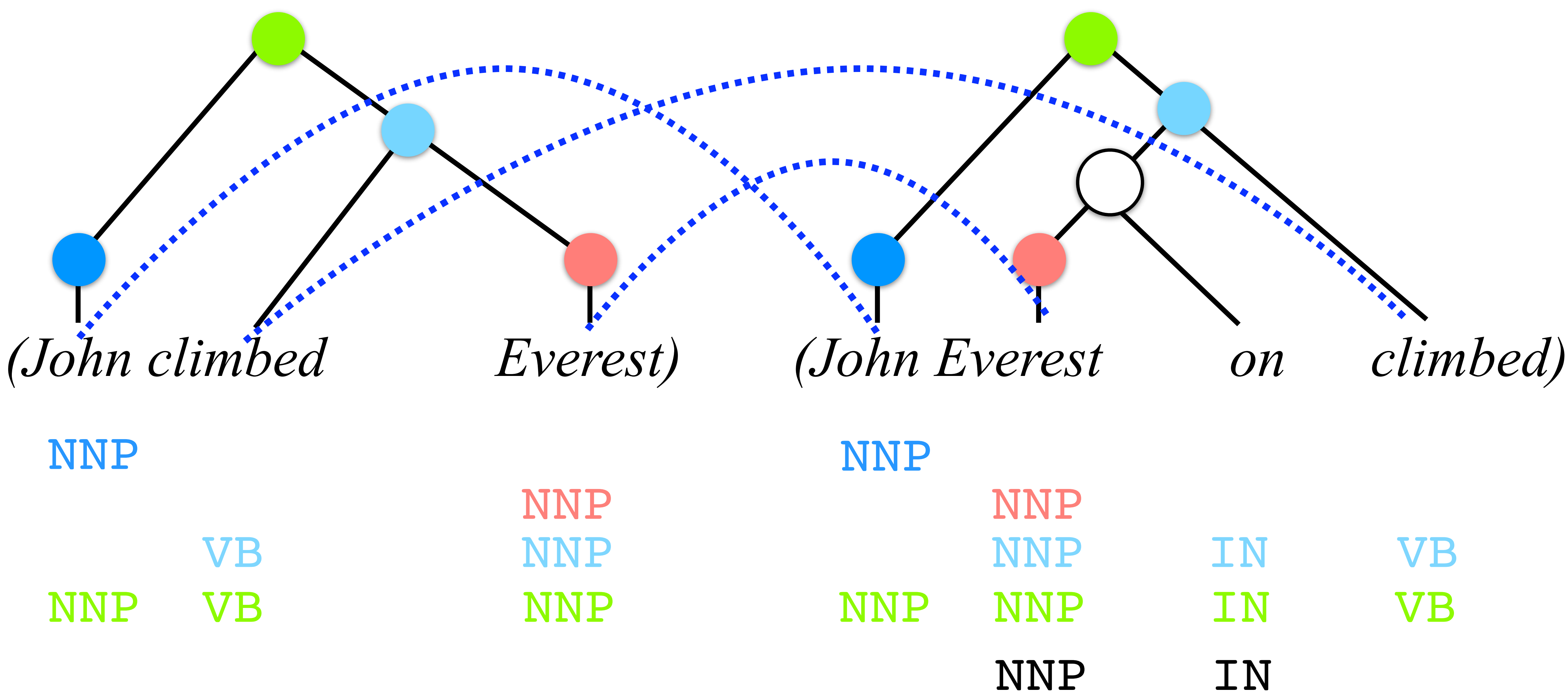
Generative Story

Draw word alignments between *aligned* and *unaligned* nodes according to ϕ^+ and ϕ^-:



Generative Story

Draw word alignments between *aligned* and *unaligned* nodes according to ϕ^{+} and ϕ^{-}:



Inference: Gibbs Sampling

- Sample each aligned tree pair conditioned on others:

$$P\left((T_1, T_2, A)_i \middle| (\mathbf{T_1}, \mathbf{T_2}, \mathbf{A})_{-i}\right)$$

- Marginalize over all parameter values using standard closed forms (accumulated counts + hyperparameters)

Sampling Aligned Trees

Sampling Aligned Trees

- Hard to sample aligned tree pair: (T_1, T_2, A)

Sampling Aligned Trees

- Hard to sample aligned tree pair: (T_1, T_2, A)
- Use *proposal distribution* Q, which assumes *no* nodes are aligned, to separately sample T_1^*, T_2^*

Sampling Aligned Trees

- Hard to sample aligned tree pair: (T_1, T_2, A)
- Use *proposal distribution* Q, which assumes *no* nodes are aligned, to separately sample T_1^*, T_2^*
- Accept with probability:

$$\min\left\{1, \frac{P(T_1^*, T_2^*)\ Q(T_1, T_2)}{P(T_1, T_2)\ Q(T_1^*, T_2^*)}\right\}$$ (Metropolis-Hastings)

Sampling Aligned Trees

- Hard to sample aligned tree pair: (T_1, T_2, A)
- Use *proposal distribution* Q, which assumes *no* nodes are aligned, to separately sample T_1^*, T_2^*
- Accept with probability:

$$\min\left\{1, \frac{P(T_1^*, T_2^*)\ Q(T_1, T_2)}{P(T_1, T_2)\ Q(T_1^*, T_2^*)}\right\}$$ (Metropolis-Hastings)

- Conditionally sample tree alignment: $A|T_1, T_2$

Sampling Aligned Trees

- Hard to sample aligned tree pair: (T_1, T_2, A)
- Use *proposal distribution* Q, which assumes *no* nodes are aligned, to $\boxed{\text{separately sample } T_1^*, T_2^*}$
- Accept with probability:

$$\min\left\{1, \frac{\boxed{\begin{matrix}P(T_1^*, T_2^*)\\ \hline P(T_1, T_2)\end{matrix}}\ }{}\frac{Q(T_1, T_2)}{Q(T_1^*, T_2^*)}\right\}$$ (Metropolis-Hastings)

- Conditionally sample tree alignment: $\boxed{A|T_1, T_2}$

separately sample T_1^*, T_2^*

Sampling each Tree: Inside-Outside

- Recursively sample split-points from the top down
- Calculate probability of each split-point by marginalizing over all possible subtrees ("inside" table of inside-outside)

DT	NN	VB	IN	DT	JJ	NN
The	*boy*	*ran*	*through*	*the*	*haunted*	*house*

$\boxed{\text{computing} \quad P(T_1, T_2)}$ $\Rightarrow$ need to marginalize over all possible alignments A

