Hierarchical and Syntax-Based Translation

David Chiang

UNIVERSITY OF NOTRE DAME

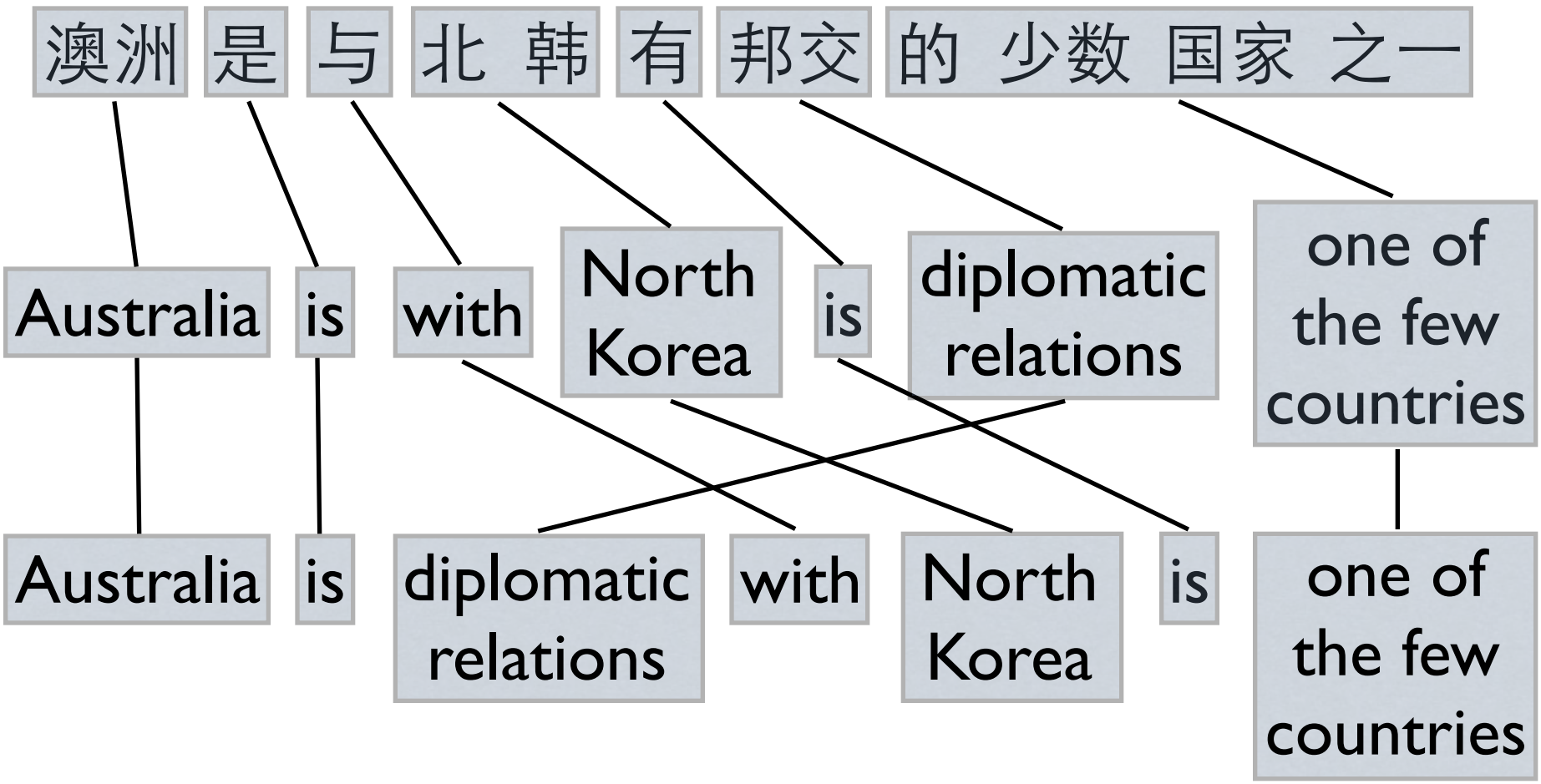
Word-based translation

澳洲 是 与 北 韩 有 邦交 的 少数 国家 之一

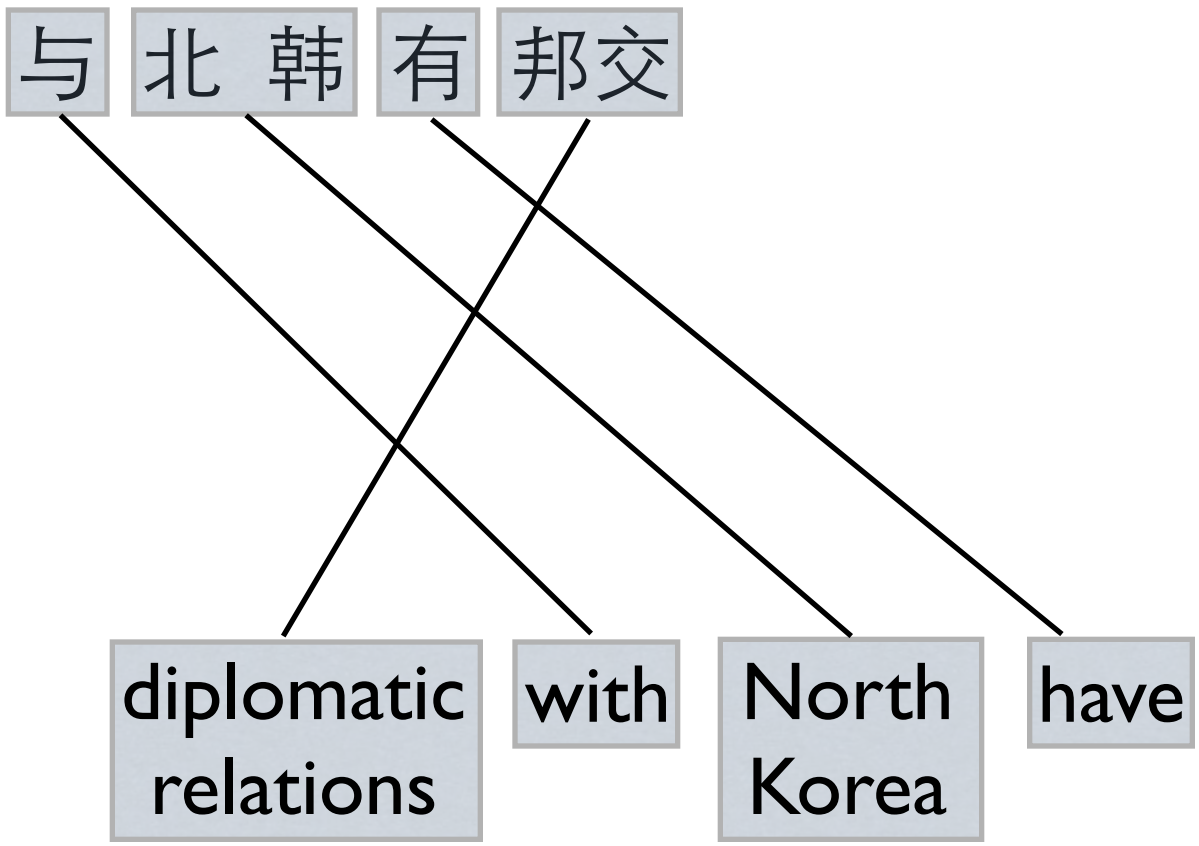
Australia is with North Korea have diplomatic relations that few countries one of

Human translation:
Australia is one of the few countries that have diplomatic relations with North Korea.

