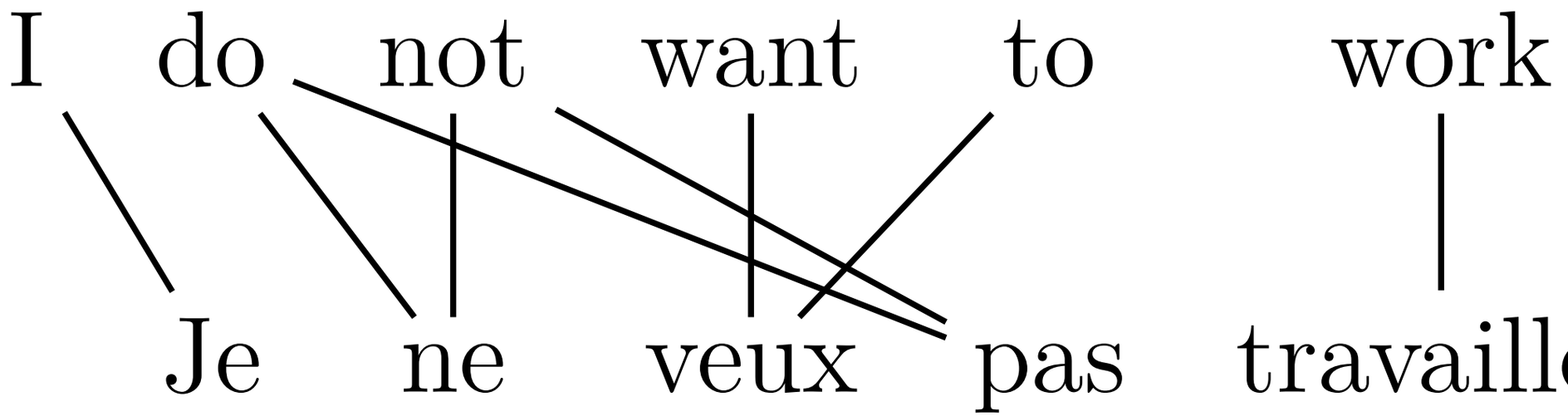
Tree-based Models for SMT

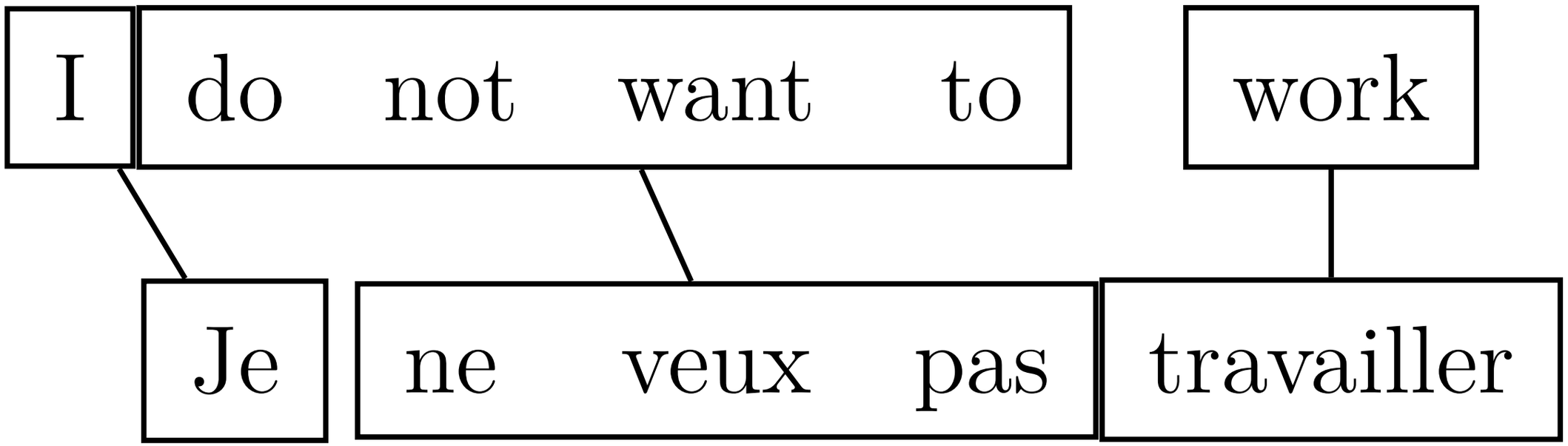
Taro Watanabe

Word-based MT



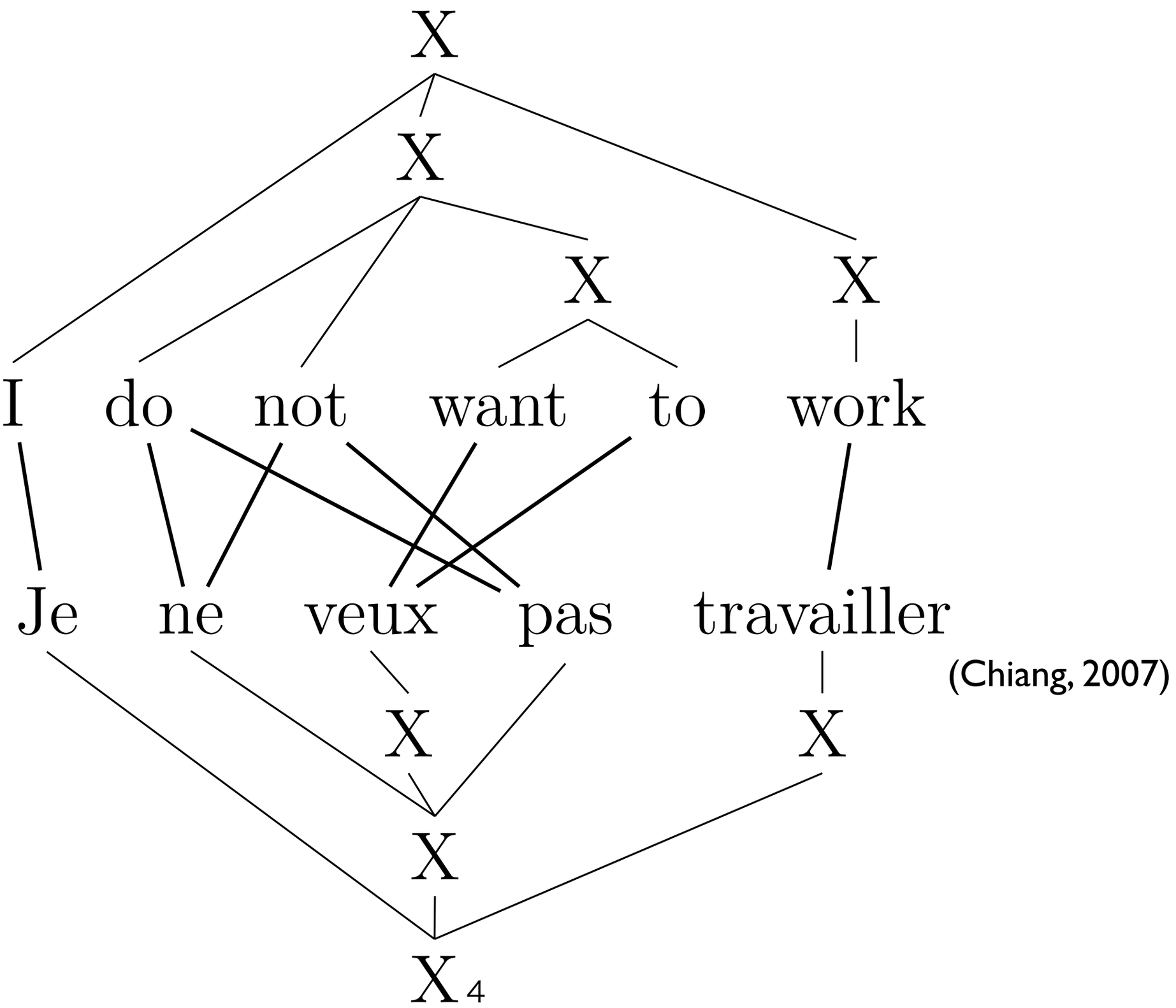
(Brown et al., 1993)

Phrase-based MT



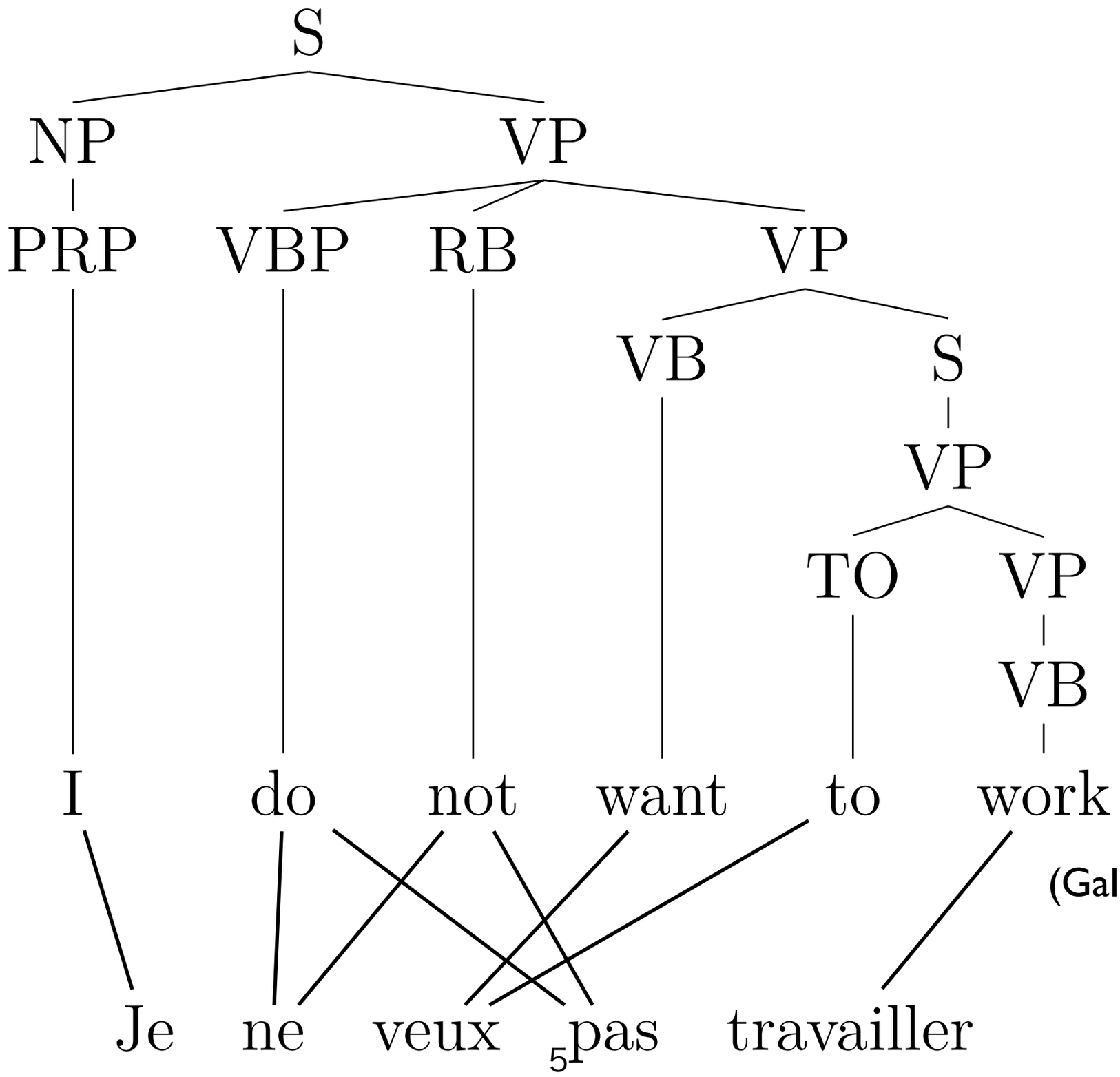
(Koehn et al., 2003)

Hierarchical PBMT



(Chiang, 2007)

Syntax-based MT



(Galley et al., 2004)

Variations

tree	(partial) examples
none	Chiang (2007), Zollman and Venugopal (2006)
source	Huang et al. (2006), Liu et al. (2006), Quirk et al. (2005)
target	Galley et al. (2004), Shen et al. (2008)
both	Ding and Palmer (2005), Liu et al. (2009)

- formally syntactical, linguistically syntactical
- dependency structure and constituency structure
- {tree,string}-to-{tree,string}
- In this talk, we will summarize them as "tree-based MT"

Overview

- Backgrounds
  - CFG, parsing, hypergraph, deductive system semirings
- Tree-based SMT
  - Synchronous-CFG
  - String-to-Tree, Tree-to-String
  - Bitext parsing

Overview

- Backgrounds
  - CFG, parsing, hypergraph, deductive system semirings
- Tree-based SMT
  - Synchronous-CFG
  - String-to-Tree, Tree-to-String
  - Bitext parsing

Backgrounds: CFG

- parsing = intersection of CFG with a string (regular grammar)

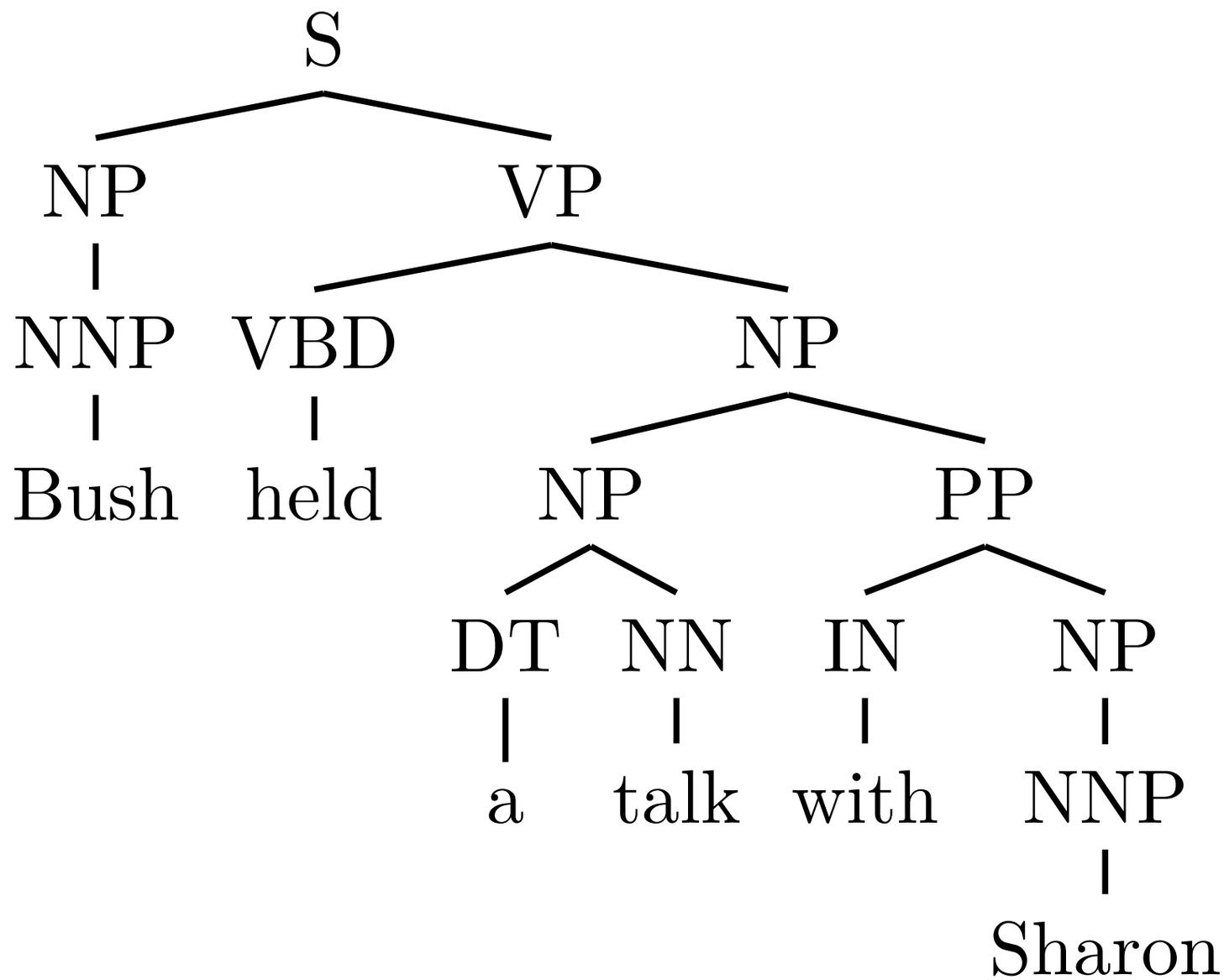
Backgrounds: CFG

$$
\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{NNP} \\
\text{NP} &\rightarrow \text{NP PP} \\
\text{NP} &\rightarrow \text{DP NN} \\
\text{NP} &\rightarrow \text{DT NN} \\
\text{VP} &\rightarrow \text{VBD NP} \\
\text{NNP} &\rightarrow \text{Bush} \\
\text{VBD} &\rightarrow \text{held} \\
&\vdots
\end{aligned}
$$

- parsing = intersection of CFG with a string (regular grammar)

Backgrounds: CFG

$$
\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{NNP} \\
\text{NP} &\rightarrow \text{NP PP} \\
\text{NP} &\rightarrow \text{DP NN} \\
\text{NP} &\rightarrow \text{DT NN} \\
\text{VP} &\rightarrow \text{VBD NP} \\
\text{NNP} &\rightarrow \text{Bush} \\
\text{VBD} &\rightarrow \text{held} \\
&\vdots
\end{aligned}
$$

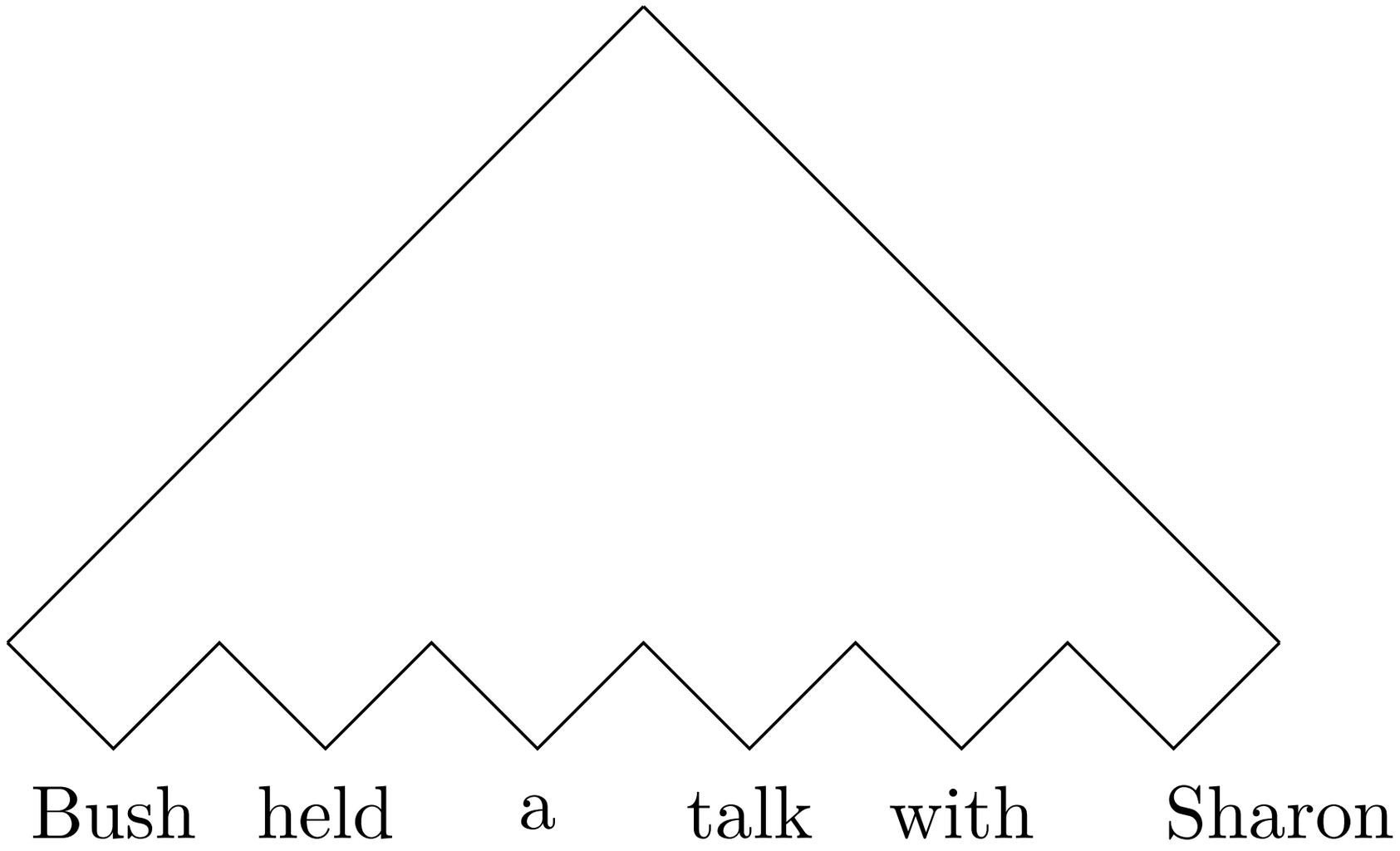


- parsing = intersection of CFG with a string (regular grammar)

Parsing: CKY

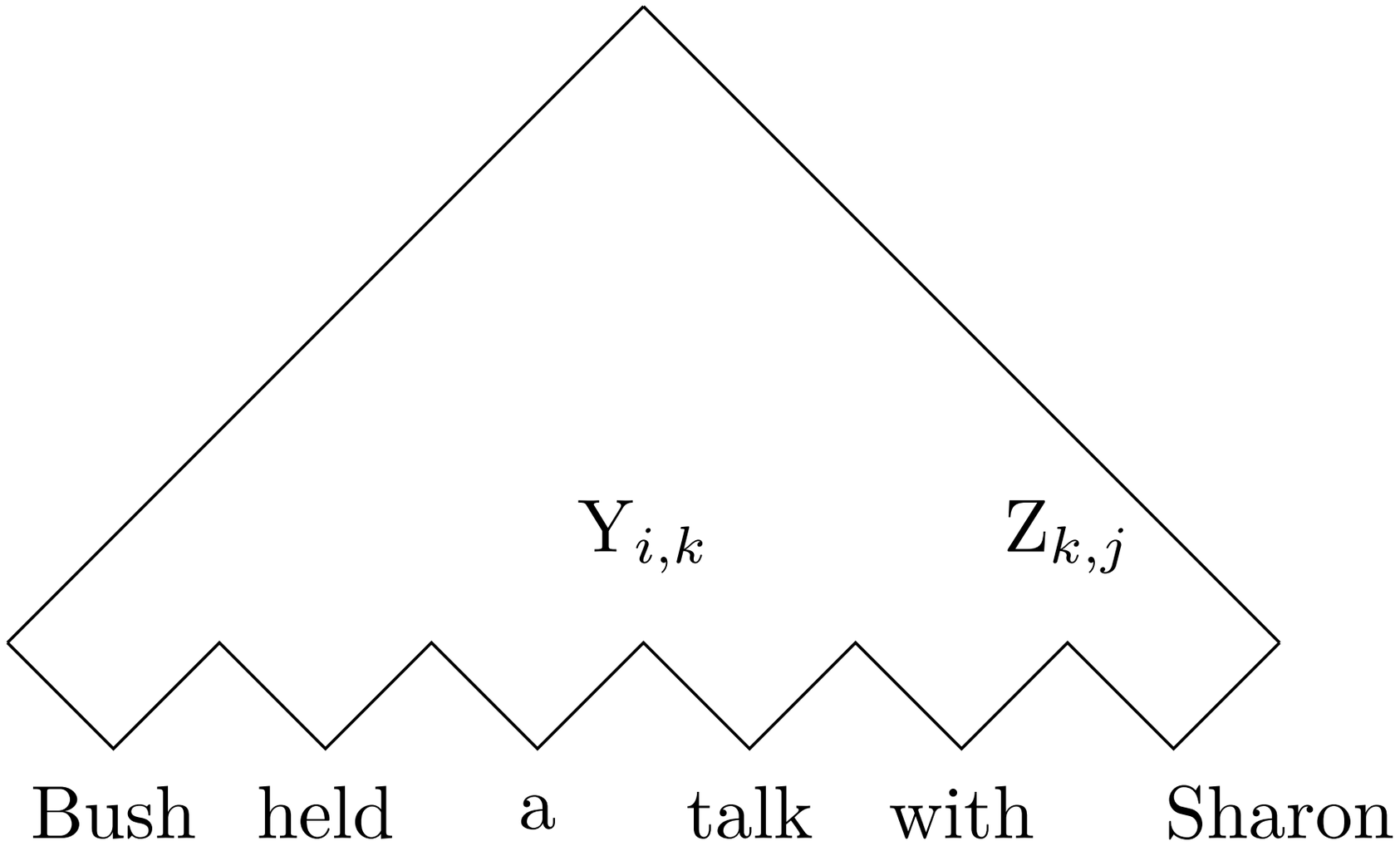
- $O(n^3)$: For each length n, for each position i, for each rule $X \rightarrow Y\ Z$, for each split point k
- (Bottom-up) topological order

Parsing: CKY



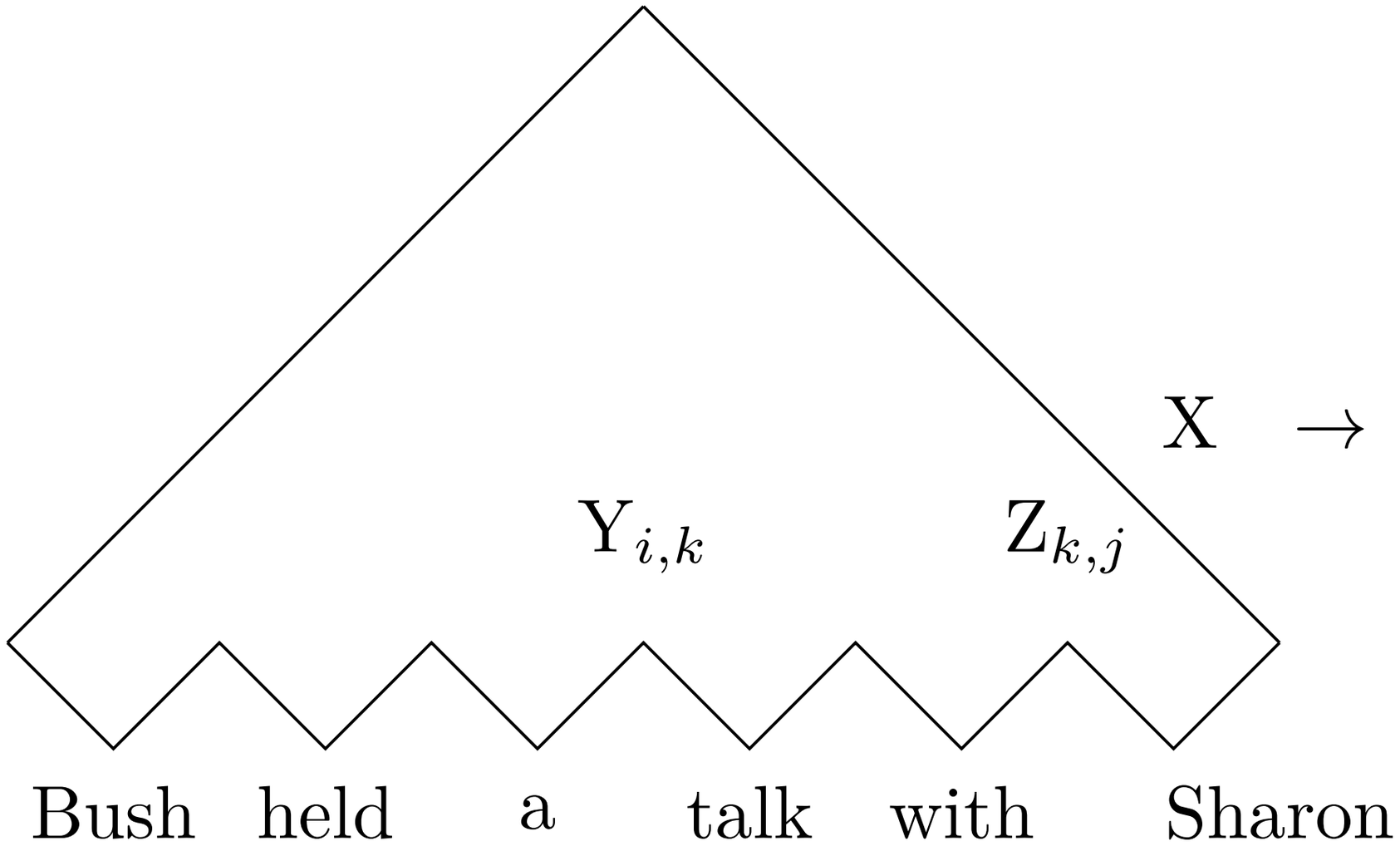
- $O(n^3)$: For each length n, for each position i, for each rule $X \rightarrow Y\ Z$, for each split point k
- (Bottom-up) topological order

Parsing: CKY



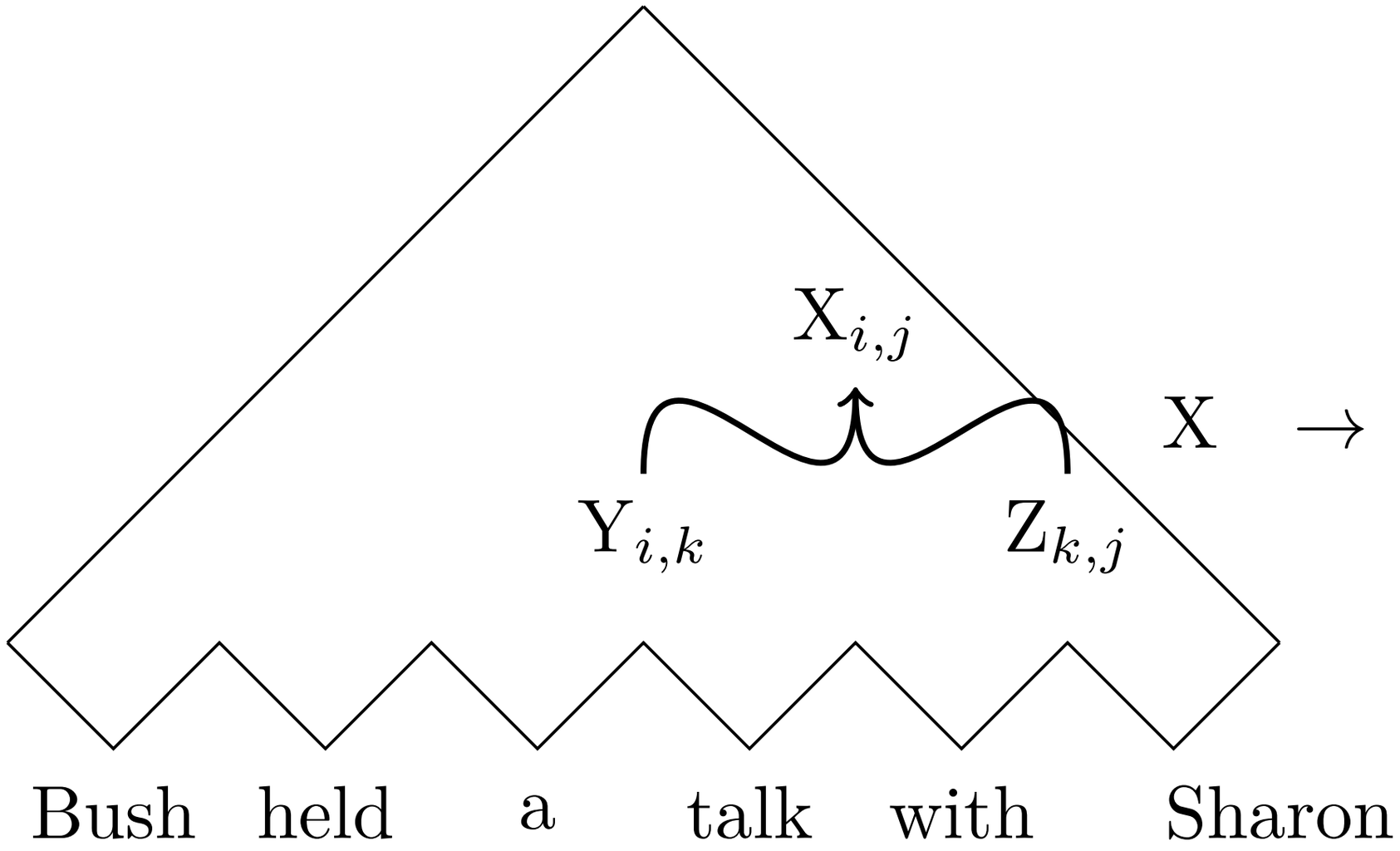
- $O(n^3)$: For each length n, for each position i, for each rule $X \rightarrow Y\ Z$, for each split point k
- (Bottom-up) topological order

Parsing: CKY



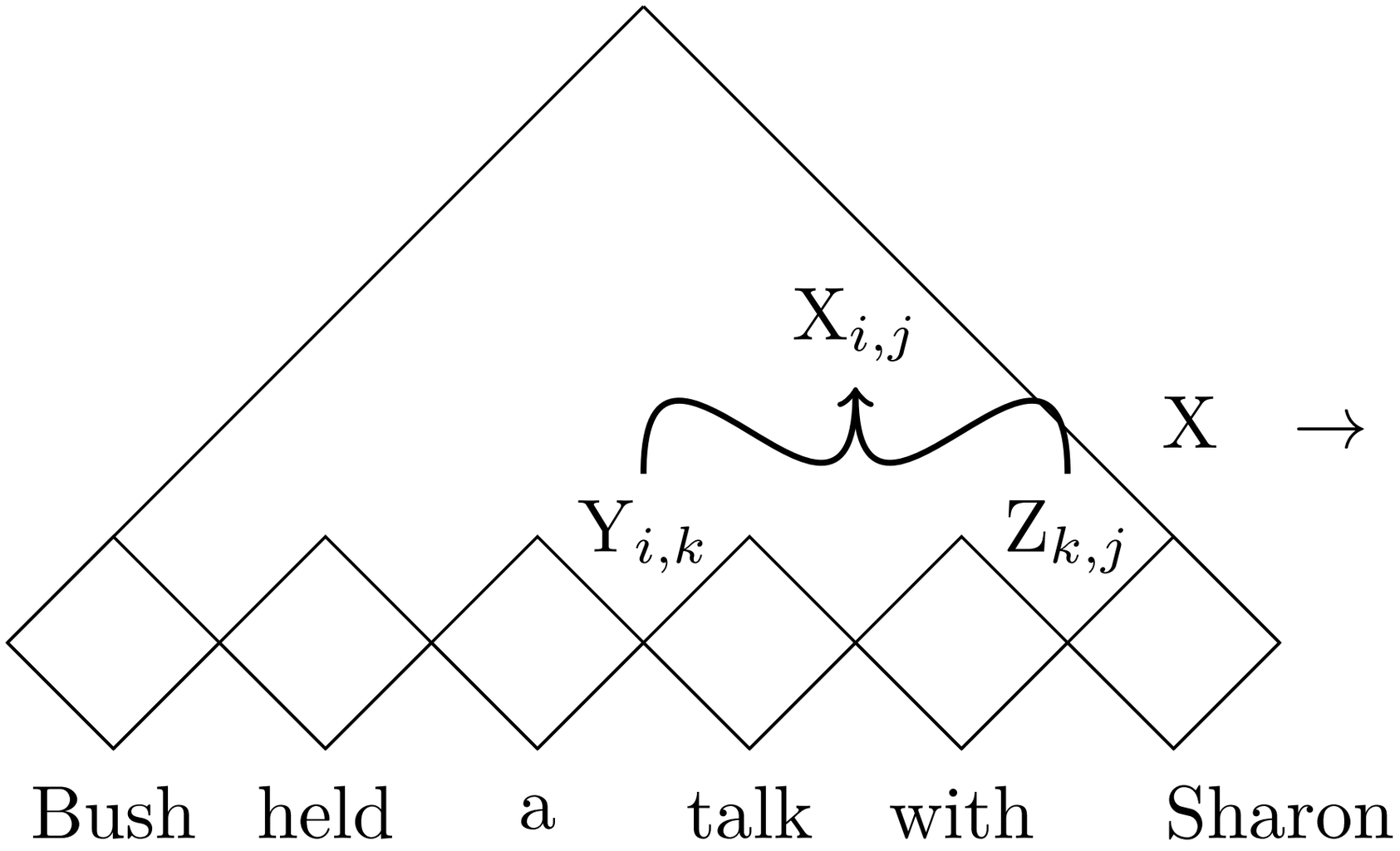
- $O(n^3)$: For each length n, for each position i, for each rule X → Y Z, for each split point k
- (Bottom-up) topological order

Parsing: CKY



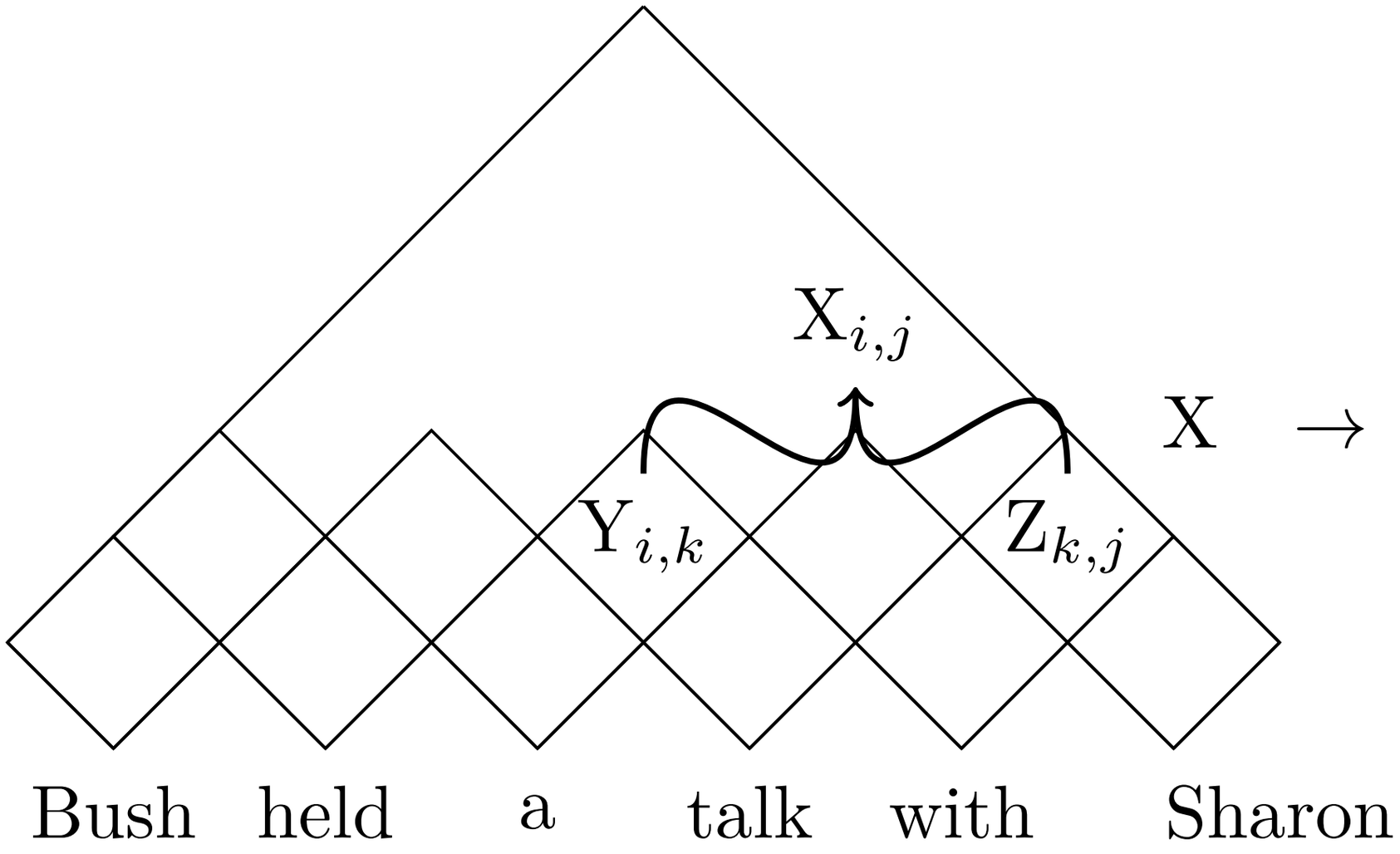
- $O(n^3)$: For each length n, for each position i, for each rule $X \rightarrow Y\ Z$, for each split point k
- (Bottom-up) topological order

Parsing: CKY



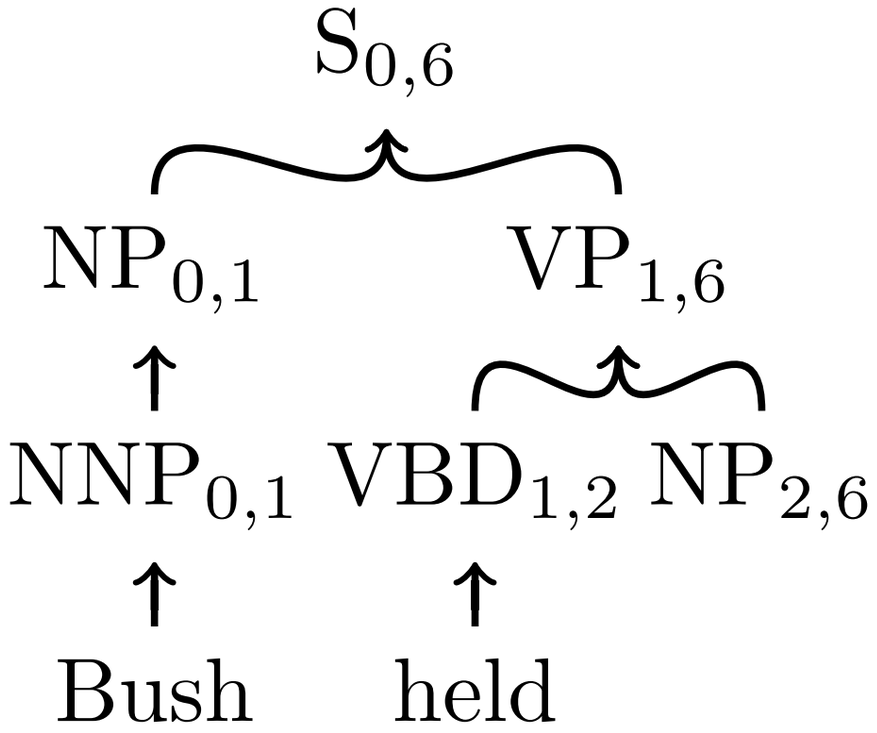
- $O(n^3)$: For each length n, for each position i, for each rule $X \rightarrow Y\ Z$, for each split point k
- (Bottom-up) topological order

Parsing: CKY



- $O(n^3)$: For each length n, for each position i, for each rule X → Y Z, for each split point k
- (Bottom-up) topological order

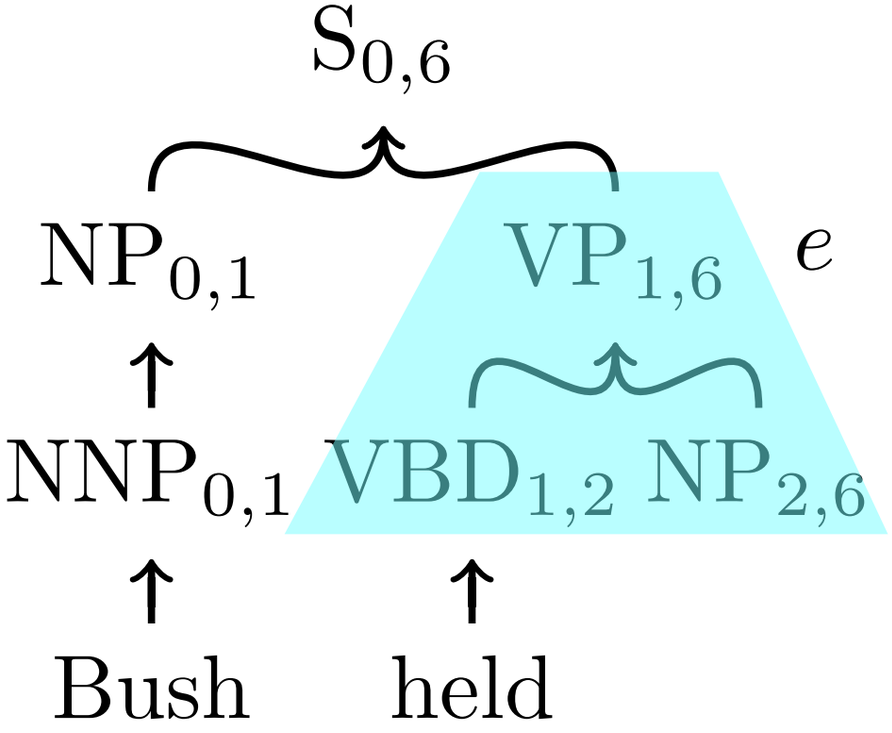
Hypergraph



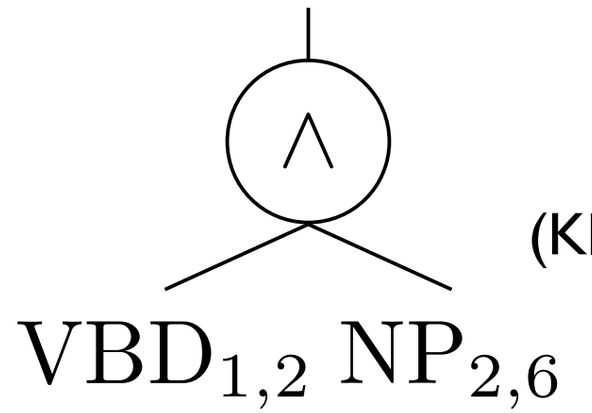
(Klein and Manning, 2001)