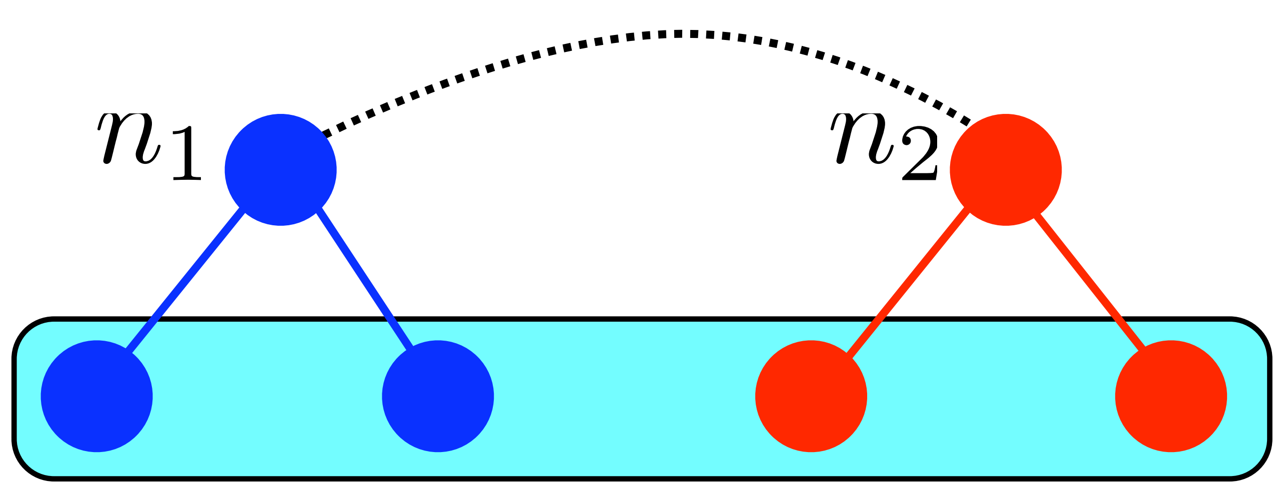
$\boxed{\text{computing } P(T_1, T_2)}$ $\Rightarrow$ need to marginalize over all possible alignments A

- For $n_1 \in T_1, n_2 \in T_2$ table D stores marginal probability of subtrees rooted at n_1, n_2
- Bottom-up dynamic program computes D in time $O(|T_1||T_2|)$

$\boxed{\text{computing } P(T_1, T_2)}$ $\Rightarrow$ need to marginalize over all possible alignments A

- For $n_1 \in T_1, n_2 \in T_2$ table D stores marginal probability of subtrees rooted at n_1, n_2
- Bottom-up dynamic program computes D in time $O(|T_1||T_2|)$

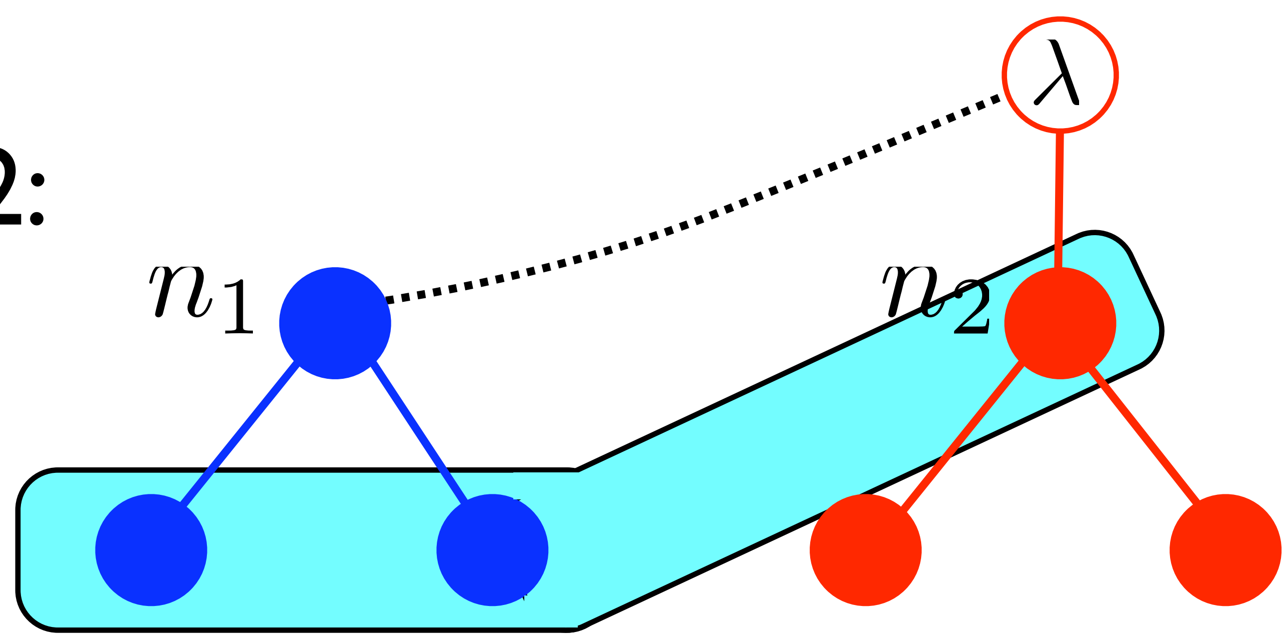
case 1:



$\boxed{\text{computing } P(T_1, T_2)}$ $\Rightarrow$ need to marginalize over all possible alignments A

- For $n_1 \in T_1, n_2 \in T_2$ table D stores marginal probability of subtrees rooted at n_1, n_2
- Bottom-up dynamic program computes D in time $O(|T_1||T_2|)$

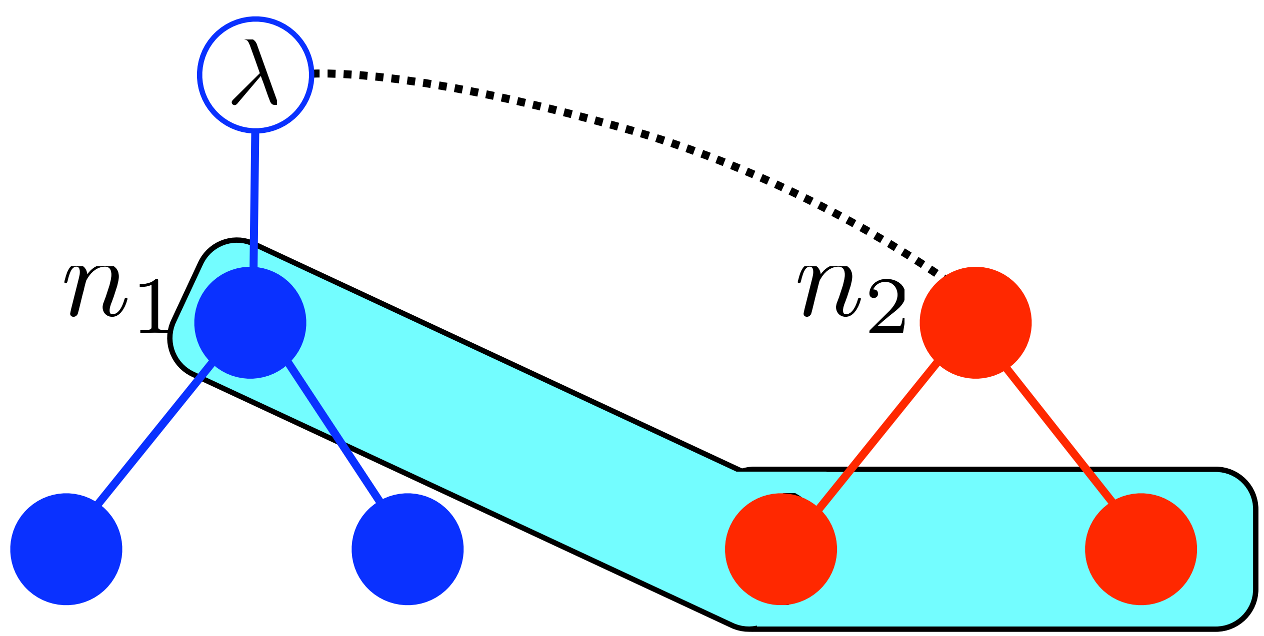
case 2:



$\boxed{\text{computing } P(T_1, T_2)}$ $\Rightarrow$ need to marginalize over all possible alignments A

- For $n_1 \in T_1, n_2 \in T_2$ table D stores marginal probability of subtrees rooted at n_1, n_2
- Bottom-up dynamic program computes D in time $O(|T_1||T_2|)$

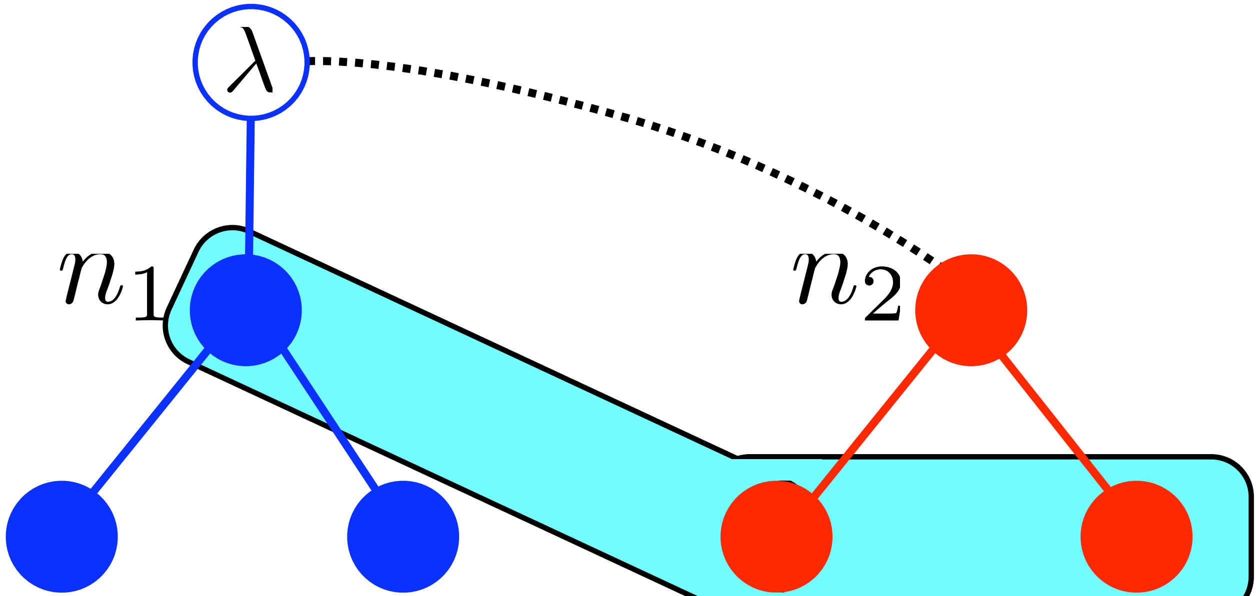
case 3:



computing $P(T_1, T_2)$ $\Rightarrow$ need to marginalize over all possible alignments A

- For $n_1 \in T_1, n_2 \in T_2$ table D stores marginal probability of subtrees rooted at n_1, n_2
- Bottom-up dynamic program computes D in time $O(|T_1||T_2|)$

case 3:



similar for sampling $A|T_1, T_2$

Experiments

<u>Input:</u> Bilingual POS sequences (w/ giza alignments)

<u>Output:</u> Binary tree bracketings

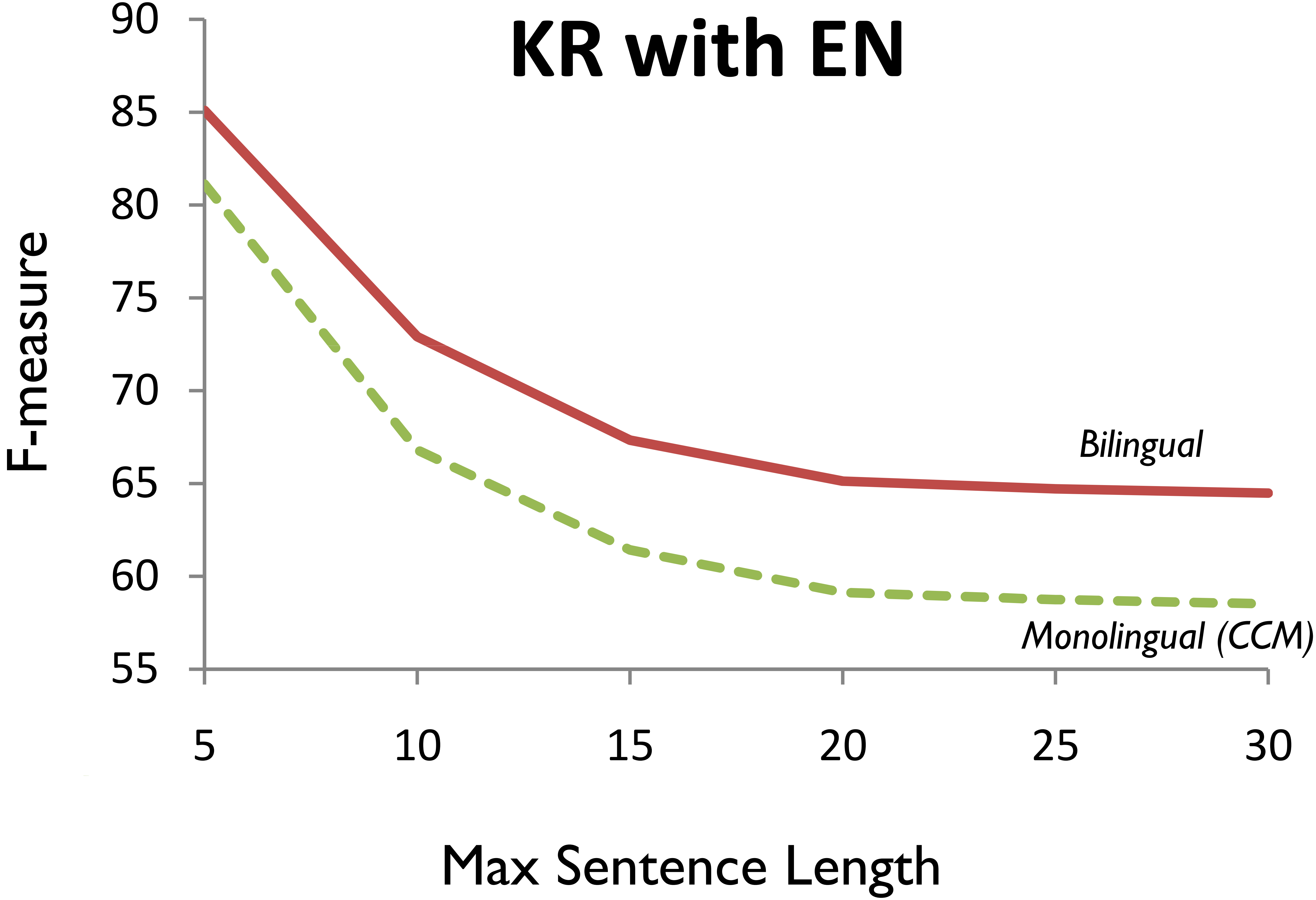
<u>Evaluate:</u> Bracket precision, recall, F-measure, on held-out *monolingual* test data.