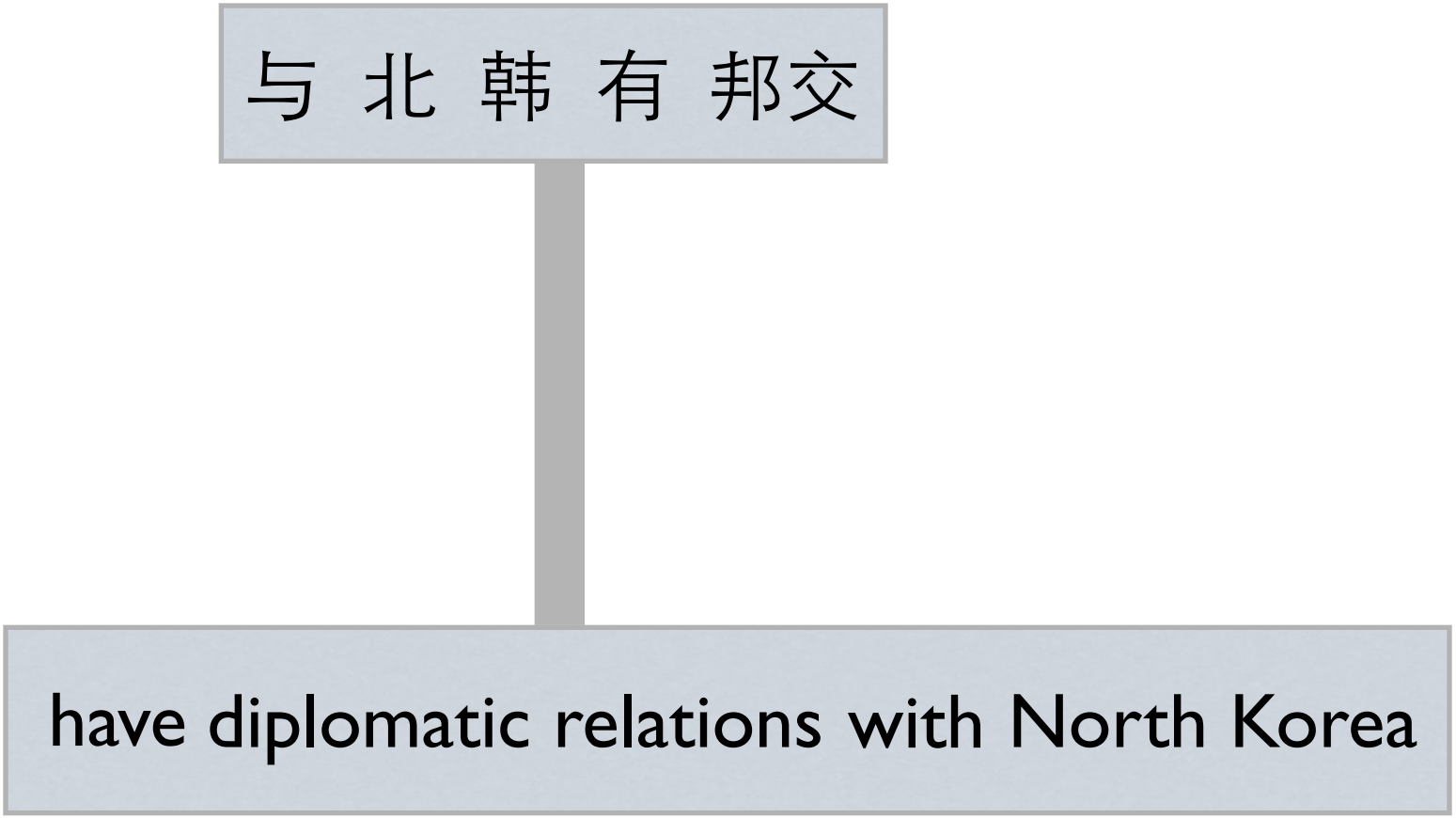
Phrase-based translation



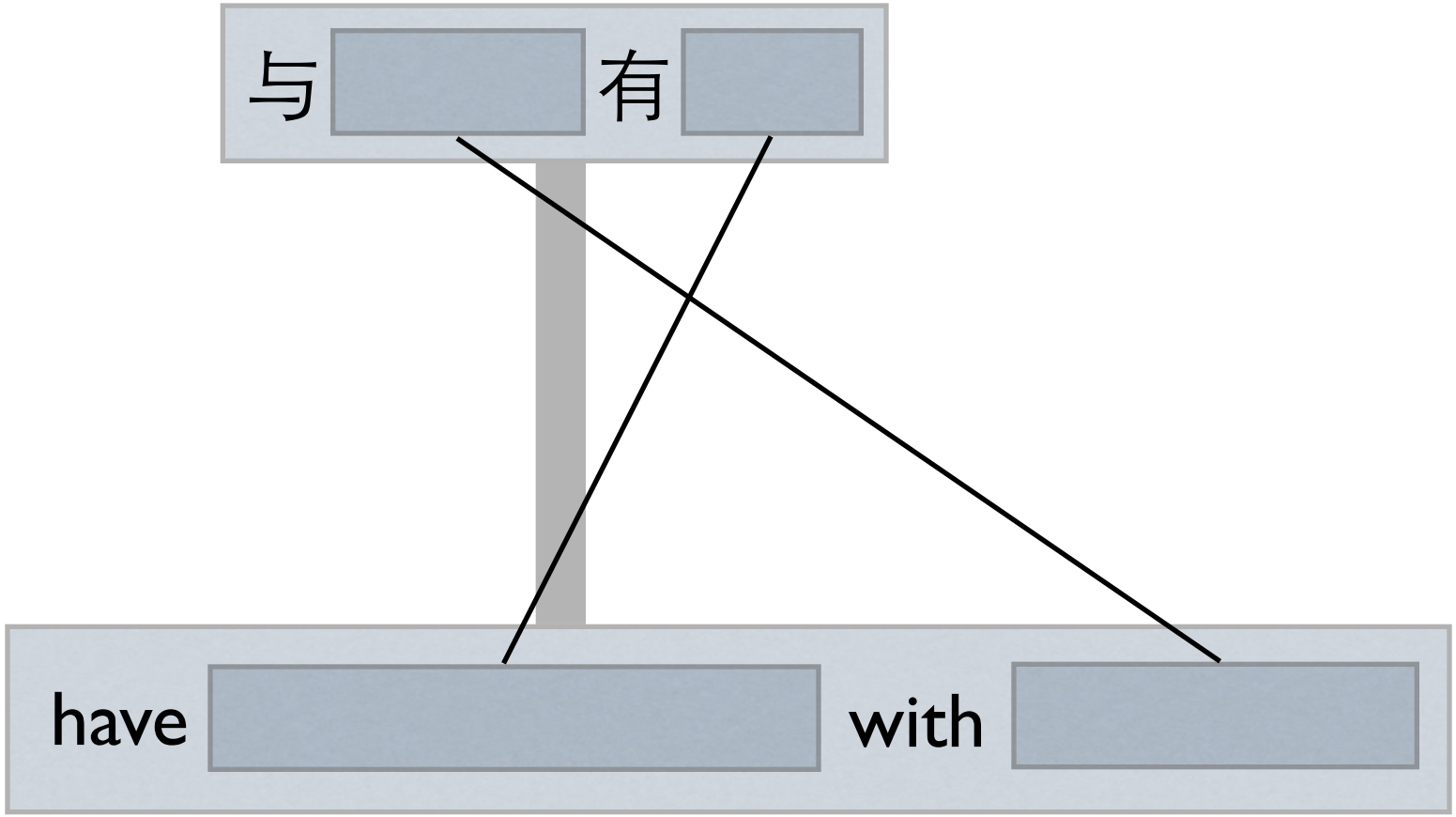
Phrases of phrases



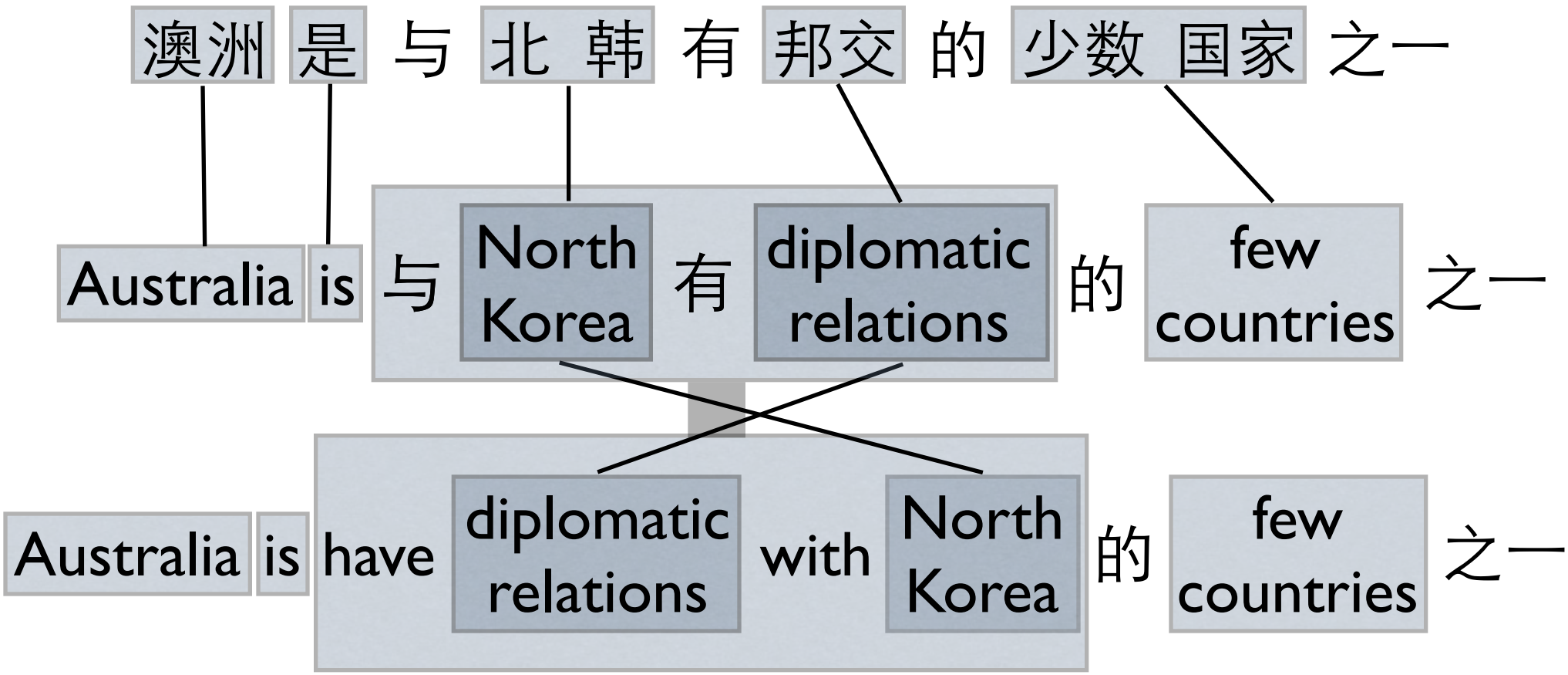
Phrases of phrases



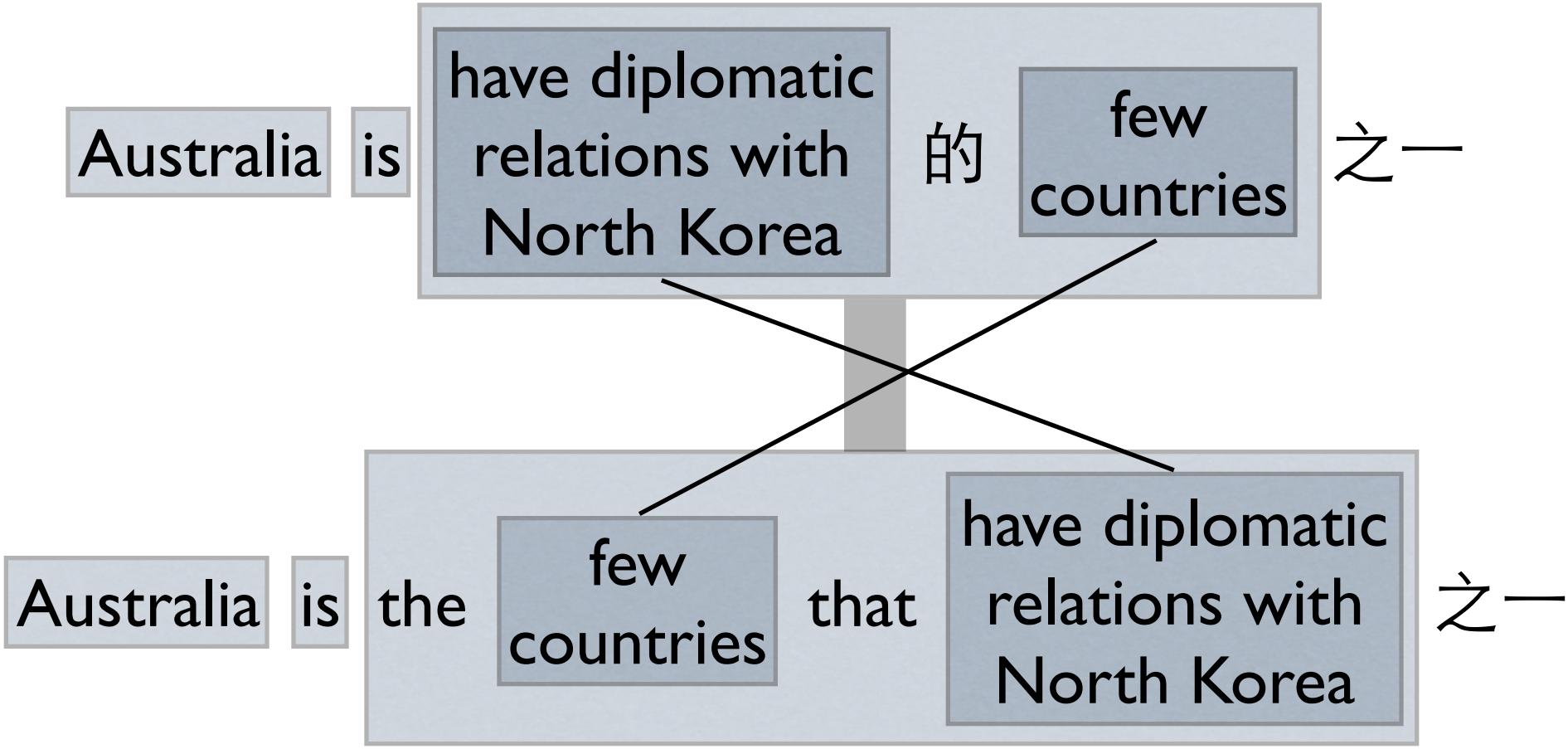
Phrases of phrases