- Generalization of graphs:
  - h(e): head node of hyperedge e
  - T(e): tail node(s) of hyperedge e, arity = |T(e)|
  - hyperedge = instantiated rule
- Represented as and-or graphs

Hypergraph



$$e = \langle \underbrace{\mathrm{VP}_{1,6}}_{h(e)}, \underbrace{\{\mathrm{VBD}_{1,2}, \mathrm{NP}_{2,6}\}}_{T(e)} \rangle$$



(Klein and Manning, 2001)

- Generalization of graphs:
  - h(e): head node of hyperedge e
  - T(e): tail node(s) of hyperedge e, arity = |T(e)|
  - hyperedge = instantiated rule
- Represented as and-or graphs

Deductive System

(Shieber et al., 1995)

- Parsing algorithm as a deductive system
- We start from initial items (axioms) until we reach a goal item
- If antecedents are proved, its consequent is proved
- deduction = hyperedge

Deductive System

$$\frac{\overbrace{\mathrm{VBD}_{1,2}\ \mathrm{NP}_{2,6}}^{\text{antecedents}}}{\underbrace{\mathrm{VP}_{1,6}}_{\text{consequent}}}\ \mathrm{VP}_{[i,j]} \rightarrow \mathrm{VBZ}_{[j,k]}\ \mathrm{NP}_{[i,k]}$$

(Shieber et al., 1995)

- Parsing algorithm as a deductive system
- We start from initial items (axioms) until we reach a goal item
- If antecedents are proved, its consequent is proved
- deduction = hyperedge

Deductive System

$$\overbrace{\mathrm{VBD}_{1,2} \quad \mathrm{NP}_{2,6}}^{\mathrm{VP}_{1,6}}$$

$$\frac{\overbrace{\mathrm{VBD}_{1,2}\ \mathrm{NP}_{2,6}}^{\text{antecedents}}}{\underbrace{\mathrm{VP}_{1,6}}_{\text{consequent}}}\ \mathrm{VP}_{[i,j]} \rightarrow \mathrm{VBZ}_{[j,k]}\ \mathrm{NP}_{[i,k]}$$

(Shieber et al., 1995)

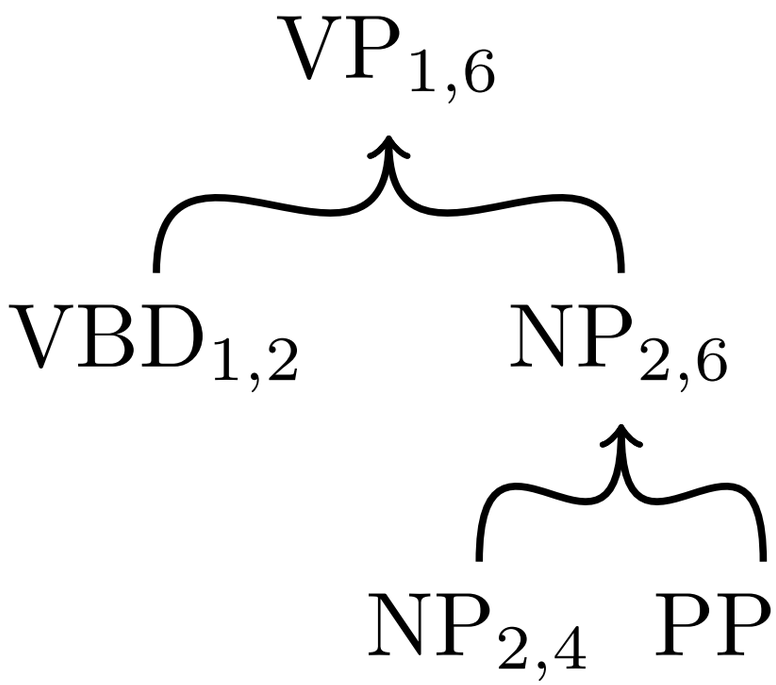
- Parsing algorithm as a deductive system
- We start from initial items (axioms) until we reach a goal item
- If antecedents are proved, its consequent is proved
- deduction = hyperedge

Packed Forest

(Klein and Manning, 2001; Huang and Chiang, 2005)

- A polynomial space encoding of exponentially many parses by sharing common sub-derivations
- Single derivation = tree

Packed Forest

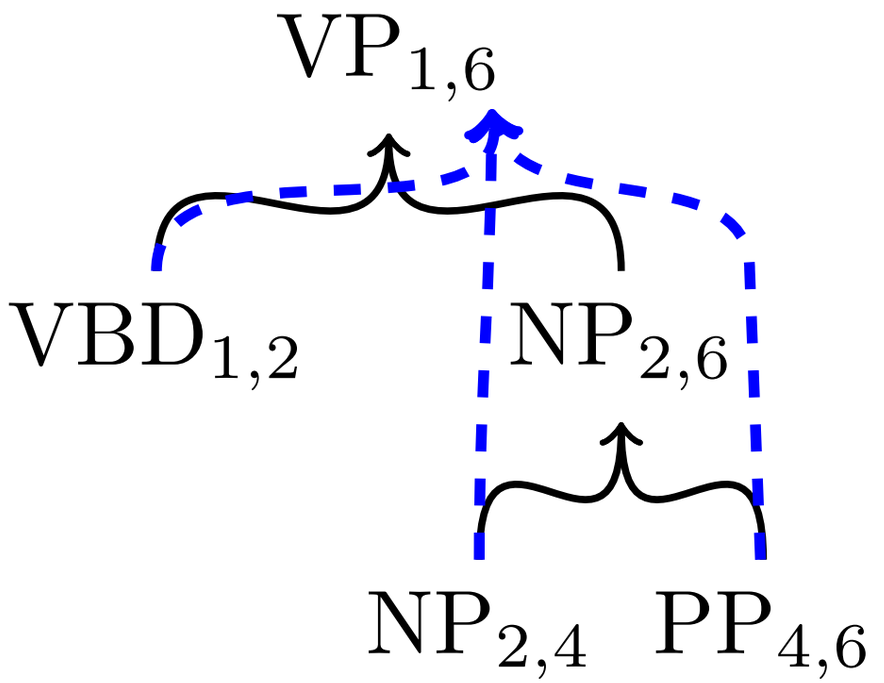


$$\frac{\mathrm{VBD}_{1,2} \quad \dfrac{\mathrm{NP}_{2,4} \ \mathrm{PP}_{4,6}}{\mathrm{NP}_{2,6}}}{\mathrm{VP}_{1,6}}$$

(Klein and Manning, 2001; Huang and Chiang, 2005)

- A polynomial space encoding of exponentially many parses by sharing common sub-derivations
- Single derivation = tree

Packed Forest



$$\frac{\mathrm{VBD}_{1,2} \quad \frac{\mathrm{NP}_{2,4} \ \mathrm{PP}_{4,6}}{\mathrm{NP}_{2,6}}}{\mathrm{VP}_{1,6}}$$

$$\frac{\mathrm{VBD}_{1,2} \ \mathrm{NP}_{2,4} \ \mathrm{PP}_{4,6}}{\mathrm{VP}_{1,6}}$$

(Klein and Manning, 2001; Huang and Chiang, 2005)

- A polynomial space encoding of exponentially many parses by sharing common sub-derivations
- Single derivation = tree

Summary of Formalisms

Summary of Formalisms

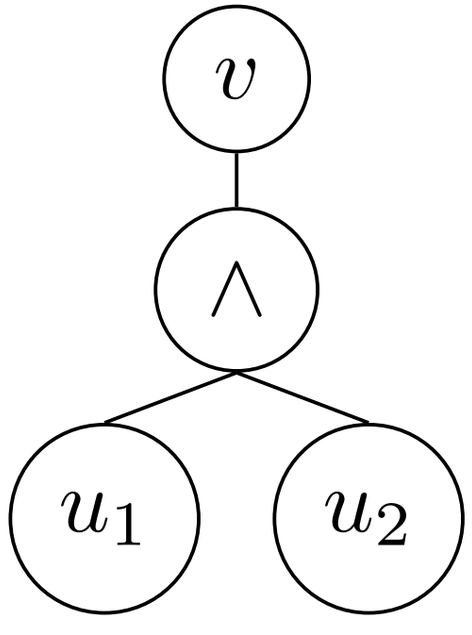
hypergraph
vertex source-vertex target-vertex
hyperedge

$\langle v, \{u_1, u_2\} \rangle$

Summary of Formalisms

hypergraph	AND/OR graph
vertex	OR-node
source-vertex	leaf OR-node
target-vertex	root OR-node
hyperedge	AND-node

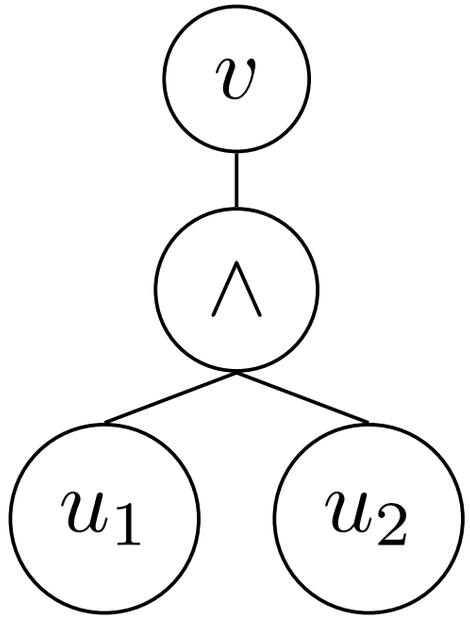
$\langle v, \{u_1, u_2\} \rangle$



Summary of Formalisms

hypergraph	AND/OR graph	CFG
vertex	OR-node	symbol
source-vertex	leaf OR-node	terminal
target-vertex	root OR-node	start symbol
hyperedge	AND-node	production

$\langle v, \{u_1, u_2\} \rangle$

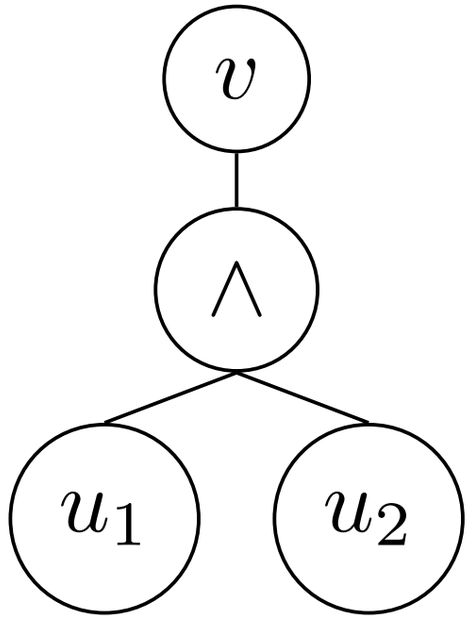


$v \rightarrow u_1 \ u_2$

Summary of Formalisms

hypergraph	AND/OR graph	CFG	deductive system
vertex	OR-node	symbol	item
source-vertex	leaf OR-node	terminal	axiom
target-vertex	root OR-node	start symbol	goal item
hyperedge	AND-node	production	instantiated deduction

$\langle v, \{u_1, u_2\} \rangle$



$v \rightarrow u_1 \ u_2$

$$\frac{u_1 \ u_2}{v}$$

Weights and Semirings

- Associate weights as in WFST
- $\otimes$: extension (multiplicative), $\oplus$: summary (additive)

Weights and Semirings

$$\text{VP} \xrightarrow{w_1} \text{VBD NP}$$

$$\text{NP} \xrightarrow{w_2} \text{NP PP}$$

- Associate weights as in WFST
- $\otimes$: extension (multiplicative), $\oplus$: summary (additive)

Weights and Semirings

$$\mathrm{VP} \xrightarrow{w_1} \mathrm{VBD\ NP}$$

$$\mathrm{NP} \xrightarrow{w_2} \mathrm{NP\ PP}$$

$$\mathrm{VP}_{1,6} : w_1 \otimes c \otimes d$$

$$\mathrm{VBD}_{1,2} : c \quad \mathrm{NP}_{2,6} : d$$

$$\frac{\mathrm{VBD}_{1,2} : c \ \mathrm{NP}_{2,6} : d}{\mathrm{VP}_{1,6} : w_1 \otimes c \otimes d} : w_1$$

- Associate weights as in WFST
- $\otimes$: extension (multiplicative), $\oplus$: summary (additive)

Weights and Semirings

$$\mathrm{VP} \xrightarrow{w_1} \mathrm{VBD\ NP}$$

$$\mathrm{NP} \xrightarrow{w_2} \mathrm{NP\ PP}$$

$$\mathrm{VP}_{1,6} : w_1 \otimes c \otimes d$$

$$\mathrm{VBD}_{1,2} : c \quad \mathrm{NP}_{2,6} : d$$

$$\frac{\mathrm{VBD}_{1,2} : c\ \mathrm{NP}_{2,6} : d}{\mathrm{VP}_{1,6} : w_1 \otimes c \otimes d} : w_1$$

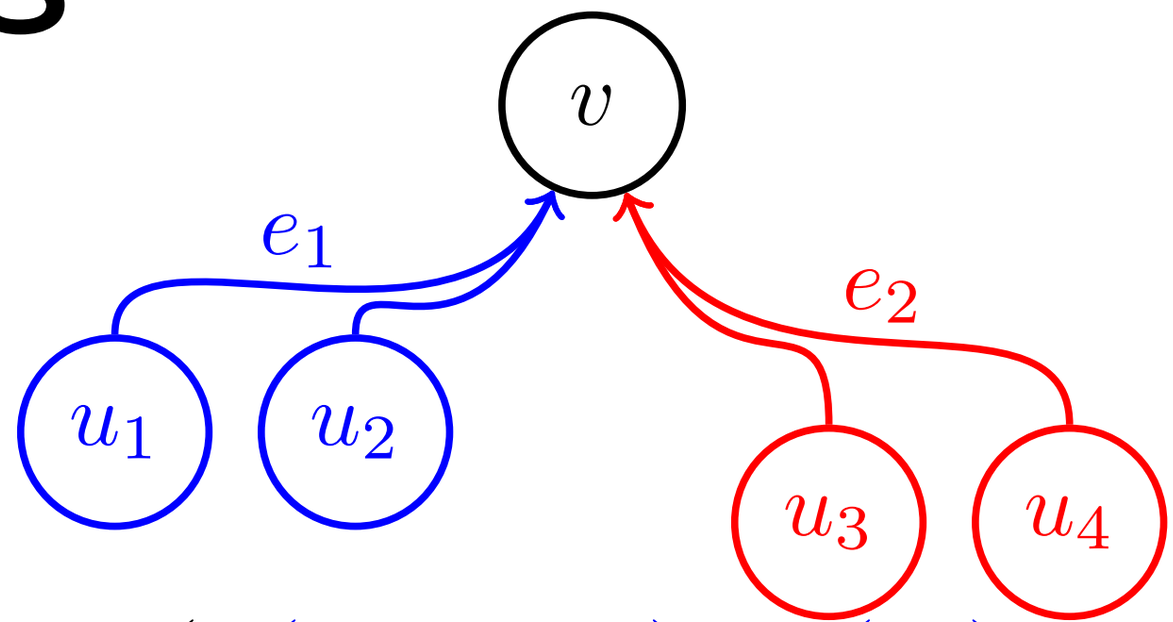
$$\mathrm{NP}_{2,6} : w_2 \otimes a \otimes b$$

$$\mathrm{NP}_{2,4} : a \quad \mathrm{PP}_{4,6} : b$$

$$\frac{\mathrm{NP}_{2,4} : a\ \mathrm{PP}_{4,6} : b}{\mathrm{NP}_{2,6} : w_2 \otimes a \otimes b} : w_2$$

- Associate weights as in WFST
- $\otimes$: extension (multiplicative), $\oplus$: summary (additive)

Weights and Semirings



$$d(v) = (w(e_1, u_1, u_2) \otimes d(u_1) \otimes d(u_2)) \oplus (w(e_2, u_3, u_4) \otimes d(u_3) \otimes d(u_4))$$

- The weight of a hyperedge is dependent on antecedents (non-monotonic)
- The weight of a derivation is the product of hyperedge weights
- The weight of a vertex is the summary of (sub-) derivation weights

Semirings

$$\mathbf{K} = \langle K, \oplus, \otimes, \mathbf{0}, \mathbf{1} \rangle$$

semiring	K	⊕	⊗	0	1
Viterbi	[0,1]	max	×	0	1
Real	R	+	x	0	1
Log	R	logsumexp	+	+∞	0
Tropical	R	min	+	+∞	0

Conclusion

- Review important concepts from “parsing”
  - CFG, parsing, hypergraph, deductive system, weights, semirings

Overview

- Backgrounds
  - CFG, parsing, hypergraph, deductive system semirings
- Tree-based SMT
  - Synchronous-CFG
  - String-to-Tree, Tree-to-String
  - Bitext parsing

Overview

- Backgrounds
  - CFG, parsing, hypergraph, deductive system semirings
- Tree-based SMT
  - Synchronous-CFG
  - String-to-Tree, Tree-to-String
  - Bitext parsing

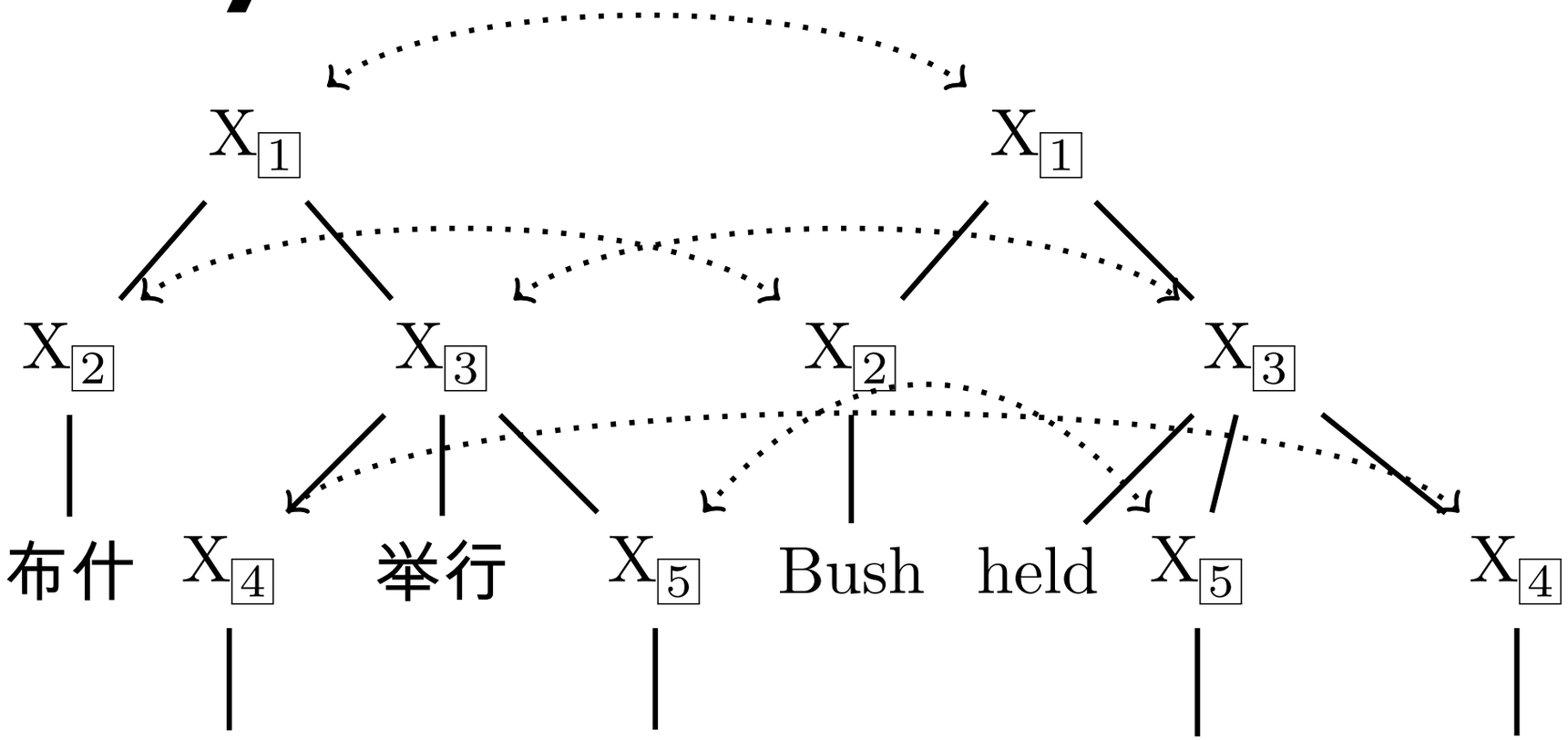
Synchronous-CFG

$$\begin{aligned}\hat{\mathbf{e}} &= \operatorname*{argmax}_{\mathbf{e}} \frac{\exp\left(\mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}, D, \mathbf{f})\right)}{\sum_{\mathbf{e}', D'} \exp\left(\mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}', D', \mathbf{f})\right)} \\ &= \operatorname*{argmax}_{\mathbf{e}} \mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}, D, \mathbf{f})\end{aligned}$$

(Chiang, 2007)

- D: a single derivation constructed by intersecting SCFG with input string

Synchronous-CFG



$$\hat{\mathbf{e}} = \operatorname*{argmax}_{\mathbf{e}} \frac{\exp\left(\mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}, D, \mathbf{f})\right)}{\sum_{\mathbf{e}', D'} \exp\left(\mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}', D', \mathbf{f})\right)} \quad \text{(Chiang, 2007)}$$

$$= \operatorname*{argmax}_{\mathbf{e}} \mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}, D, \mathbf{f})$$

- D: a single derivation constructed by intersecting SCFG with input string

Synchronous-CFG: Model

- We use two categories, S and X (Chiang, 2007)
- Or, borrow linguistic categories from syntactic parse (Zollman and Venugopal, 2006)

Synchronous-CFG: Model

$$
\begin{aligned}
\mathrm{S} &\rightarrow \langle \mathrm{S}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}}, \mathrm{S}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}} \rangle \\
\mathrm{S} &\rightarrow \langle \mathrm{X}_{\boxed{1}}, \mathrm{X}_{\boxed{1}} \rangle \\
\mathrm{X} &\rightarrow \langle \mathrm{X}_{\boxed{1}}\ \text{举行}\ \mathrm{X}_{\boxed{2}}, \text{hold}\ \mathrm{X}_{\boxed{2}}\ \mathrm{X}_{\boxed{1}} \rangle \\
\mathrm{X} &\rightarrow \langle \text{与 沙龙}, \text{with Sharon} \rangle
\end{aligned}
$$

- We use two categories, S and X (Chiang, 2007)
- Or, borrow linguistic categories from syntactic parse (Zollman and Venugopal, 2006)

Synchronous-CFG: Model

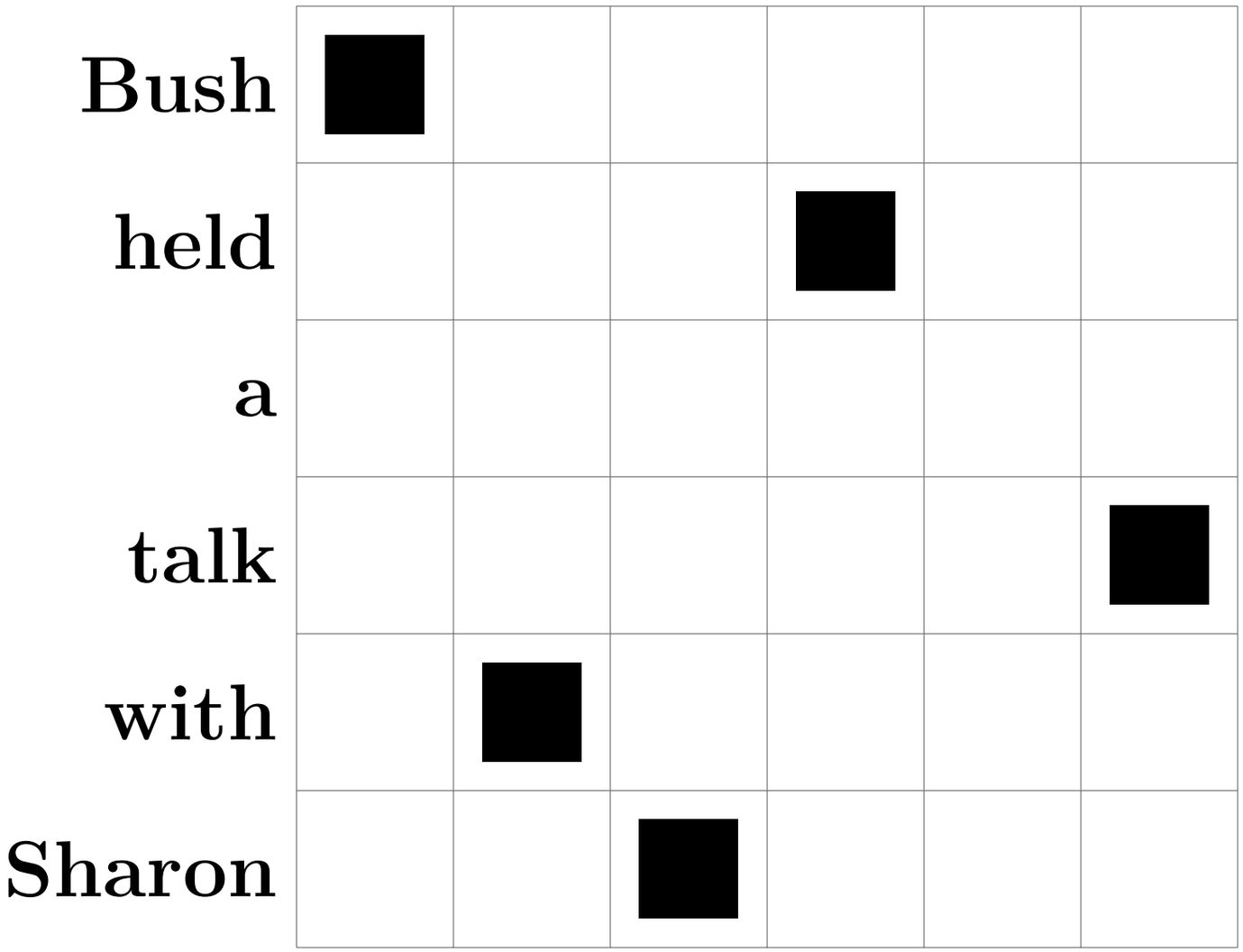
$$
\begin{aligned}
\mathrm{S} &\rightarrow \langle \mathrm{S}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}}, \mathrm{S}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}} \rangle \\
\mathrm{S} &\rightarrow \langle \mathrm{X}_{\boxed{1}}, \mathrm{X}_{\boxed{1}} \rangle \\
\mathrm{X} &\rightarrow \langle \mathrm{X}_{\boxed{1}}\ \text{举行}\ \mathrm{X}_{\boxed{2}}, \text{hold}\ \mathrm{X}_{\boxed{2}}\ \mathrm{X}_{\boxed{1}} \rangle \\
\mathrm{X} &\rightarrow \langle \text{与 沙龙}, \text{with Sharon} \rangle
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{VP} &\rightarrow \langle \mathrm{VBD}_{\boxed{1}}\ \mathrm{NP}_{\boxed{2}},\ \mathrm{NP}_{\boxed{2}}\ \mathrm{VBD}_{\boxed{1}} \rangle \\
\mathrm{NP} &\rightarrow \langle \mathrm{NP}_{\boxed{1}}\ \mathrm{PP}_{\boxed{2}},\ \mathrm{NP}_{\boxed{1}}\ \mathrm{PP}_{\boxed{2}} \rangle \\
\mathrm{VP} &\rightarrow \langle \mathrm{VBD}_{\boxed{1}}\ \mathrm{NP}_{\boxed{2}}\ \mathrm{PP}_{\boxed{3}},\ \mathrm{NP}_{\boxed{2}}\ \mathrm{PP}_{\boxed{3}}\ \mathrm{VBD}_{\boxed{1}} \rangle
\end{aligned}
$$

- We use two categories, S and X (Chiang, 2007)
- Or, borrow linguistic categories from syntactic parse (Zollman and Venugopal, 2006)

Rule Extraction

布什 与 沙龙举行 了 会谈

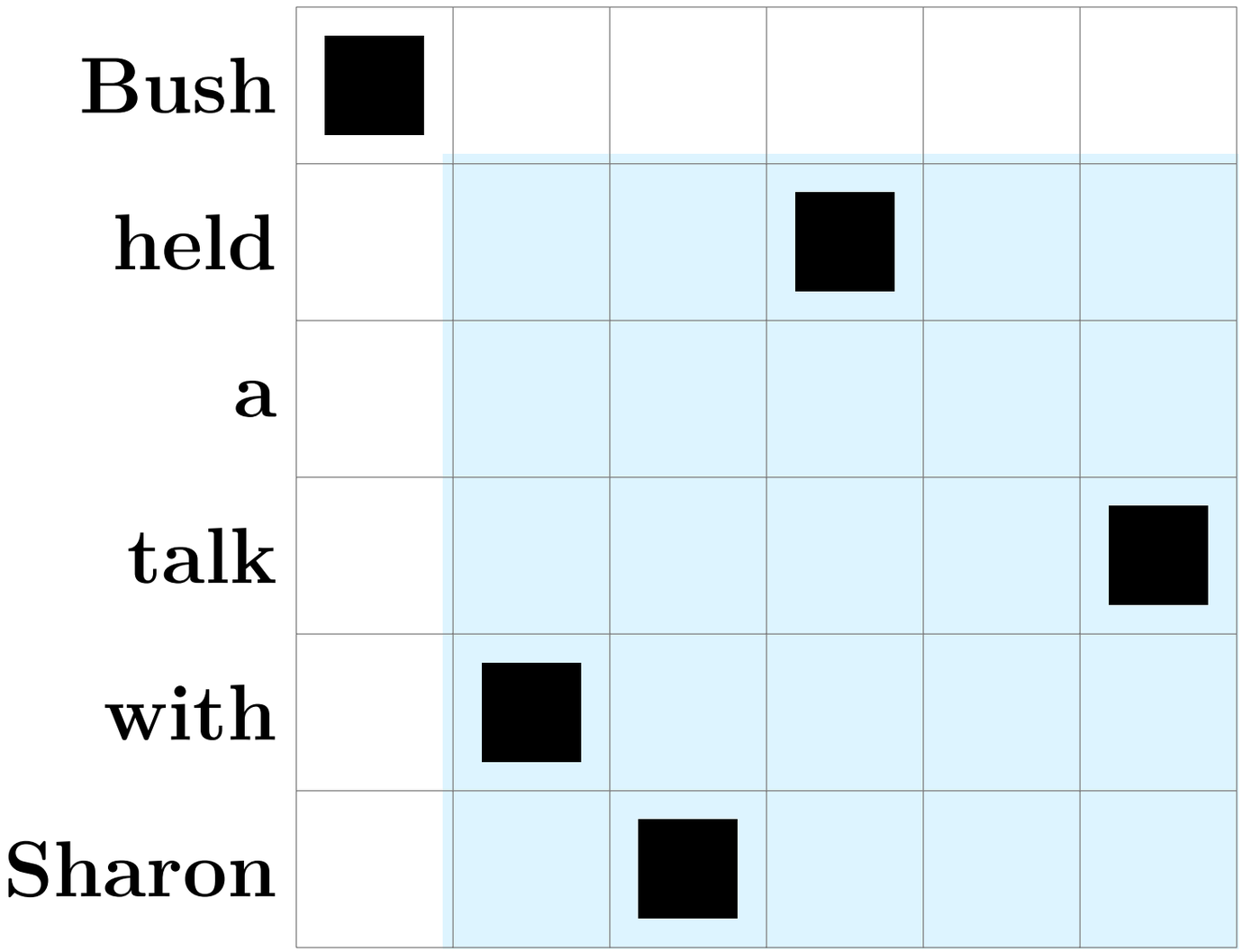


(Example from Huang and Chiang, 2007)

- As in phrase-based models, extract phrases then, use sub-phrases as non-terminals, aka Hiero (Chiang, 2007)

Rule Extraction

布什 与 沙龙举行 了 会谈



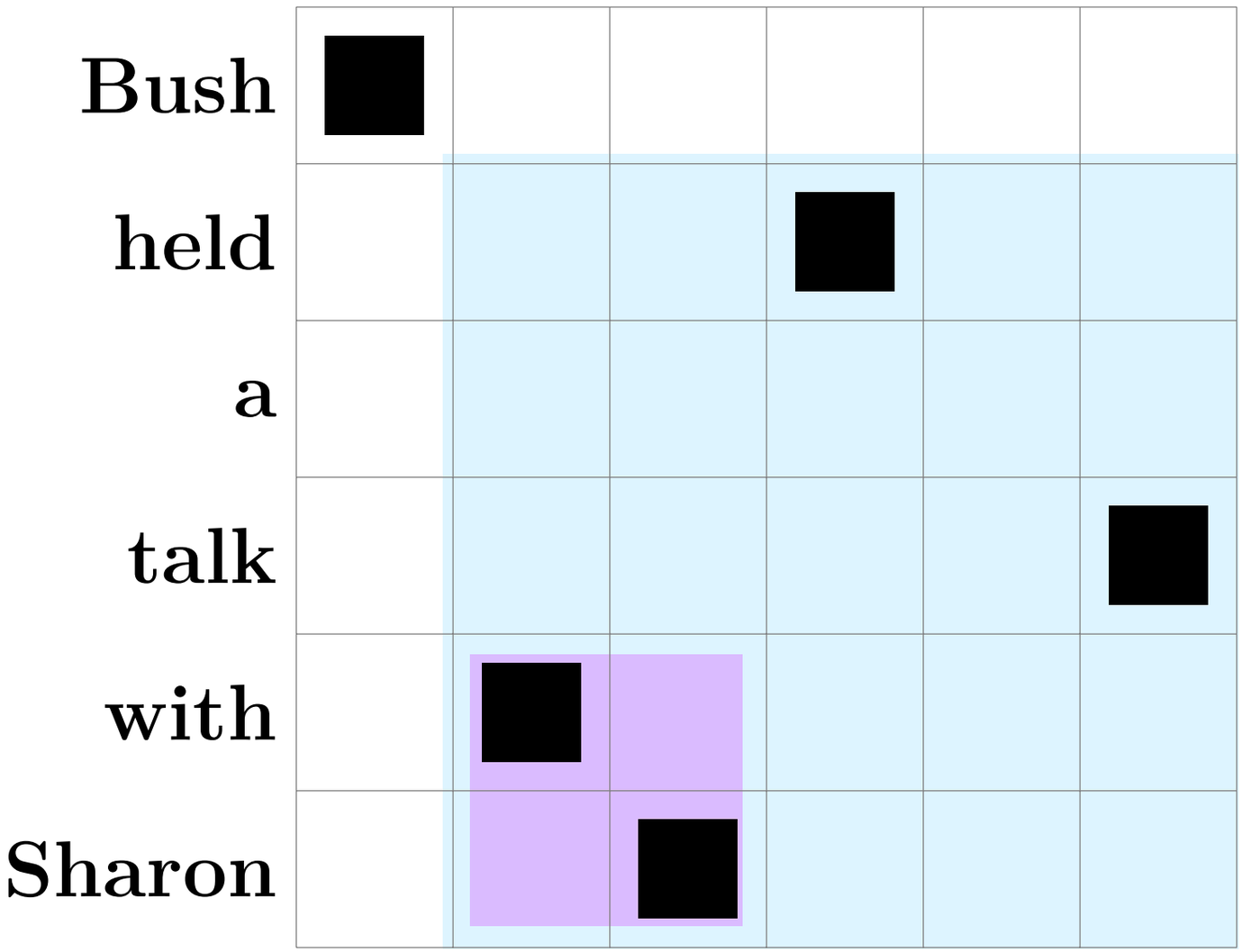
⟨held a talk with Sharon,
与 沙龙 举行 了 会谈⟩

(Example from Huang and Chiang, 2007)

- As in phrase-based models, extract phrases then, use sub-phrases as non-terminals, aka Hiero (Chiang, 2007)

Rule Extraction

布什 与 沙龙举行 了 会谈



⟨held a talk with Sharon, 与 沙龙 举行 了 会谈⟩

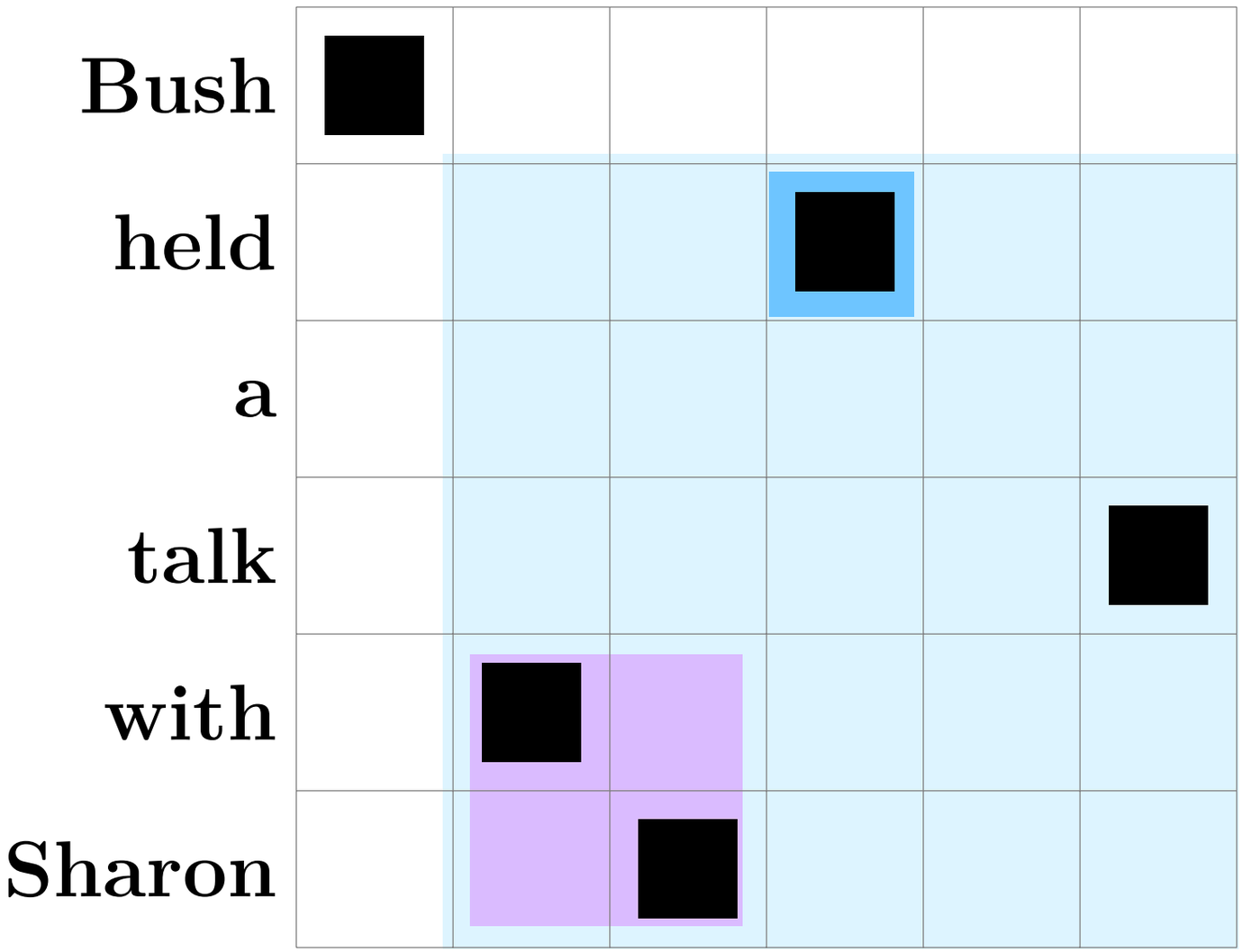
⟨with Sharon, 与 沙龙⟩

(Example from Huang and Chiang, 2007)

- As in phrase-based models, extract phrases then, use sub-phrases as non-terminals, aka Hiero (Chiang, 2007)

Rule Extraction

布什 与 沙龙举行 了 会谈



⟨held a talk with Sharon, 与 沙龙 举行 了 会谈⟩

⟨with Sharon, 与 沙龙⟩

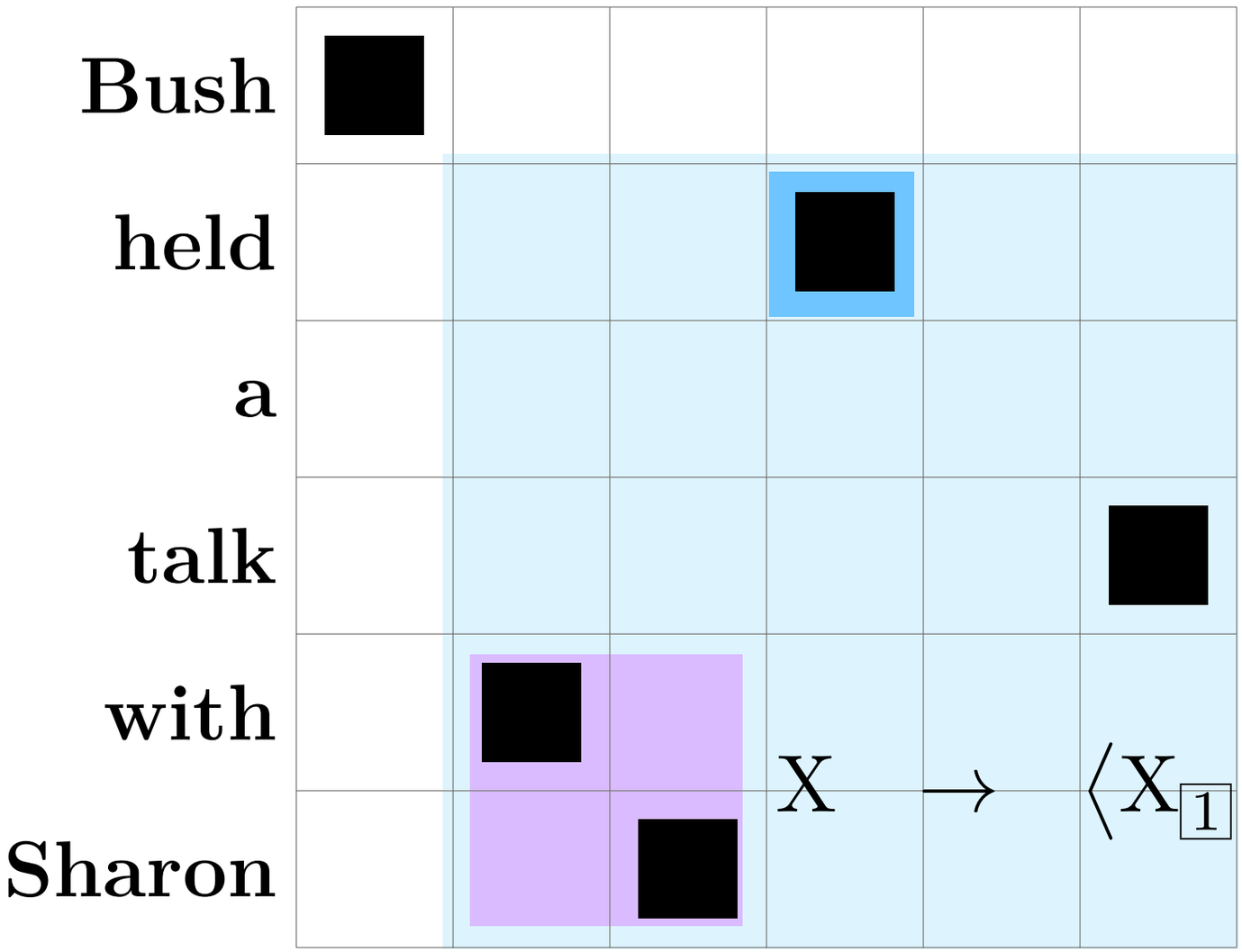
⟨held, 举行⟩

(Example from Huang and Chiang, 2007)