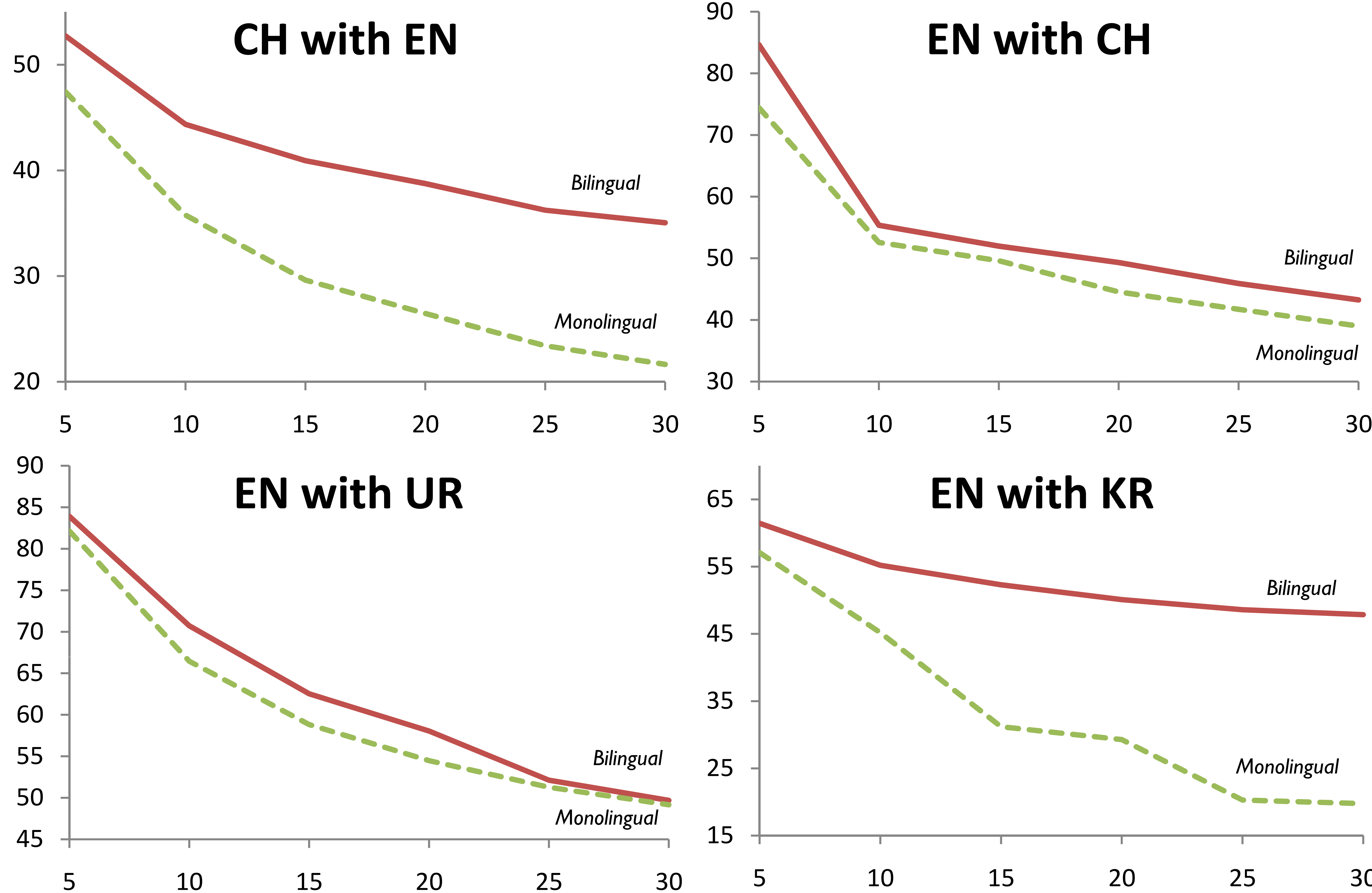
<u>Baseline:</u> (Bayesian) CCM [Klein & Manning 2002]

Corpora

- Korean-English Treebank: 5,000 sentences
- Urdu translation of WSJ: 4,300 sentences
  - no Urdu gold brackets
- English-Chinese Treebank: 3,850 sentences

Evaluate on various maximum sentence lengths (5 - 30)


KR with EN
F-measure
90
85
80
75
70
65
60
55
5
10
15
20
25
30
Max Sentence Length
Bilingual
Monolingual (CCM)


CH with EN
Bilingual
Monolingual
50
40
30
20
5
10
15
20
25
30
EN with CH
Bilingual
Monolingual
90
80
70
60
50
40
30
5
10
15
20
25
30
EN with UR
Bilingual
Monolingual
90
85
80
75
70
65
60
55
50
45
5
10
15
20
25
30
EN with KR
Bilingual
Monolingual
65
55
45
35
25
15
5
10
15
20
25
30

Results

- Average improvement across all scenarios:

  Precision: +10

  Recall: +8

  F-measure: +9

- Average reduction in error relative to binary tree oracle: 19%

Analysis

Percentage of tree nodes aligned

CH-EN	
UR-EN	
KR-EN	

Analysis

Percentage of tree nodes aligned

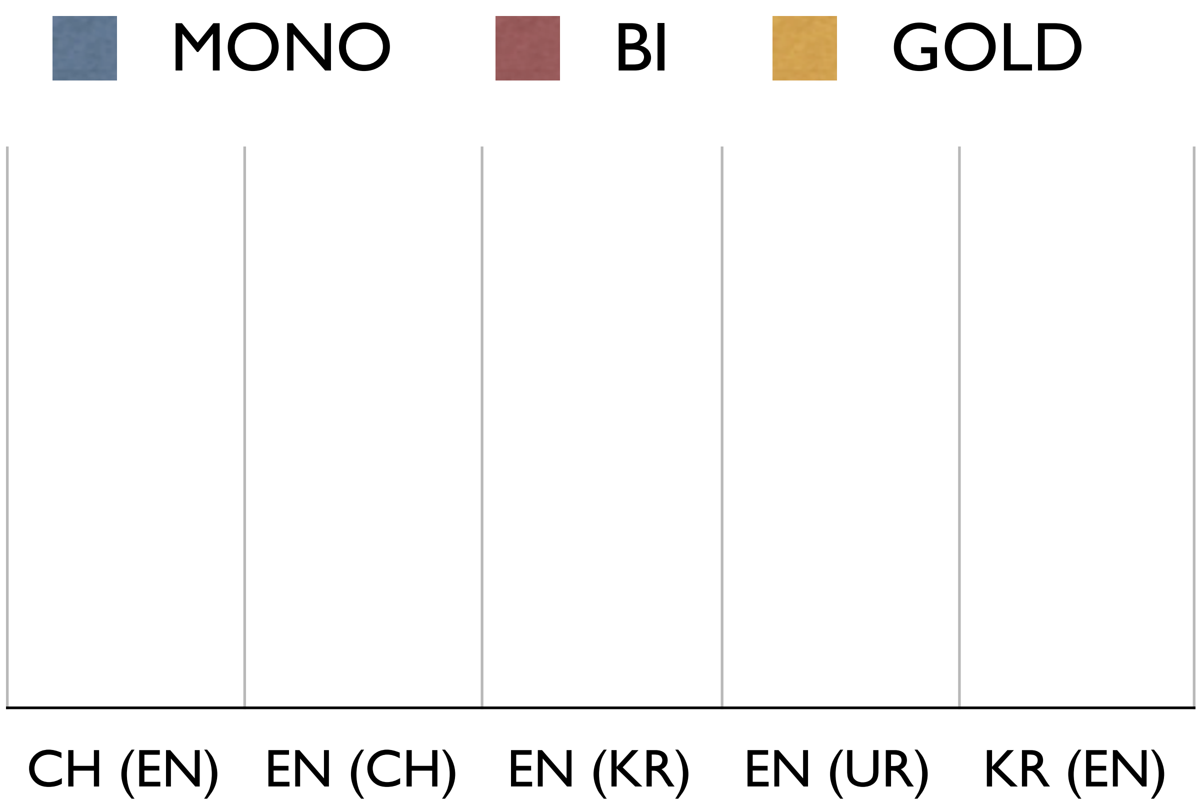
CH-EN	71.6%
UR-EN	68.8%
KR-EN	60.2%

Analysis

Percentage of tree nodes aligned

CH-EN	71.6%
UR-EN	68.8%
KR-EN	60.2%

Entropy of bracketed POS sequences

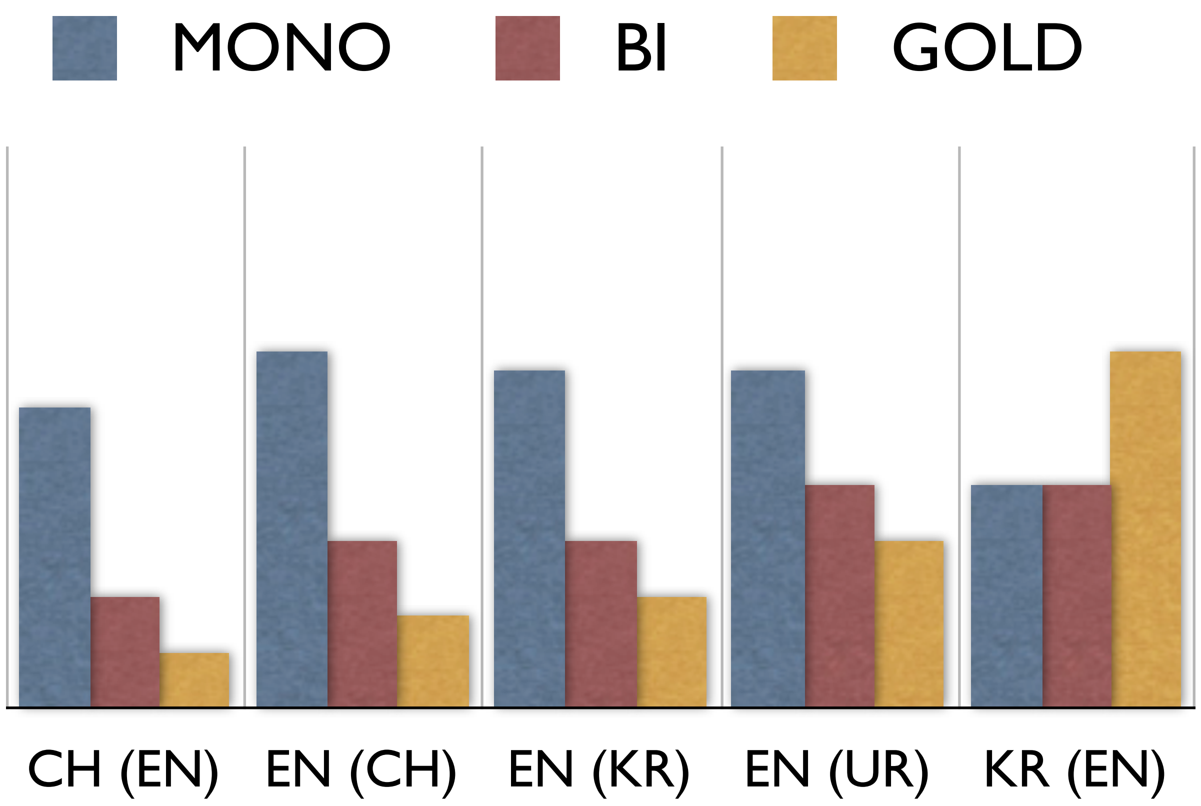


Analysis

Percentage of tree nodes aligned

CH-EN	71.6%
UR-EN	68.8%
KR-EN	60.2%

Entropy of bracketed POS sequences

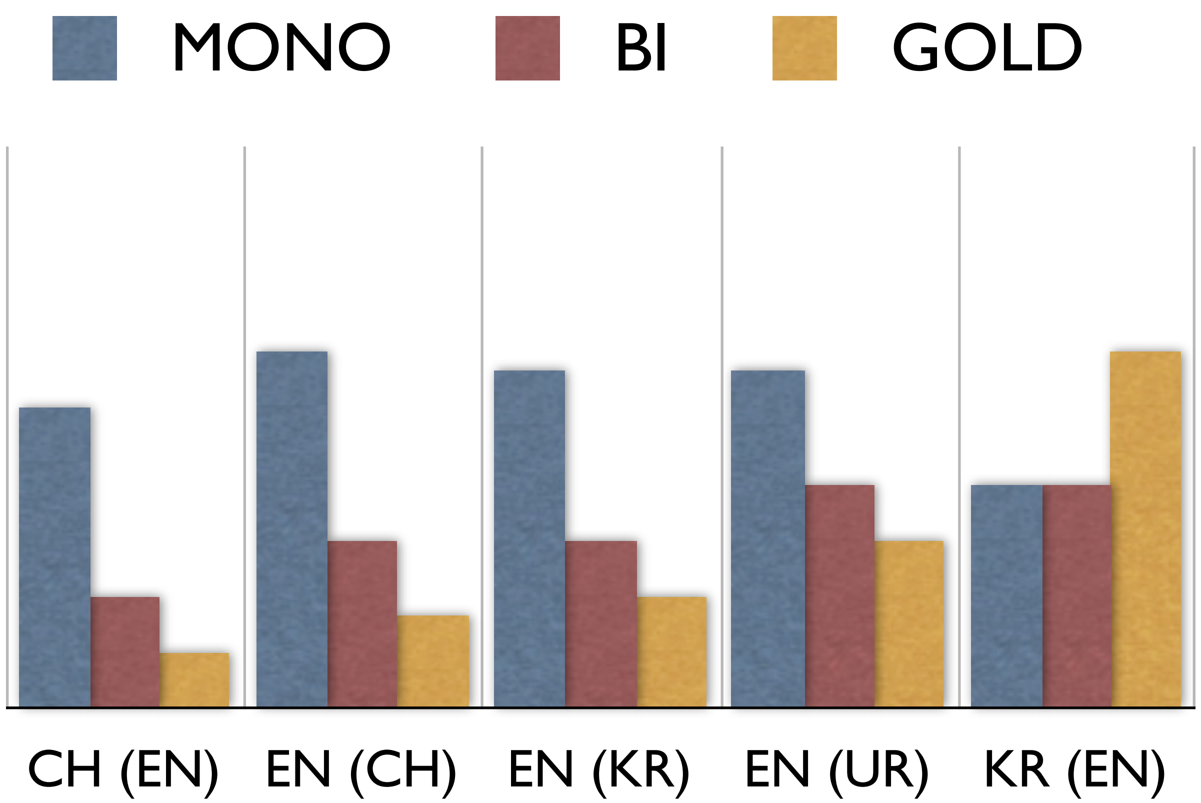


Analysis

Percentage of tree nodes aligned