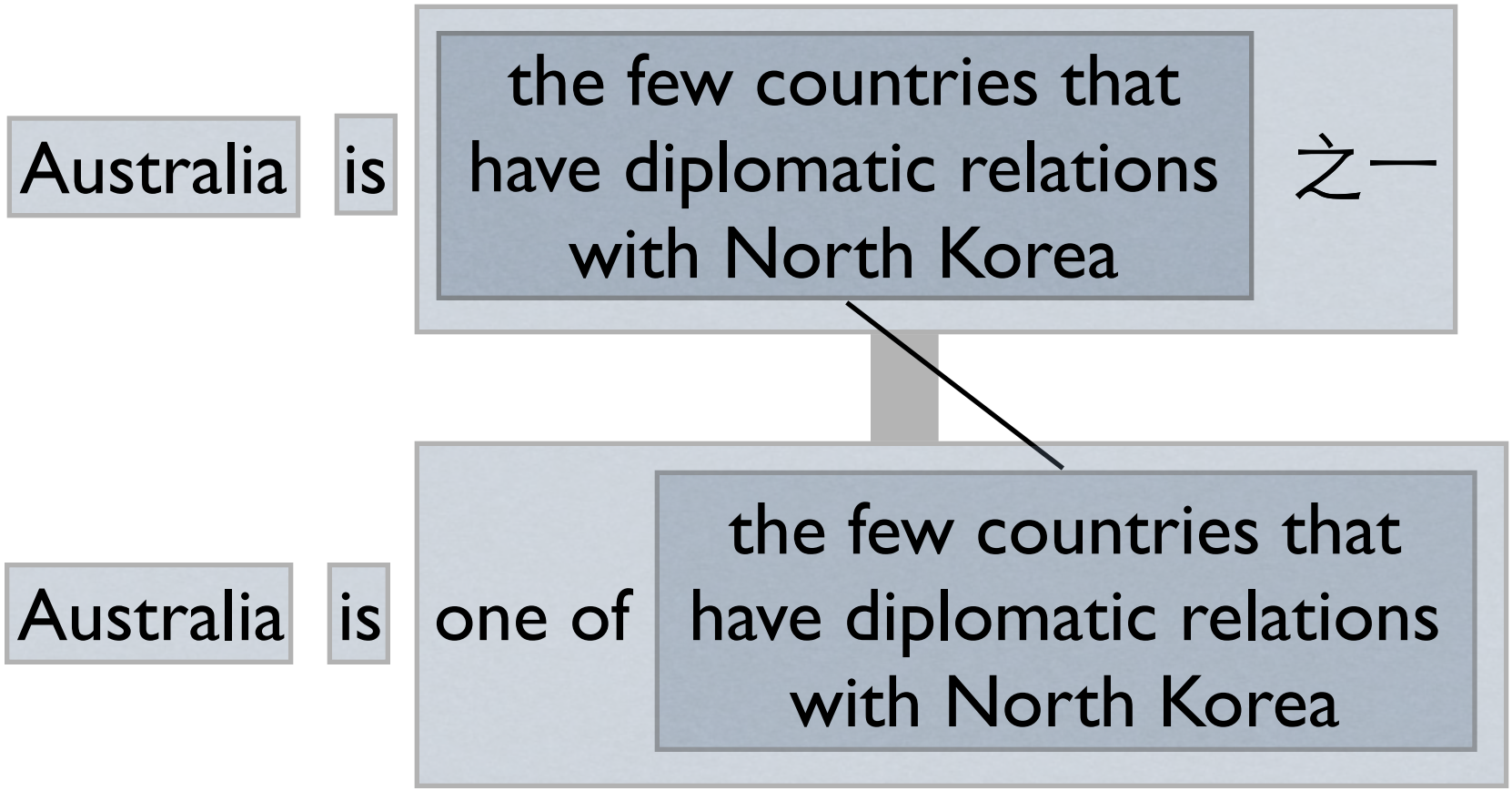
Hierarchical phrases



Hierarchical phrases



Hierarchical phrases



Definition

Context-free grammars

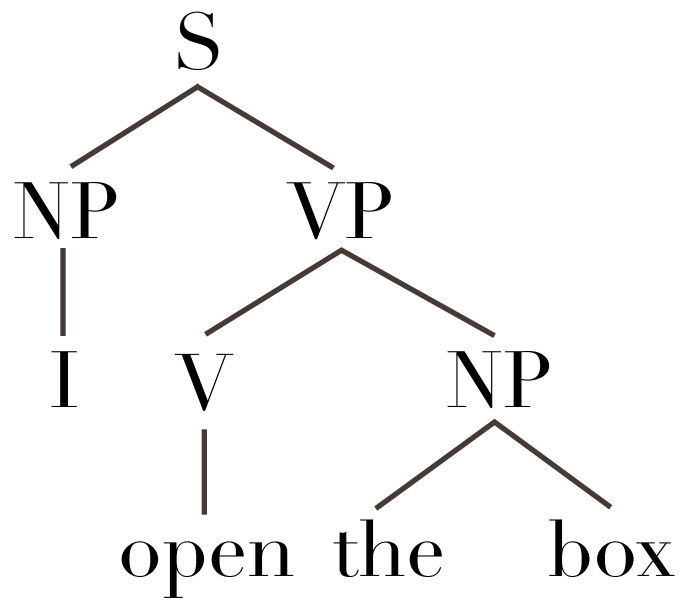
S → NP VP

NP → I

NP → the box

VP → V NP

V → open