- As in phrase-based models, extract phrases then, use sub-phrases as non-terminals, aka Hiero (Chiang, 2007)

Rule Extraction

布什 与 沙龙举行 了 会谈



$\langle$held a talk with Sharon, 与 沙龙 举行 了 会谈$\rangle$

$\langle$with Sharon, 与 沙龙$\rangle$

$\langle$held, 举行$\rangle$

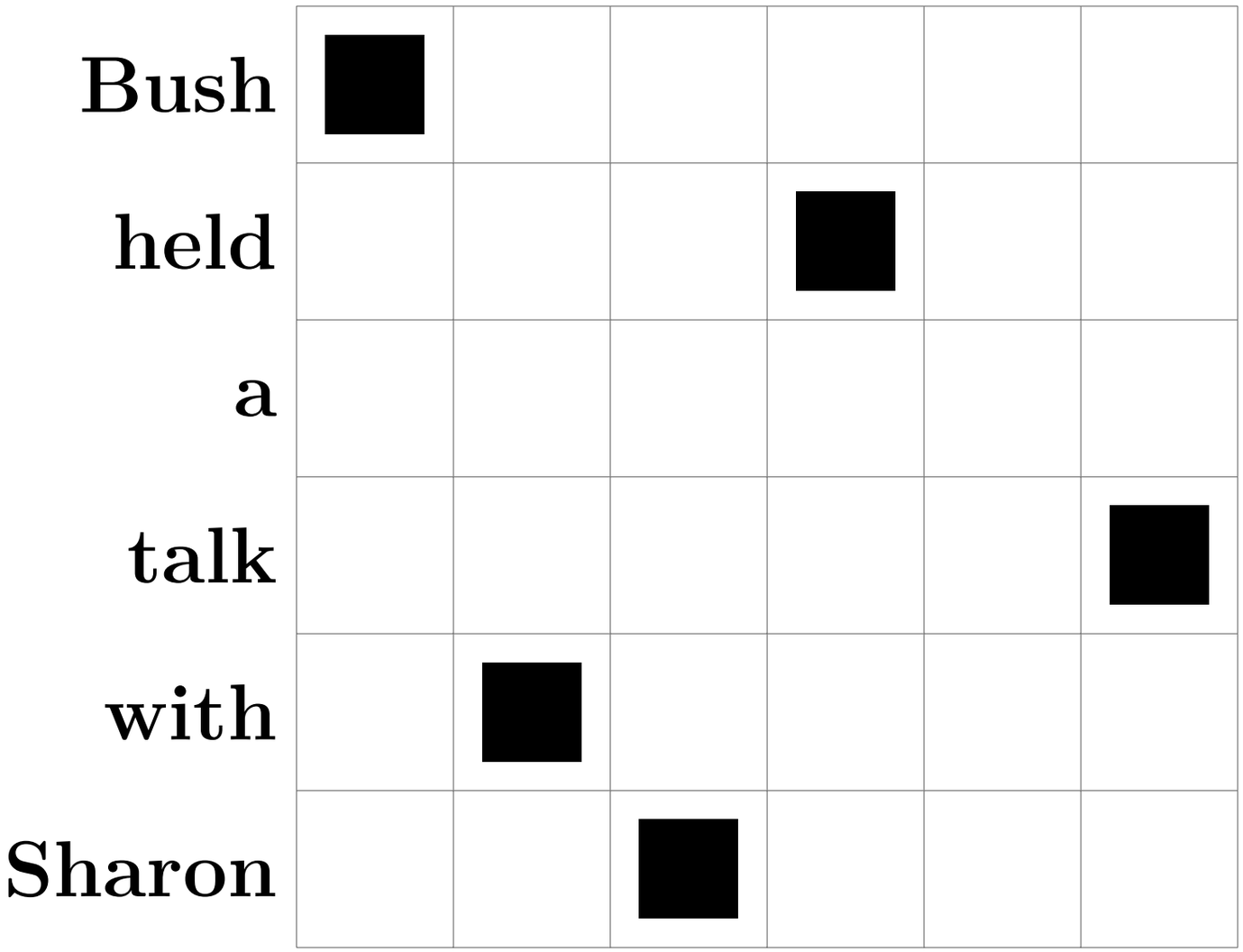
$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}}$ 了 会谈, $X_{\boxed{2}}$ a talk $X_{\boxed{1}}\rangle$

(Example from Huang and Chiang, 2007)

- As in phrase-based models, extract phrases then, use sub-phrases as non-terminals, aka Hiero (Chiang, 2007)

Syntactic Categories

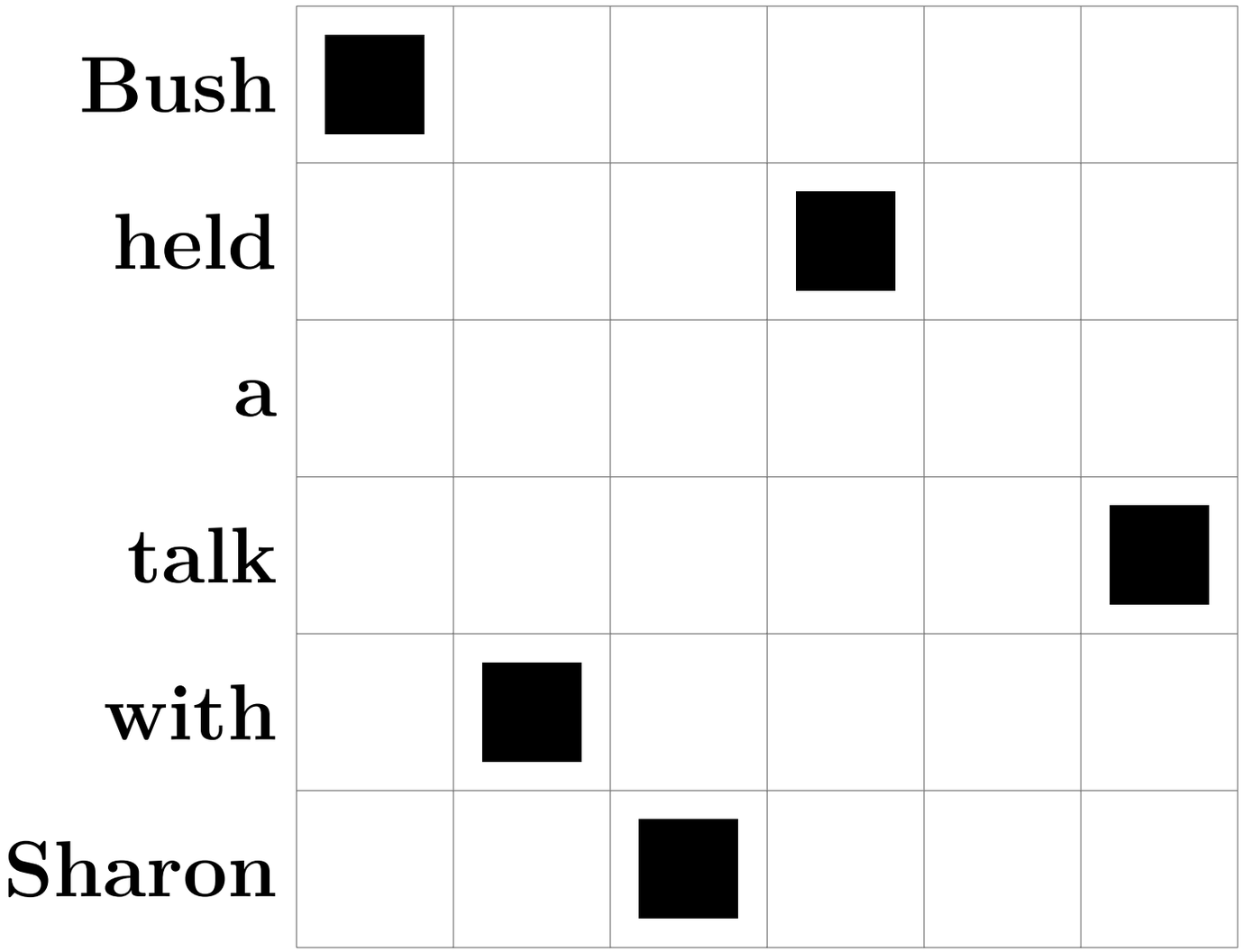
布什 与 沙龙举行 了 会谈



- Borrow syntactic categories either from source/ target side, aka SAMT (Zollman and Venugopal, 2006)

Syntactic Categories

布什 与 沙龙举行 了 会谈

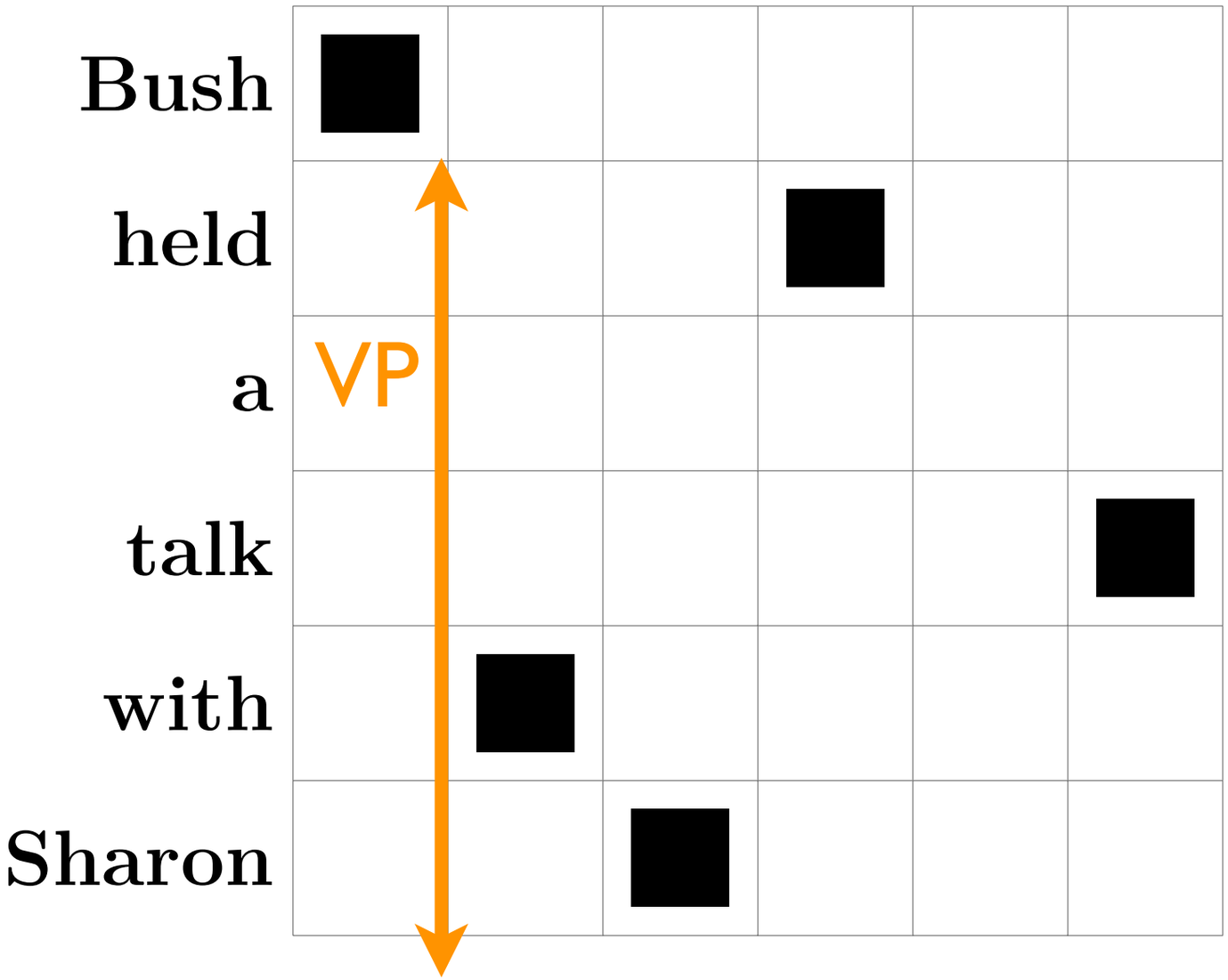


⟨held a talk with Sharon,
与 沙龙 举行 了 会谈⟩

- Borrow syntactic categories either from source/ target side, aka SAMT (Zollman and Venugopal, 2006)

Syntactic Categories

布什 与 沙龙举行 了 会谈

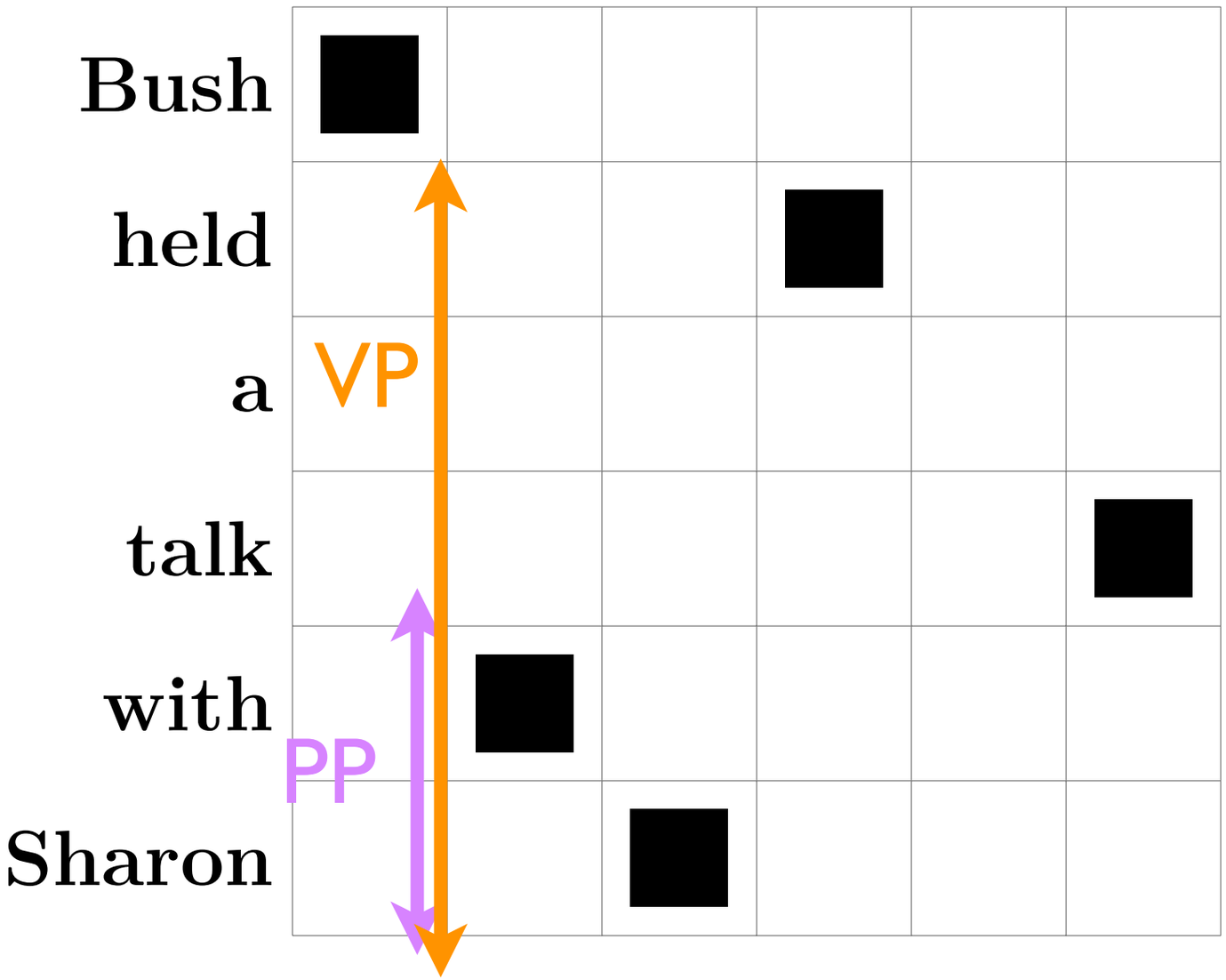


⟨held a talk with Sharon,
与 沙龙 举行 了 会谈⟩

- Borrow syntactic categories either from source/target side, aka SAMT (Zollman and Venugopal, 2006)

Syntactic Categories

布什 与 沙龙举行 了 会谈



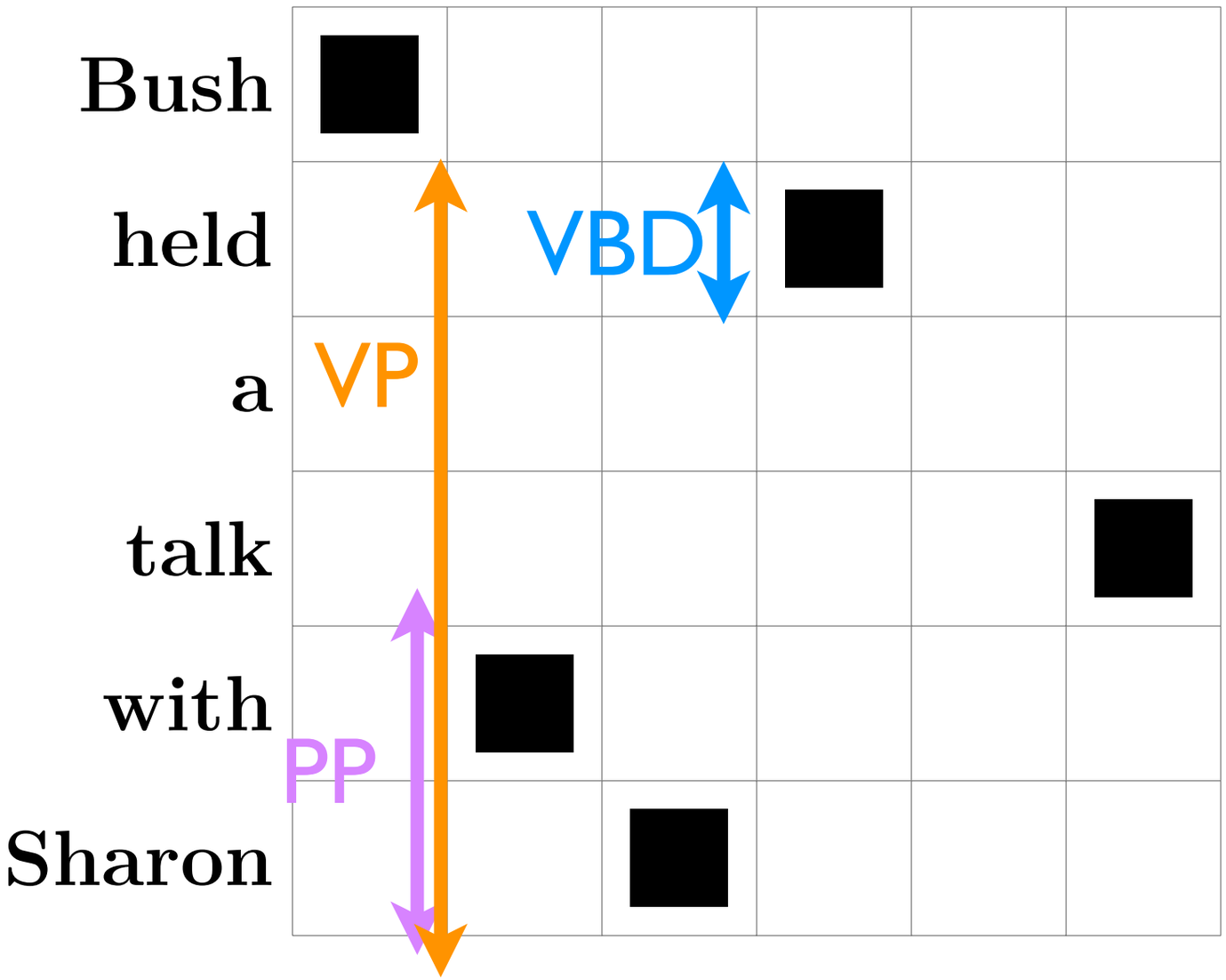
⟨held a talk with Sharon, 与 沙龙 举行 了 会谈⟩

⟨with Sharon, 与 沙龙⟩

- Borrow syntactic categories either from source/ target side, aka SAMT (Zollman and Venugopal, 2006)

Syntactic Categories

布什 与 沙龙举行 了 会谈



⟨held a talk with Sharon, 与 沙龙 举行 了 会谈⟩

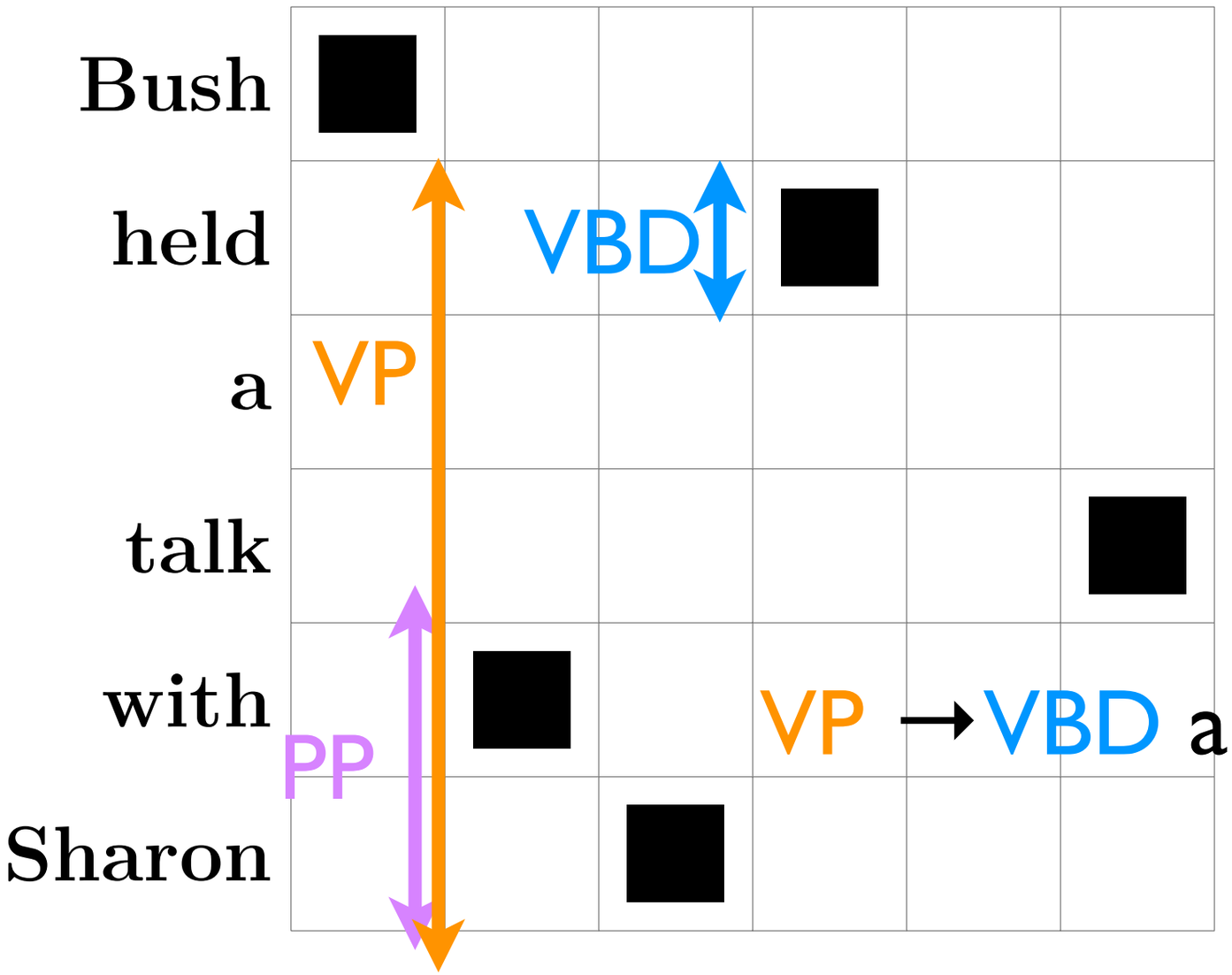
⟨with Sharon, 与 沙龙⟩

⟨held, 举行⟩

- Borrow syntactic categories either from source/ target side, aka SAMT (Zollman and Venugopal, 2006)

Syntactic Categories

布什 与 沙龙举行 了 会谈



⟨held a talk with Sharon, 与 沙龙 举行 了 会谈⟩

⟨with Sharon, 与 沙龙⟩

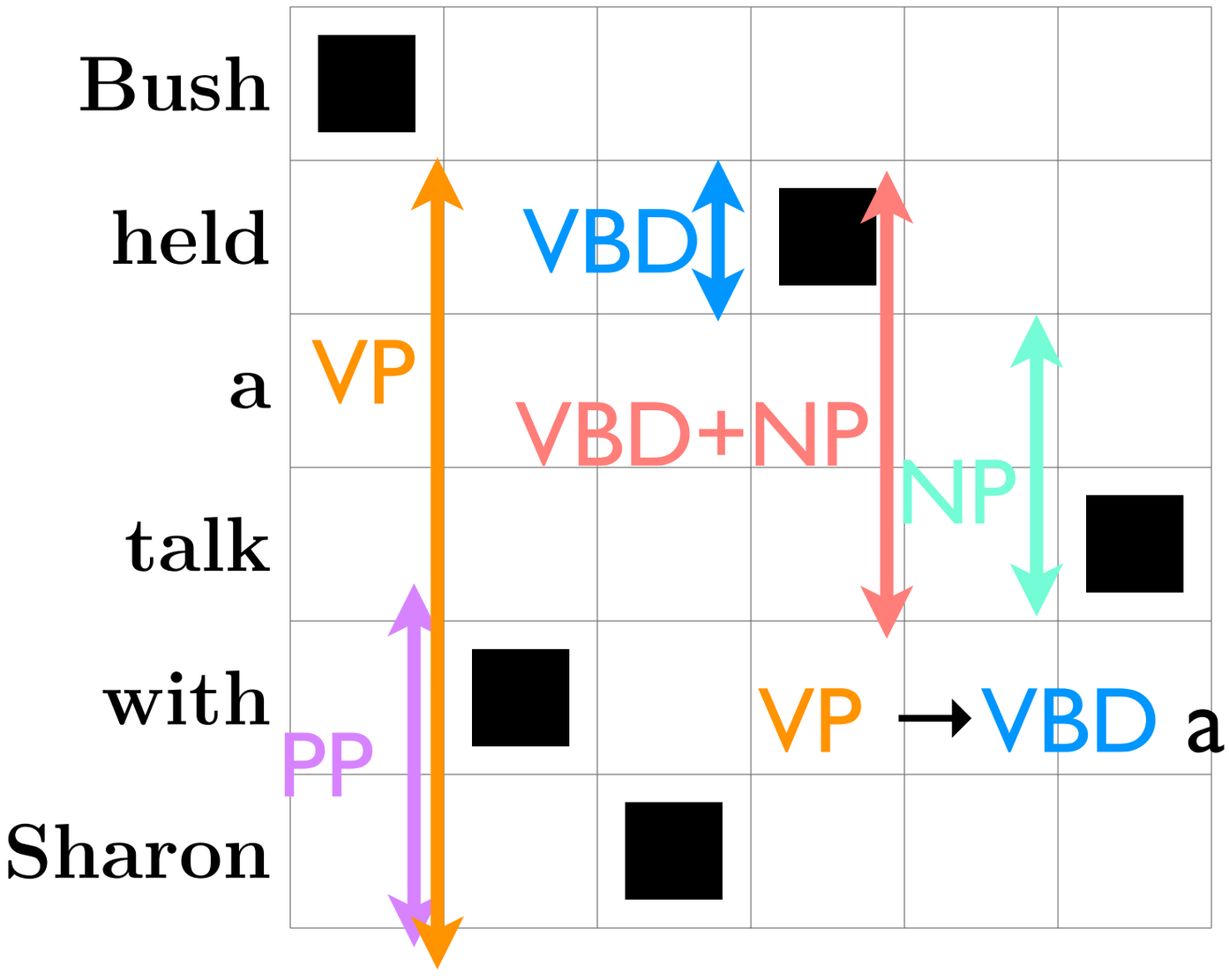
⟨held, 举行⟩

VP → VBD a talk PP, PP VBD 了 会谈

- Borrow syntactic categories either from source/target side, aka SAMT (Zollman and Venugopal, 2006)

Syntactic Categories

布什 与 沙龙举行 了 会谈



⟨held a talk with Sharon, 与 沙龙 举行 了 会谈⟩

⟨with Sharon, 与 沙龙⟩

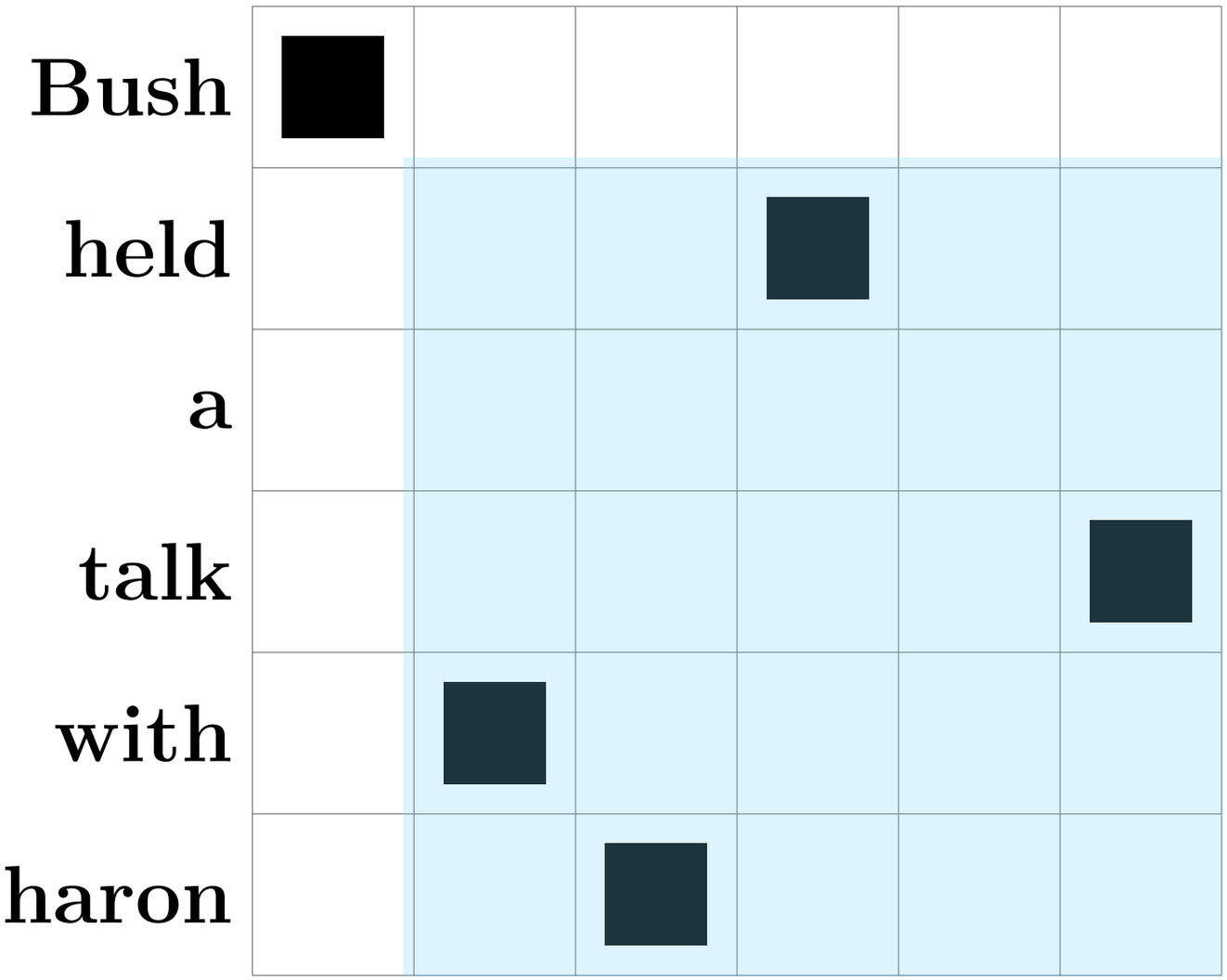
⟨held, 举行⟩

VP → VBD a talk PP, PP VBD 了 会谈

- Borrow syntactic categories either from source/target side, aka SAMT (Zollman and Venugopal, 2006)

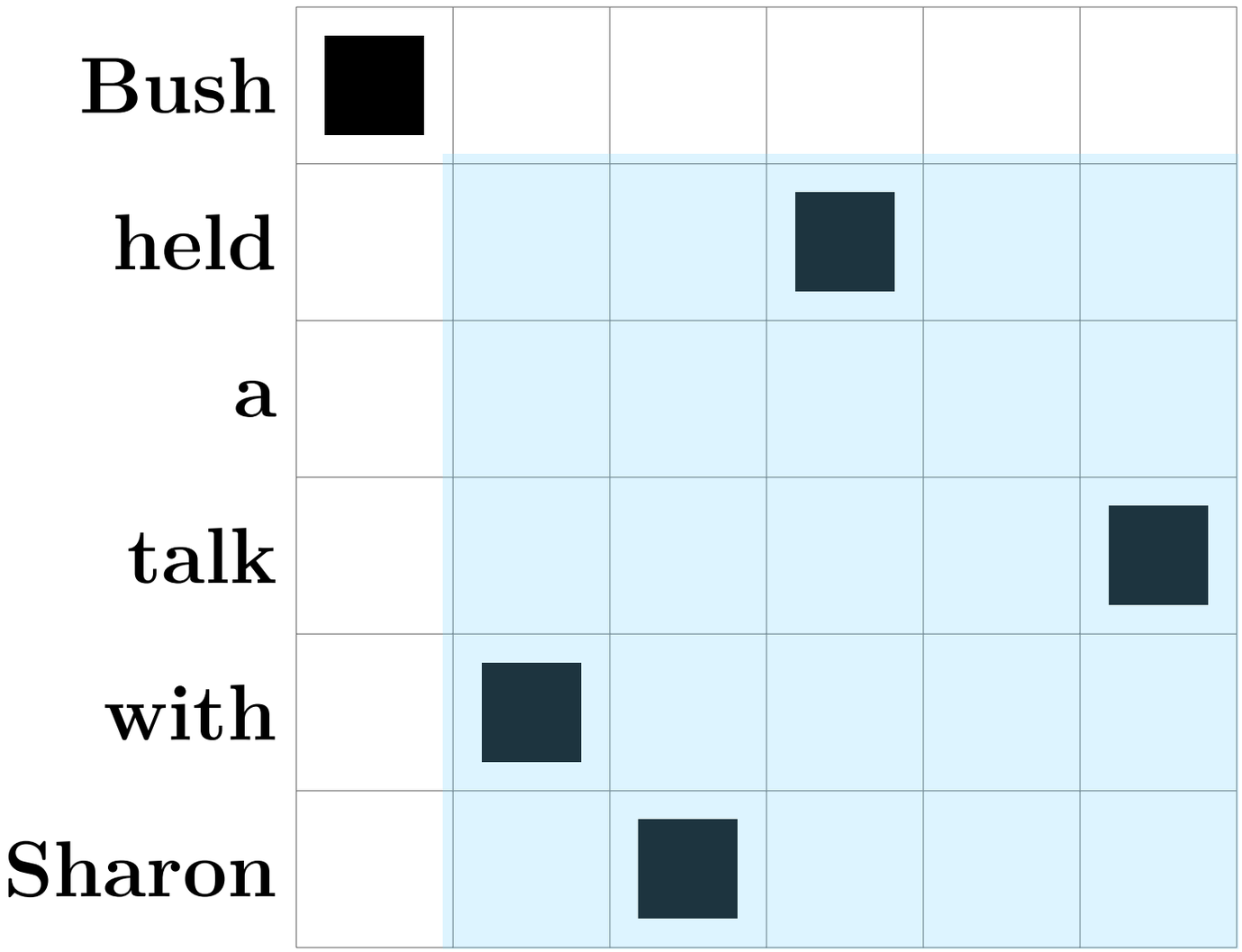
Exhaustive Extraction

布什 与 沙龙举行 了 会谈



Exhaustive Extraction

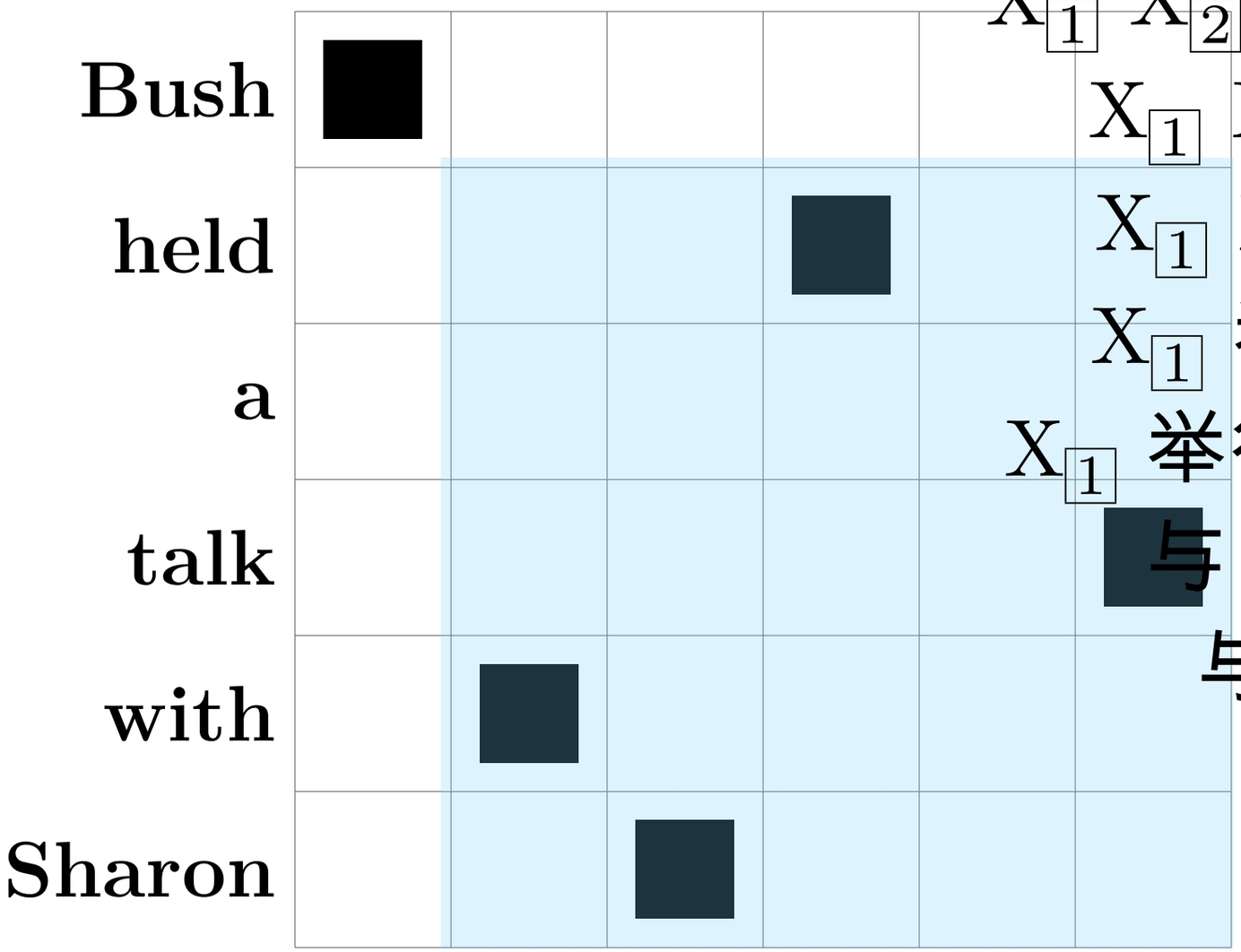
布什 与 沙龙举行 了 会谈



- Exhaustively extract rules as in phrase-based MT

Exhaustive Extraction

布什 与 沙龙举行 了 会谈



$X_{[1]}$ $X_{[2]}$ 了 会谈 $X_{[2]}$ a talk $X_{[1]}$

$X_{[1]}$ $X_{[2]}$ 会谈 $X_{[2]}$ a talk $X_{[1]}$

$X_{[1]}$ $X_{[2]}$ 会谈 $X_{[2]}$ talk $X_{[1]}$

$X_{[1]}$ 举行 $X_{[2]}$ held $X_{[2]}$ $X_{[1]}$

$X_{[1]}$ 举行 了 $X_{[2]}$ held a $X_{[2]}$ $X_{[1]}$

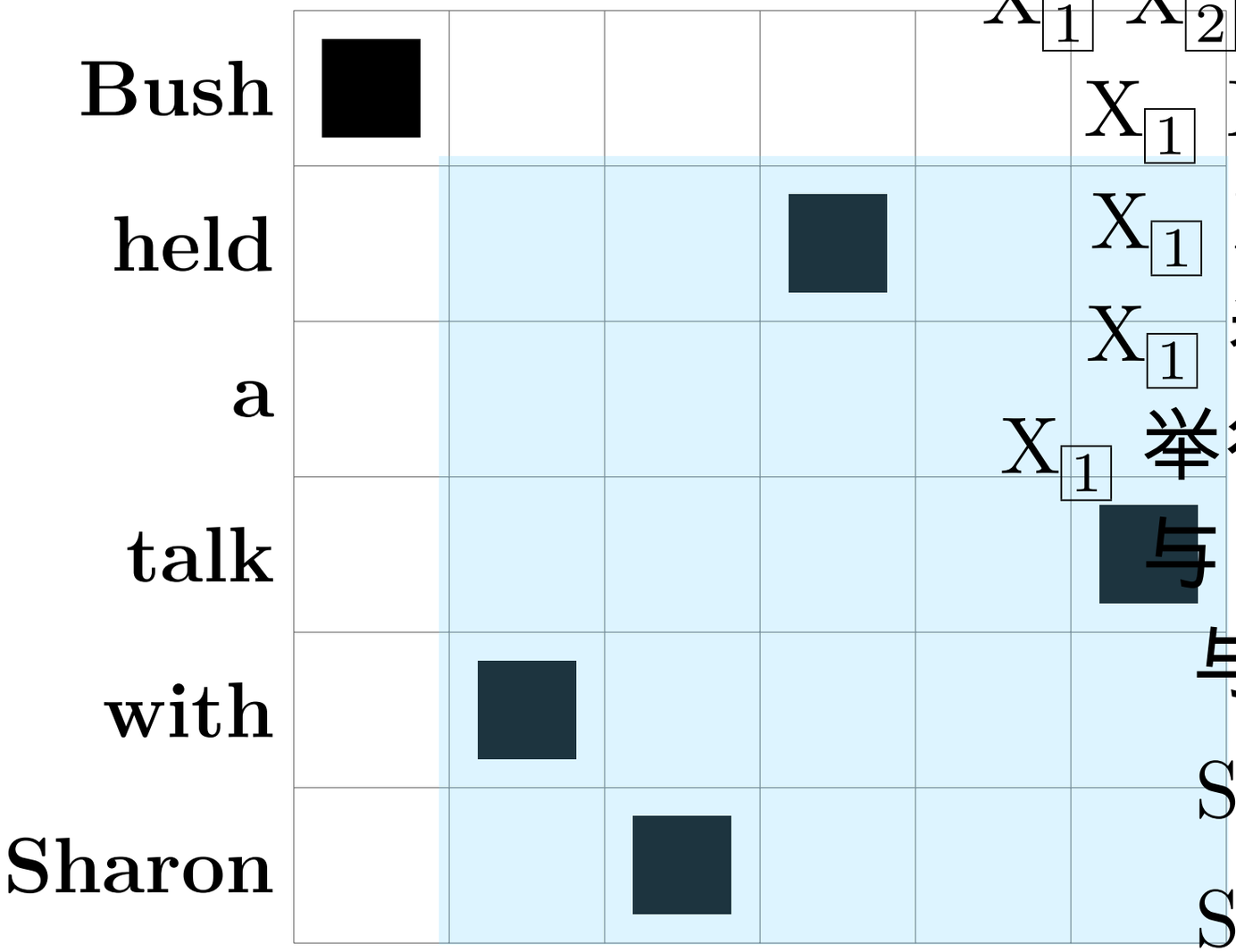
与 沙龙 $X_{[1]}$ $X_{[1]}$ with Sharon

与 $X_{[1]}$ $X_{[2]}$ $X_{[2]}$ with $X_{[1]}$

- Exhaustively extract rules as in phrase-based MT

Exhaustive Extraction

布什 与 沙龙举行 了 会谈



$X_{\boxed{1}}$ $X_{\boxed{2}}$ 了 会谈 $X_{\boxed{2}}$ a talk $X_{\boxed{1}}$

$X_{\boxed{1}}$ $X_{\boxed{2}}$ 会谈 $X_{\boxed{2}}$ a talk $X_{\boxed{1}}$

$X_{\boxed{1}}$ $X_{\boxed{2}}$ 会谈 $X_{\boxed{2}}$ talk $X_{\boxed{1}}$

$X_{\boxed{1}}$ 举行 $X_{\boxed{2}}$ held $X_{\boxed{2}}$ $X_{\boxed{1}}$

$X_{\boxed{1}}$ 举行 了 $X_{\boxed{2}}$ held a $X_{\boxed{2}}$ $X_{\boxed{1}}$

与 沙龙 $X_{\boxed{1}}$ $X_{\boxed{1}}$ with Sharon

与 $X_{\boxed{1}}$ $X_{\boxed{2}}$ $X_{\boxed{2}}$ with $X_{\boxed{1}}$

$S \rightarrow \langle S_{\boxed{1}}\ X_{\boxed{2}}, S_{\boxed{1}}\ X_{\boxed{2}} \rangle$