CH-EN	71.6%
UR-EN	68.8%
KR-EN	60.2%

Entropy of bracketed POS sequences

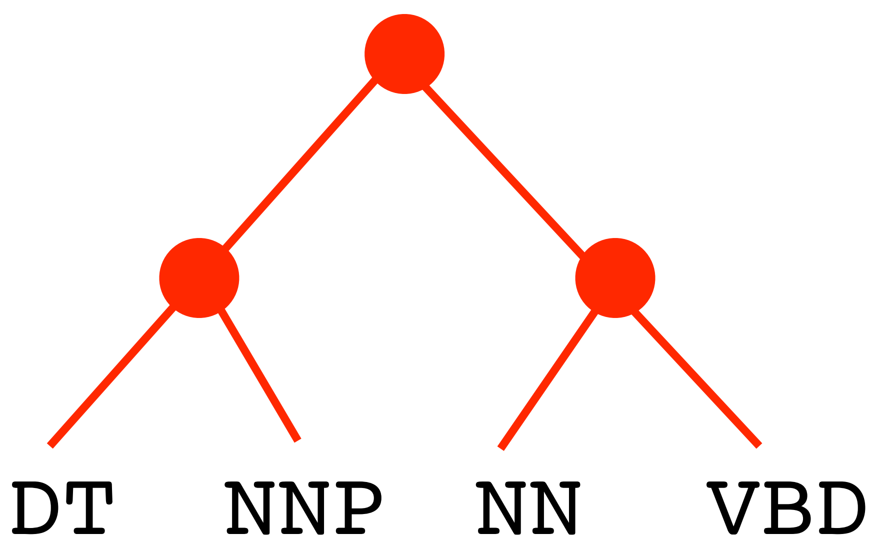


MONO	BI	GOLD
6.7	6.0	5.8

The FCC effort Collapsed

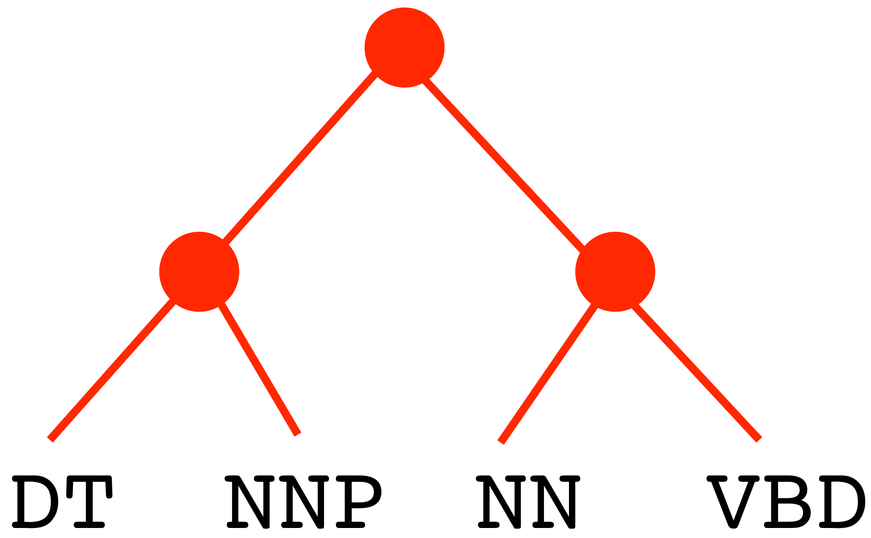
The FCC effort Collapsed

Monolingual X

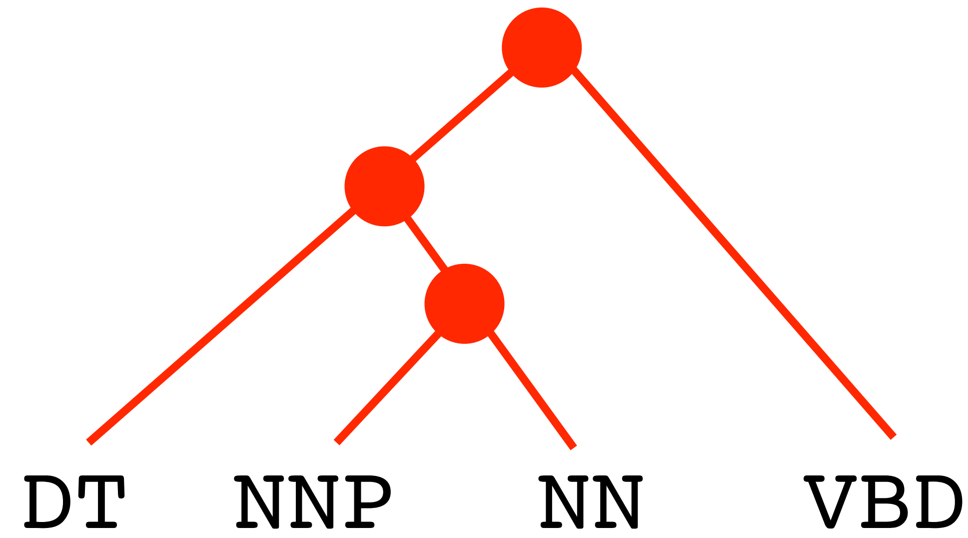


The FCC effort Collapsed

Monolingual X

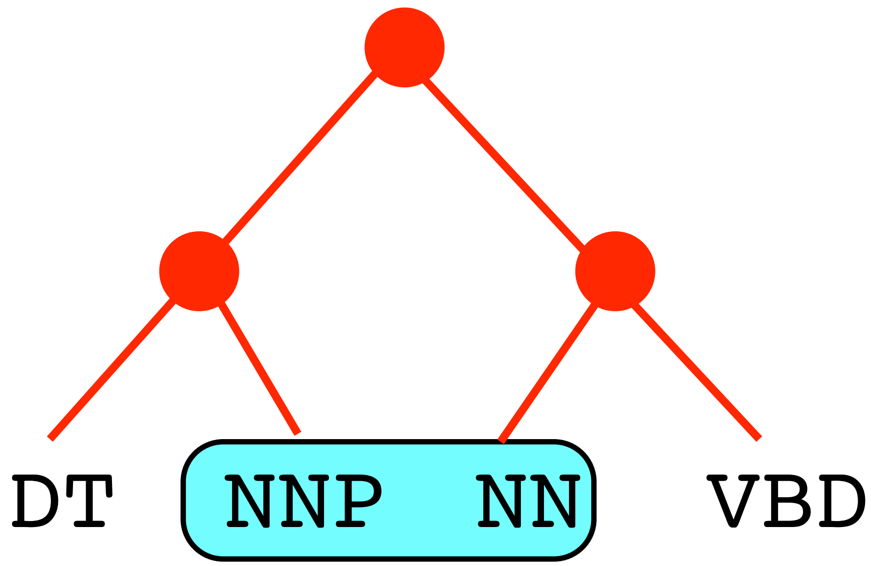


Bilingual (EN-UR) ✓

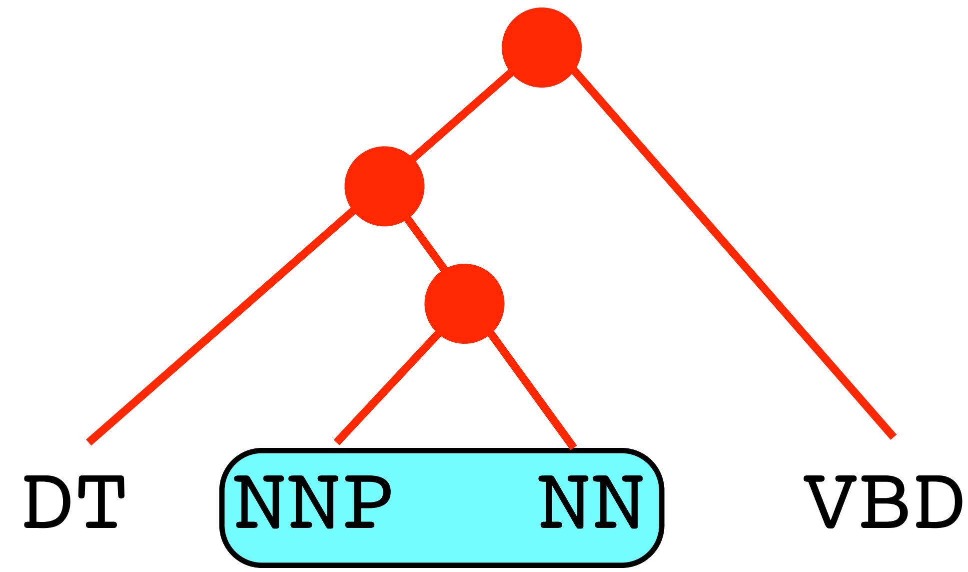


The $\boxed{\text{FCC effort}}$ Collapsed

Monolingual X



Bilingual (EN-UR) ✓



$$\Pr_{mono}(\texttt{NNP NN}) < \Pr_{bi}(\texttt{NNP NN})$$

The FCC effort Collapsed

Monolingual X　　　　　　*Bilingual (EN-UR)* ✓

$$Pr_{mono}(\texttt{NNP}\;\;\texttt{NN}) < Pr_{bi}(\texttt{NNP}\;\;\texttt{NN})$$

English: NNP NN
Urdu: NNP OF NN

Conclusions

Key idea: Use bilingual cues to learn better unsupervised monolingual models of grammar

- Adapt *Tree Alignment* to probabilistic setting:
  - Discover partial shared structure
  - Allow language-specific divergence
  - Computationally tractable
- Achieve improved performance on five corpora, across all sentence lengths

Thank you!

Analysis

Entropy of constituent tag sequences

Percentage of a
tree node

CH-EN	
UR-EN	
KR-EN	

	MONO	*BI*	*GOLD*
CH_{EN}	6.6	5.6	5.3
EN_{CH}	6.9	5.9	5.5
KR_{EN}	6.2	6.2	6.9
EN_{KR}	6.8	5.9	5.6
EN_{UR}	6.8	6.2	5.9
avg	6.7	6.0	5.8