Context-free grammars

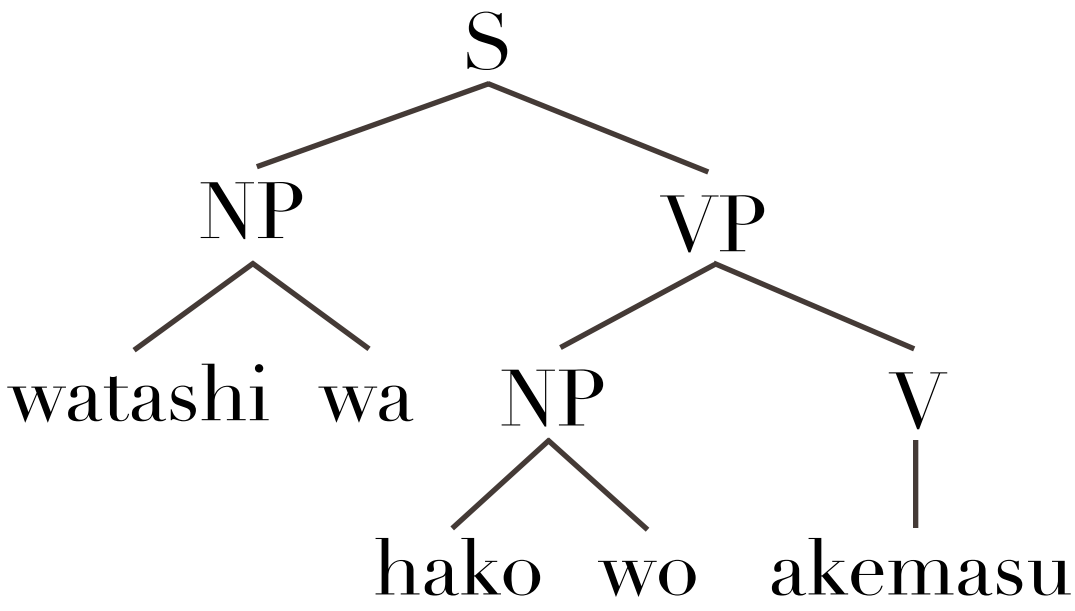
S → NP VP

NP → watashi wa

NP → hako wo

VP → NP V

V → akemasu



Synchronous CFGs

S → NP VP	S → NP VP
NP → I	NP → watashi wa
NP → the box	NP → hako wo
VP → V NP	VP → NP V
V → open	V → akemasu