$S \rightarrow \langle X_{\boxed{1}}, X_{\boxed{1}} \rangle$

- Exhaustively extract rules as in phrase-based MT
- + glue rules

Features from Rules

$$\log p_r(\bar{\alpha}|\bar{\beta}) = \log \frac{\text{count}(\bar{\beta}, \bar{\alpha})}{\sum_{\bar{\alpha'}} \text{count}(\bar{\beta}, \bar{\alpha'})}$$

$$\log p_r(\bar{\beta}|\bar{\alpha}) = \log \frac{\text{count}(\bar{\beta}, \bar{\alpha})}{\sum_{\bar{\beta'}} \text{count}(\bar{\beta'}, \bar{\alpha})}$$

- Collect all the rules (α, β) from the data:
  - α = source side string, β = target side string
- Maximum likelihood estimates by relative frequencies
- Employ scores in two directions

Remarks on Rules

- Too many rules extracted (Chiang, 2007):
  - at most two non-terminal symbols
  - at least one terminal between non-terminals in the source side
    - Span at most 15 words for “holes”
- Fractional counts (Chiang, 2007):
  - Each phrases counted in phrase-based MT
  - Fractional counts for rules sharing the same source/target span

Other Features

- Lexical weights as used in phrase-based MT
- ngram language model(s)
- word count: bias for ngram language model(s)
- rule count: shorter or longer phrases
- glue-rule counts: bias for monotonic glue rules

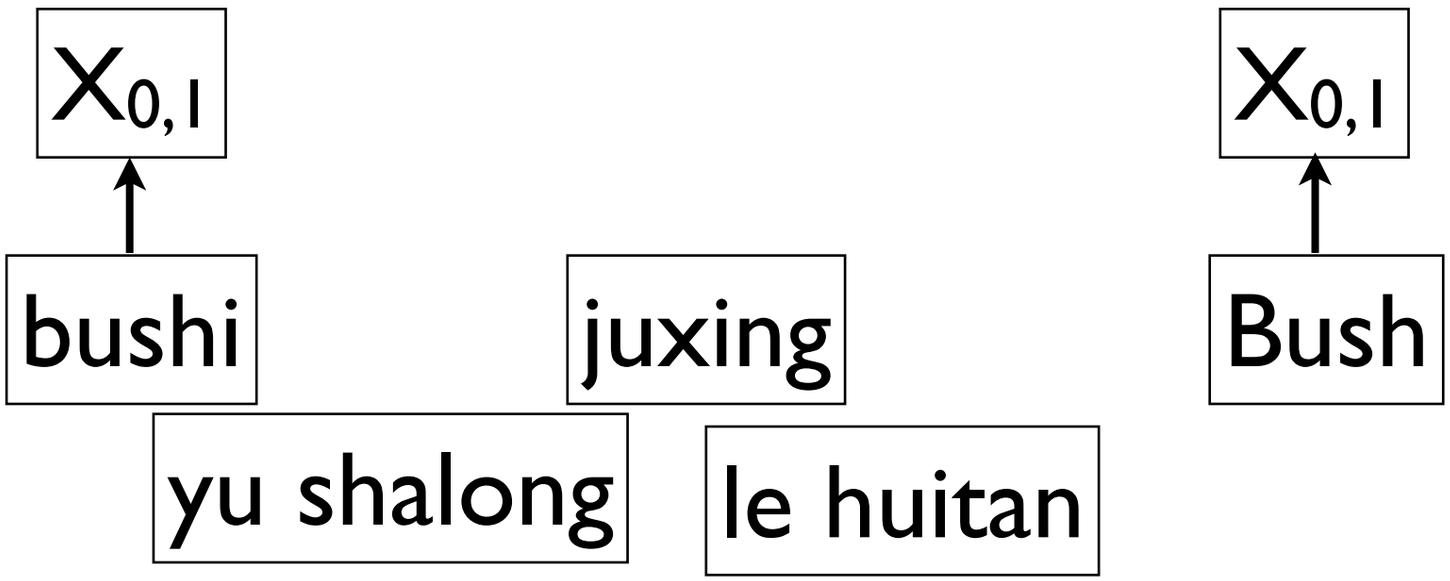
Synchronous-CFG: Parsing

bushi juxing

yu shalong le huitan

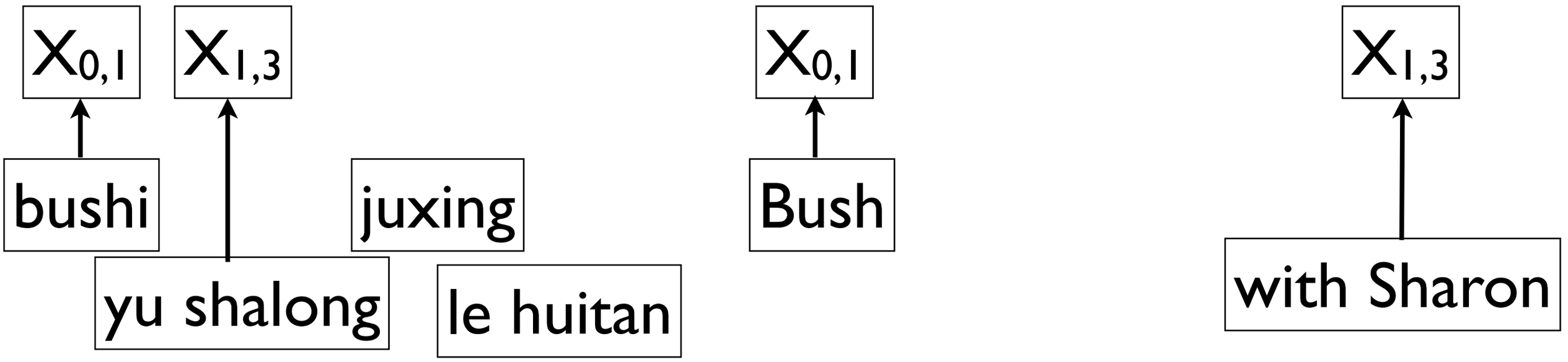
- Parse input sentence using the source side, and construct a translation forest by target side

Synchronous-CFG: Parsing



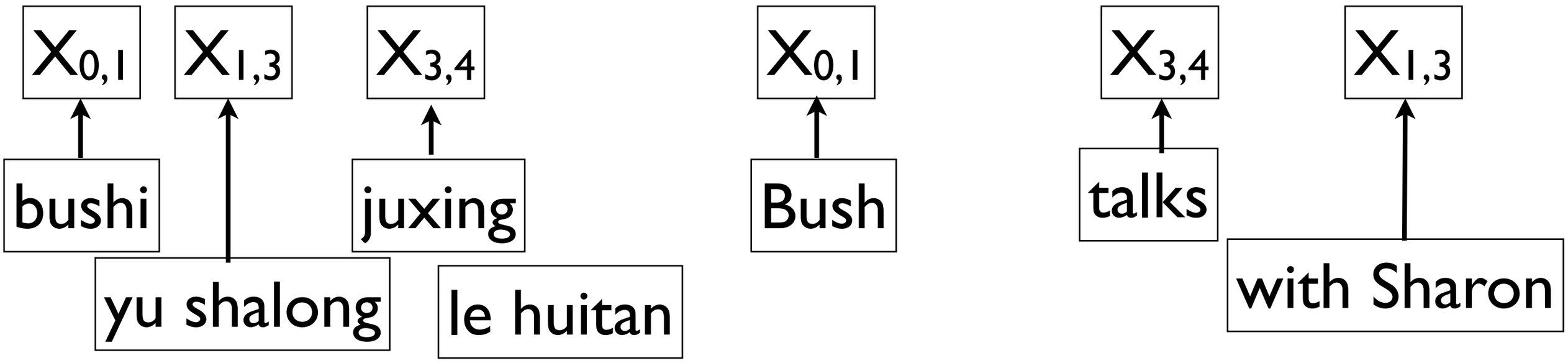
- Parse input sentence using the source side, and construct a translation forest by target side

Synchronous-CFG: Parsing



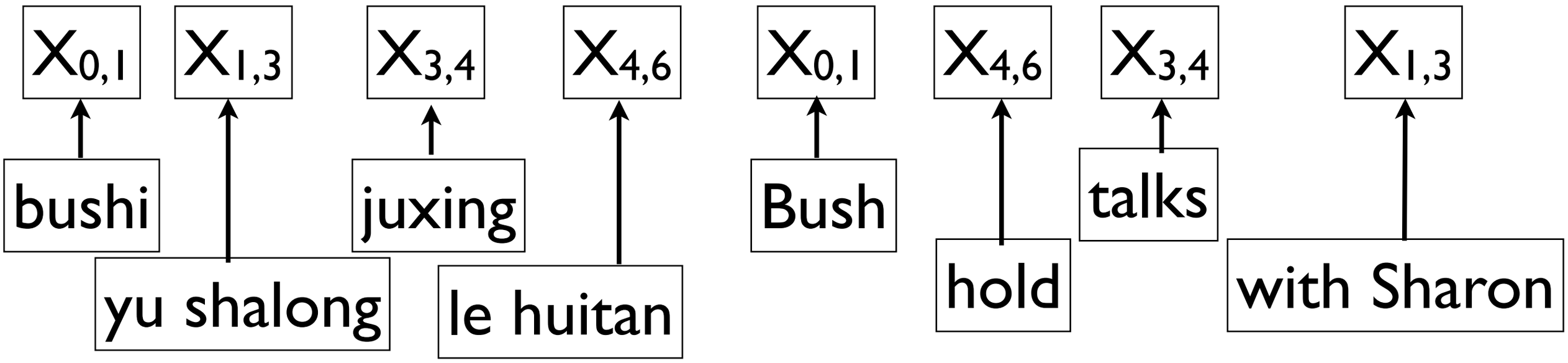
- Parse input sentence using the source side, and construct a translation forest by target side

Synchronous-CFG: Parsing



- Parse input sentence using the source side, and construct a translation forest by target side

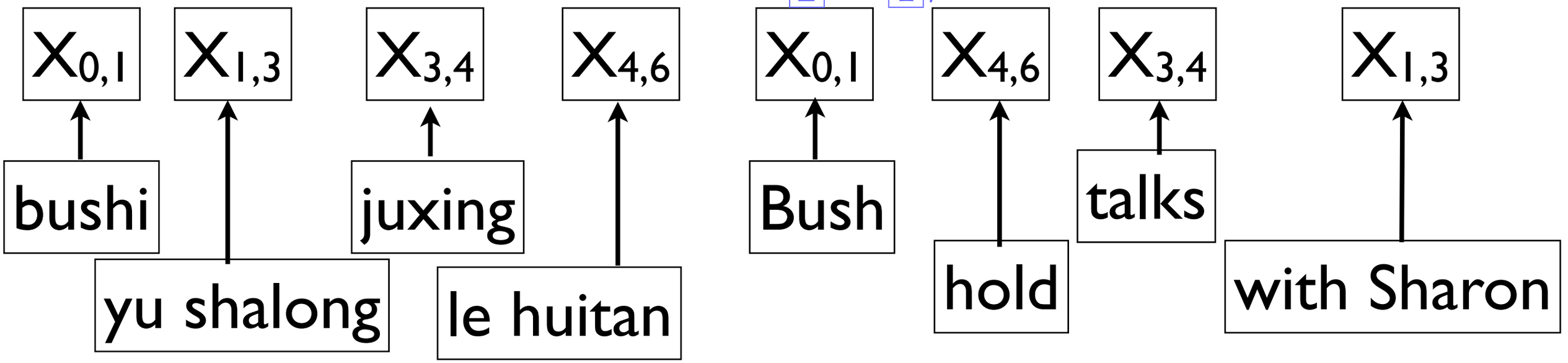
Synchronous-CFG: Parsing



- Parse input sentence using the source side, and construct a translation forest by target side

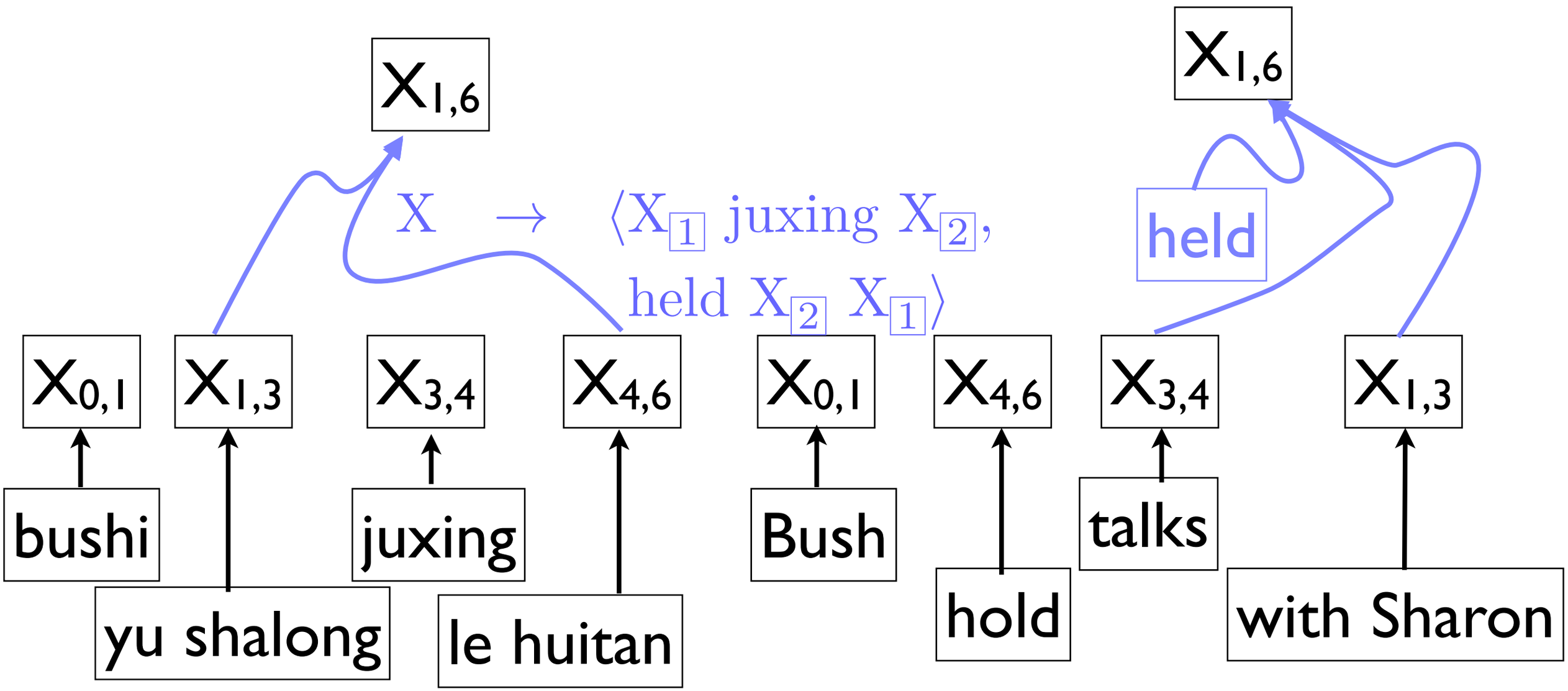
Synchronous-CFG: Parsing

$$\mathrm{X} \rightarrow \langle \mathrm{X}_{\boxed{1}} \text{ juxing } \mathrm{X}_{\boxed{2}}, \text{held } \mathrm{X}_{\boxed{2}} \ \mathrm{X}_{\boxed{1}} \rangle$$



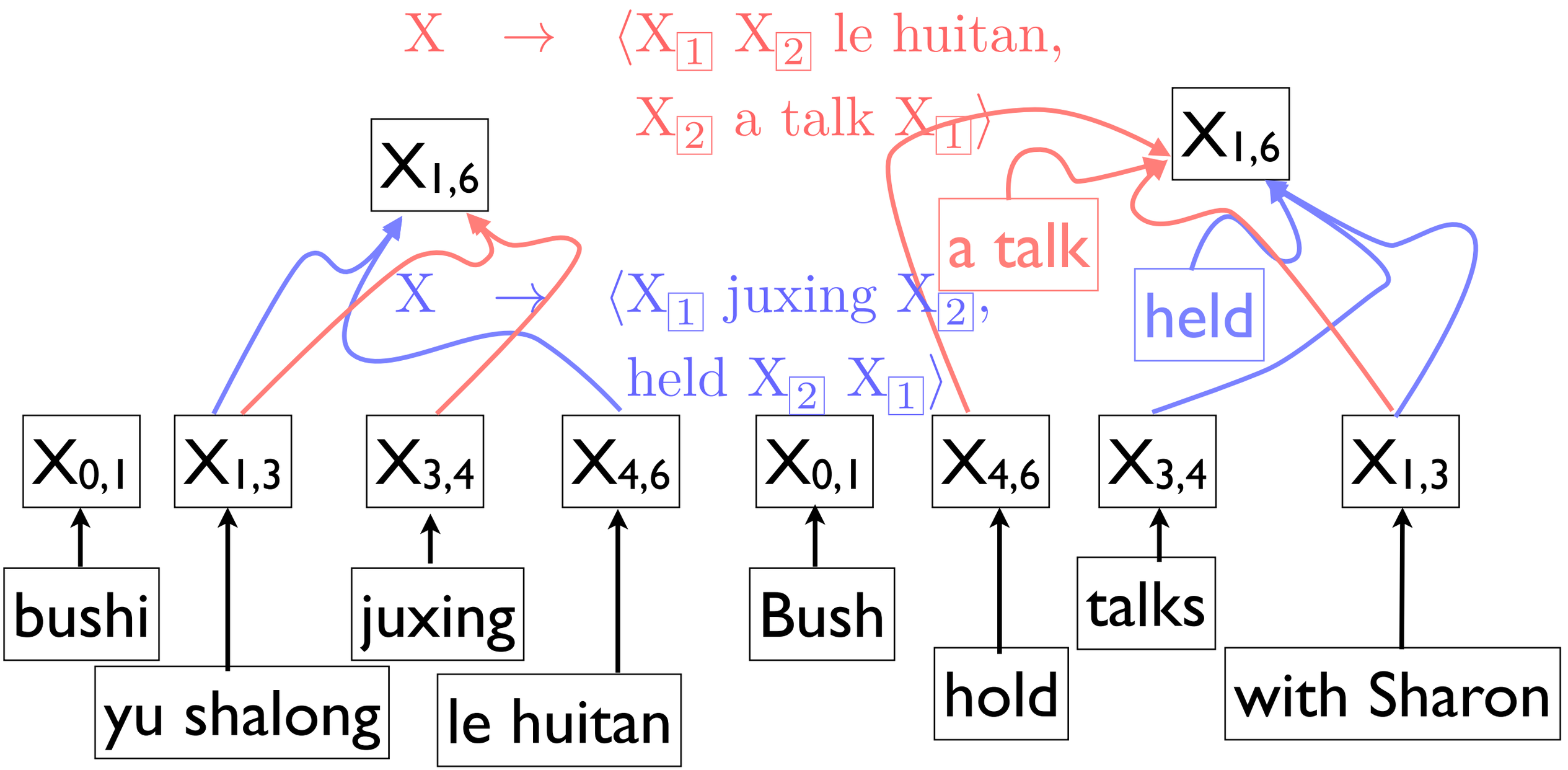
- Parse input sentence using the source side, and construct a translation forest by target side

Synchronous-CFG: Parsing



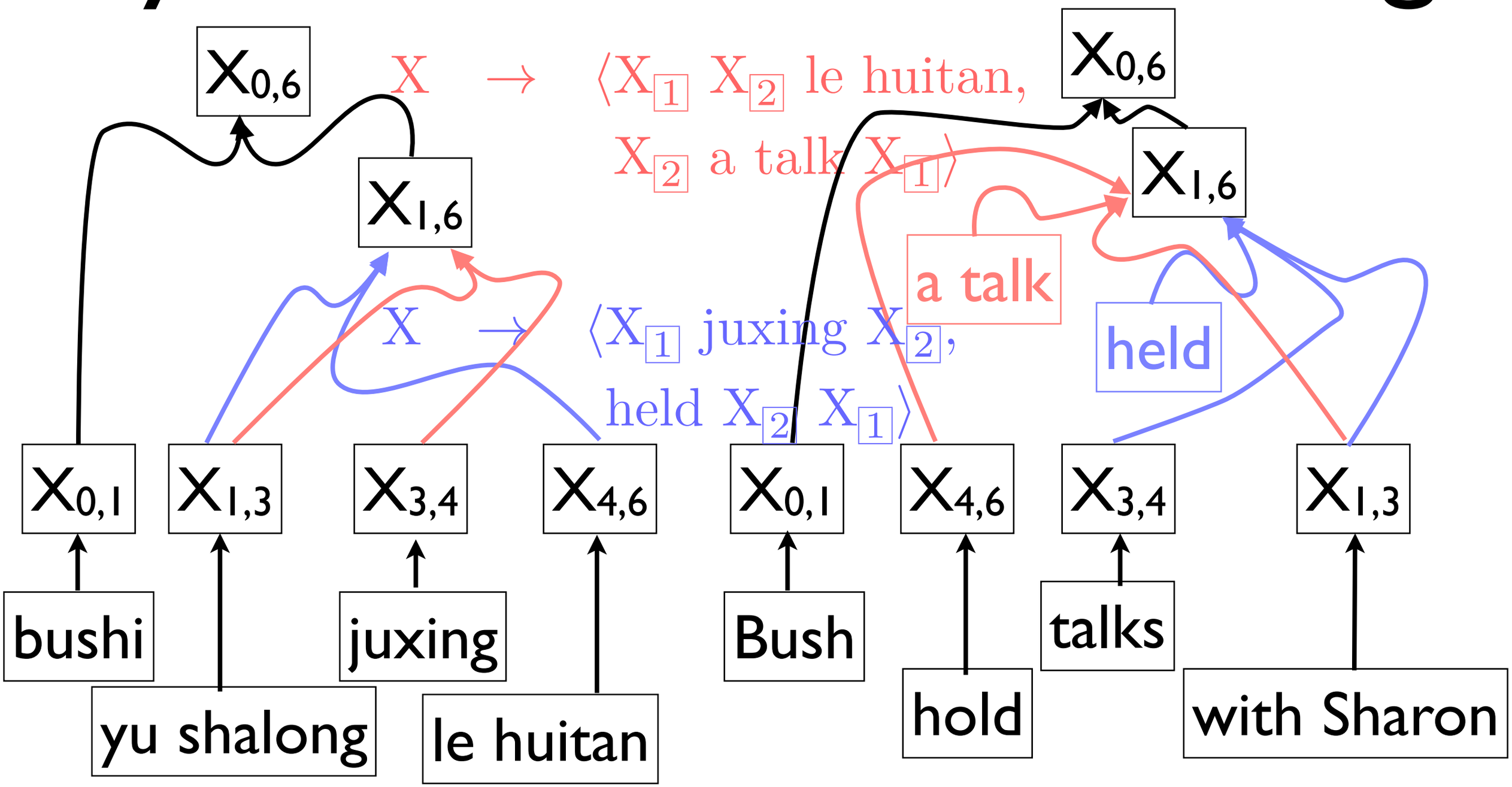
- Parse input sentence using the source side, and construct a translation forest by target side

Synchronous-CFG: Parsing



- Parse input sentence using the source side, and construct a translation forest by target side

Synchronous-CFG: Parsing



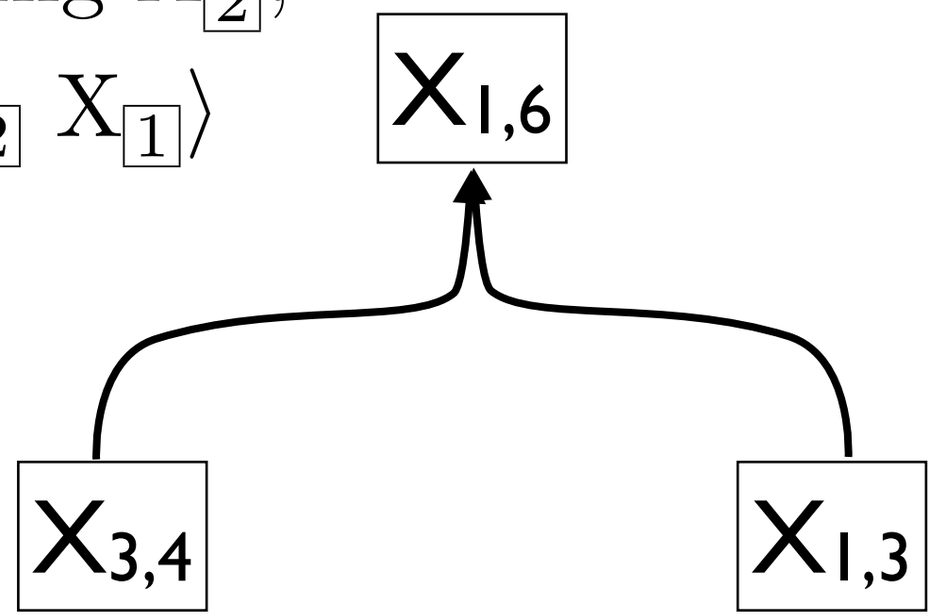
- Parse input sentence using the source side, and construct a translation forest by target side

Synchronous-CFG: Parsing

- Translation by SCFG = monolingual parsing using the source side grammar
- Complexity: $O(n^3)$ as in monolingual CKY

Non-Local Features

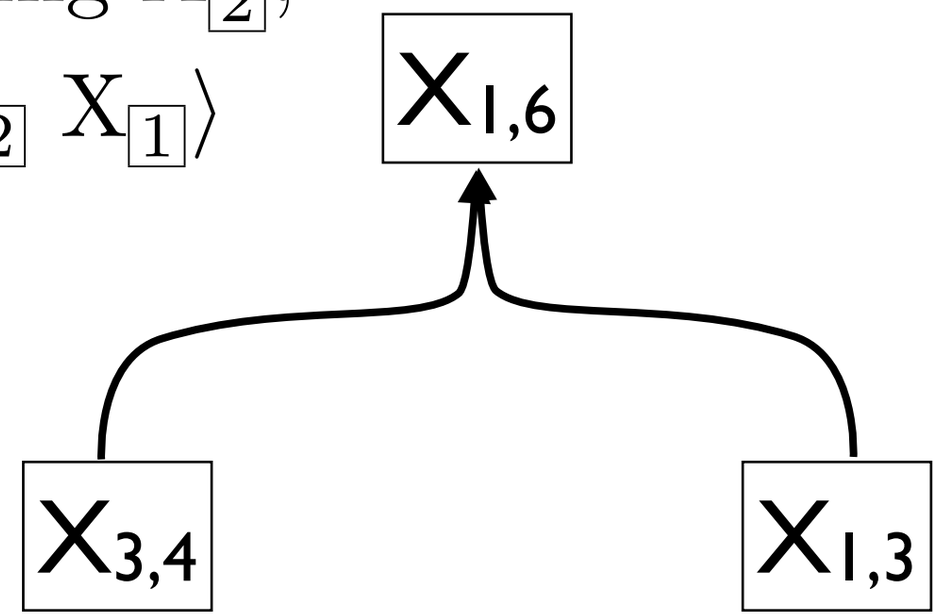
$$X \rightarrow \langle X_{\boxed{1}} \text{ juxing } X_{\boxed{2}}, \text{ held } X_{\boxed{2}}\ X_{\boxed{1}} \rangle$$



- non-local features which requires out-of-span context, i.e. bigram LM

Non-Local Features

$$X \rightarrow \langle X_{\boxed{1}} \text{ juxing } X_{\boxed{2}}, \text{ held } X_{\boxed{2}}\ X_{\boxed{1}} \rangle$$



a talk

talks

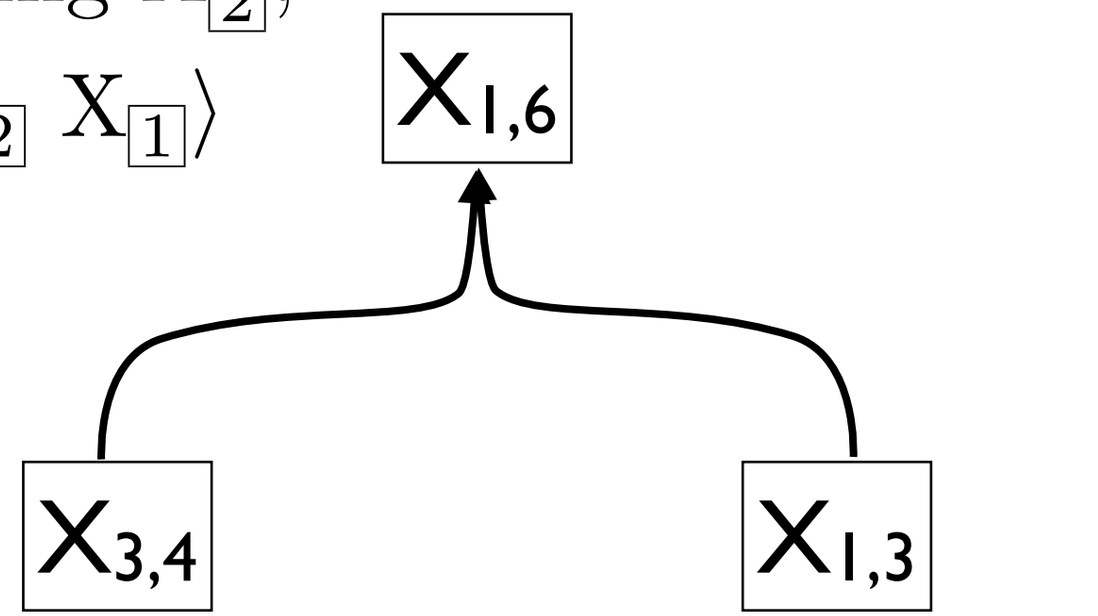
meeting

meetings

- non-local features which requires out-of-span context, i.e. bigram LM

Non-Local Features

$$X \rightarrow \langle X_{\boxed{1}} \text{ juxing } X_{\boxed{2}}, \text{ held } X_{\boxed{2}}\ X_{\boxed{1}} \rangle$$



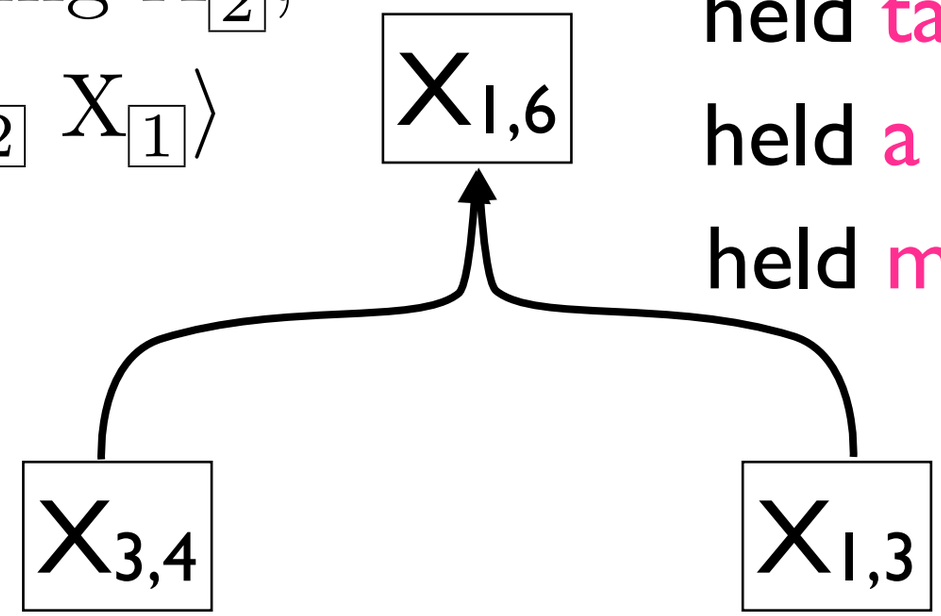
a talk
talks
meeting
meetings

with Sharon
and Sharon
Sharon with
Sharon and

- non-local features which requires out-of-span context, i.e. bigram LM

Non-Local Features

$$X \rightarrow \langle X_{\boxed{1}} \text{ juxing } X_{\boxed{2}}, \text{ held } X_{\boxed{2}} \; X_{\boxed{1}} \rangle$$



held a talk with Sharon
held talks with Sharon
held a talk and Sharon
held meeting Sharon with

a talk
talks
meeting
meetings

with Sharon
and Sharon
Sharon with
Sharon and

- non-local features which requires out-of-span context, i.e. bigram LM

Non-Local Features

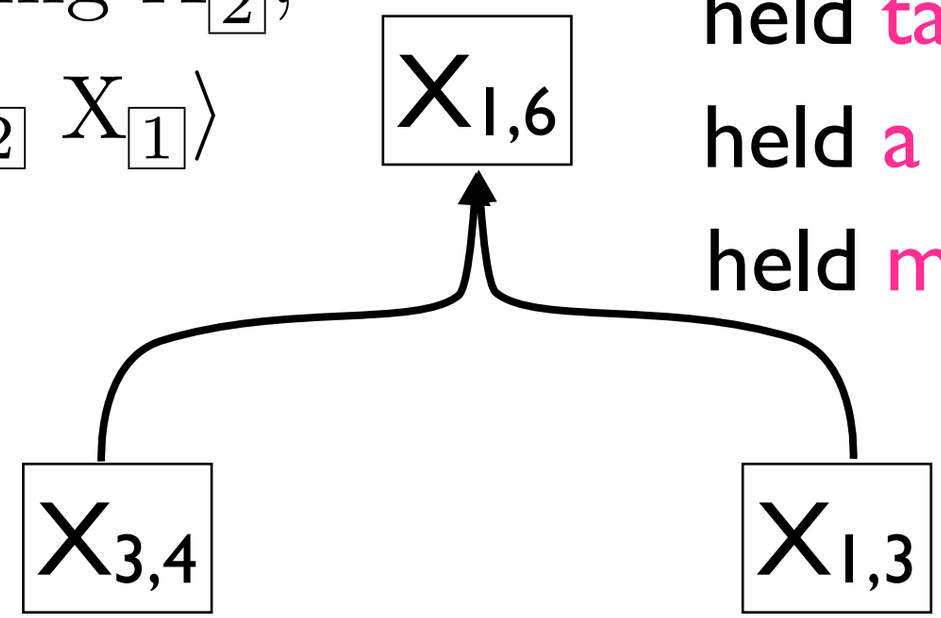
$$X \rightarrow \langle X_{\boxed{1}} \text{ juxing } X_{\boxed{2}}, \text{ held } X_{\boxed{2}} \; X_{\boxed{1}} \rangle$$

held a talk with Sharon

held talks with Sharon

held a talk and Sharon

held meeting Sharon with



p(talk | a)

a talk

talks

meeting

meetings

with Sharon

and Sharon

Sharon with

Sharon and

- non-local features which requires out-of-span context, i.e. bigram LM

Non-Local Features

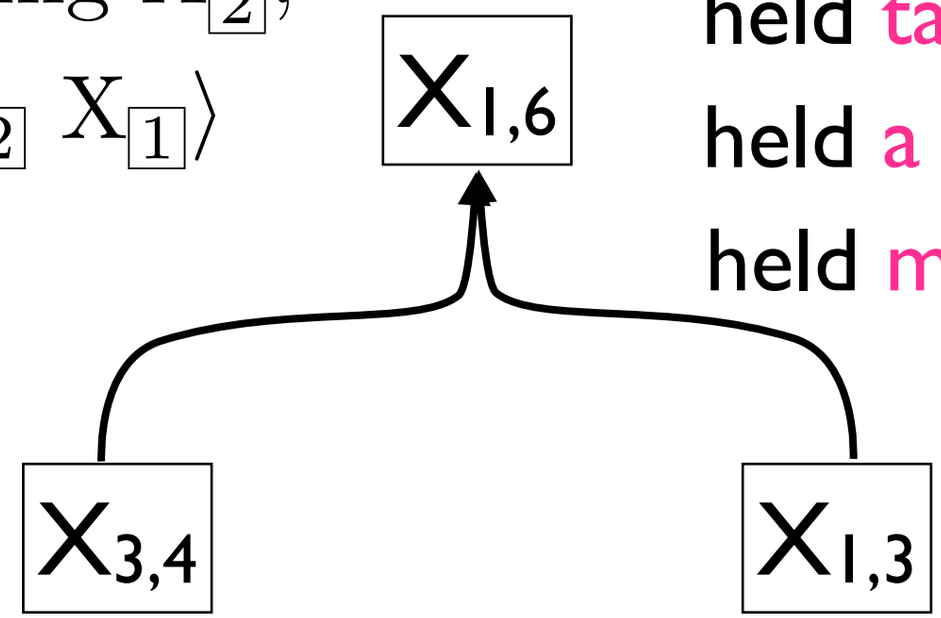
$$X \rightarrow \langle X_{\boxed{1}} \text{ juxing } X_{\boxed{2}}, \text{ held } X_{\boxed{2}} \; X_{\boxed{1}} \rangle$$

held a talk with Sharon

held talks with Sharon

held a talk and Sharon

held meeting Sharon with



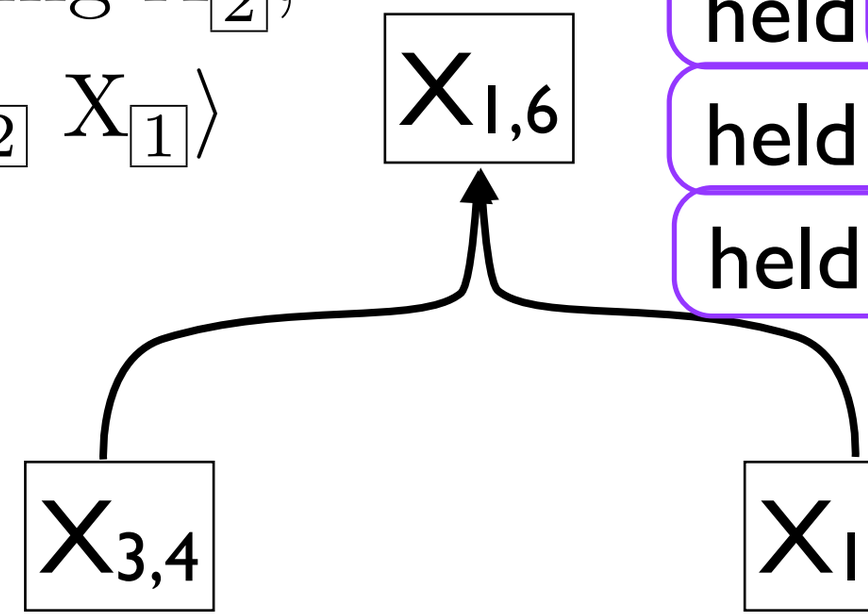
p(talk | a) a talk, talks, meeting, meetings

with Sharon	p(Sharon | with)
and Sharon	p(Sharon | and)
Sharon with	p(with | Sharon)
Sharon and	p(and | Sharon)

- non-local features which requires out-of-span context, i.e. bigram LM

Non-Local Features

$$X \rightarrow \langle X_{\boxed{1}} \text{ juxing } X_{\boxed{2}}, \text{ held } X_{\boxed{2}} \; X_{\boxed{1}} \rangle$$



held a talk with Sharon

held talks with Sharon

held a talk and Sharon

held meeting Sharon with

Update boundary words only

p(talk | a)

a talk

talks

meeting

meetings

with Sharon p(Sharon | with)

and Sharon p(Sharon | and)

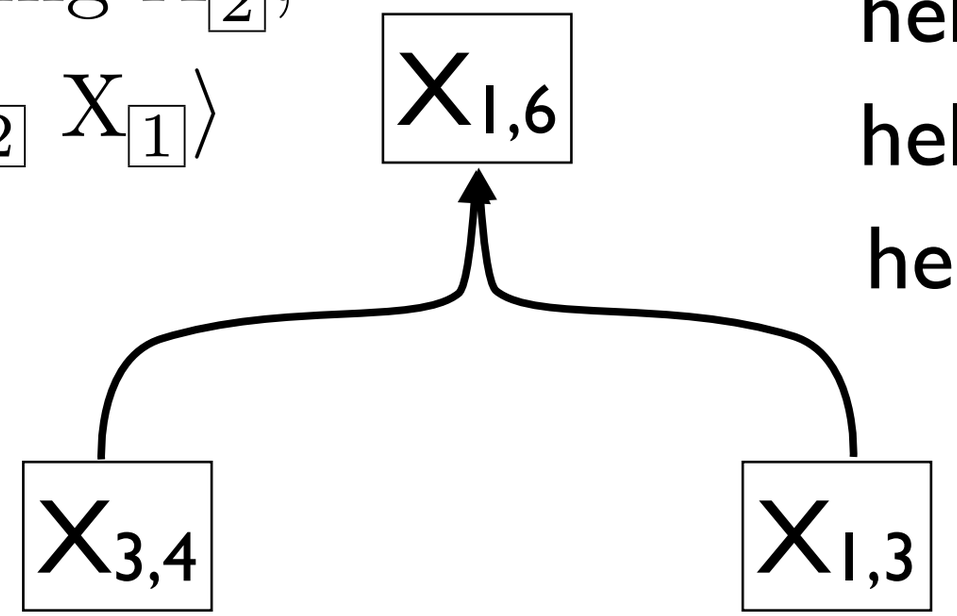
Sharon with p(with | Sharon)

Sharon and p(and | Sharon)

- non-local features which requires out-of-span context, i.e. bigram LM

Bigram Features

$$X \rightarrow \langle X_{\boxed{1}} \text{ juxing } X_{\boxed{2}}, \text{ held } X_{\boxed{2}} \ X_{\boxed{1}} \rangle$$



held * Sharon
held * Sharon
held * Sharon
held * with

a * talk
talks
meeting
meetings

with * Sharon
and * Sharon
Sharon * with
Sharon * and

- We keep only bigram states

Language Model Updates

- Each hypothesis keeps two contexts:
  - Prefix: ngrams to be scored with antecedents
  - Suffix: contexts for future ngrams (i.e. Phrase-based MT)
- Complexity: $O(n^3V^{2(m-1)})$
- Very inefficient: we need to explicitly enumerate all the hypotheses in antecedents

Cube Pruning

$$X \rightarrow \langle X_{\boxed{1}} \text{ juxing } X_{\boxed{2}}, \text{ held } X_{\boxed{2}} \; X_{\boxed{1}} \rangle$$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.5	1.7	2.6	3.2
a * talk	1.0	2.5	2.7	3.6	4.2
talks	1.3	2.8	3.0	3.9	4.5
meeting	2.2	3.7	3.9	4.8	5.4
meetings	2.6	4.1	4.3	5.2	5.8

- For each rule, create a "cube" representing combinations of antecedents (Huang and Chiang, 2007)

Cube Pruning

$$X \rightarrow \langle X_{\boxed{1}} \text{ juxing } X_{\boxed{2}}, \text{ held } X_{\boxed{2}} \; X_{\boxed{1}} \rangle$$