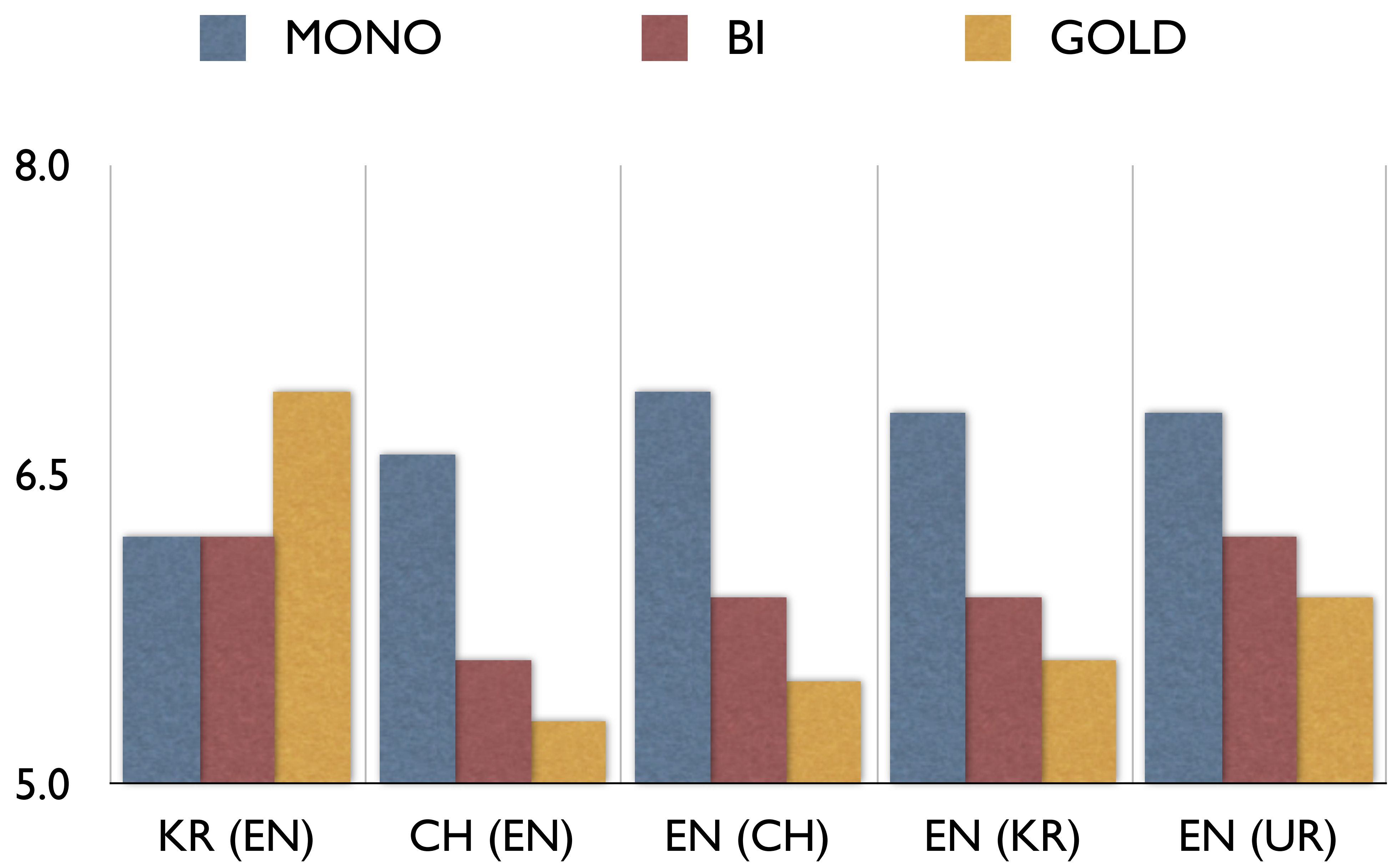
Analysis

Entropy of constituent tag sequences

Percentage of a
tree node

CH-EN	71.
UR-EN	68.
KR-EN	60.

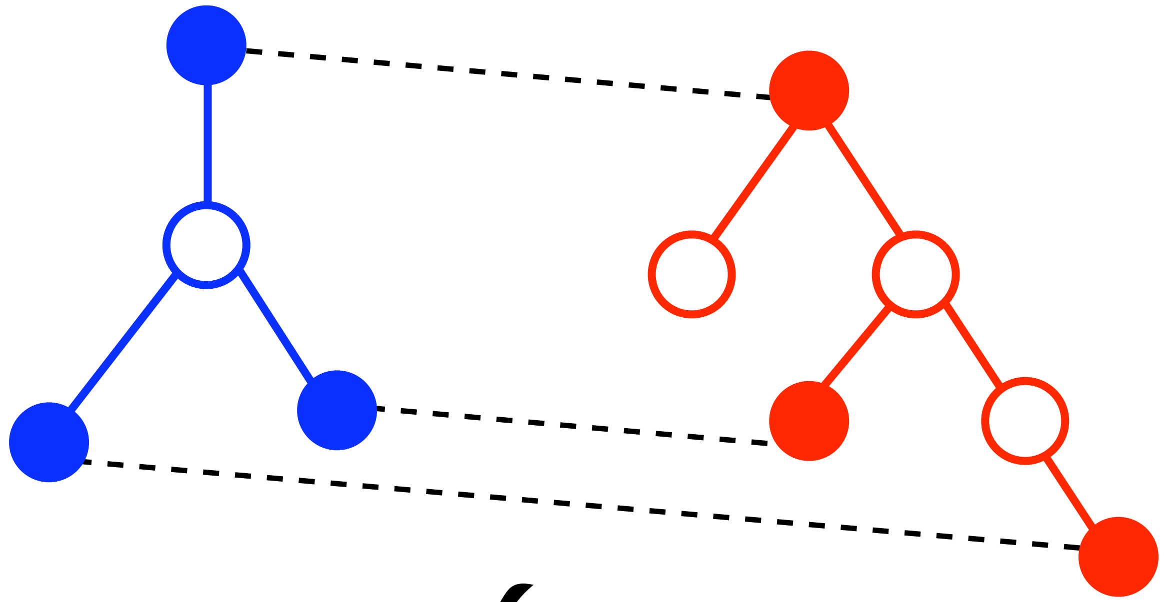
	MONO	*BI*	*GOLD*
CH$_{EN}$	6.6	5.6	5.3
EN$_{CH}$	6.9	5.9	5.5
KR$_{EN}$	6.2	6.2	6.9
EN$_{KR}$	6.8	5.9	5.6
EN$_{UR}$	6.8	6.2	5.9
avg	6.7	6.0	5.8


MONO
BI
GOLD
8.0
6.5
5.0
KR (EN)
CH (EN)
EN (CH)
EN (KR)
EN (UR)

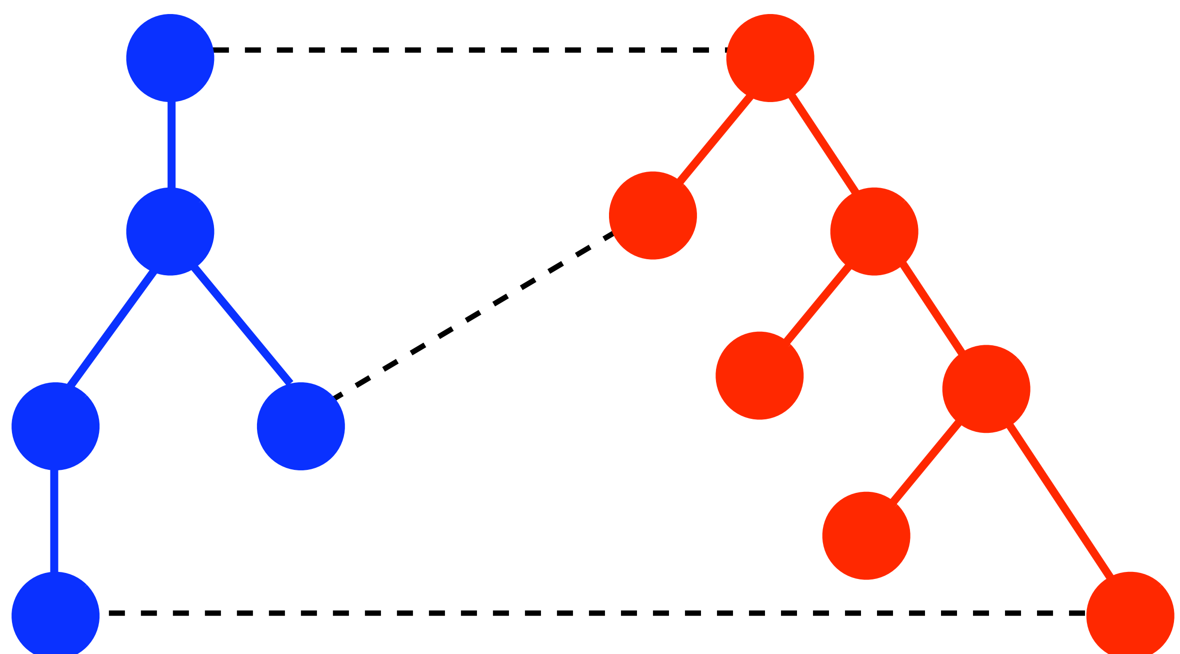
Morphology:
acl 2008

POS tagging:
emnlp 2008
naacl 2009

Syntax:
acl 2009 (this talk)



✓



✓