Synchronous CFGs

$$S \rightarrow NP_1\, VP_2,\ NP_1\, VP_2$$

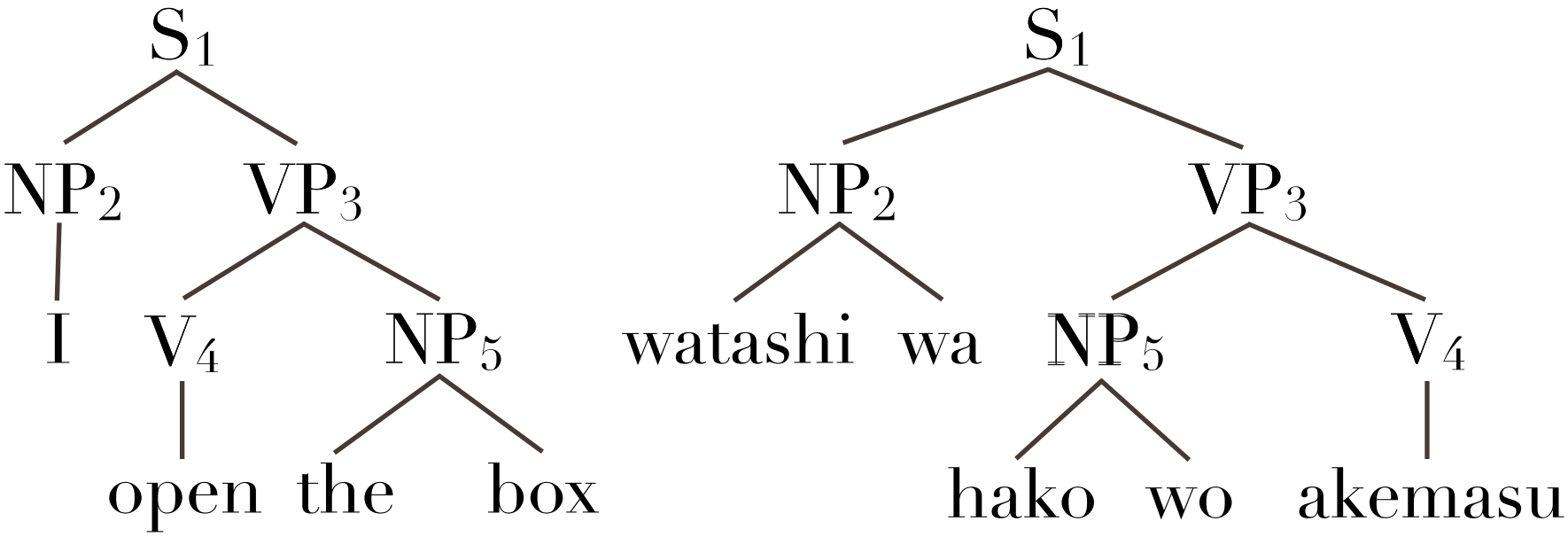
$$NP \rightarrow \text{I, watashi wa}$$

$$NP \rightarrow \text{the box, hako wo}$$

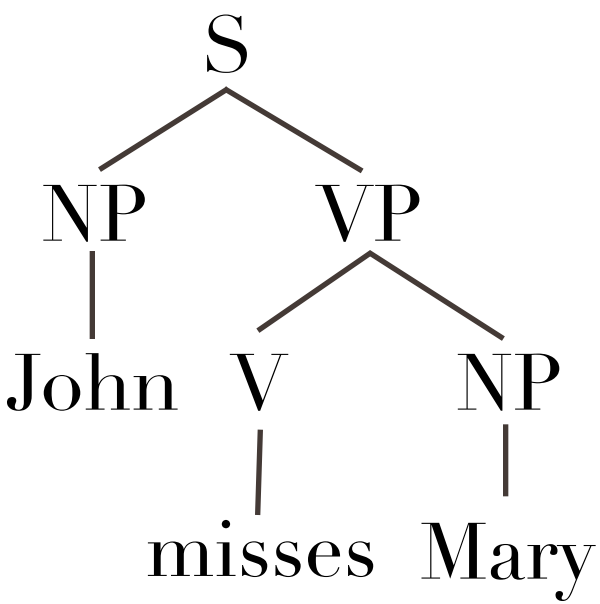
$$VP \rightarrow V_1\, NP_2,\ NP_2\, V_1$$