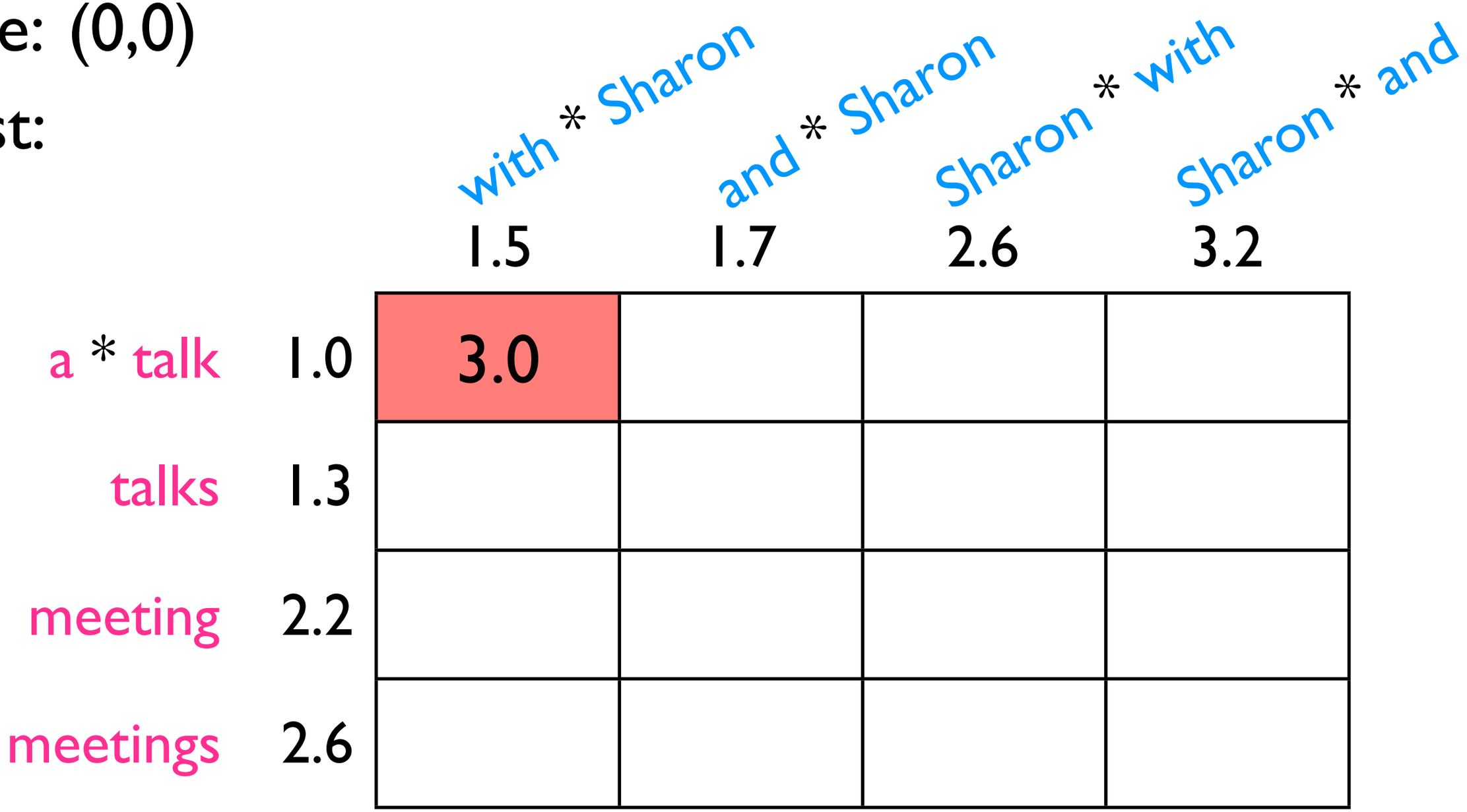
		with * Sharon 1.5	and * Sharon 1.7	Sharon * with 2.6	Sharon * and 3.2
a * talk	1.0	2.5 +0.5	2.7 +1.0	3.6 +1.5	4.2 +2.5
talks	1.3	2.8 +0.3	3.0 +1.5	3.9 +2.0	4.5 +2.0
meeting	2.2	3.7 +0.5	3.9 +1.0	4.8 +1.5	5.4 +2.5
meetings	2.6	4.1 +0.3	4.3 +1.5	5.2 +2.0	5.8 +2.0

- Bigrams require contexts from antecedents: non-monotonic scoring

Cube Pruning

queue: (0,0)

k-best:

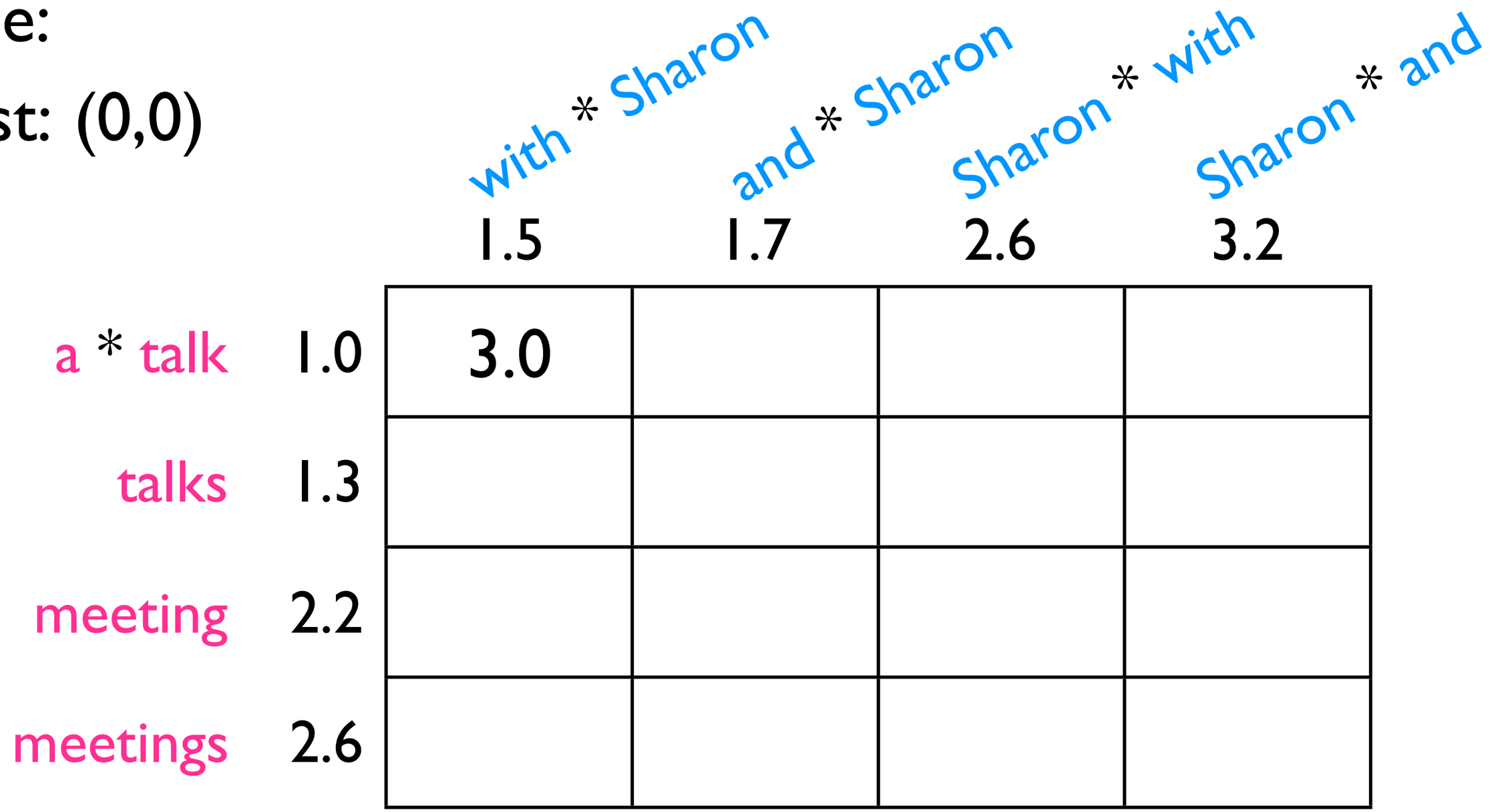


- Starting from the upper-left corner, enumerate antecedent combinations

Cube Pruning

queue:

k-best: (0,0)

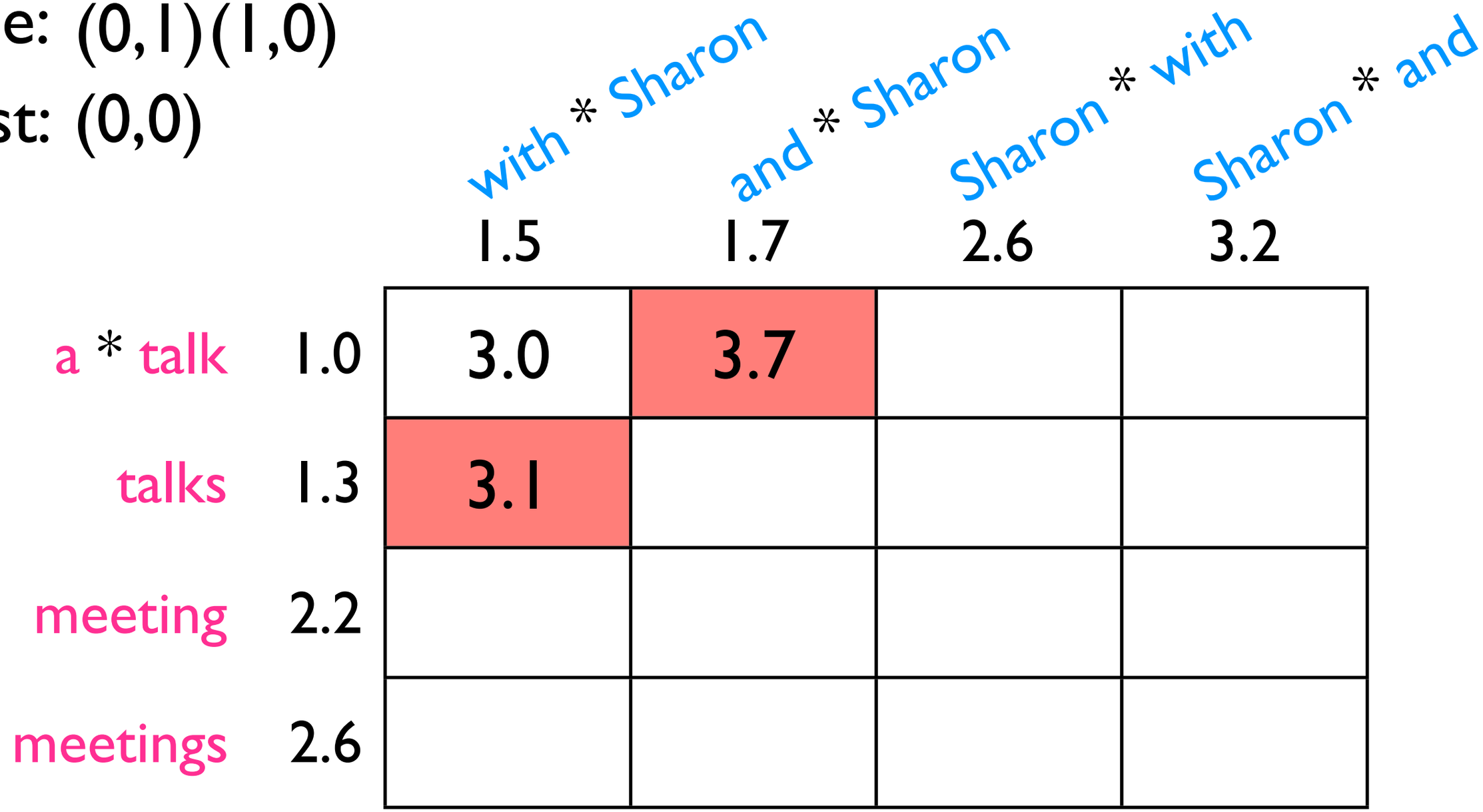


- Starting from the upper-left corner, enumerate antecedent combinations

Cube Pruning

queue: (0,1)(1,0)

k-best: (0,0)

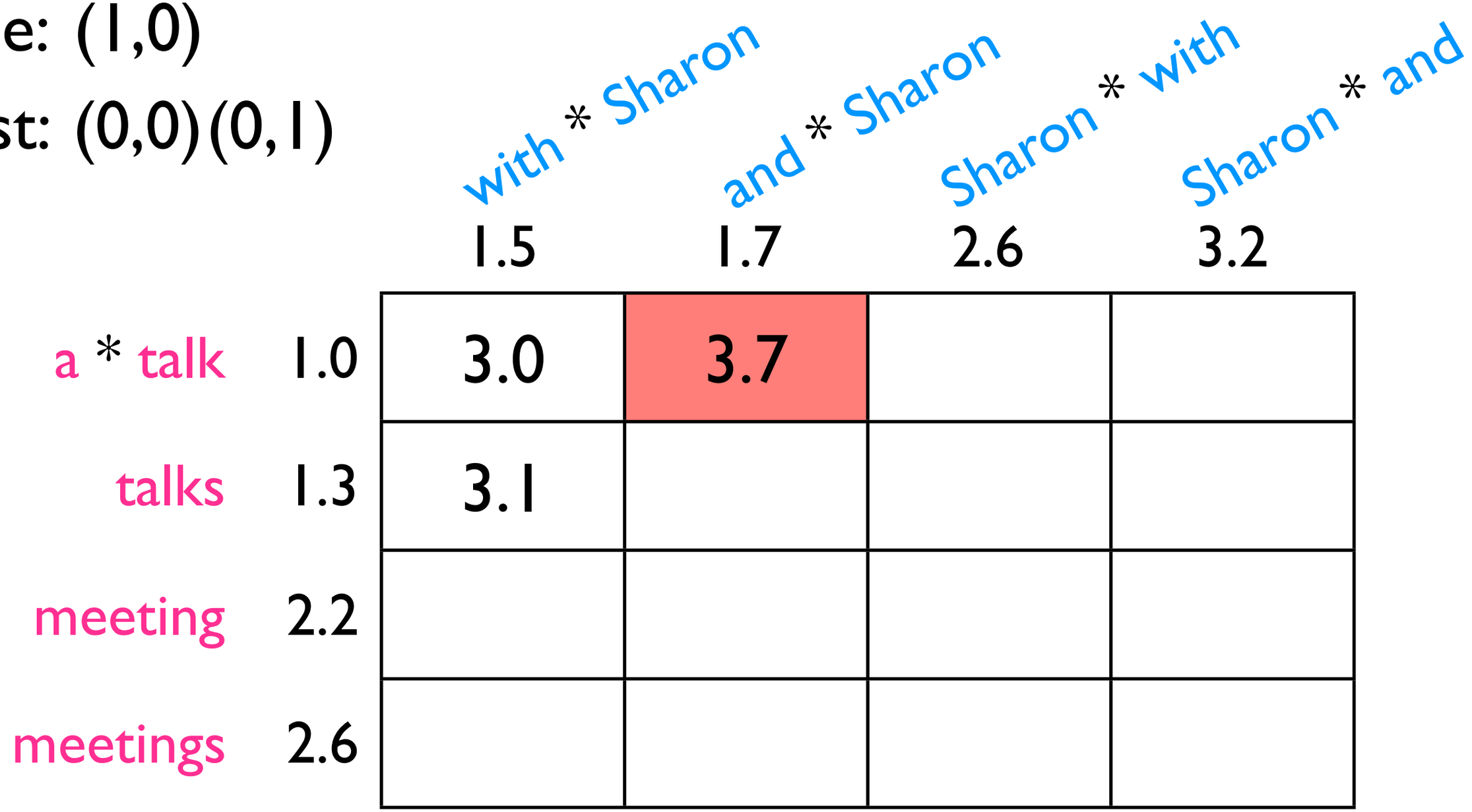


- Starting from the upper-left corner, enumerate antecedent combinations

Cube Pruning

queue: (1,0)

k-best: (0,0)(0,1)

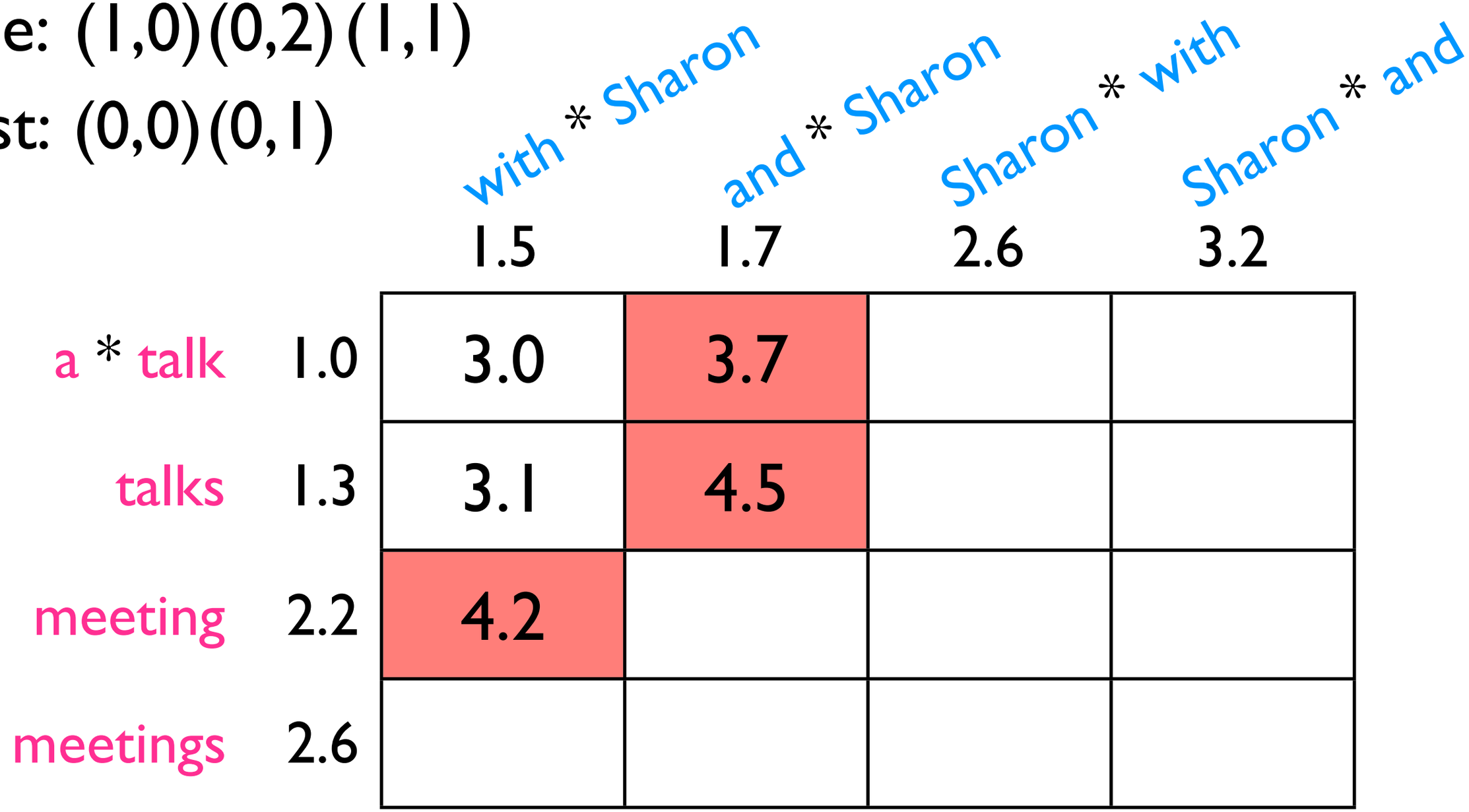


- Starting from the upper-left corner, enumerate antecedent combinations

Cube Pruning

queue: (1,0) (0,2) (1,1)

k-best: (0,0) (0,1)



- Starting from the upper-left corner, enumerate antecedent combinations

Cube Pruning

queue: (0,2) (1,1)

k-best: (0,0) (0,1) (1,0)

		with * Sharon 1.5	and * Sharon 1.7	Sharon * with 2.6	Sharon * and 3.2
a * talk	1.0	3.0	3.7		
talks	1.3	3.1	4.5		
meeting	2.2	4.2			
meetings	2.6				

- Starting from the upper-left corner, enumerate antecedent combinations

Cube Pruning

queue: (0,2) (1,1) (3,0)

k-best: (0,0) (0,1) (1,0)

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.5	1.7	2.6	3.2
a * talk	1.0	3.0	3.7	5.1	
talks	1.3	3.1	4.5		
meeting	2.2	4.2			
meetings	2.6				

- Starting from the upper-left corner, enumerate antecedent combinations

Cube Pruning

queue: (1,1)(3,0)

k-best: (0,0)(0,1)(1,0) (0,2)

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.5	1.7	2.6	3.2
a * talk	1.0	3.0	3.7	5.1	
talks	1.3	3.1	4.5		
meeting	2.2	4.2			
meetings	2.6				

- Starting from the upper-left corner, enumerate antecedent combinations

Cube Pruning

queue: (0,4) (1,1) (1,2) (3,0)

k-best: (0,0) (0,1) (1,0) (0,2)

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.5	1.7	2.6	3.2
a * talk	1.0	3.0	3.7	5.1	
talks	1.3	3.1	4.5		
meeting	2.2	4.2	4.9		
meetings	2.6	4.4			

- Starting from the upper-left corner, enumerate antecedent combinations

Multiple Rules

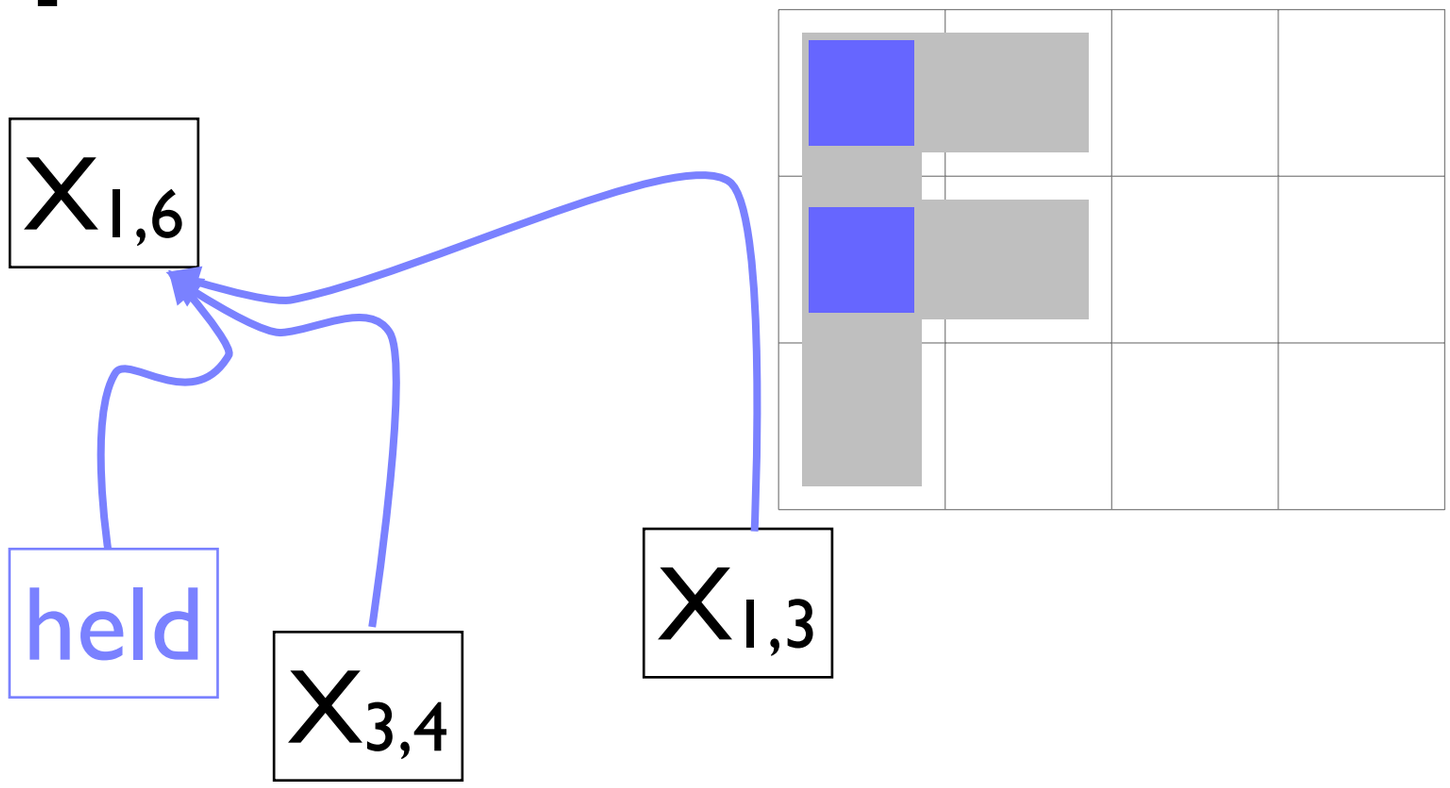
- Multiple rules sharing the same span are queued
  - Each rule is associated with a cube
  - hypothesis = rule + cube-position

Multiple Rules

$X_{1,6}$

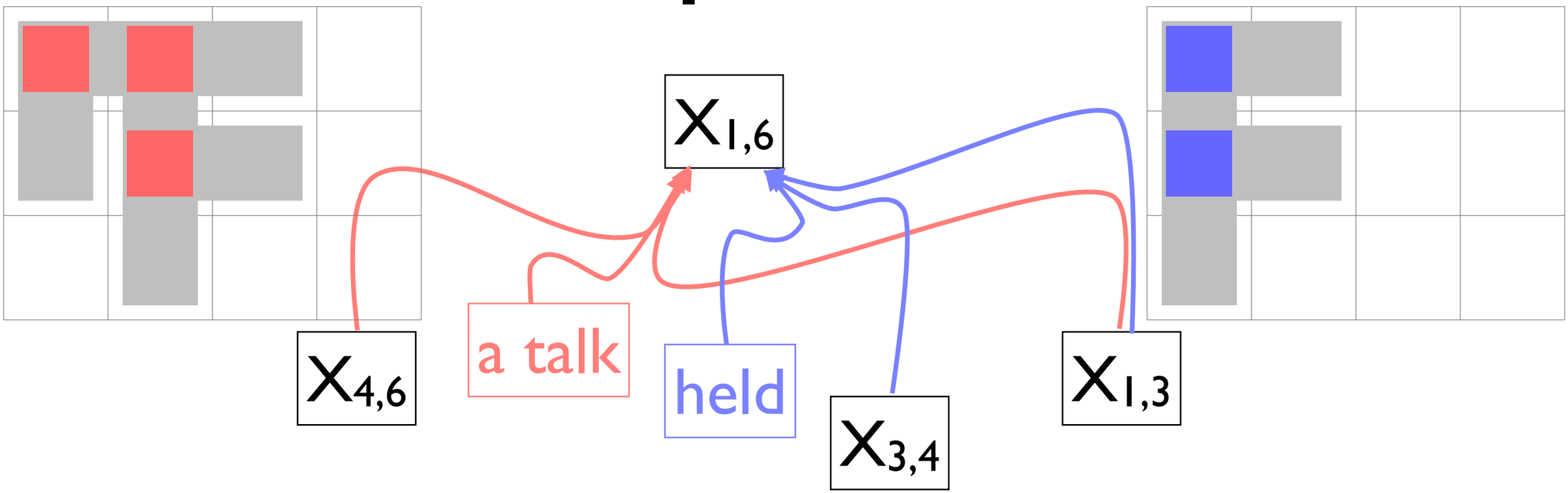
- Multiple rules sharing the same span are queued
  - Each rule is associated with a cube
  - hypothesis = rule + cube-position

Multiple Rules



- Multiple rules sharing the same span are queued
  - Each rule is associated with a cube
  - hypothesis = rule + cube-position

Multiple Rules



- Multiple rules sharing the same span are queued
  - Each rule is associated with a cube
  - hypothesis = rule + cube-position

Further Faster Pruning

- Cube Growing (Huang and Chiang, 2007)
  - Top-down pruning combined with heuristic estimates
- Faster Cube Pruning (Gesmundo and Henderson, 2010)
  - Eliminate bookkeeping for inserted hypotheses by determining the ordering of cube enumerations
  - Push minimum hypotheses by looking up ancestors

Conclusion

- Synchronous-CFG
  - paired CFG + shared non-terminal symbols
- Training is based on phrase-based MT by treating sub-phrase as a non-terminal
- Decoding: monolingual parsing
  - An efficient antecedent combination via cube-pruning

Overview

- Backgrounds
  - CFG, parsing, hypergraph, deductive system semirings
- Tree-based SMT
  - Synchronous-CFG
  - String-to-Tree, Tree-to-String
  - Bitext parsing

Overview

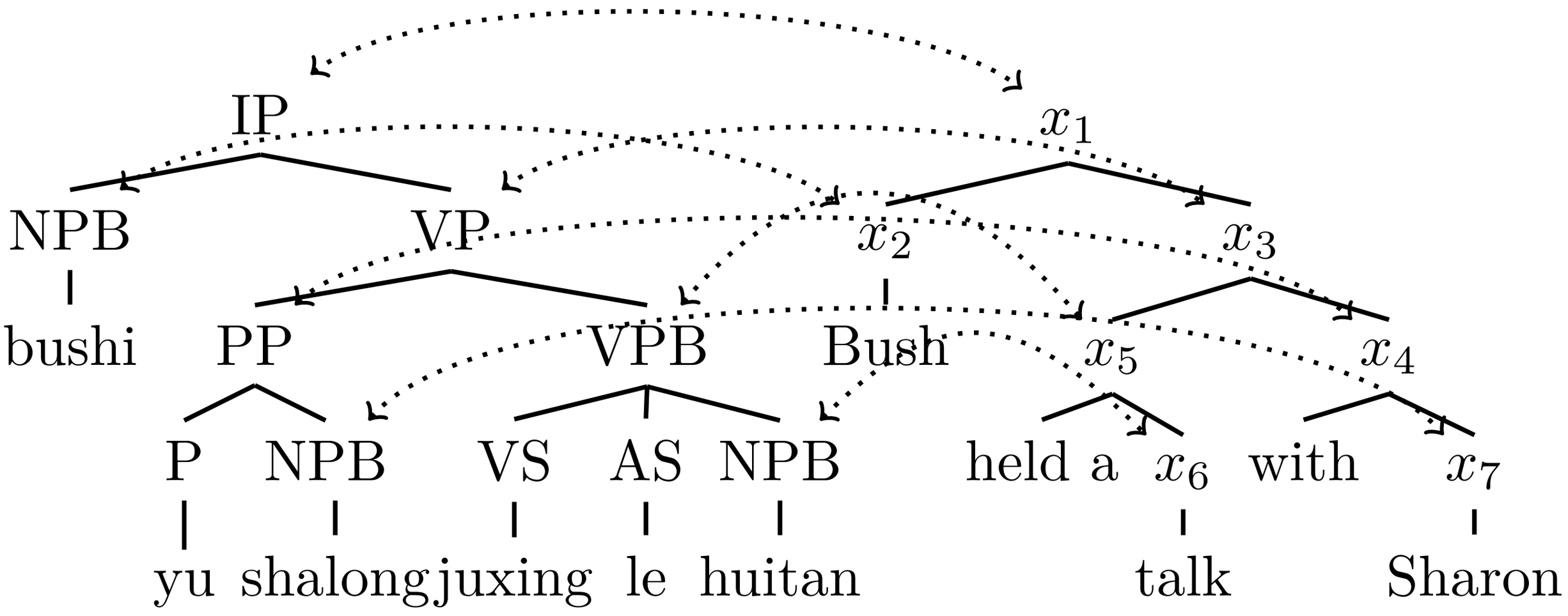
- Backgrounds
  - CFG, parsing, hypergraph, deductive system semirings
- Tree-based SMT
  - Synchronous-CFG
  - String-to-Tree, Tree-to-String
  - Bitext parsing

{Tree,String}-to-{Tree,String}

(Galley et al., 2004)

- Each synchronous rule has a subtree structure
- Flat structure + sharing the same non-terminal symbols = synchronous-CFG

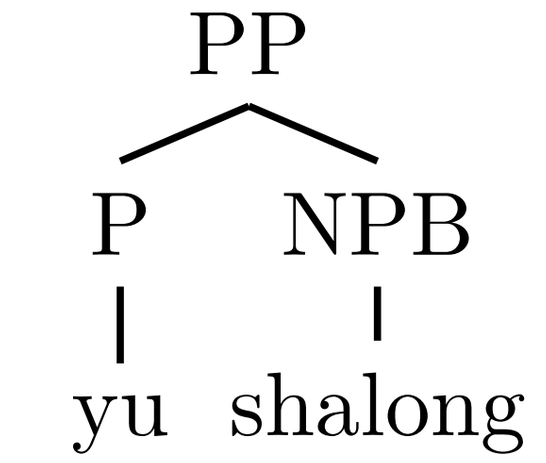
{Tree,String}-to-{Tree,String}



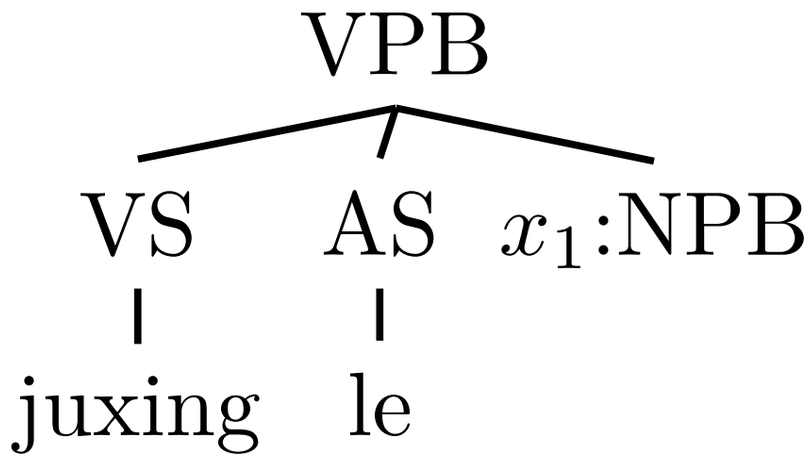
(Galley et al., 2004)

- Each synchronous rule has a subtree structure
- Flat structure + sharing the same non-terminal symbols = synchronous-CFG

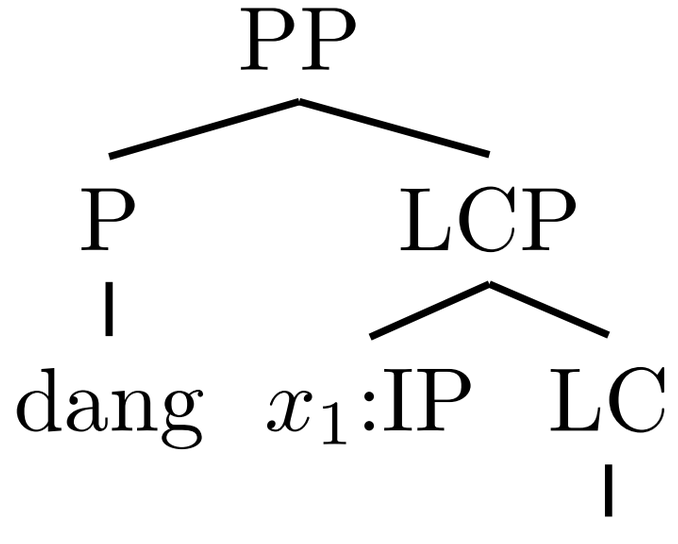
Tree-to-String Rules



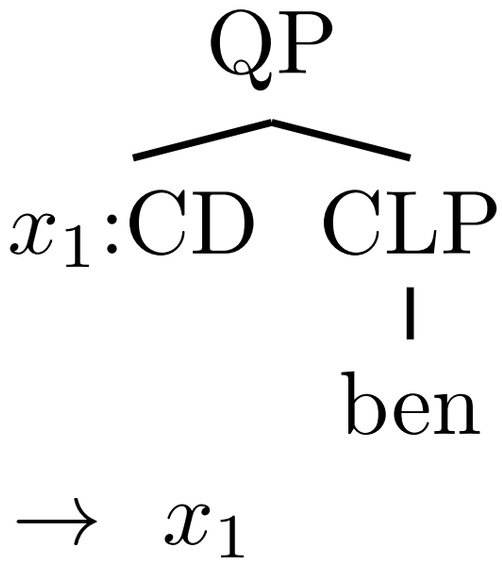
$\rightarrow$ with Sharon



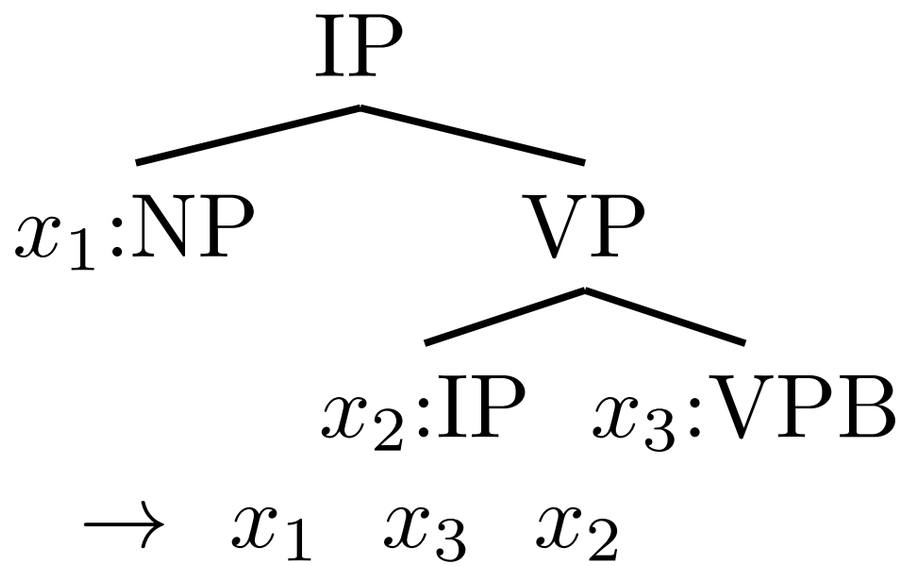
$\rightarrow$ held a x_1



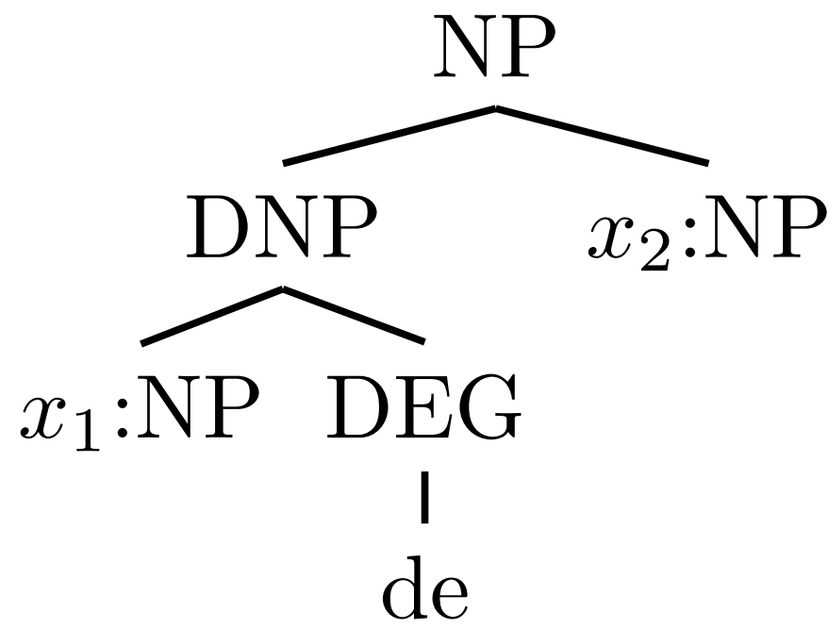
$\rightarrow$ when x_1



$\rightarrow$ x_1

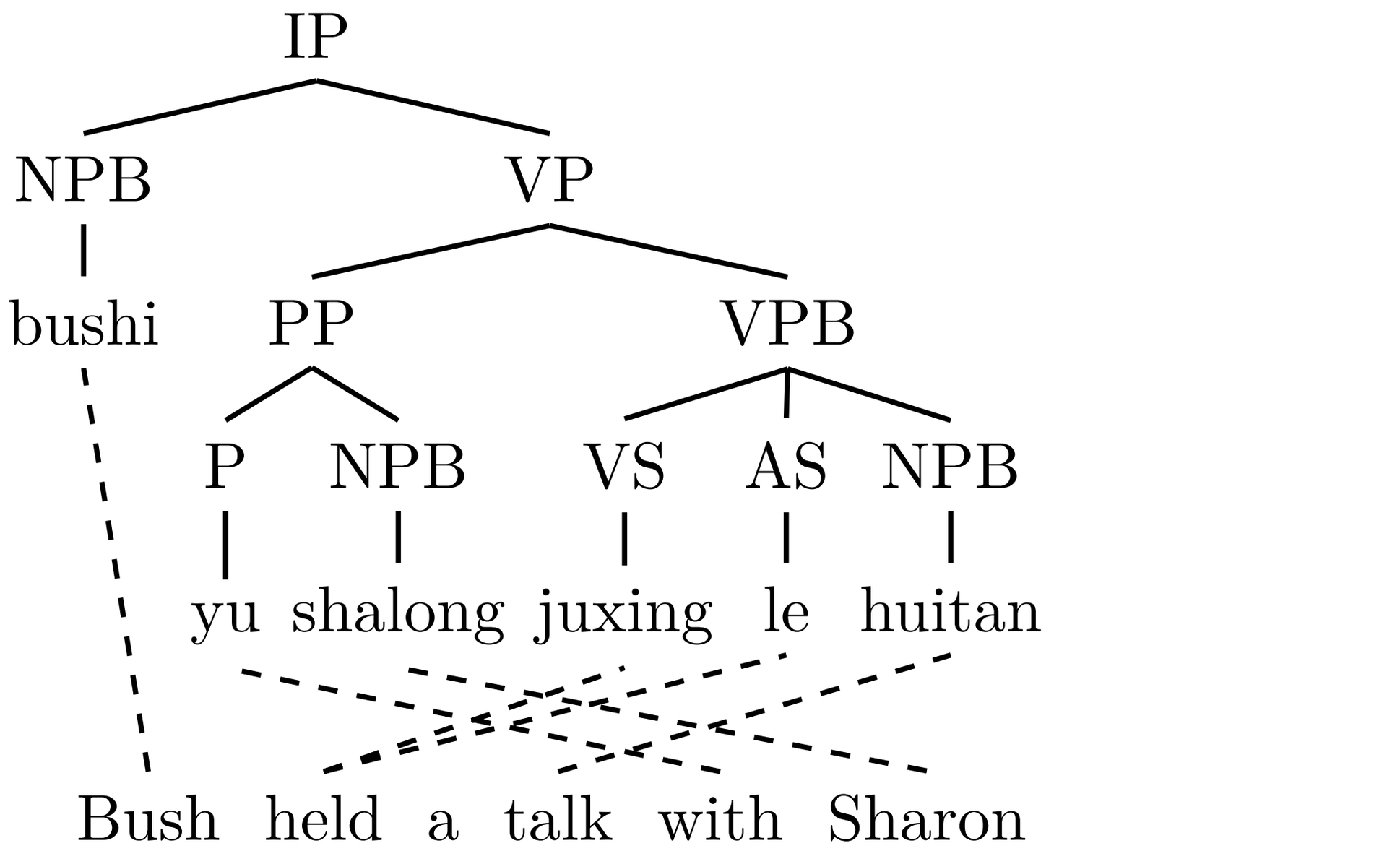


$\rightarrow$ x_1 x_3 x_2



$\rightarrow$ x_2 of x_1

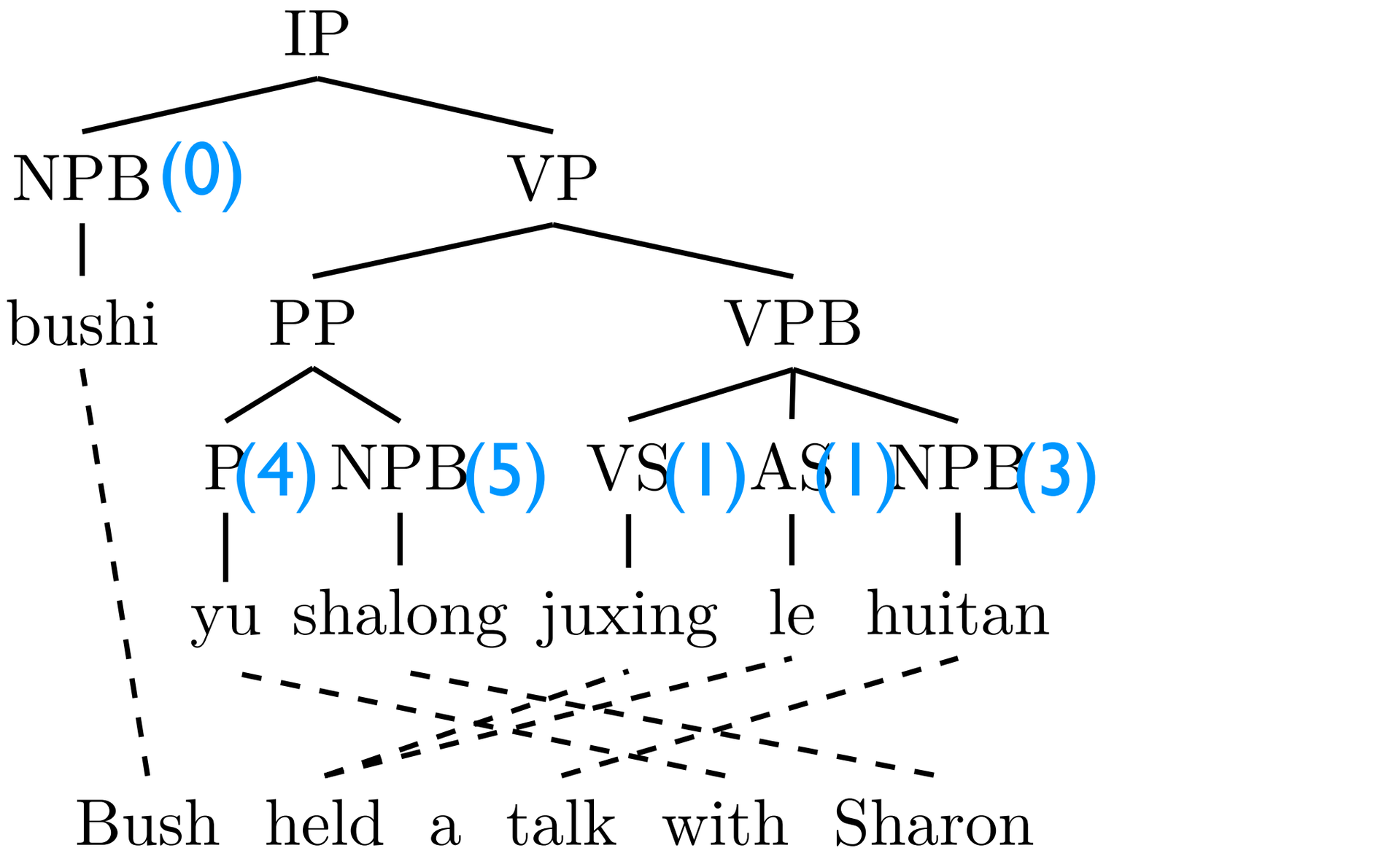
Rule Extraction



(Galley et al., 2004)

- Compute "spans" by propagating alignment in bottom-up

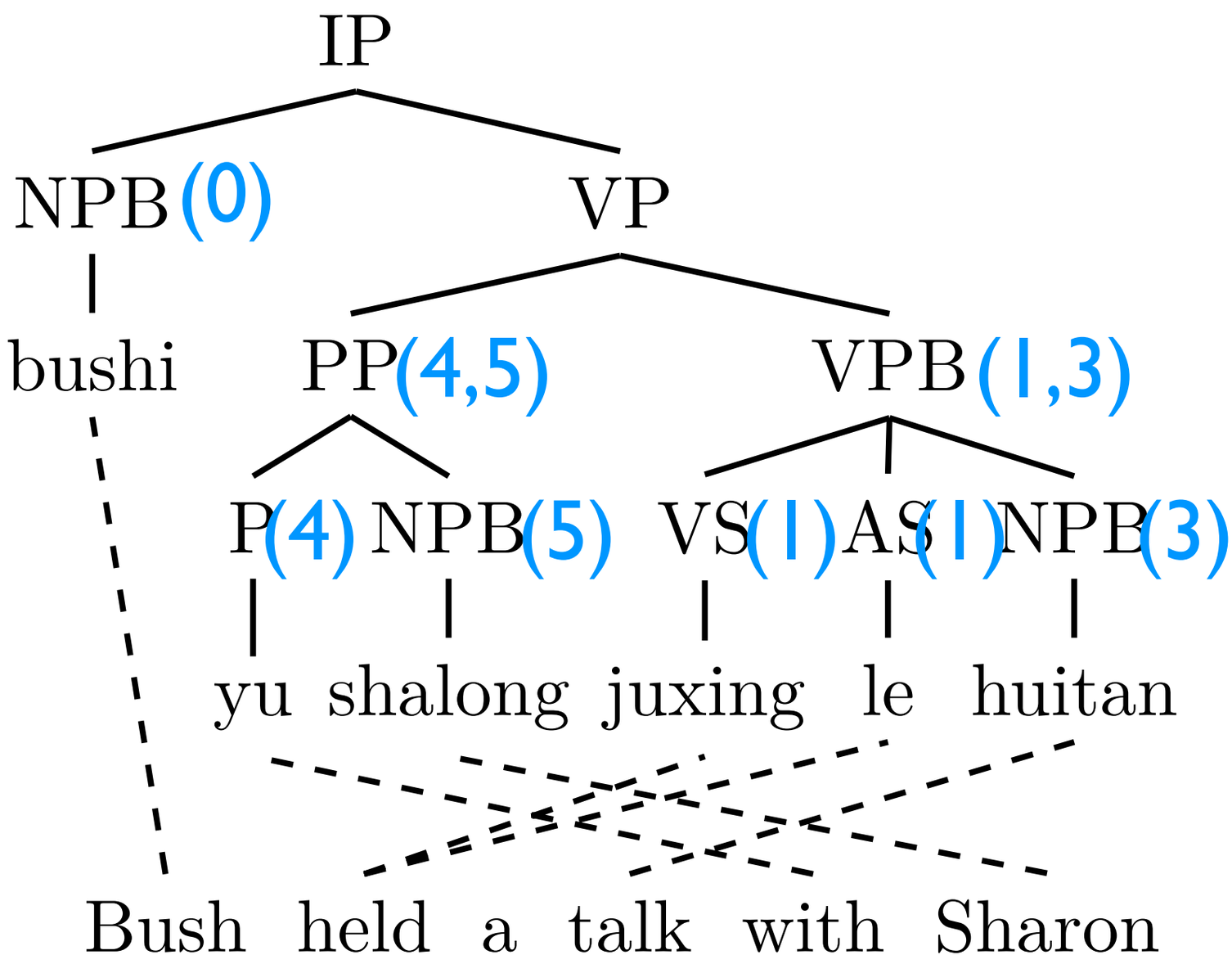
Rule Extraction



(Galley et al., 2004)

- Compute "spans" by propagating alignment in bottom-up

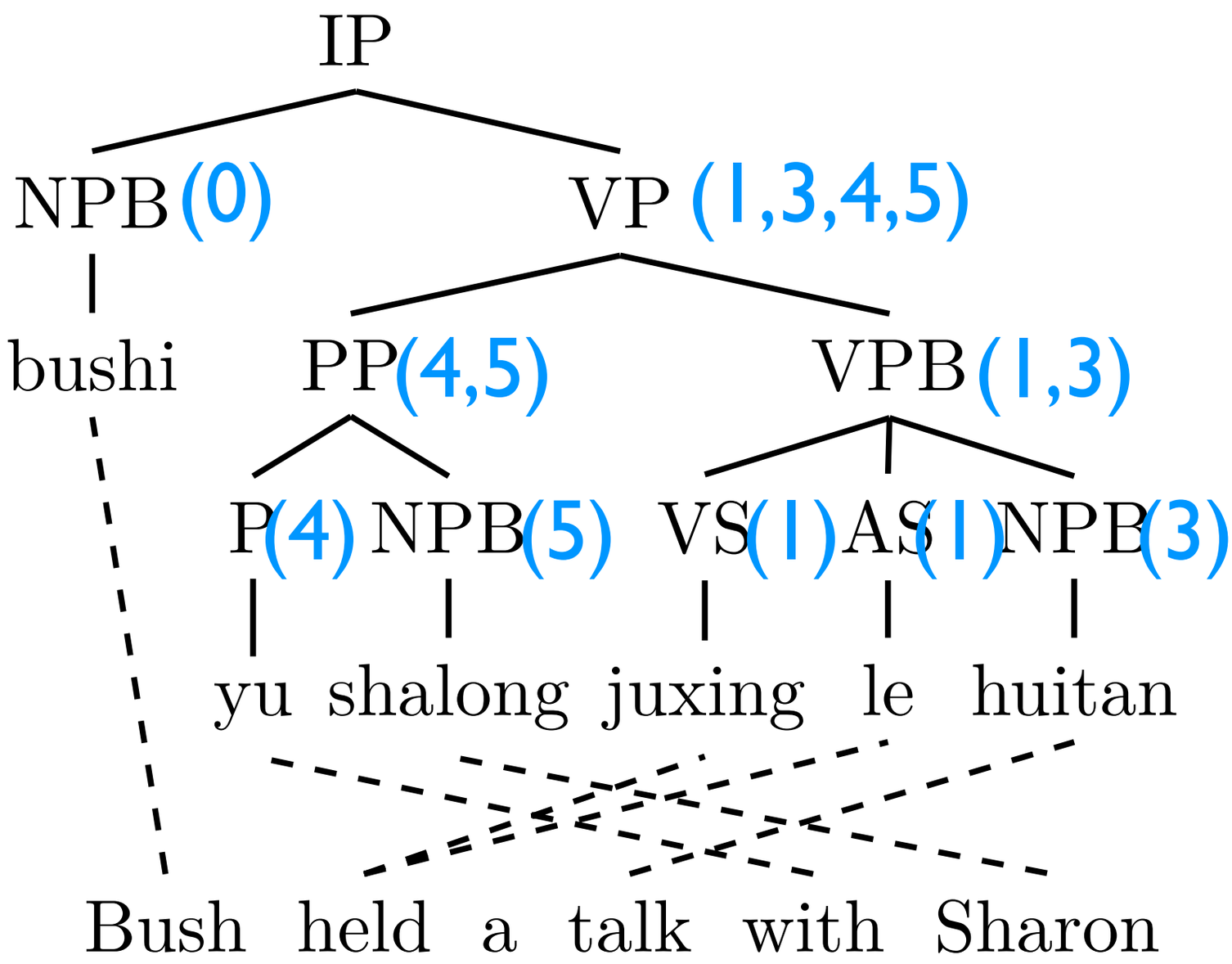
Rule Extraction



(Galley et al., 2004)

- Compute "spans" by propagating alignment in bottom-up

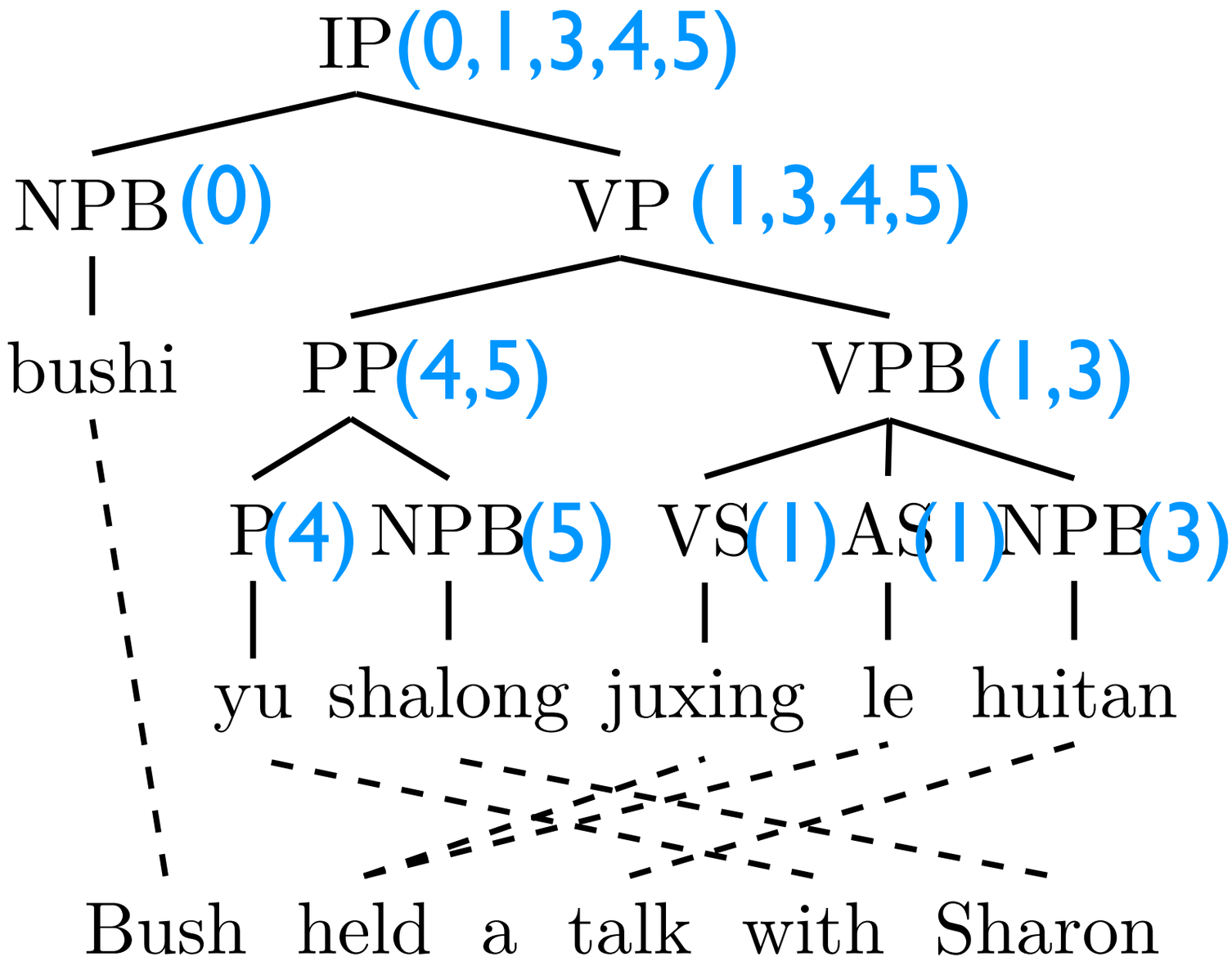
Rule Extraction



(Galley et al., 2004)

- Compute "spans" by propagating alignment in bottom-up

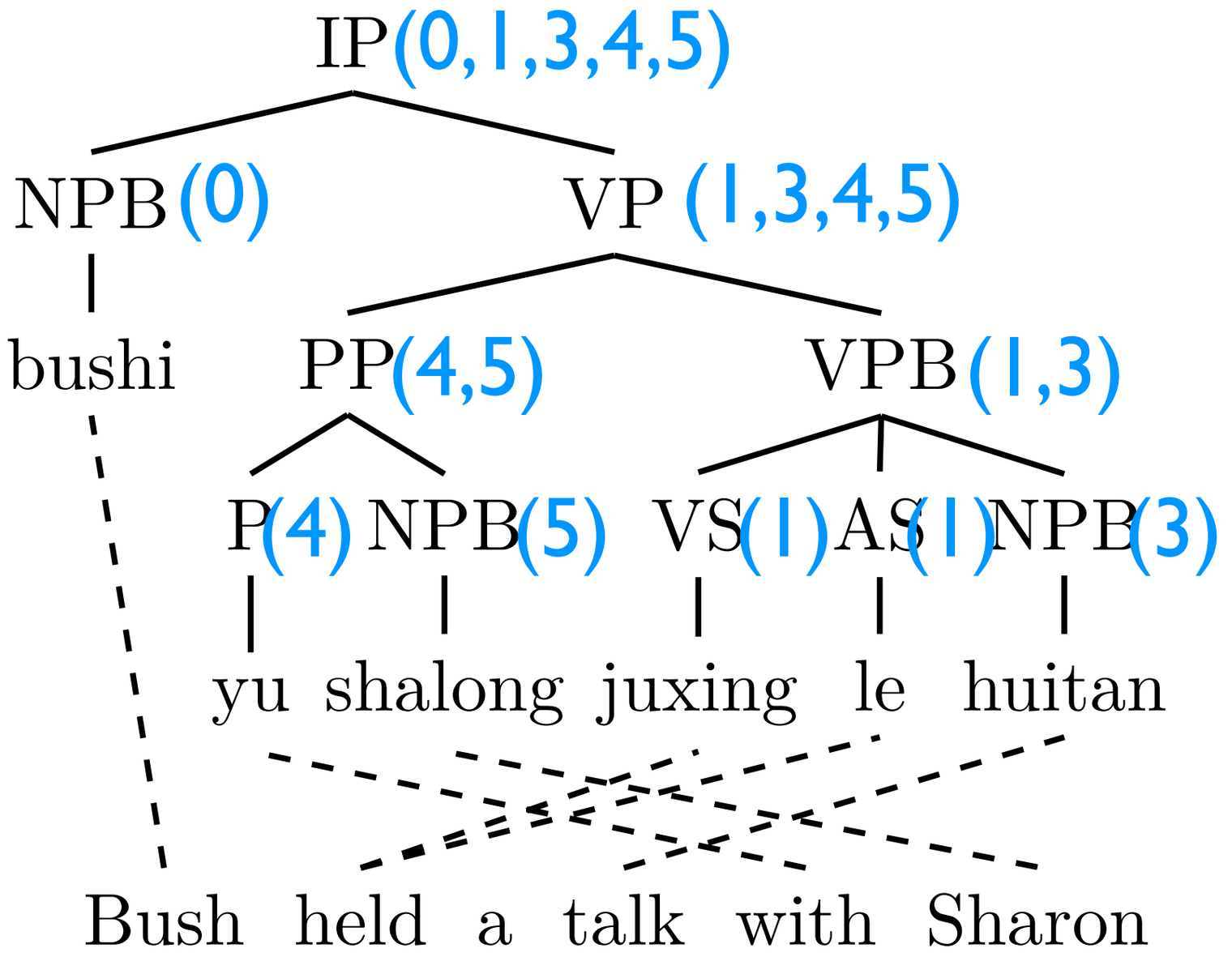
Rule Extraction



(Galley et al., 2004)

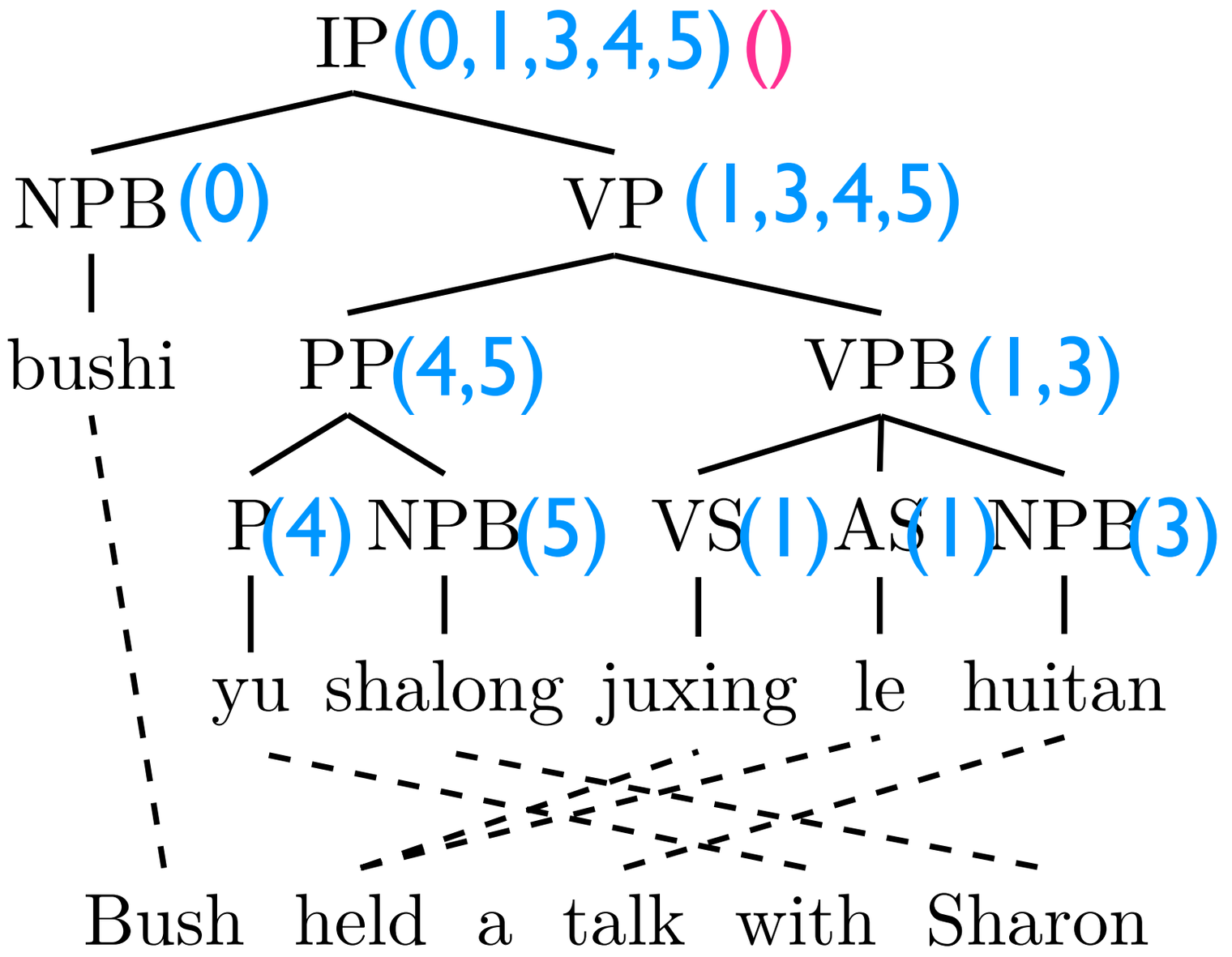
- Compute "spans" by propagating alignment in bottom-up

Rule Extraction



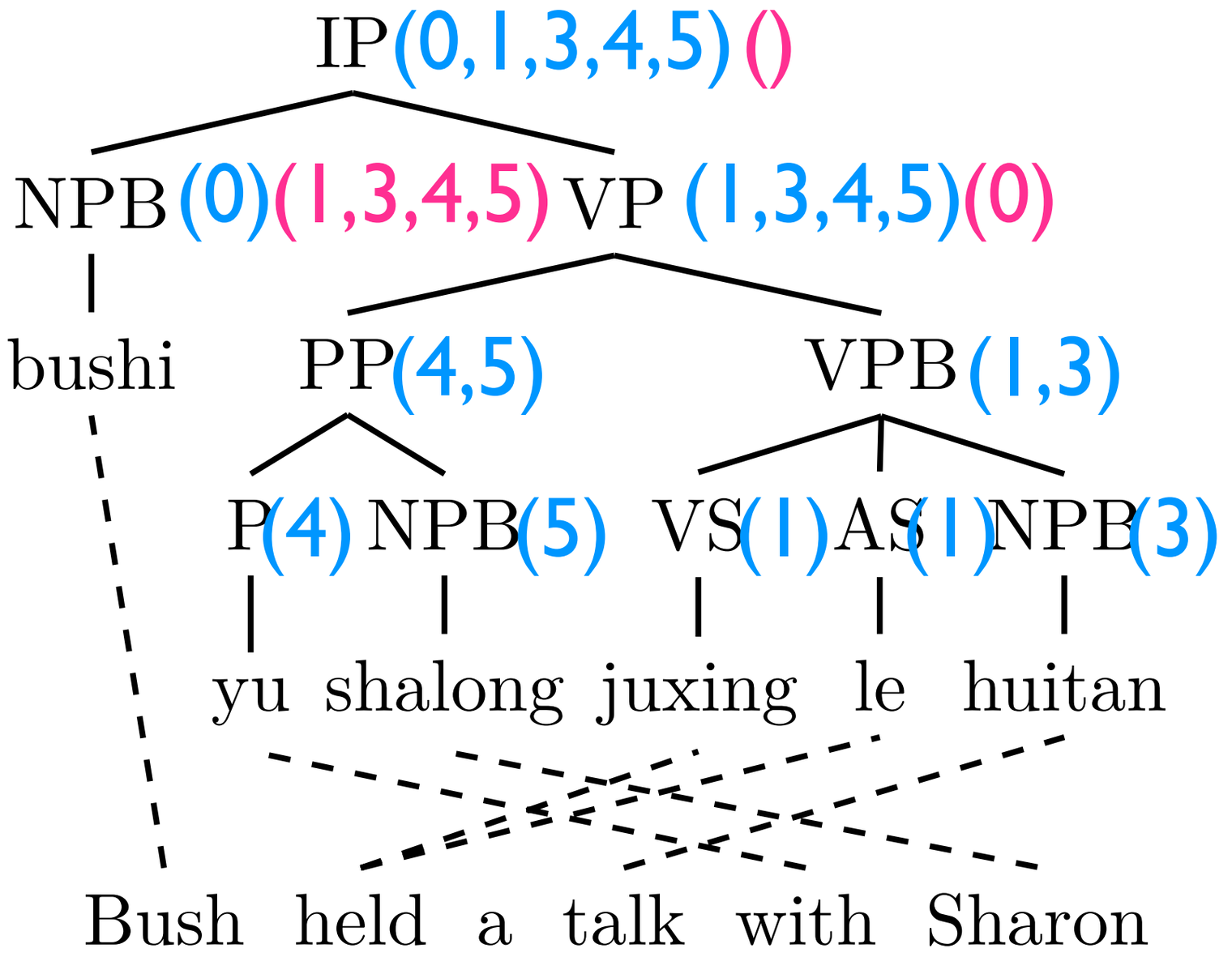
- Compute "complements" in top-down

Rule Extraction



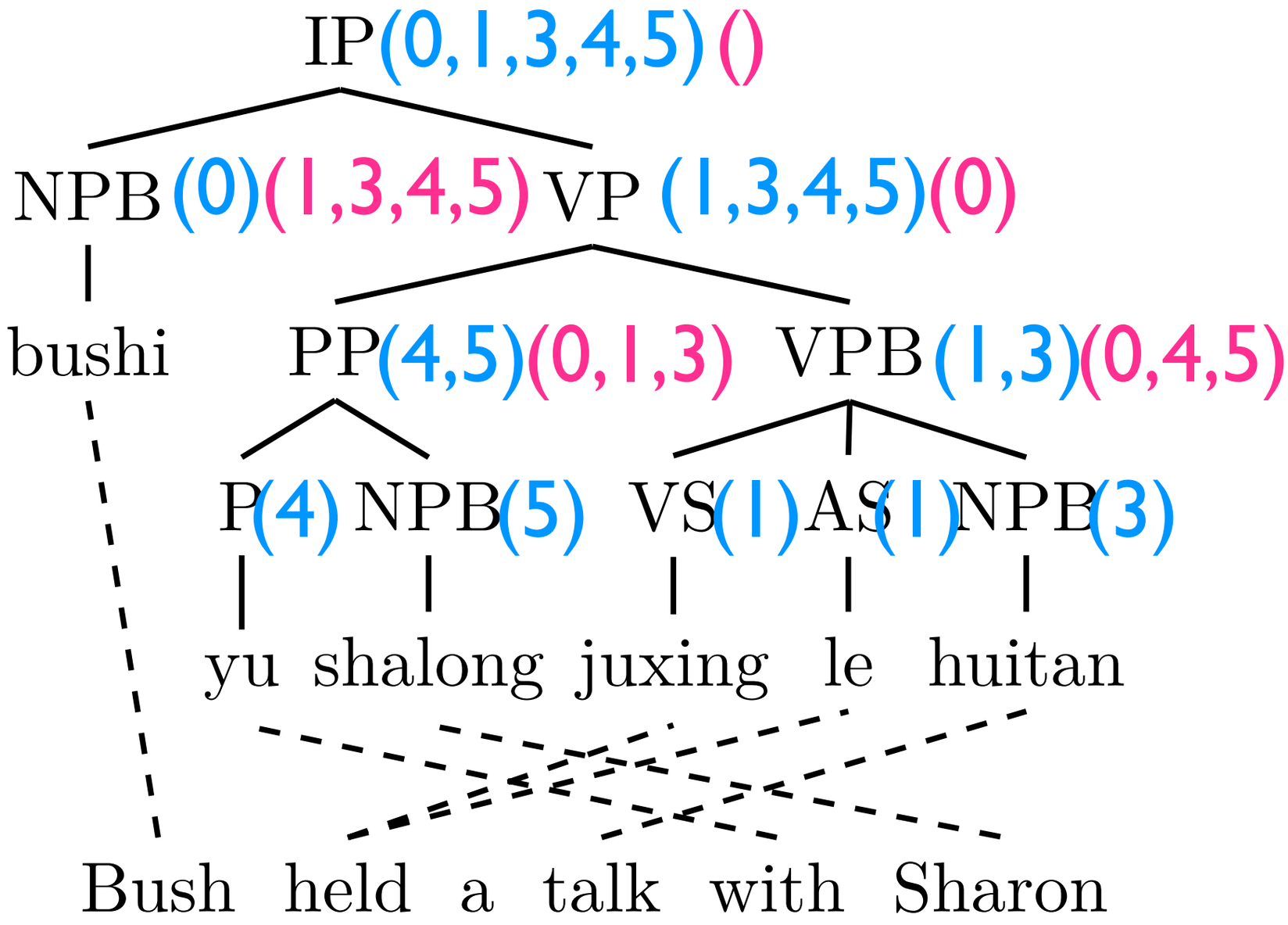
- Compute "complements" in top-down

Rule Extraction



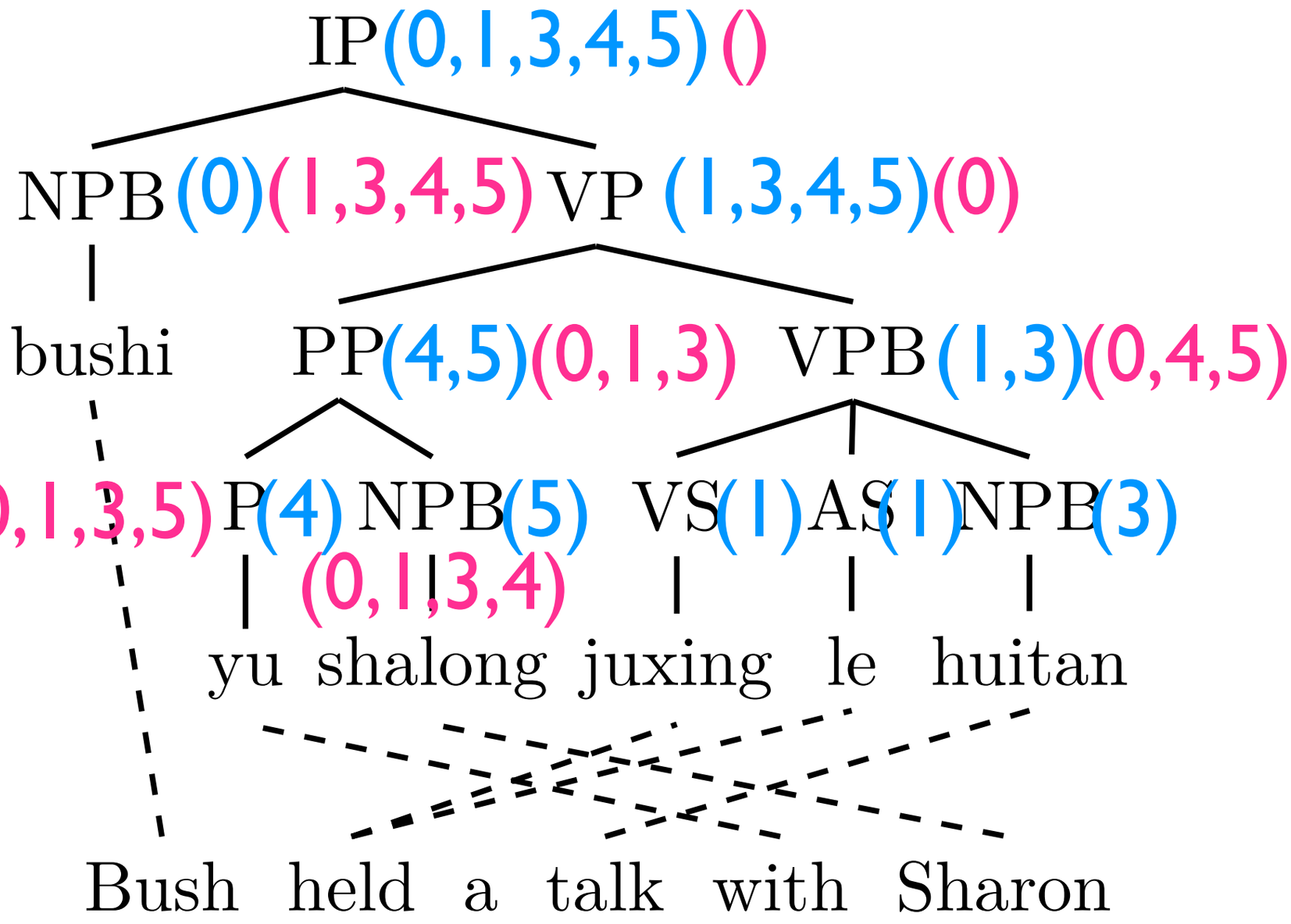
- Compute "complements" in top-down

Rule Extraction



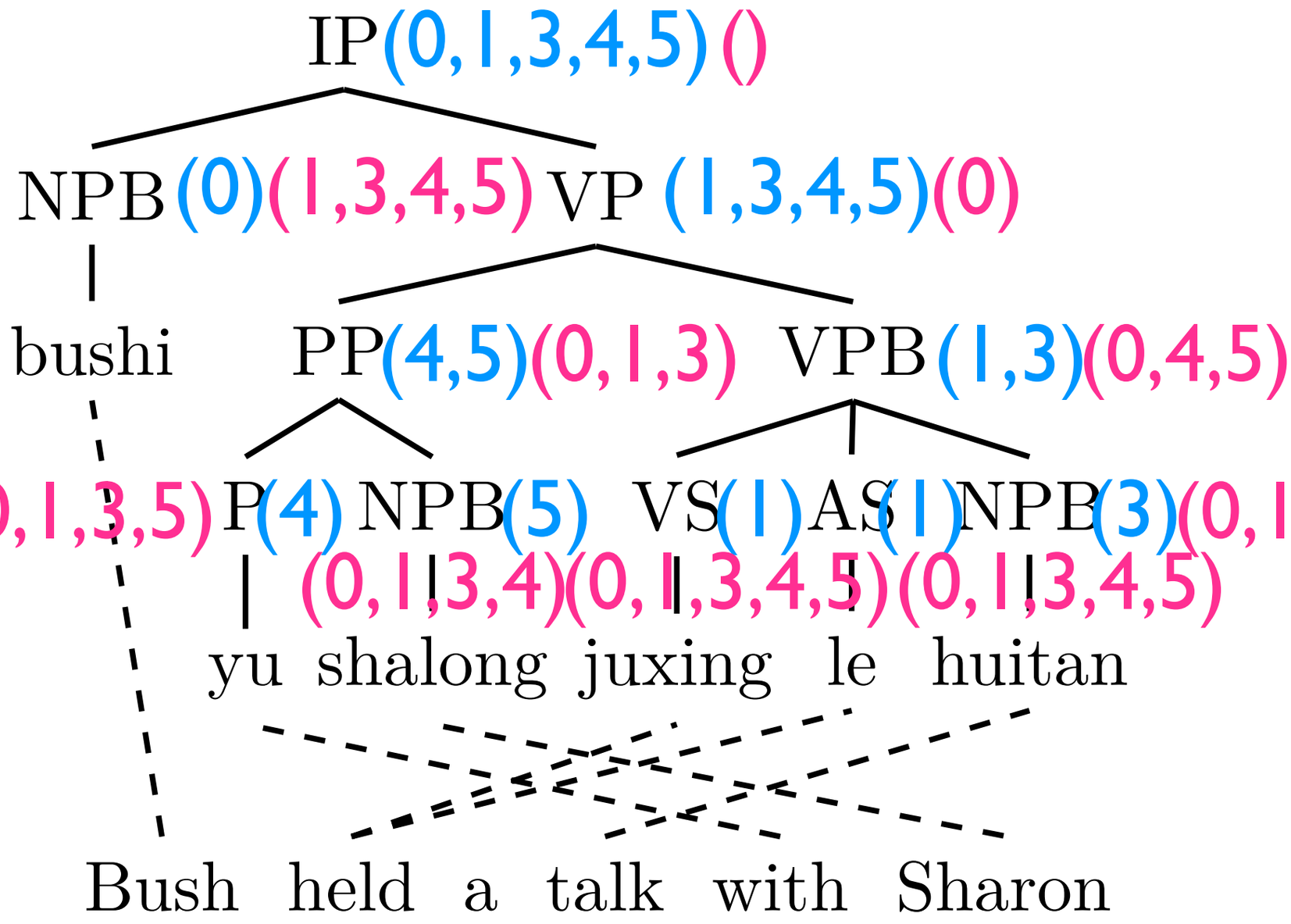
- Compute "complements" in top-down

Rule Extraction



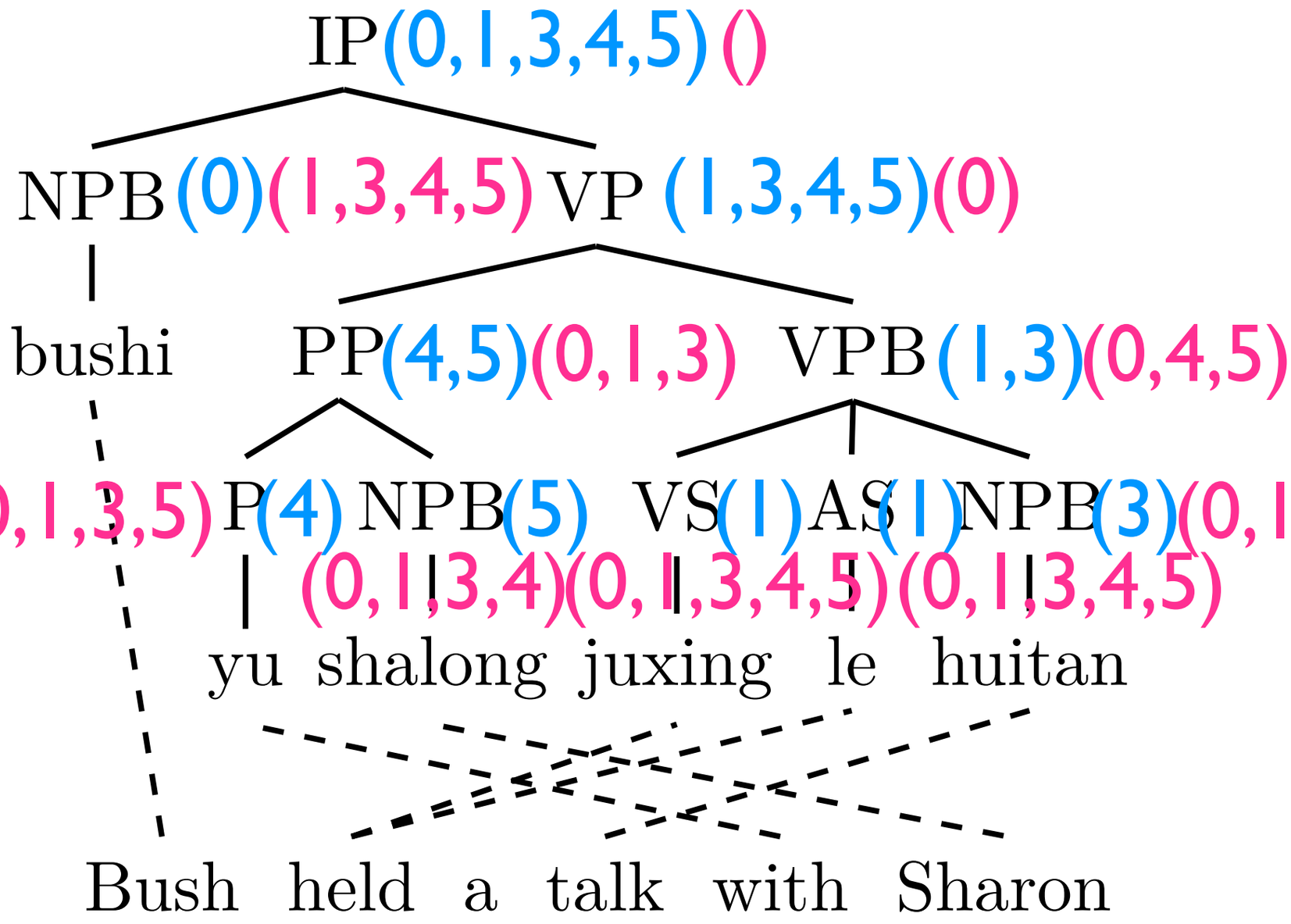
- Compute "complements" in top-down

Rule Extraction



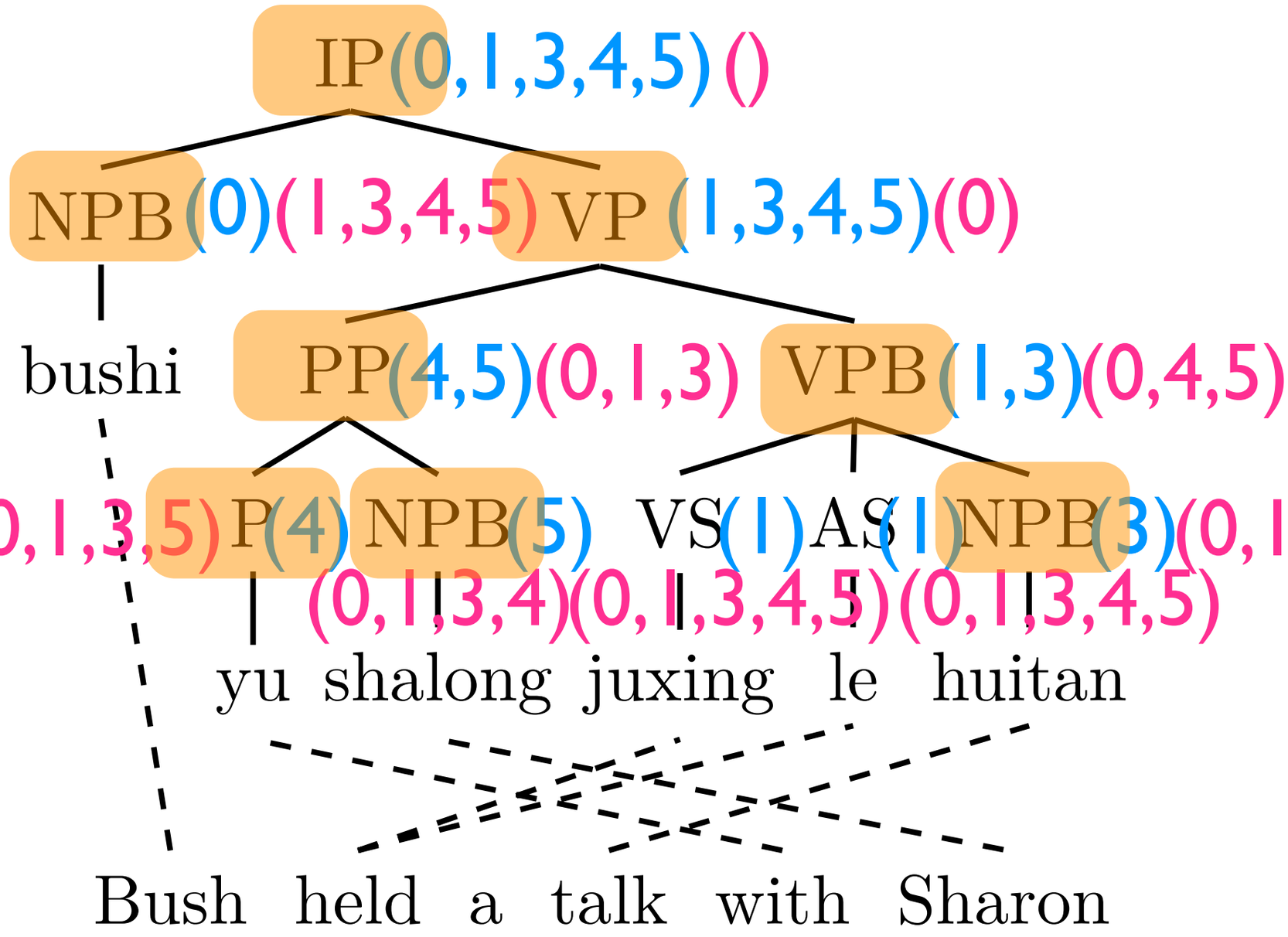
- Compute "complements" in top-down

Rule Extraction



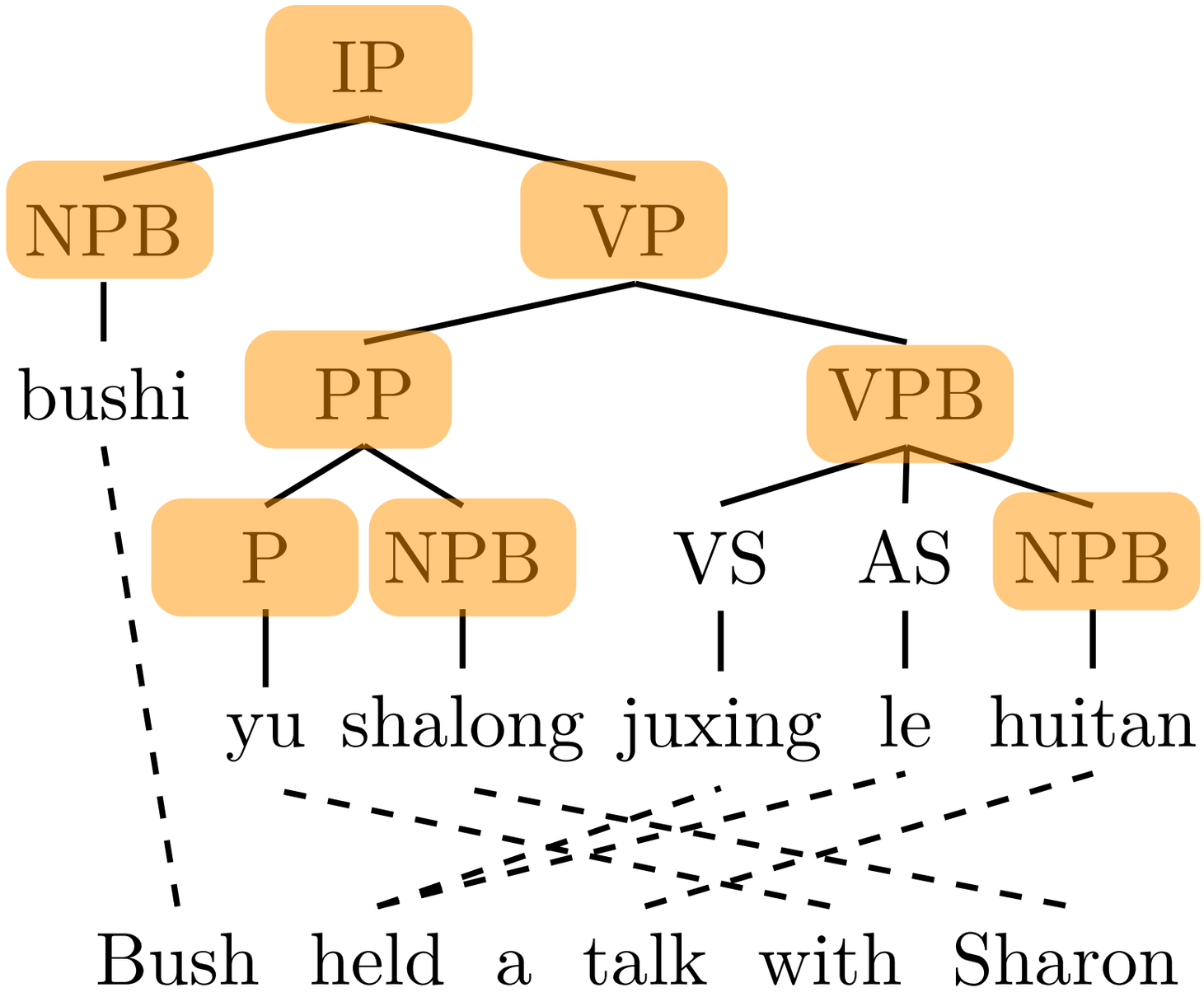
- Compute "frontiers": The nodes in which the intersection of "spans" and "complements" is empty