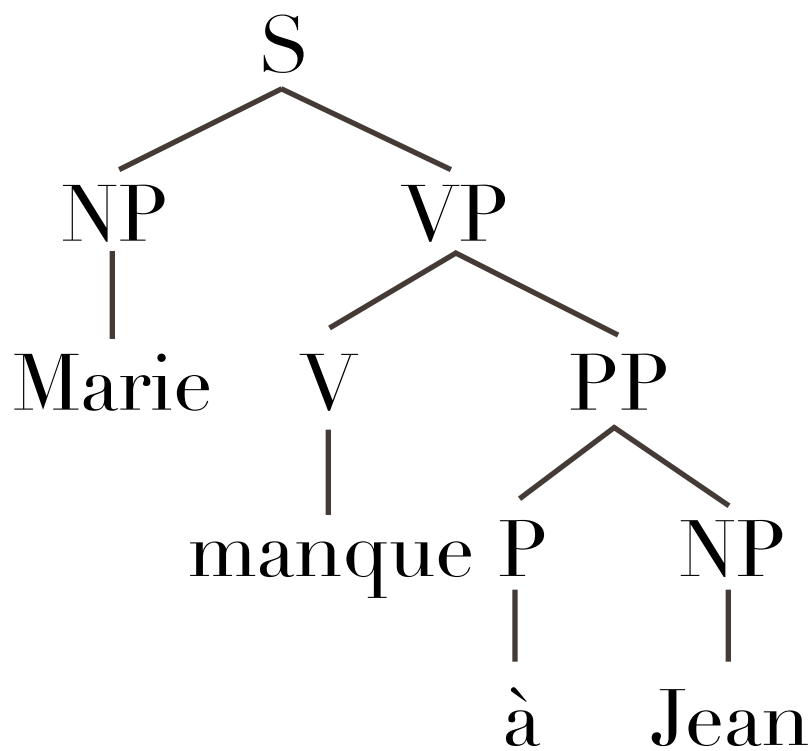
$$V \rightarrow \text{open, akemasu}$$

Synchronous CFGs

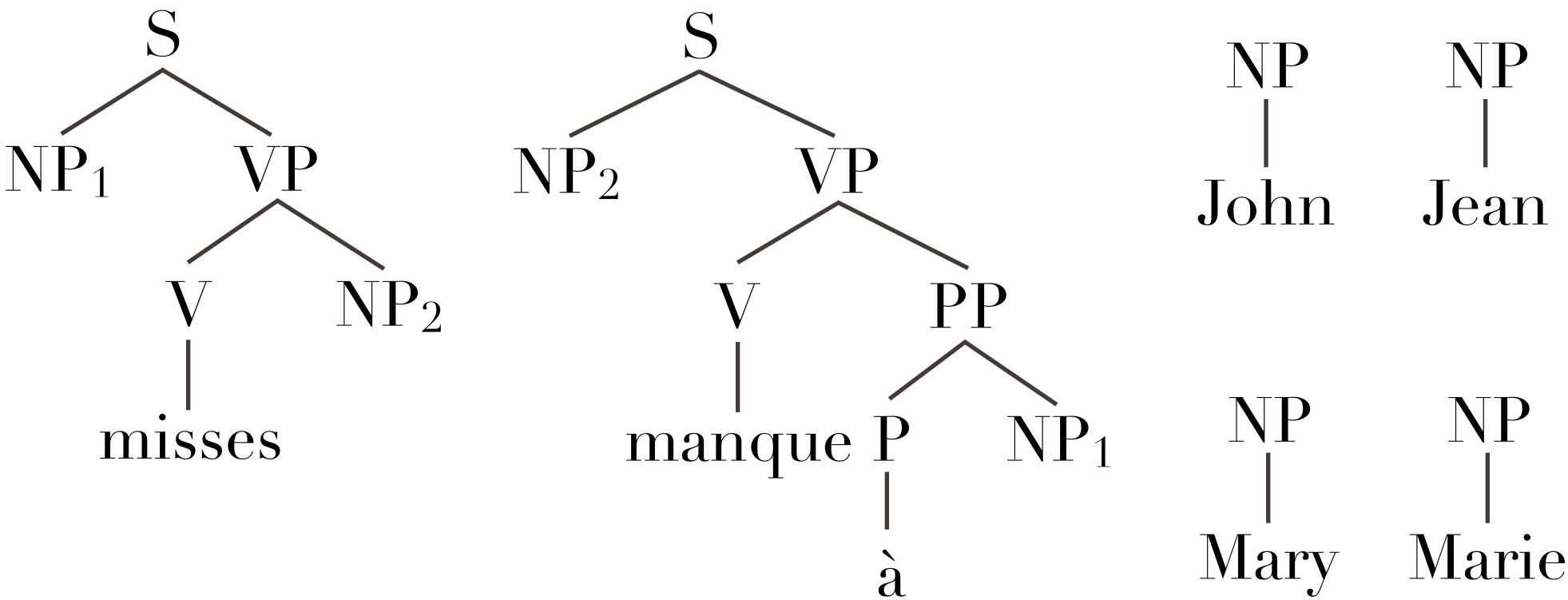


Limitations of synchronous CFGs

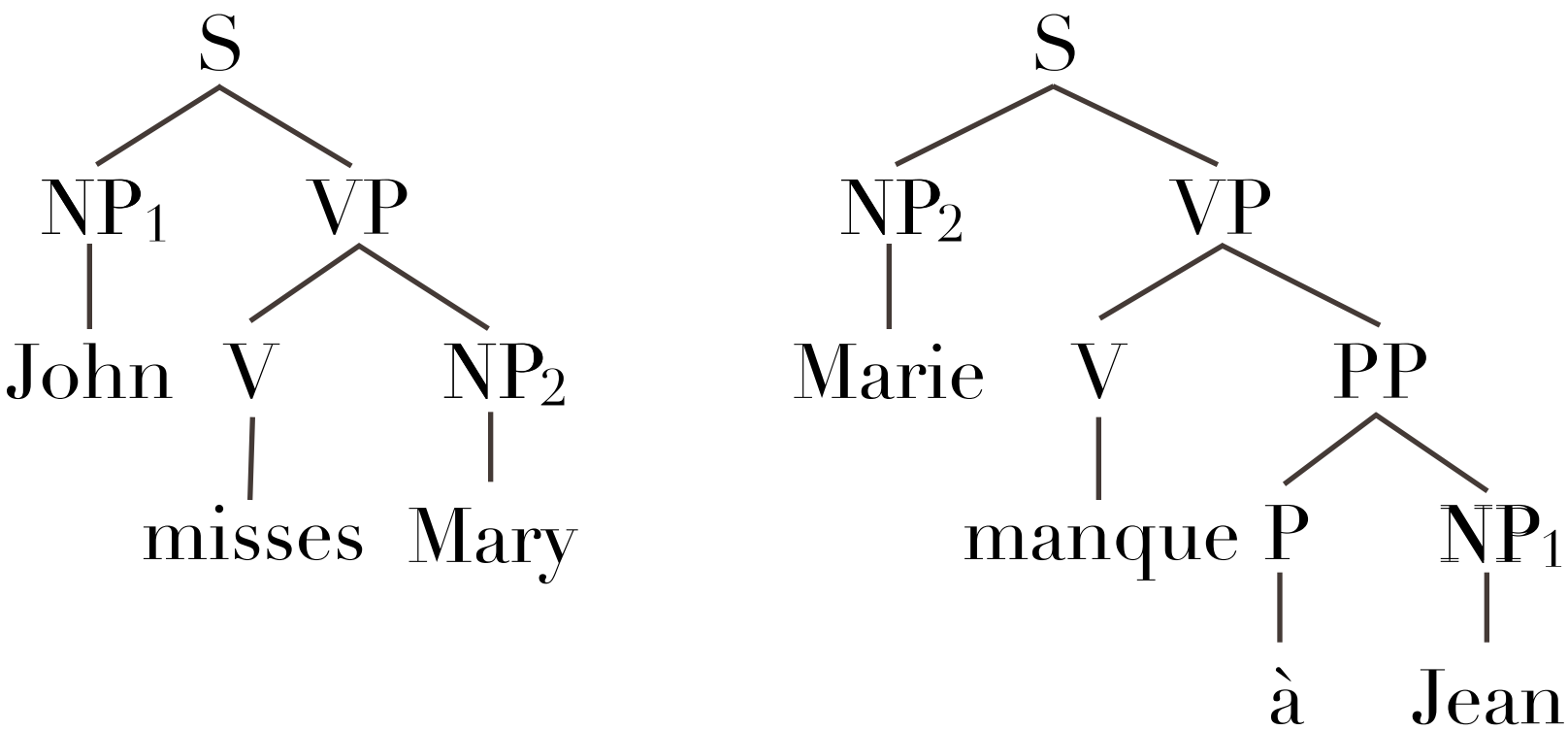




Synchronous tree substitution grammars



Synchronous tree substitution grammars



Properties

Chomsky normal form

$$X \rightarrow Y\,Z$$

$$X \rightarrow a$$

Chomsky normal form

$$A \rightarrow \quad B \quad C \quad D \quad E \quad F \qquad\qquad \textit{rank } 5$$

Chomsky normal form

$A \rightarrow [[[B\ \ C]\ D]\ E]\ F$ rank 5

$A \rightarrow V1\ F$

$V1 \rightarrow V2\ E$

$V2 \rightarrow V3\ D$

$V3 \rightarrow B\ C$

rank 2

A hierarchy of synchronous CFGs

$$\text{1-CFG} \subsetneq \text{2-CFG} = \text{3-CFG} = \text{4-CFG} = \ldots$$

$$\text{1-SCFG} \subsetneq \text{2-SCFG} = \text{3-SCFG} \subsetneq \text{4-SCFG} \subsetneq \ldots$$

$$\text{ITG}$$

(Wu, 1997)

Synchronous CNF?

$$A \rightarrow (B_1 \;\; C_2 \;\; D_3 \;, \;\; C_2 \;\; D_3 \;\; B_1)$$ rank 3

Synchronous CNF?