Rule Extraction



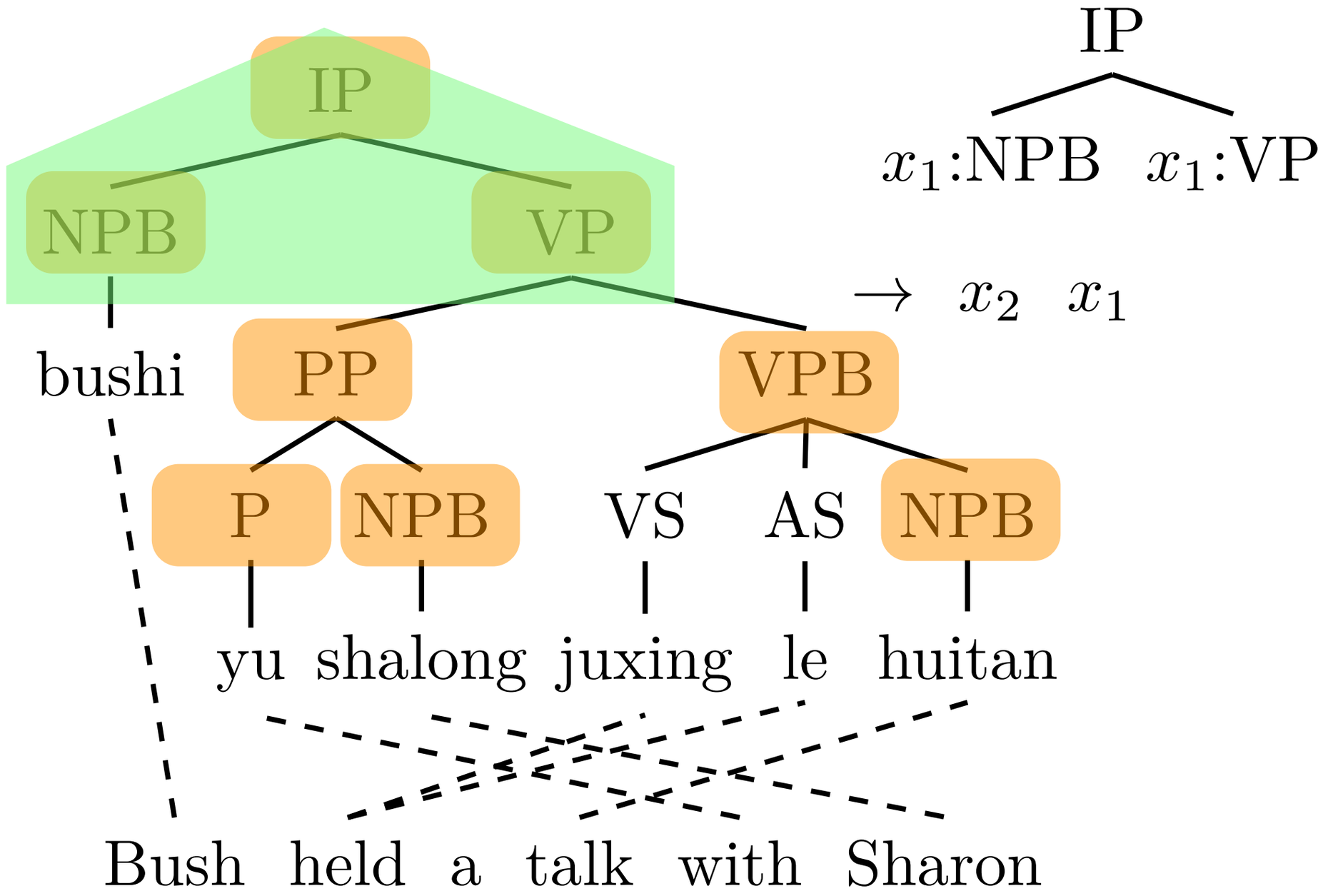
- Compute "frontiers": The nodes in which the intersection of "spans" and "complements" is empty

Rule Extraction



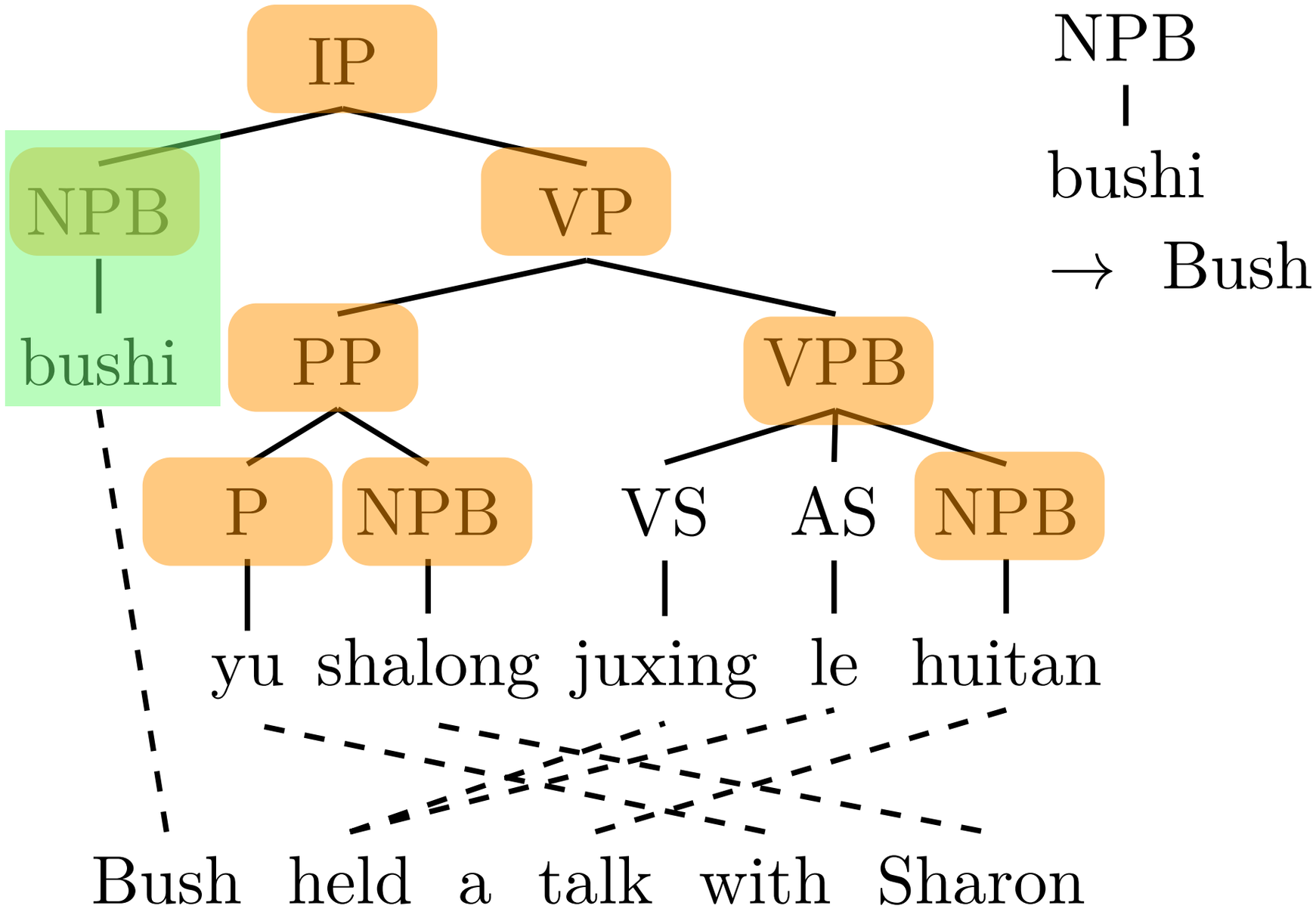
- Extract minimum rules using frontiers

Rule Extraction



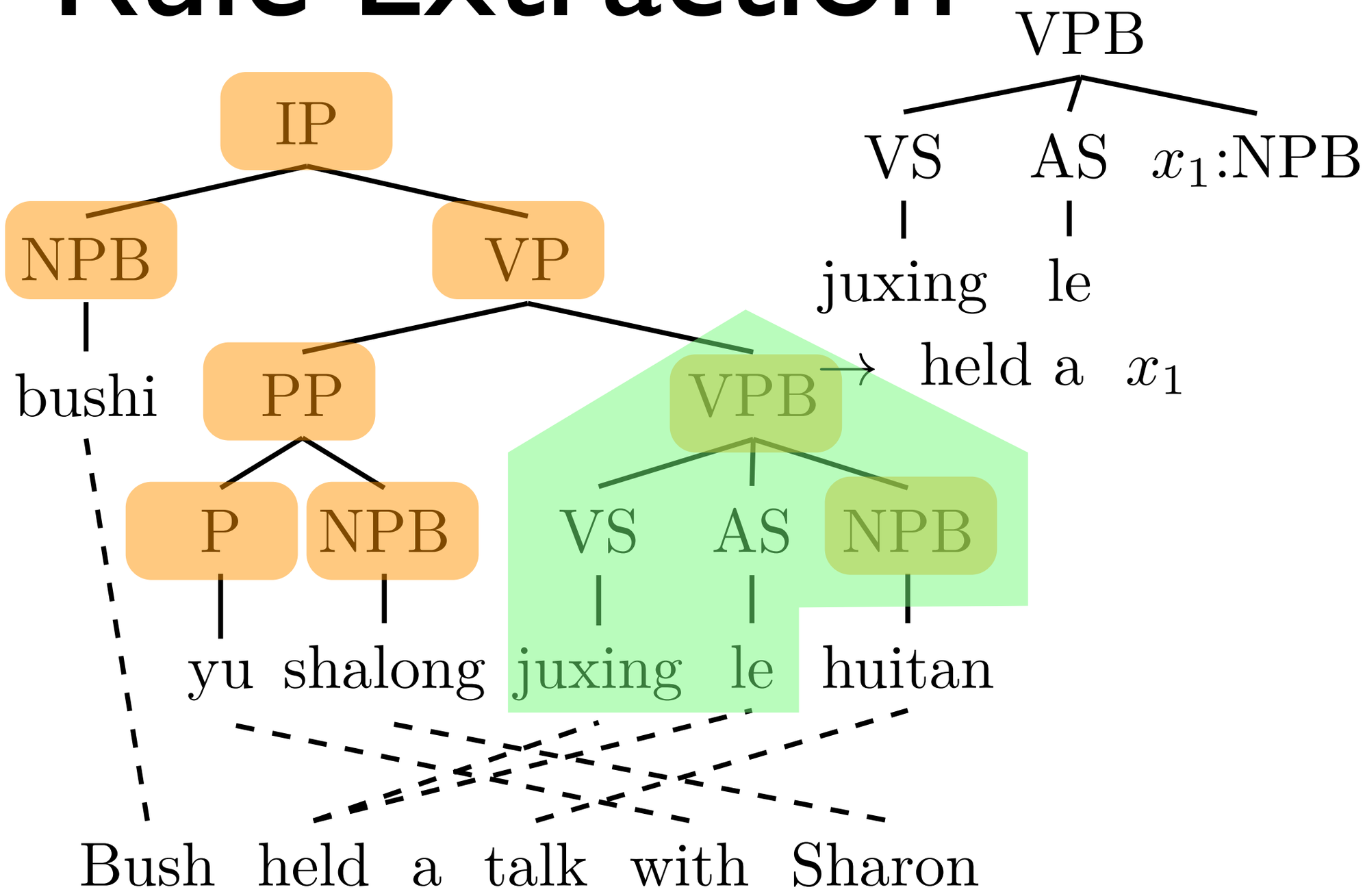
- Extract minimum rules using frontiers

Rule Extraction



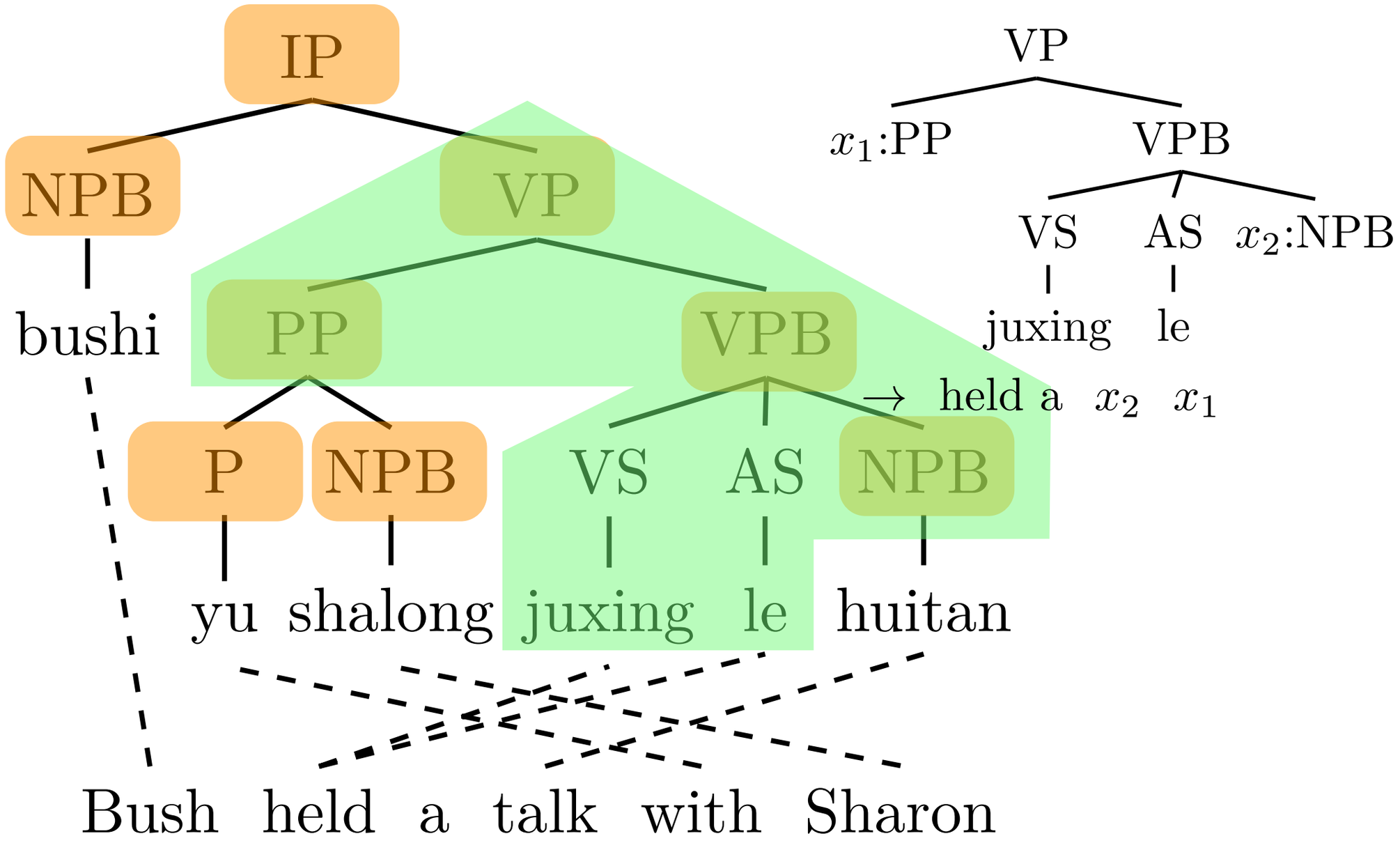
- Extract minimum rules using frontiers

Rule Extraction



- Extract minimum rules using frontiers

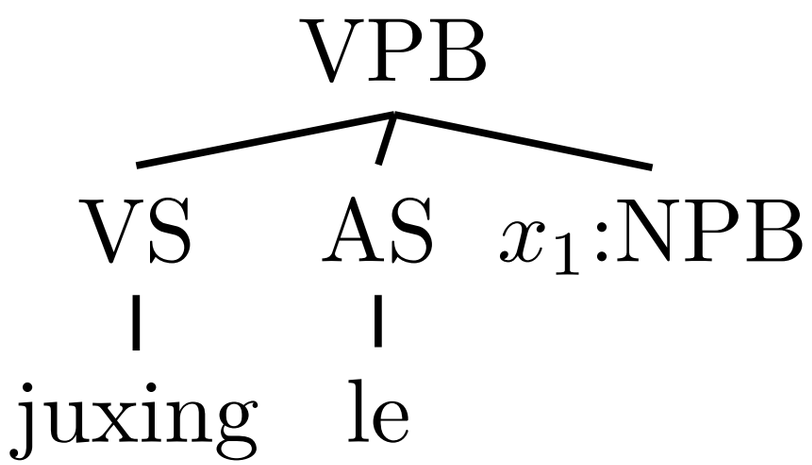
Rule Extraction



(Galley et al., 2006)

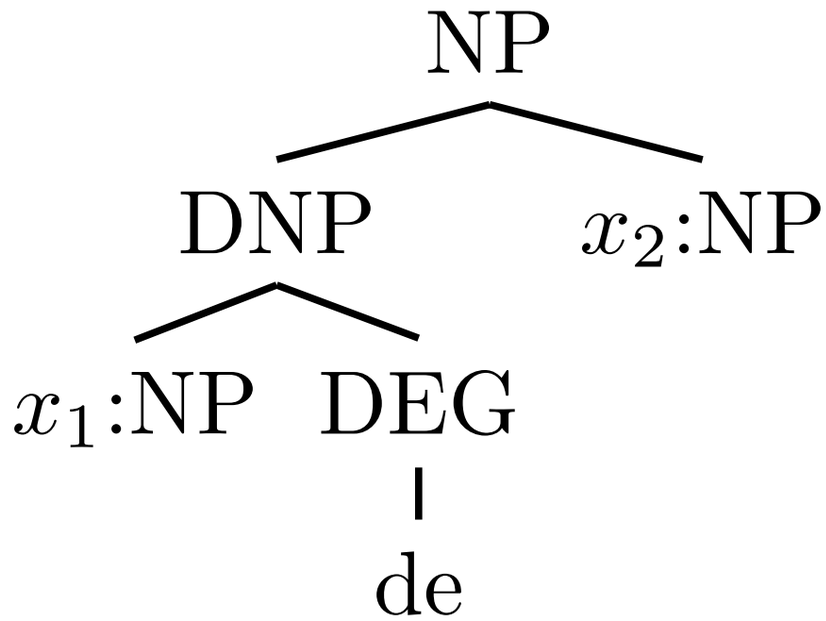
- Extract "compound rules" by combining minimum rules (i.e. longer phrases)

Decoding: String-{String, Tree}



$\rightarrow$ held a x_1

$\langle \text{VPB} \rightarrow \text{juxing le NPB}_1, \; x \rightarrow \text{hold a } x_1 \rangle$



$\rightarrow$ x_2 of x_1

$\langle \text{NP} \rightarrow \text{NP}_1 \text{ de NP}_2, \; x \rightarrow x_2 \text{ of } x_1 \rangle$

(Galley et al., 2004)

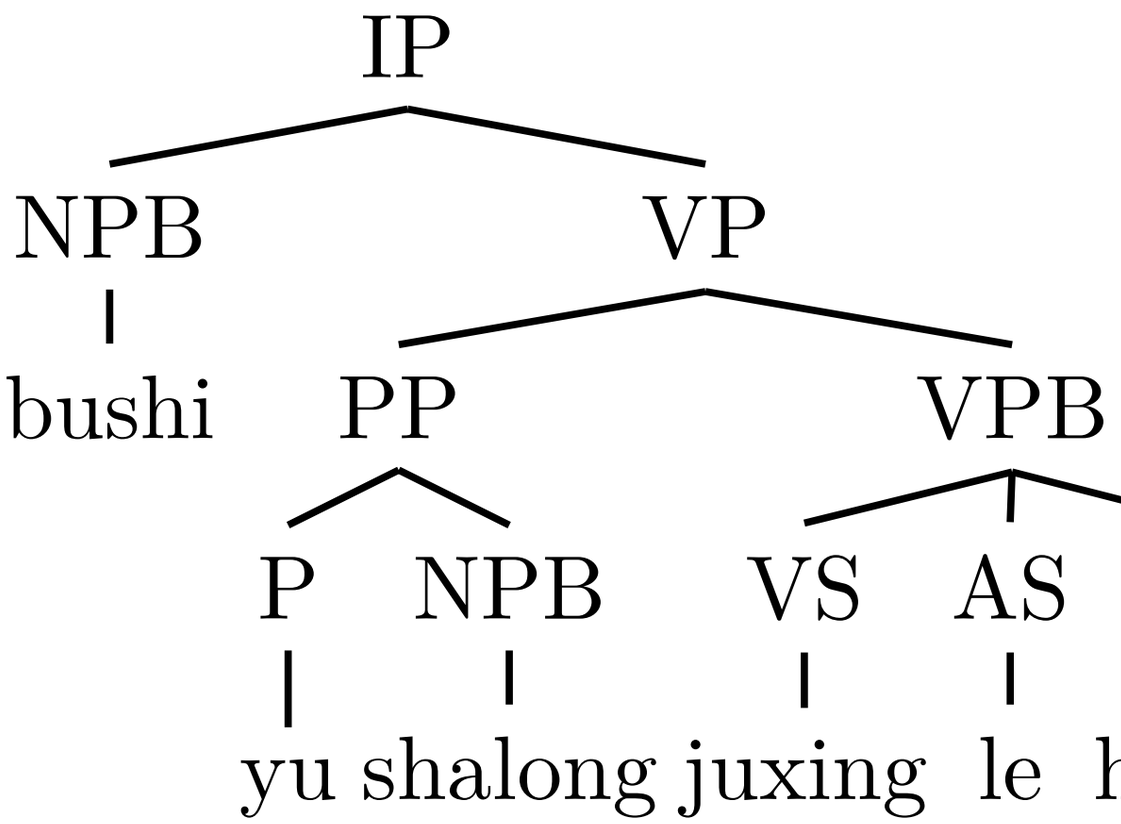
- Similar to SCFG decoding: Use the "collapsed" source side rule to perform CKY parsing
- Construct a translation forest using the target side

Decoding: Tree-{String, Tree}

(Huang et al., 2006)

- First, an input sentence is parsed
- Input tree is transformed into a translation forest by tree rewriting

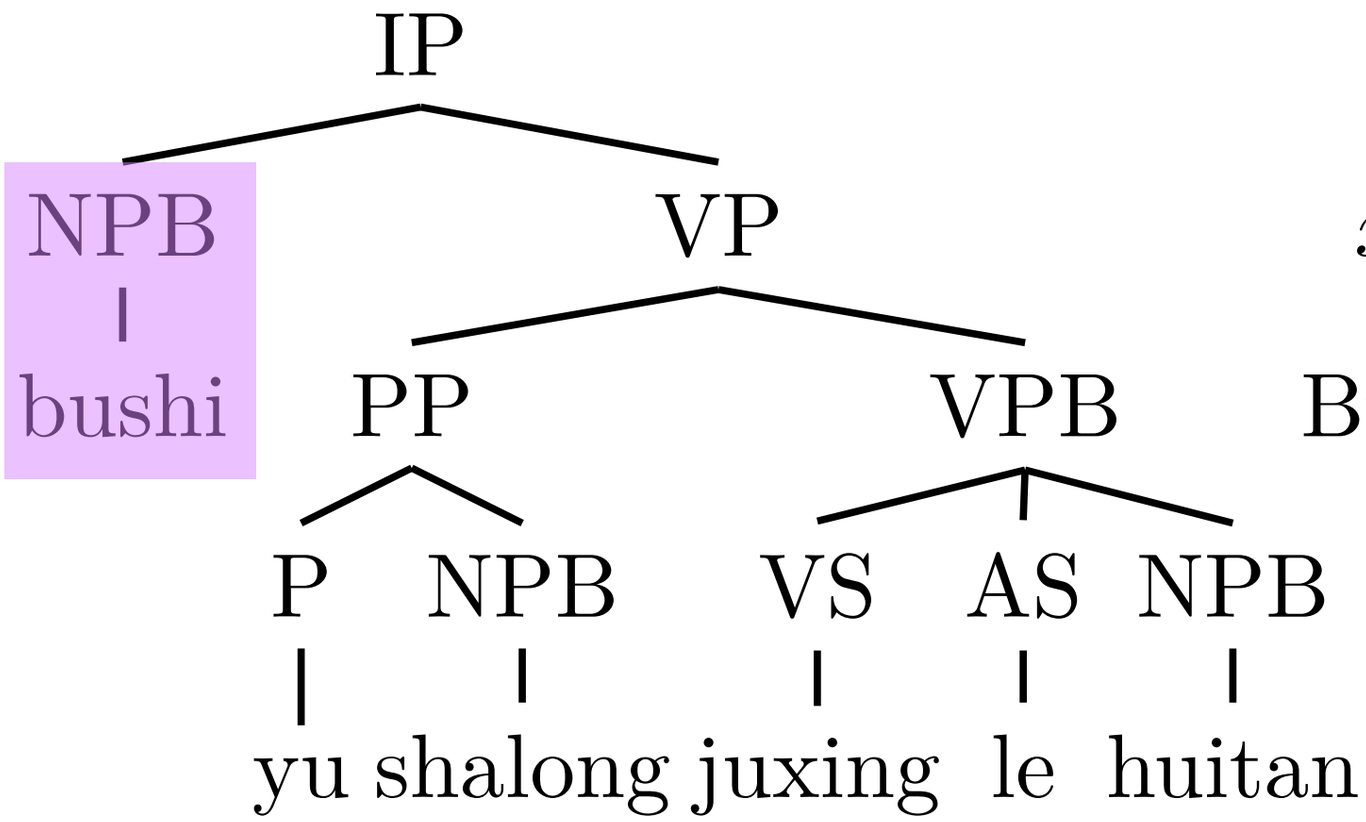
Decoding: Tree-{String, Tree}



(Huang et al., 2006)

- First, an input sentence is parsed
- Input tree is transformed into a translation forest by tree rewriting

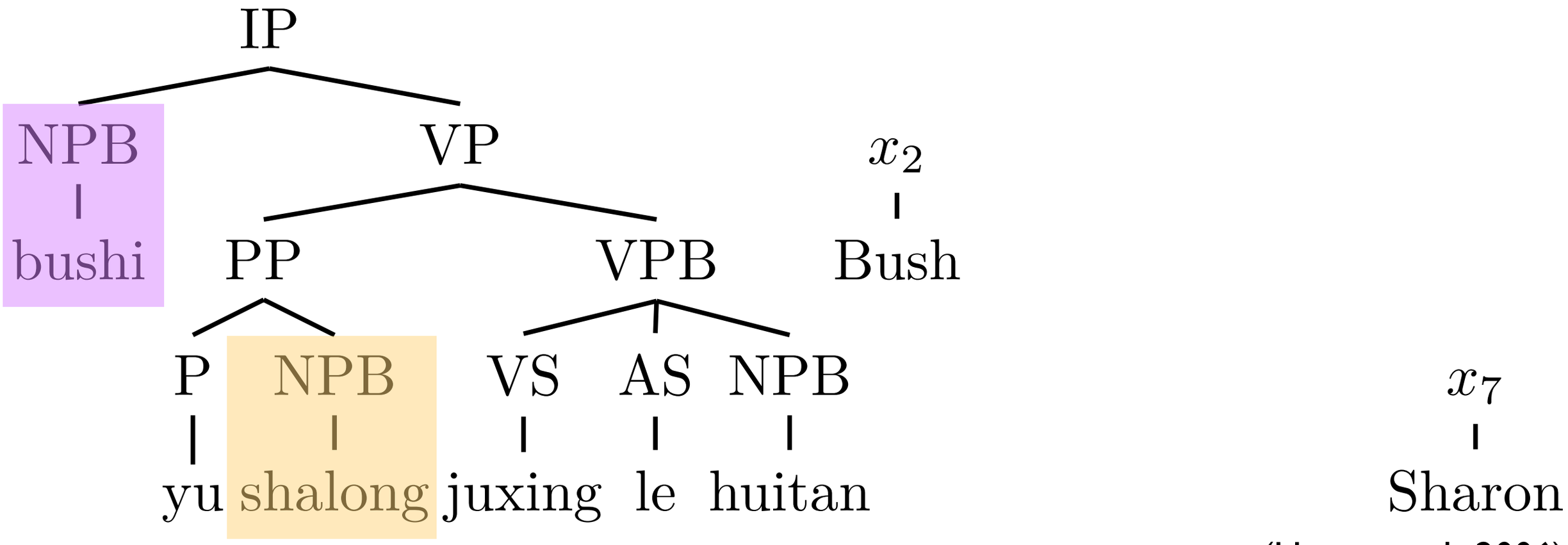
Decoding: Tree-{String, Tree}



(Huang et al., 2006)

- First, an input sentence is parsed
- Input tree is transformed into a translation forest by tree rewriting

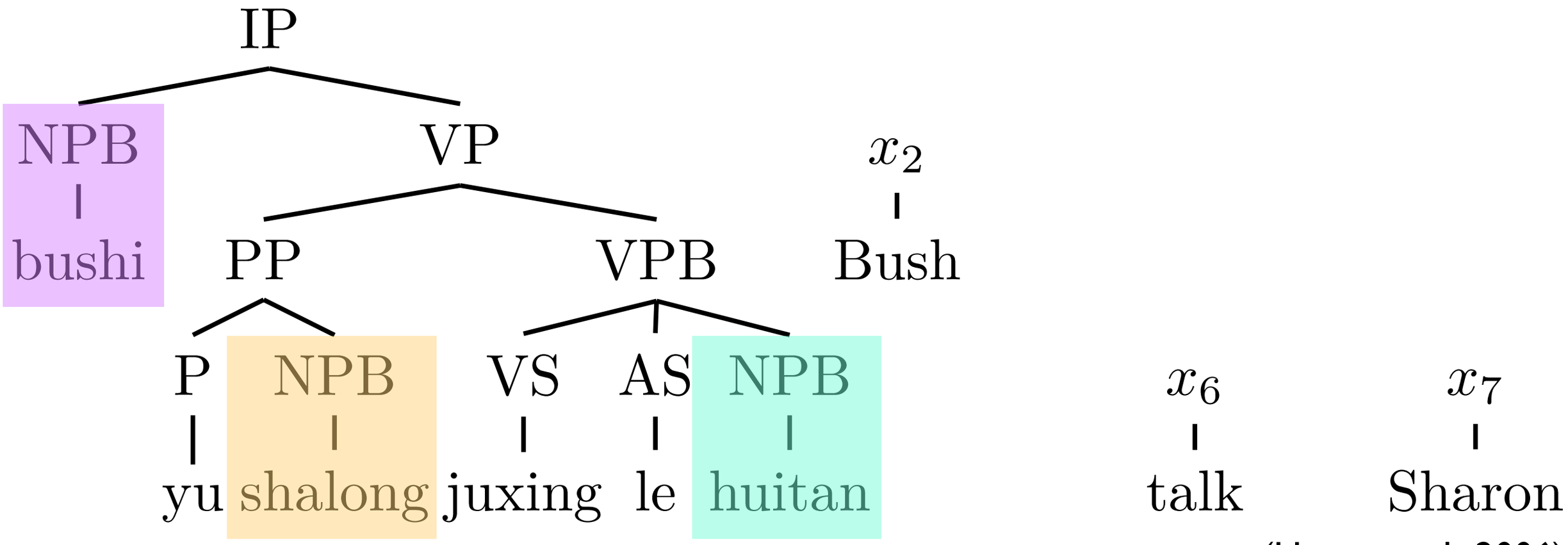
Decoding: Tree-{String, Tree}



(Huang et al., 2006)

- First, an input sentence is parsed
- Input tree is transformed into a translation forest by tree rewriting

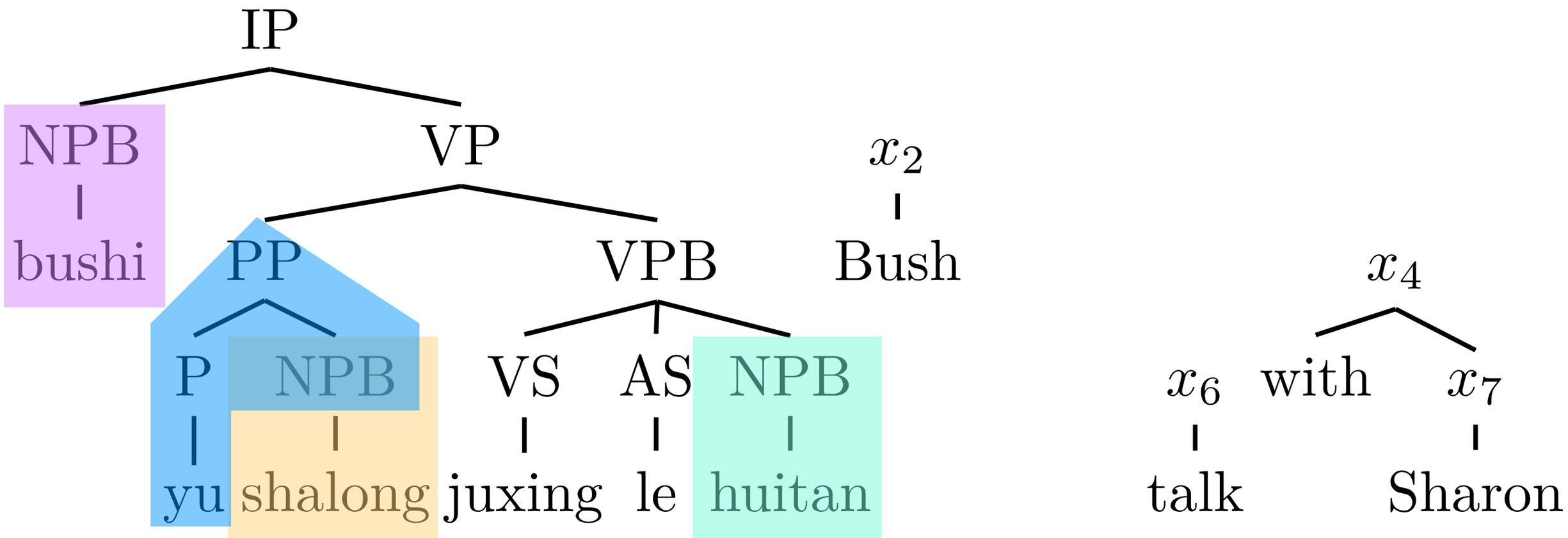
Decoding: Tree-{String, Tree}



(Huang et al., 2006)

- First, an input sentence is parsed
- Input tree is transformed into a translation forest by tree rewriting

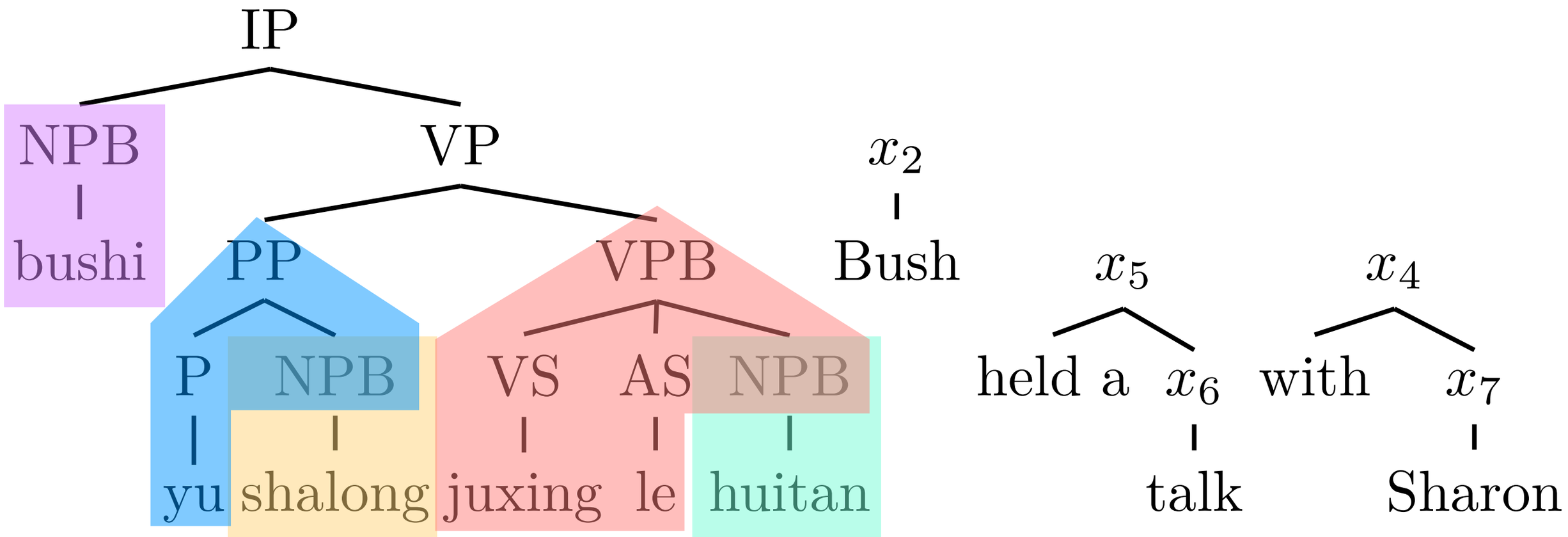
Decoding: Tree-{String, Tree}



(Huang et al., 2006)

- First, an input sentence is parsed
- Input tree is transformed into a translation forest by tree rewriting

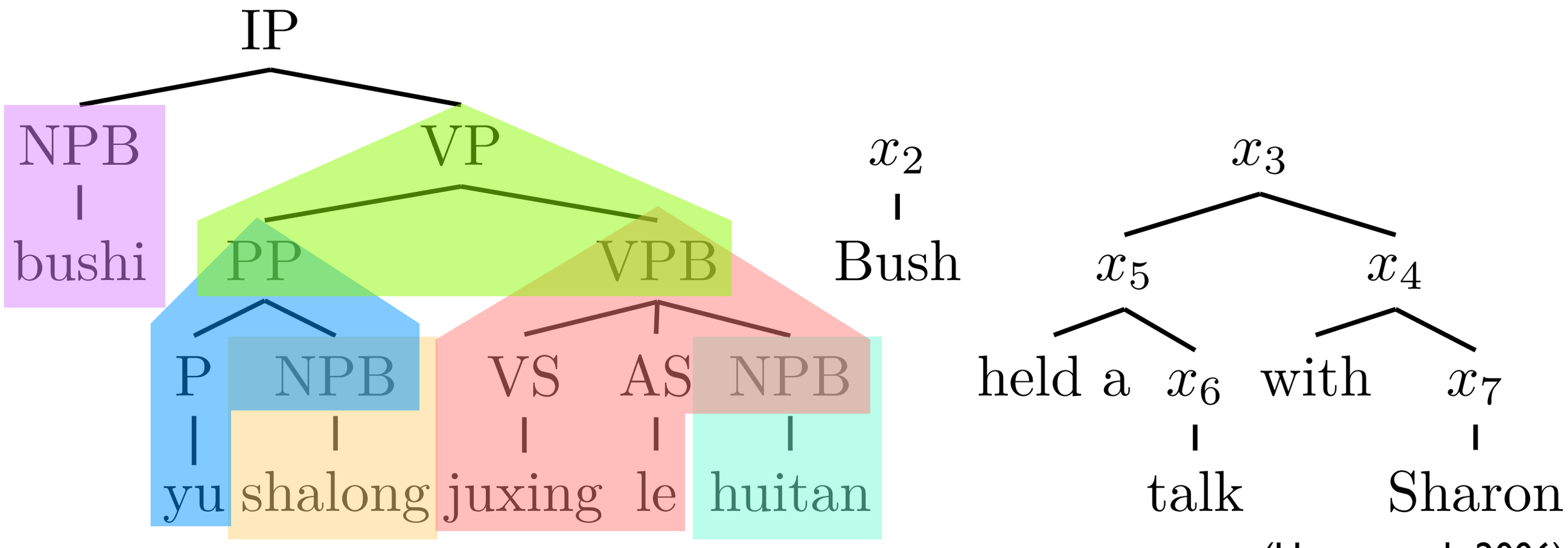
Decoding: Tree-{String, Tree}



(Huang et al., 2006)

- First, an input sentence is parsed
- Input tree is transformed into a translation forest by tree rewriting

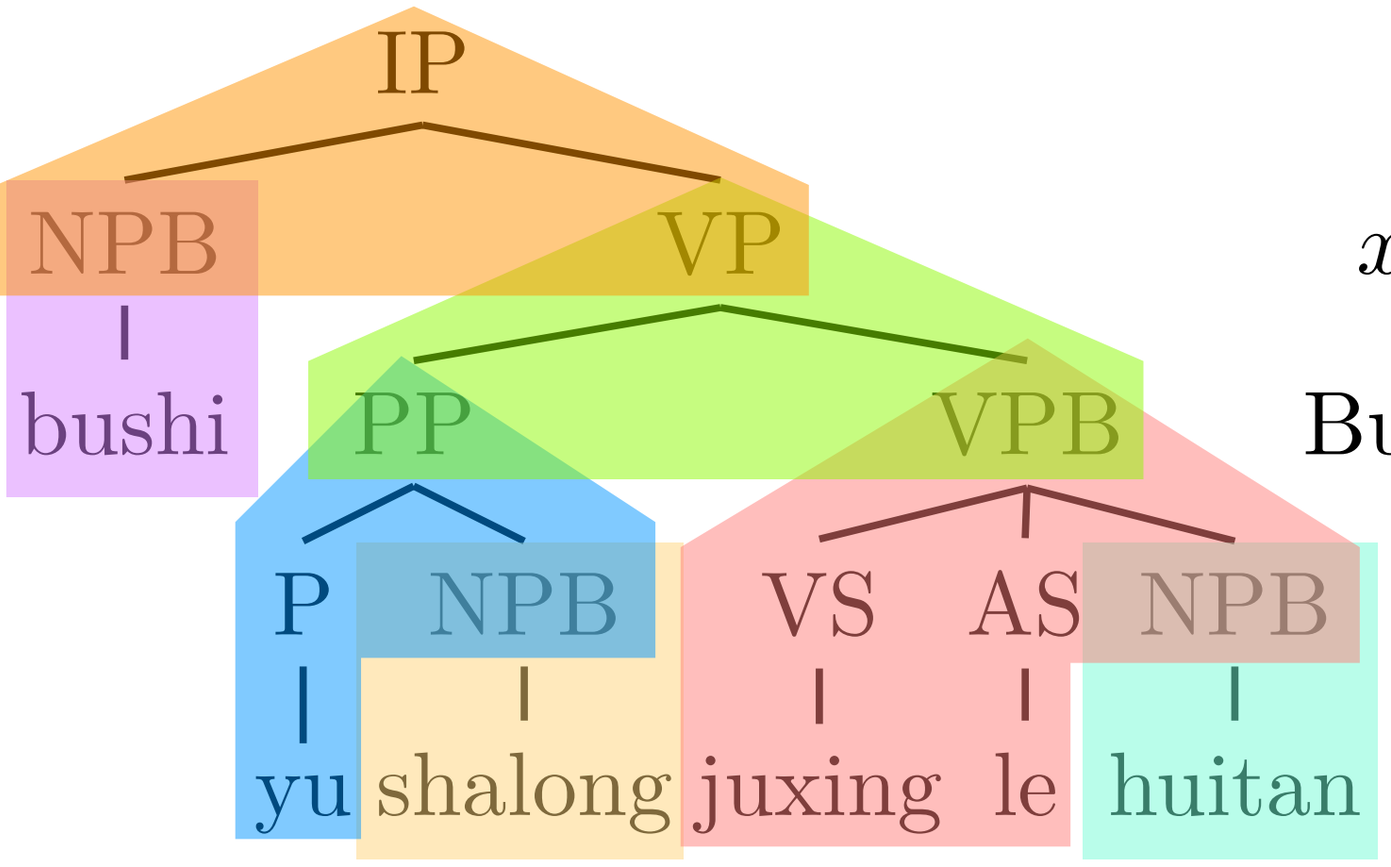
Decoding: Tree-{String, Tree}

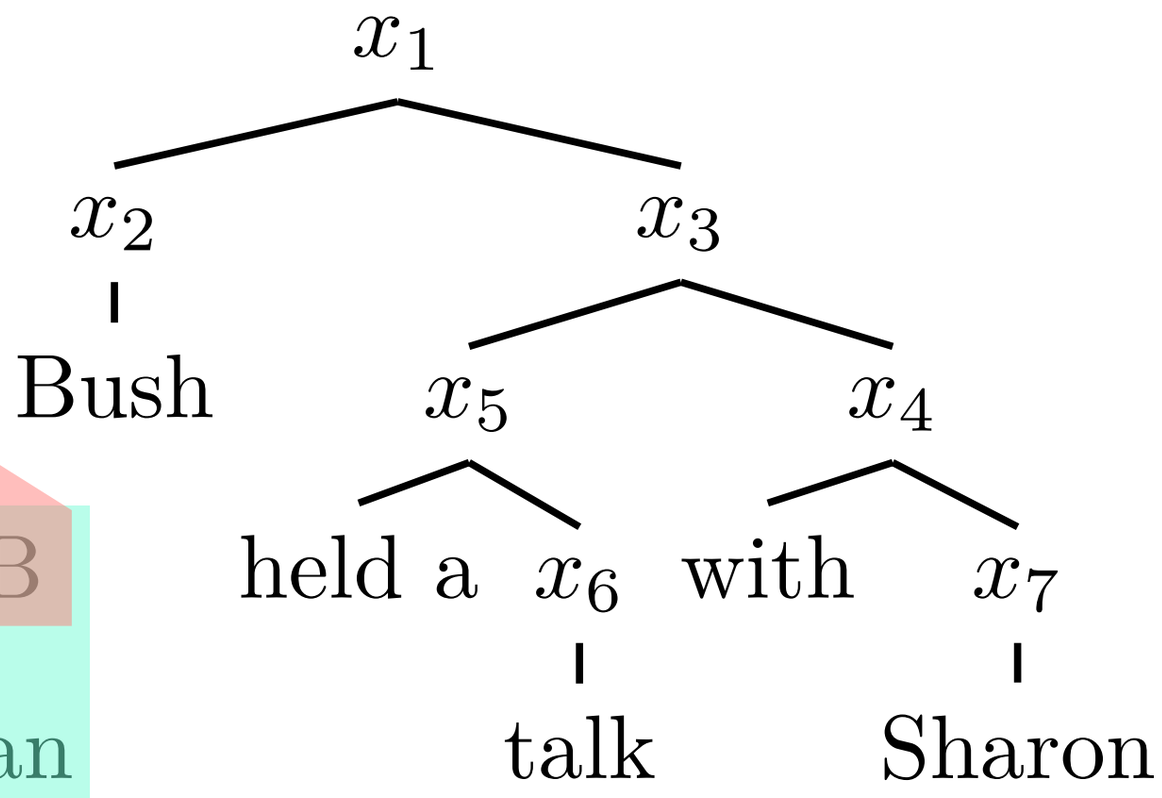


(Huang et al., 2006)

- First, an input sentence is parsed
- Input tree is transformed into a translation forest by tree rewriting

Decoding: Tree-{String, Tree}





(Huang et al., 2006)

- First, an input sentence is parsed
- Input tree is transformed into a translation forest by tree rewriting

Conclusion

- {String, Tree}-to-{String, Tree} translation models
- Rules extraction by GHKM (Galley et al., 2004)
  - Galley M, Hopkins M, Knight K, Marcu D, 2004
- Decoding:
  - String-to-{String, Tree} by CKY
  - Tree-to-{String, Tree} by tree-rewrite

More on Tree-based Models

- Forest-based approach: instead of 1-best parse, use forest encoding k-bests (Mi and Huang, 2008; Mi et al., 2008)
- "Binarized forest" as an alternative to represent multiple parses (Zhang et al., 2011)
- Fuzzy tree-to-tree as a way to overcome "stricktness" of tree-based models (Chiang, 2010)
- Use of dependency (Mi and Liu, 2010; Xie et al., 2011)

State-of-the-art (?)

		BLEU	
		dev	test
English-Chinese	pb	29.7	39.4
	hier	31.7	38.9
	bf2s	31.9	40.7**
English-Czech	*wmt best*	-	*15.4*
	pb	14.3	15.5
	hier	14.7	16.0
	bf2s	14.8	16.3*
English-French	*wmt best*	-	*27.6*
	pb	24.1	26.1
	hier	23.9	26.1
	bf2s	24.5	26.6**
English-German	*wmt best*	-	*16.3*
	pb	14.5	15.5
	hier	14.9	15.9
	bf2s	15.2	16.3**
English-Spanish	*wmt best*	-	*28.4*
	pb	24.1	27.9
	hier	24.2	28.4
	bf2s	24.9	28.9**

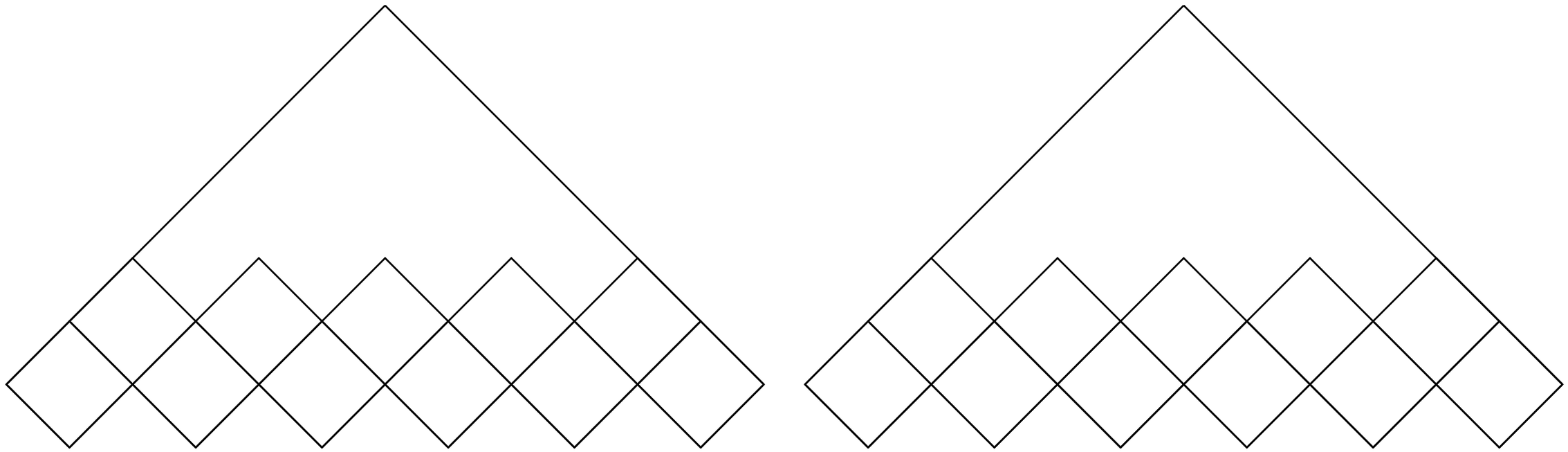
Overview

- Backgrounds
  - CFG, parsing, hypergraph, deductive system semirings
- Tree-based SMT
  - Synchronous-CFG
  - String-to-Tree, Tree-to-String
  - Bitext parsing

Overview

- Backgrounds
  - CFG, parsing, hypergraph, deductive system semirings
- Tree-based SMT
  - Synchronous-CFG
  - String-to-Tree, Tree-to-String
  - Bitext parsing

Bitext Parsing



- Given bitext (and a synchronous grammar), compute the best paired derivation
- Bitext parsing takes $O(N^3 M^3)$ for ITG (Wu, 1997)
  - For each length n and m, for each position i and j, for each rule X → LHS, for each split point k and l

ITG

$$
\begin{aligned}
\mathrm{X} &\rightarrow \langle \mathrm{X}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}}, \mathrm{X}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}} \rangle \\
\mathrm{X} &\rightarrow \langle \mathrm{X}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}}, \mathrm{X}_{\boxed{2}}\ \mathrm{X}_{\boxed{1}} \rangle \\
\mathrm{X} &\rightarrow \langle f, e \rangle
\end{aligned}
$$

- Inversion Transduction Grammar (ITG) (Wu, 1997), an instance of SCFG
- Frequently, we use an abbreviated form

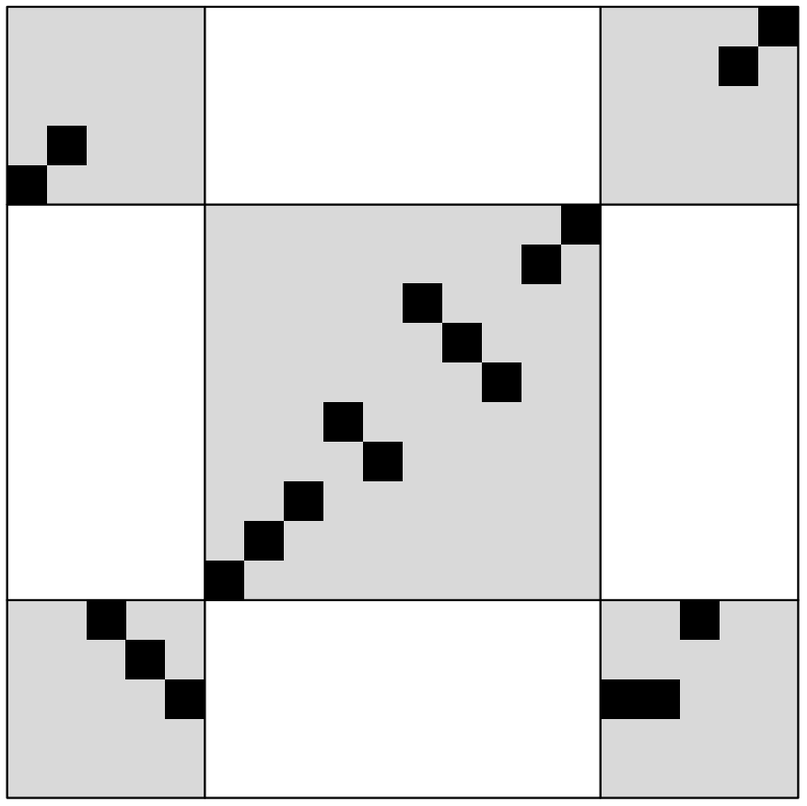
ITG

$$
\begin{aligned}
\mathrm{X} &\rightarrow \langle \mathrm{X}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}}, \mathrm{X}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}} \rangle \\
\mathrm{X} &\rightarrow \langle \mathrm{X}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}}, \mathrm{X}_{\boxed{2}}\ \mathrm{X}_{\boxed{1}} \rangle \\
\mathrm{X} &\rightarrow \langle f, e \rangle
\end{aligned}
$$

$$\mathrm{X} \rightarrow [X\ X] \mid \langle X\ X \rangle \mid f/e$$

- Inversion Transduction Grammar (ITG) (Wu, 1997), an instance of SCFG
- Frequently, we use an abbreviated form

Span Pruning



- You do not have to visit all the span pairs
- Use figure-of-merit to prune spans
  - $O(n^4)$ for a naive algorithm (Zhang and Gildea, 2005)
  - $O(n^3)$ for a DP-based algorithm (Zhang et al., 2008)

Beam Pruning

- Re-organize the search space by the cardinality (Saers et al., 2009)
  - Cardinality = # of source/target words parsed
- Prune by the cardinality: Complexity $O(bn^3)$

Beam Pruning

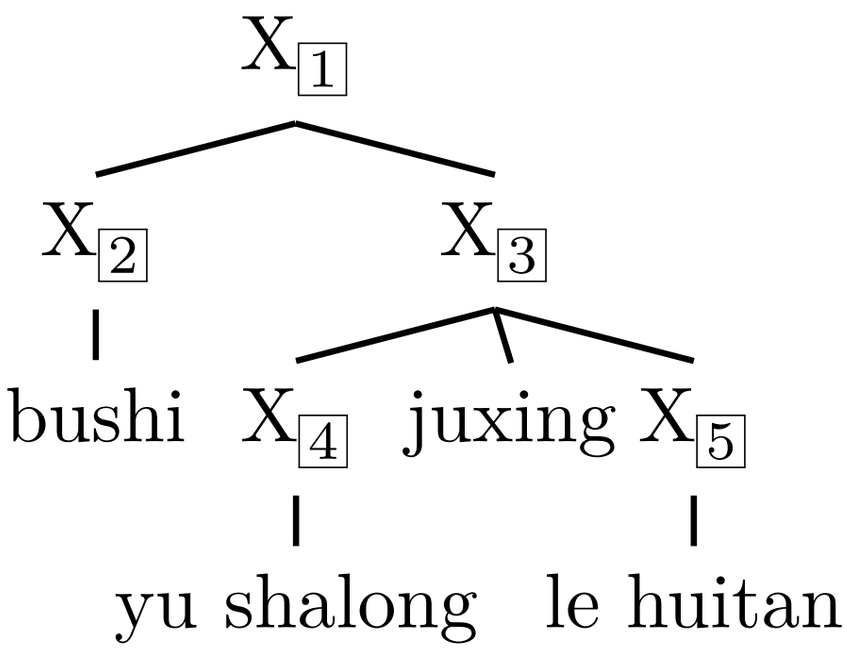
--●---,------	--●---,--●---	--●---,--●●--	--●●--,--●●--
-●----,------	-●----,--●---	-●----,--●●--	●●----,--●●--
------,----●-	---●--,----●-	---●●-,----●-	---●●-,---●●-
------,---●--	------,--●●--	---●--,--●●--	---●●-,--●●--
------,-●----	---●--,-●----	--●●--,-●----	--●●--,●●----

- Re-organize the search space by the cardinality (Saers et al., 2009)
  - Cardinality = # of source/target words parsed
- Prune by the cardinality: Complexity $O(bn^3)$

Two Parse

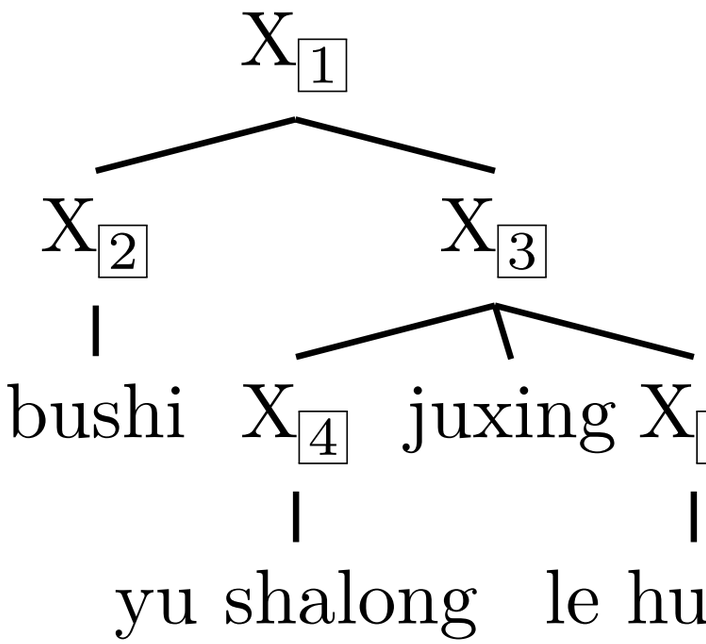
- In practice, we do not have to enumerate all the rules for the Hiero-like grammar (Dyer, 2010)
  - First, parsing using the source side rules
  - Then, parse the target by the "instantiated" rules

Two Parse



- In practice, we do not have to enumerate all the rules for the Hiero-like grammar (Dyer, 2010)
  - First, parsing using the source side rules
  - Then, parse the target by the "instantiated" rules

Two Parse

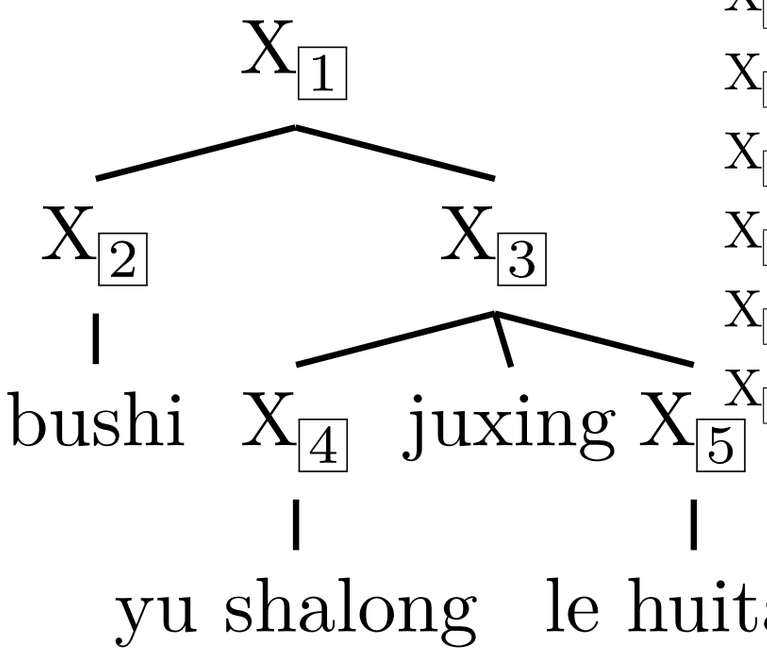


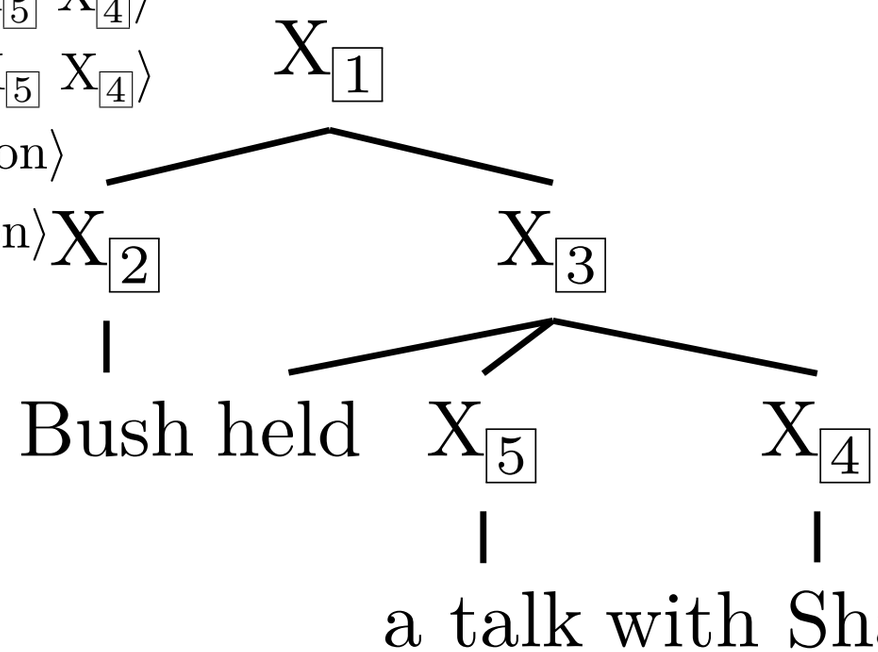
$$
\begin{aligned}
X_{\boxed{1}} &\rightarrow \langle X_{\boxed{2}}\ X_{\boxed{3}}, X_{\boxed{2}}\ X_{\boxed{3}} \rangle \\
X_{\boxed{1}} &\rightarrow \langle X_{\boxed{2}}\ X_{\boxed{3}}, X_{\boxed{3}}\ X_{\boxed{2}} \rangle \\
X_{\boxed{2}} &\rightarrow \langle \text{bushi}, \text{Bush} \rangle \\
X_{\boxed{3}} &\rightarrow \langle X_{\boxed{4}}\ \text{juxing}\ X_{\boxed{5}}, \text{held}\ X_{\boxed{5}}\ X_{\boxed{4}} \rangle \\
X_{\boxed{3}} &\rightarrow \langle X_{\boxed{4}}\ \text{juxing}\ X_{\boxed{5}}, \text{hold}\ X_{\boxed{5}}\ X_{\boxed{4}} \rangle \\
X_{\boxed{4}} &\rightarrow \langle \text{yu shalong}, \text{with Sharon} \rangle \\
X_{\boxed{4}} &\rightarrow \langle \text{yu shalong}, \text{and Sharon} \rangle \\
X_{\boxed{5}} &\rightarrow \langle \text{le huitan}, \text{a talk} \rangle \\
X_{\boxed{5}} &\rightarrow \langle \text{le huitan}, \text{talked} \rangle
\end{aligned}
$$

- In practice, we do not have to enumerate all the rules for the Hiero-like grammar (Dyer, 2010)
  - First, parsing using the source side rules
  - Then, parse the target by the "instantiated" rules

Two Parse

$$
\begin{aligned}
X_{\boxed{1}} &\rightarrow \langle X_{\boxed{2}}\ X_{\boxed{3}}, X_{\boxed{2}}\ X_{\boxed{3}} \rangle \\
X_{\boxed{1}} &\rightarrow \langle X_{\boxed{2}}\ X_{\boxed{3}}, X_{\boxed{3}}\ X_{\boxed{2}} \rangle \\
X_{\boxed{2}} &\rightarrow \langle \text{bushi}, \text{Bush} \rangle \\
X_{\boxed{3}} &\rightarrow \langle X_{\boxed{4}}\ \text{juxing}\ X_{\boxed{5}}, \text{held}\ X_{\boxed{5}}\ X_{\boxed{4}} \rangle \\
X_{\boxed{3}} &\rightarrow \langle X_{\boxed{4}}\ \text{juxing}\ X_{\boxed{5}}, \text{hold}\ X_{\boxed{5}}\ X_{\boxed{4}} \rangle \\
X_{\boxed{4}} &\rightarrow \langle \text{yu shalong}, \text{with Sharon} \rangle \\
X_{\boxed{4}} &\rightarrow \langle \text{yu shalong}, \text{and Sharon} \rangle \\
X_{\boxed{5}} &\rightarrow \langle \text{le huitan}, \text{a talk} \rangle \\
X_{\boxed{5}} &\rightarrow \langle \text{le huitan}, \text{talked} \rangle
\end{aligned}
$$



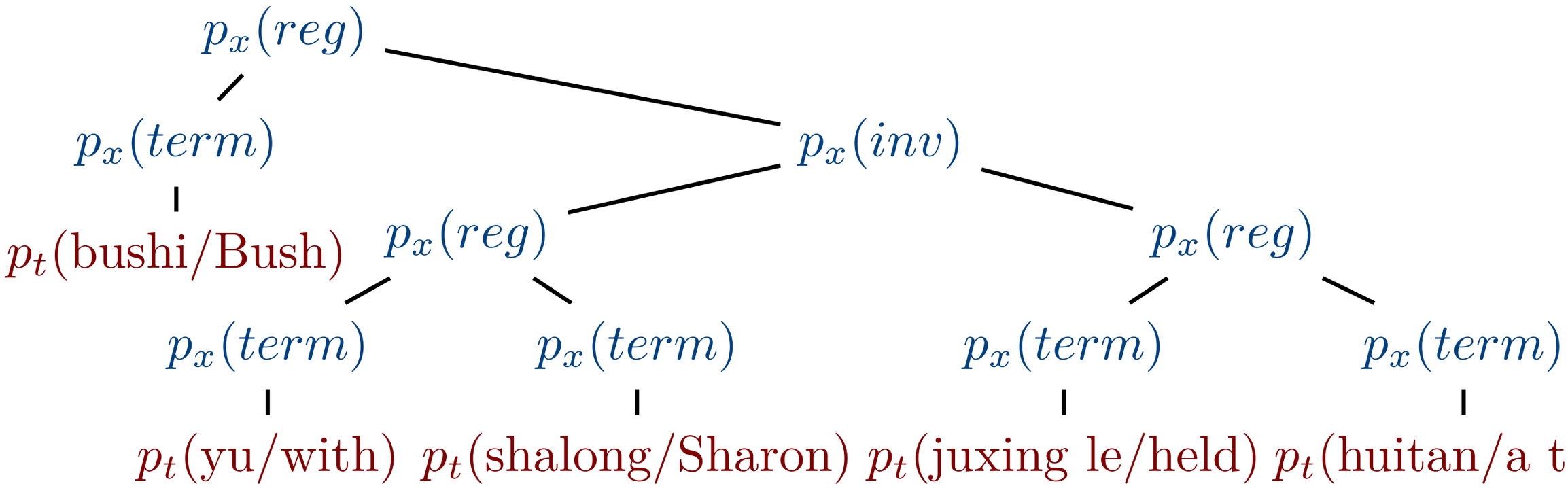


- In practice, we do not have to enumerate all the rules for the Hiero-like grammar (Dyer, 2010)
  - First, parsing using the source side rules
  - Then, parse the target by the "instantiated" rules

ITG for Phrase Induction

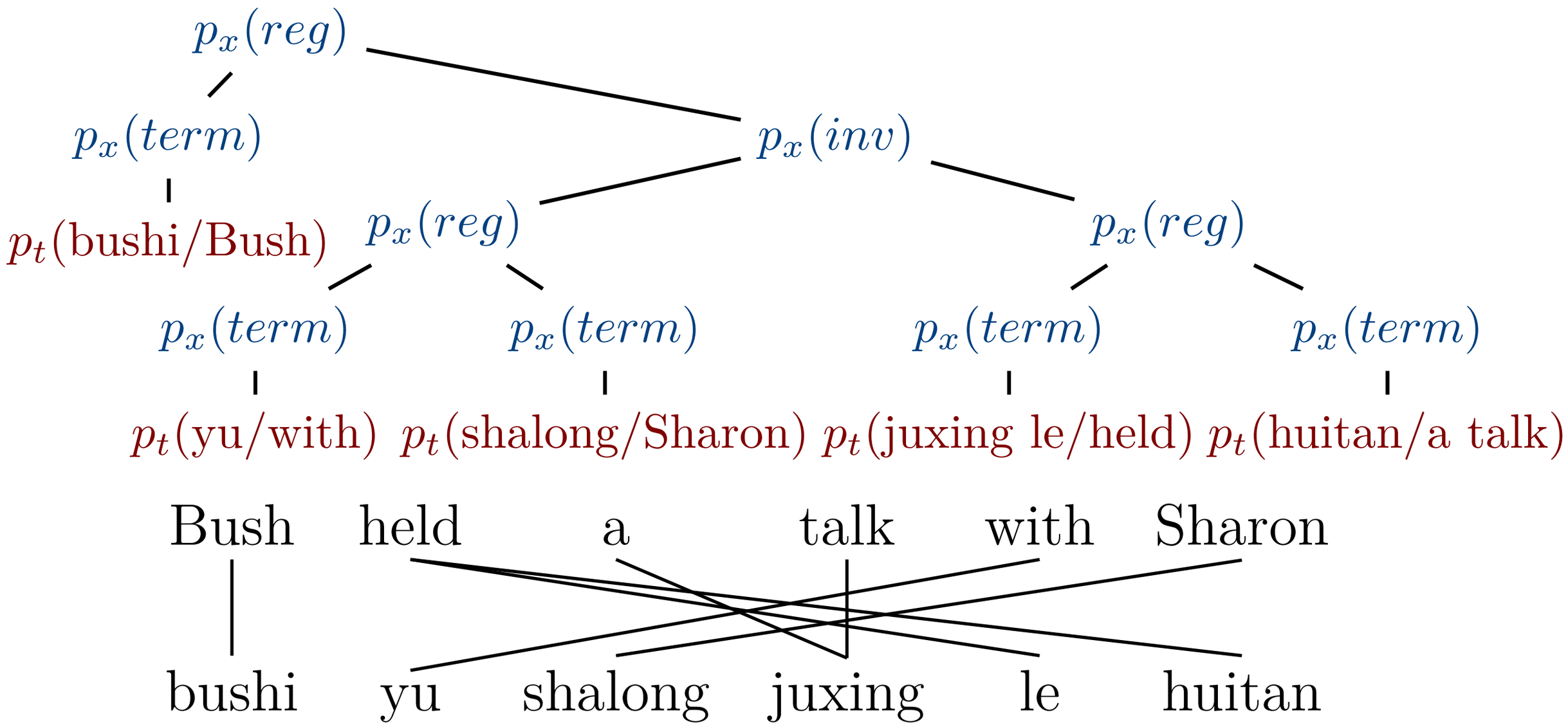
- Phrasal alignment by ITG (Cherry and Lin, 2007)

ITG for Phrase Induction



- Phrasal alignment by ITG (Cherry and Lin, 2007)

ITG for Phrase Induction

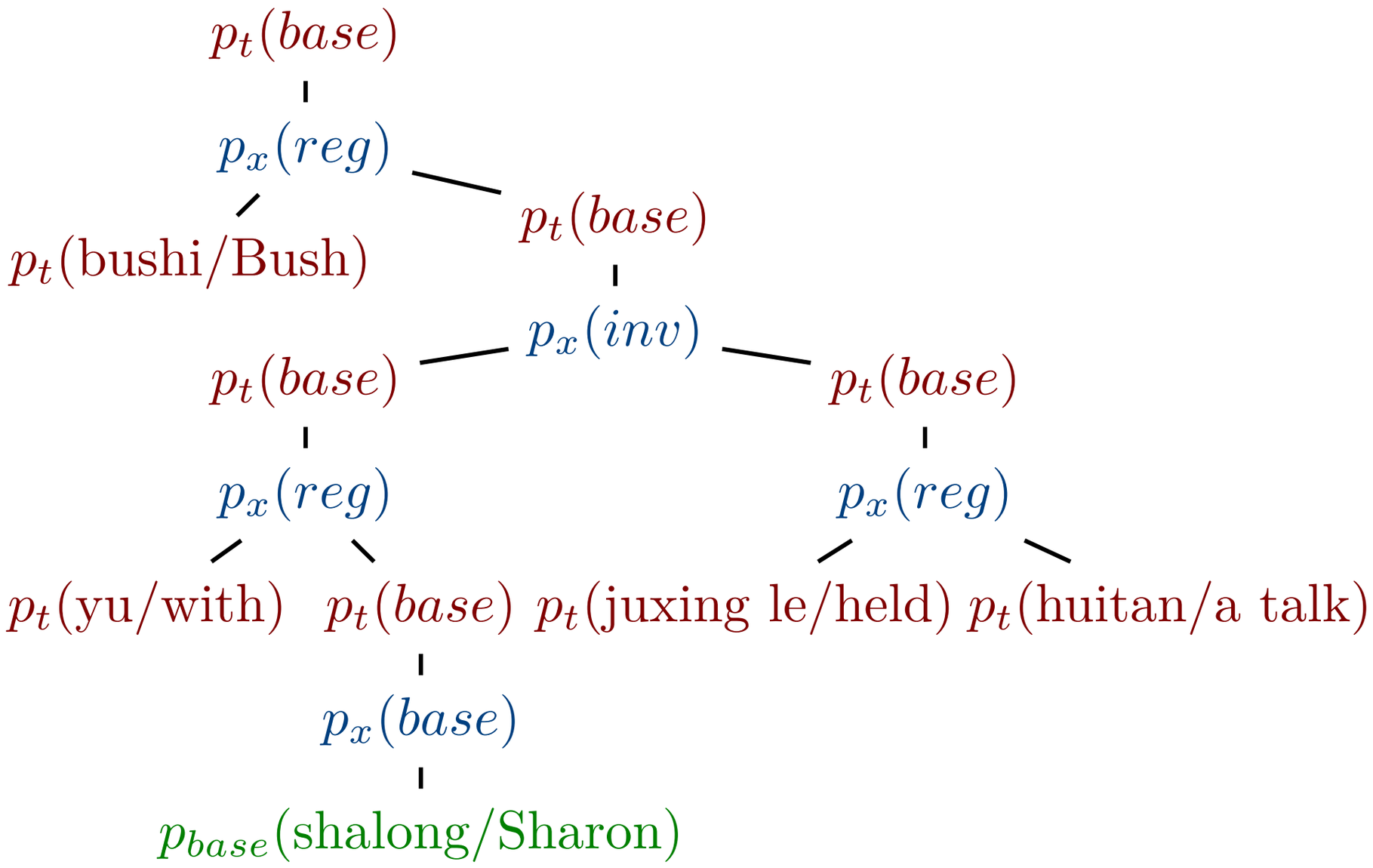


- Phrasal alignment by ITG (Cherry and Lin, 2007)

Exhaustive ITG Phrases

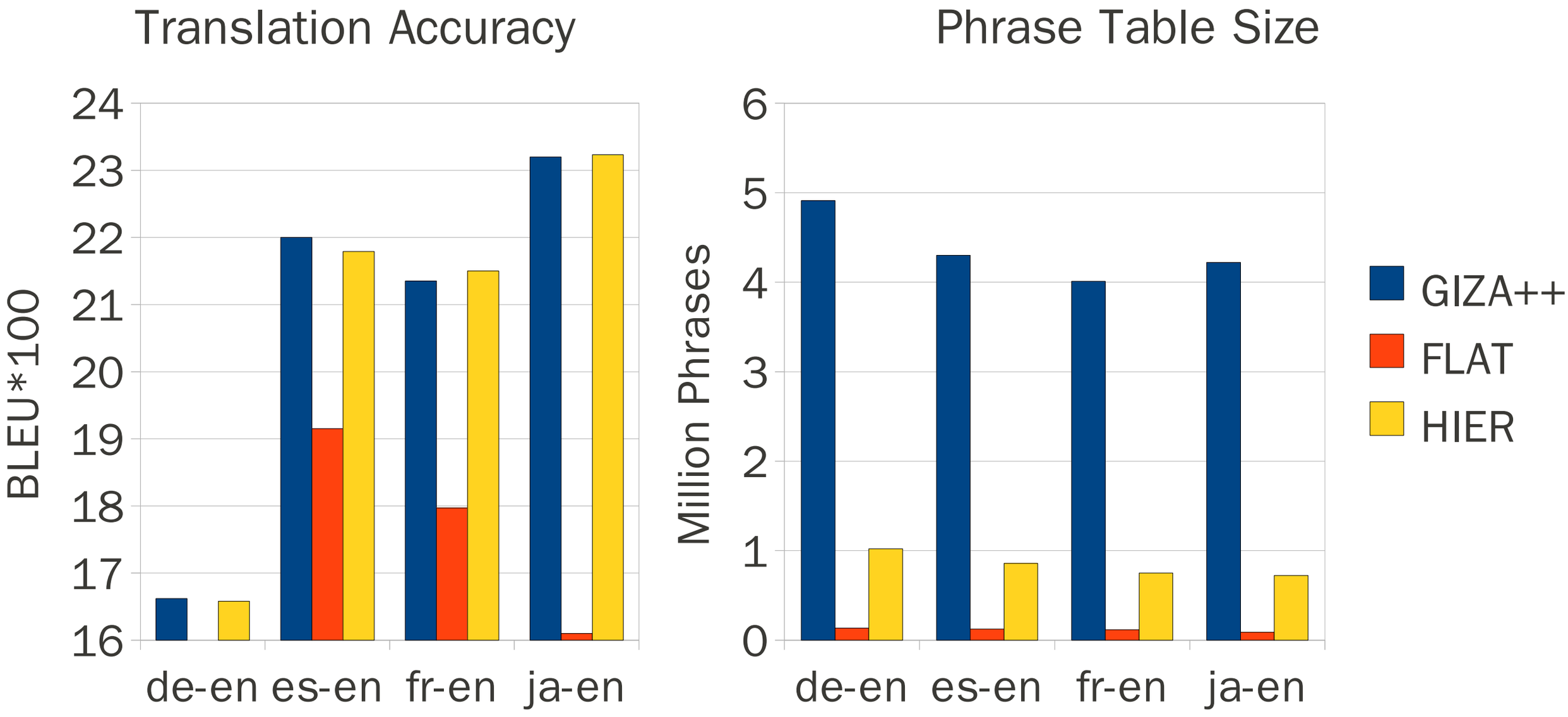
- Recursively divide-and-conquer
- Multiple granularities are included in the model (Neubig et al., 2011)

Exhaustive ITG Phrases



- Recursively divide-and-conquer
- Multiple granularities are included in the model (Neubig et al., 2011)

Compact Model



(Neubig et al., 2011)

Conclusion

- Bitext parsing with ITG
- Span pruning, beam pruning, two-parse
- ITG is a simple, yet powerful grammar for bilingual knowledge induction

Summary

- Backgrounds
  - CFG, parsing, hypergraph, deductive system semirings
- Tree-based SMT
  - Synchronous-CFG
  - String-to-Tree, Tree-to-String
  - Bitext parsing

Implementations

- synchronous-CFG
  - Cdec: http://cdec-decoder.org
  - Jane: http://www-i6.informatik.rwth-aachen.de/jane/
  - Joshua: http://joshua.sourceforge.net
  - Moses: http://www.statmt.org/moses/
- {Tree,String}-to-{Tree,String}: Usually, closed source, like my private implementation:-)
- ITG-model induction
  - Pialign: http://www.phontron.com/pialign/

References

- P. F. Brown, J. Cocke, S. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, ``A statistical approach to machine translation,'' *Computational Linguistics*, vol. 16, no. 2, pp. 79--85, 1990.

- P. Koehn, F. J. Och, and D. Marcu, ``Statistical pharse-based translation,'' in *Proc. of HLT-NAACL 2003*, (Edmonton), pp. 48--54, May-June 2003.

- D. Chiang, ``Hierarchical phrase-based translation,'' *Comput. Linguist.*, vol. 33, no. 2, pp. 201--228, 2007.

- M. Galley, M. Hopkins, K. Knight, and D. Marcu, ``What's in a translation rule?,'' in *HLT-NAACL 2004: Main Proceedings* (D. M. Susan Dumais and S. Roukos, eds.), (Boston, Massachusetts, USA), pp. 273--280, Association for Computational Linguistics, May 2 - May 7 2004.

- A. Zollmann and A. Venugopal, ``Syntax augmented machine translation via chart parsing,'' in *StatMT '06: Proceedings of the Workshop on Statistical Machine Translation*, (Morristown, NJ, USA), pp. 138--141, Association for Computational Linguistics, 2006.

- L. Huang, K. Knight, and A. Joshi, ``Statistical syntax-directed translation with extended domain of locality,'' in *In Proc. AMTA 2006*, pp. 66--73, 2006.

References

- Y. . Liu, Q. Liu, and S. Lin, ``Tree-to-string alignment template for statistical machine translation," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, (Sydney, Australia), pp. 609--616, Association for Computational Linguistics, July 2006.

- C. Quirk, A. Menezes, and C. Cherry, ``Dependency treelet translation: syntactically informed phrasal smt," in *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, (Morristown, NJ, USA), pp. 271--279, Association for Computational Linguistics, 2005.

- L. Shen, J. Xu, and R. Weischedel, ``A new string-to-dependency machine translation algorithm with a target dependency language model," in *Proceedings of ACL-08: HLT*, (Columbus, Ohio), pp. 577--585, Association for Computational Linguistics, June 2008.

- Y. Ding and M. Palmer, ``Machine translation using probabilistic synchronous dependency insertion grammars," in *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, (Morristown, NJ, USA), pp. 541--548, Association for Computational Linguistics, 2005.

- Y. Liu, Y. Lu ¨, and Q. Liu, ``Improving tree-to-tree translation with packed forests," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, (Suntec, Singapore), pp. 558--566, Association for Computational Linguistics, August 2009.

- D. Klein and C. D. Manning, ``Parsing and hypergraphs," in *In IWPT*, pp. 123--134, 2001.

References

- S. M. Shieber, Y. Schabes, and O. C. N. Pereira, ``Principles and implementation of deductive parsing,'' *Journal of Logic Programming*, 1995.

- L. Huang and D. Chiang, ``Better k-best parsing,'' in *Proceedings of the Ninth InternationalWorkshop on ParsingTechnology*, (Vancouver, British Columbia), pp. 53--64, Association for Computational Linguistics, October 2005.

- L. Huang and D. Chiang, ``Forest rescoring: Faster decoding with integrated language models,'' in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, (Prague, Czech Republic), pp. 144--151, Association for Computational Linguistics, June 2007.

- A. Gesmundo and J. Henderson, ``Faster Cube Pruning,'' in *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)* (M. Federico, I. Lane, M. Paul, and F. Yvon, eds.), pp. 267--274, 2010.

- M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thayer, ``Scalable inference and training of context-rich syntactic translation models,'' in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, (Sydney, Australia), pp. 961-968, Association for Computational Linguistics, July 2006.

- H. Mi and L. Huang, ``Forest-based translation rule extraction,'' in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, (Honolulu, Hawaii), pp. 206--214, Association for Computational Linguistics, October 2008.

- H. Mi, L. Huang, and Q. Liu, ``Forest-based translation,'' in *Proceedings of ACL-08: HLT*, (Columbus, Ohio), pp. 192--199, Association for Computational Linguistics, June 2008.

References

- H. Zhang, L. Fang, P. Xu, and X. Wu, ``Binarized forest to string translation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, (Portland, Oregon, USA), pp. 835--845, Association for Computational Linguistics, June 2011.

- D. Chiang, ``Learning to translate with source and target syntax," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, (Uppsala, Sweden), pp. 1443--1452, Association for Computational Linguistics, July 2010.

- J. Xie, H. Mi, and Q. Liu, ``A novel dependency-to-string model for statistical machine translation," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, (Edinburgh, Scotland, UK.), pp. 216-226, Association for Computational Linguistics, July 2011.

- D. Wu, ``Stochastic inversion transduction grammars and bilingual parsing of parallel corpora," *Comput. Linguist.*, vol. 23, no. 3, pp. 377--403, 1997.

- H. Zhang and D. Gildea, ``Stochastic lexicalized inversion transduction grammar for alignment," in *Proceedings of the43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, (Stroudsburg, PA, USA), pp. 475--482, Association for Computational Linguistics, 2005.

- H. Zhang, C. Quirk, R. C. Moore, and D. Gildea, ``Bayesian learning of non-compositional phrases with synchronous parsing," in *Proceedings of ACL-08: HLT*, (Columbus, Ohio), pp. 97--105, Association for Computational Linguistics, June 2008.

References

- M. Saers, J. Nivre, and D. Wu, ``Learning stochastic bracketing inversion transduction grammars with a cubic time biparsing algorithm," in *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, (Paris, France), pp. 29--32, Association for Computational Linguistics, October 2009.

- C. Dyer, ``Two monolingual parses are better than one (synchronous parse)," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (Los Angeles, California), pp. 263--266, Association for Computational Linguistics, June 2010.

- C. Cherry and D. Lin, ``Inversion transduction grammar for joint phrasal translation modeling," in *Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation*, (Rochester, New York), pp. 17--24, Association for Computational Linguistics, April 2007.

- G. Neubig, T. Watanabe, E. Sumita, S. Mori, and T. Kawahara, ``An unsupervised model for joint phrase alignment and extraction," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, (Portland, Oregon, USA), pp. 632-641, Association for Computational Linguistics, June 2011.