$$A \rightarrow (B_1\ [C_2\ \ D_3],\ [C_2\ \ D_3]\ B_1)$$

rank 3

$$A \rightarrow (B_1\ V1_2\quad ,\ V1_2\ B_1)$$

$$V1 \rightarrow (C_1\ D_2\quad ,\ C_1\ D_2)$$

rank 2

Synchronous CNF?

$$A \rightarrow (B_1\ C_2\ D_3\ E_4,\ C_2\ E_4\ B_1\ D_3)$$ rank 4

$$A \rightarrow ([B_1\ C_2]\ D_3\ E_4,\ [C_2\ E_4\ B_1]\ D_3)$$

$$A \rightarrow (B_1\ [C_2\ D_3]\ E_4,\ [C_2\ E_4\ B_1\ D_3])$$

$$A \rightarrow (B_1\ C_2\ [D_3\ E_4],\ C_2\ [E_4\ B_1\ D_3])$$

Synchronous CNF?

$$A \rightarrow (B_1\ C_2\ \ D_3,\ C_2\ D_3\ B_1)$$

	1	2	3
1			B
2	C		
3		D	

$$A \rightarrow (B_1\ C_2\ D_3\ E_4,\ C_2\ E_4\ B_1\ D_3)$$

	1	2	3	4
1			B	
2	C			
3				D
4		E		

A hierarchy of synchronous CFGs

$$\text{1-CFG} \subsetneq \text{2-CFG} = \text{3-CFG} = \text{4-CFG} = \ldots$$

$$\text{1-SCFG} \subsetneq \text{2-SCFG} = \text{3-SCFG} \subsetneq \text{4-SCFG} \subsetneq \ldots$$

$$\text{ITG}$$

(Wu, 1997)

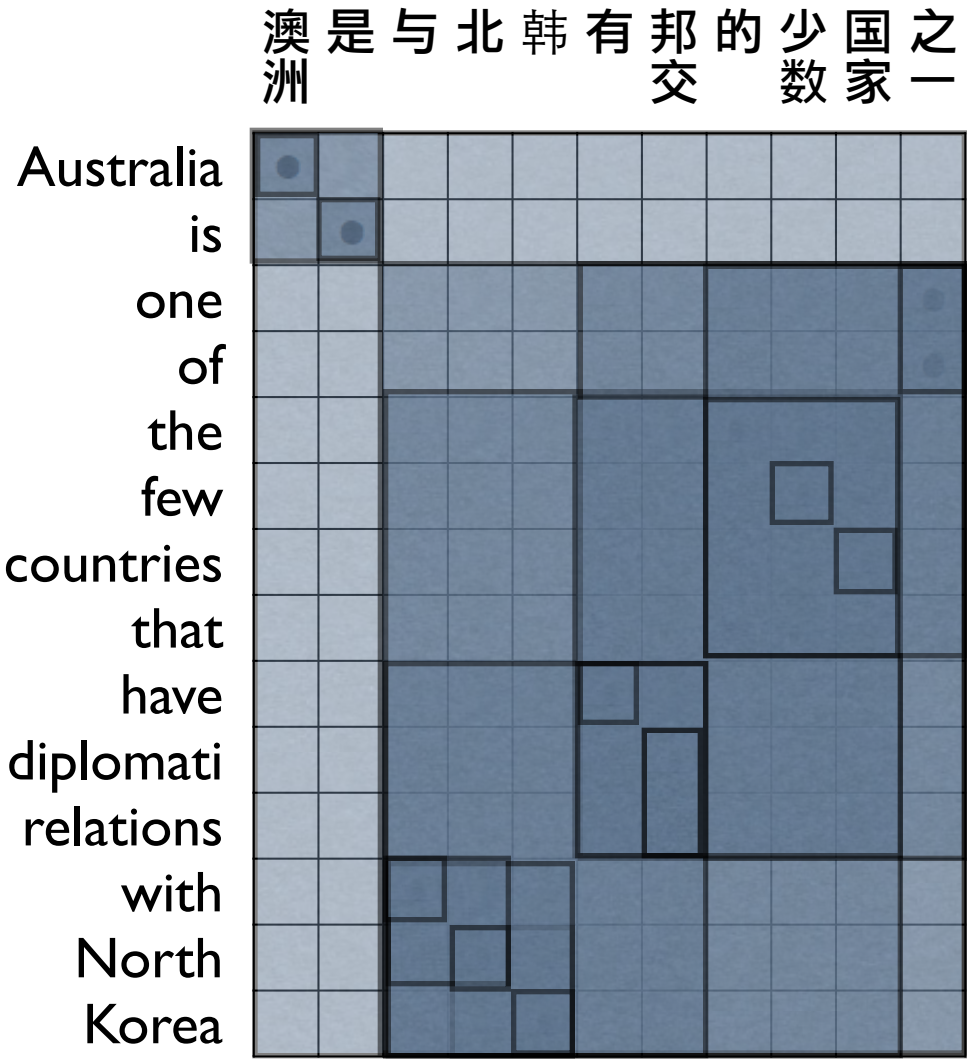
Extraction

Rule extraction

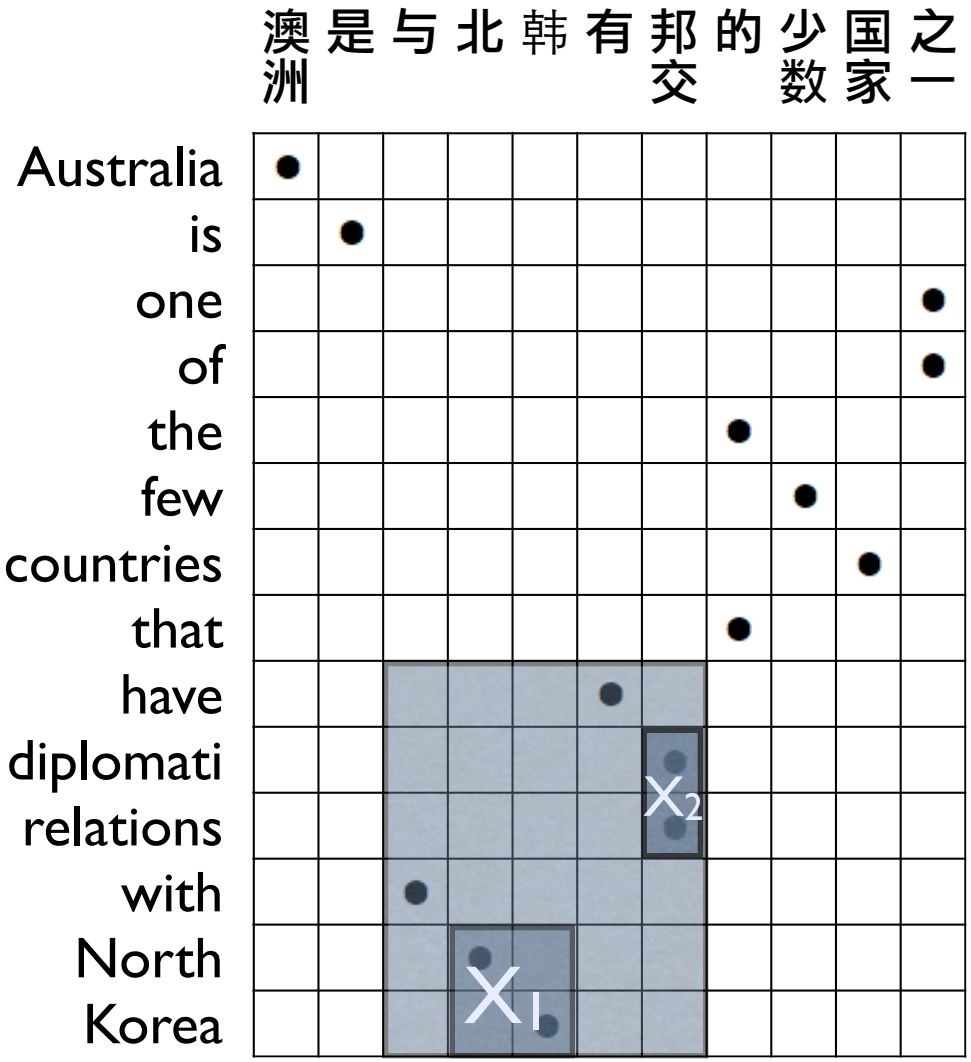
Extract *all possible* rules from each sentence pair that respect:

- Word alignments
- Syntactic structure
- Ad-hoc constraints

Grammar extraction



Grammar extraction



(与 北 韩 有 邦交, have diplomatic relations with North Korea)

(邦交, diplomatic relations)

(北 韩, North Korea)

(X → 与 X_1 有 X_2, X → have X_2 with X_1)

Ad-hoc constraints

- Hiero:
  - Initial phrases $\leq$10 words, rules $\leq$6 symbols
  - Require an aligned terminal
  - Limit to two nonterminals, nonadjacent
- Galley et al:
  - Rule must be minimal *or* have $\leq$5 English nodes (non-leaf, non-POS)

Example rules

$X \rightarrow$ 的	$X \rightarrow$'s
$X \rightarrow X_1$ 的 X_2	$X \rightarrow$ the X_2 of X_1
$X \rightarrow X_1$ 的 X_2	$X \rightarrow$ the X_2 that X_1
$X \rightarrow$ 在	$X \rightarrow$ in
$X \rightarrow$ 在 X_1 下	$X \rightarrow$ under X_1
$X \rightarrow$ 在 X_1 前	$X \rightarrow$ before X_1
$X \rightarrow$ 今年 X_1	$X \rightarrow X_1$ this year
$X \rightarrow X_1$ 之一	$X \rightarrow$ one of X_1
$X \rightarrow X_1$ 总统	$X \rightarrow$ president X_1

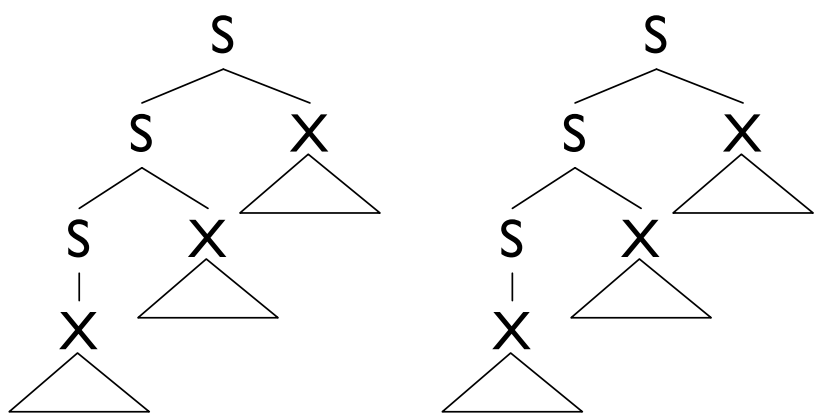
Glue rules

- Plus "glue" rules:

$$(S \rightarrow S_1 X_2, S \rightarrow S_1 X_2)$$

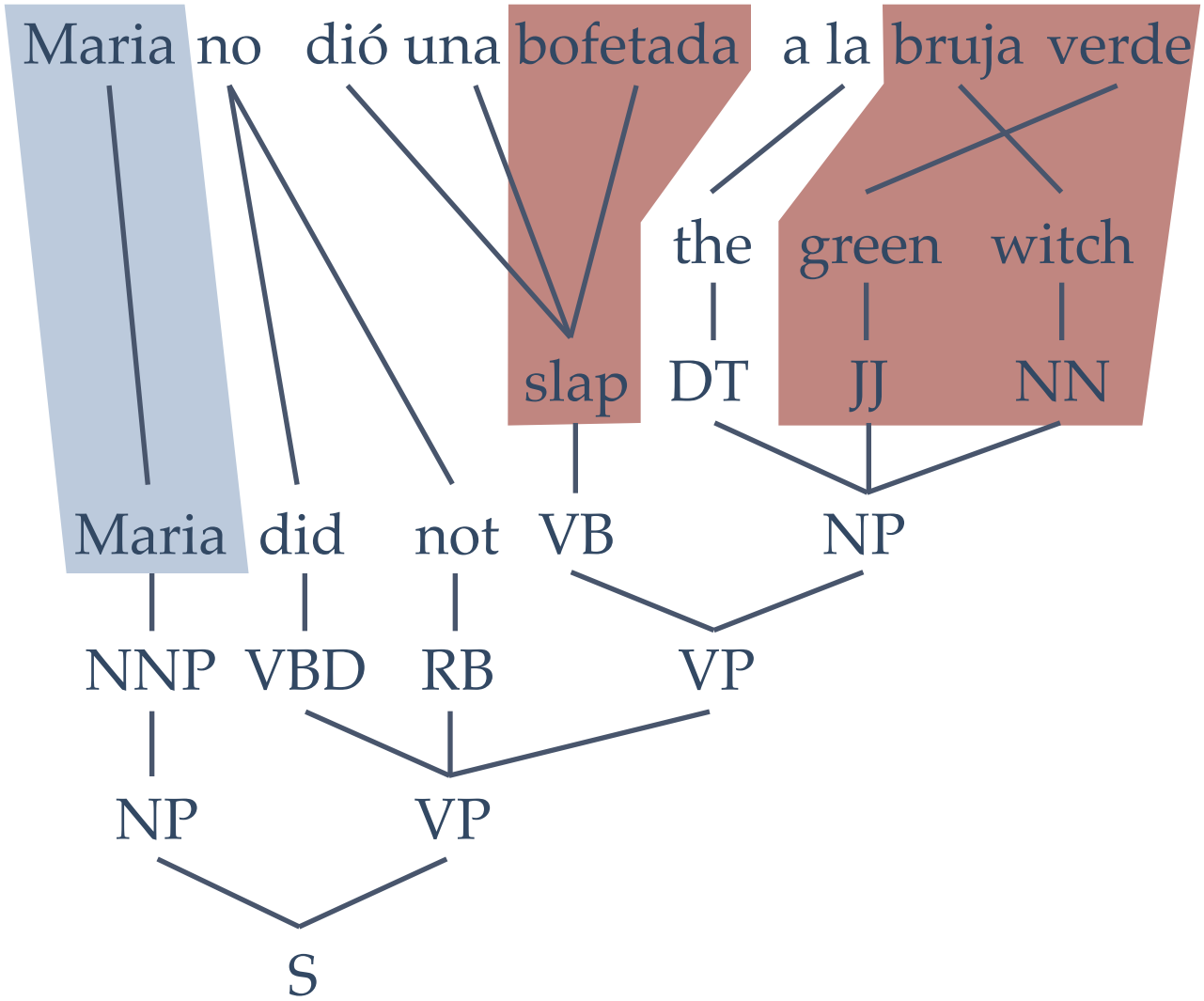
$$(S \rightarrow X_1, S \rightarrow X_1)$$

- Acts as fallback like phrase-based systems



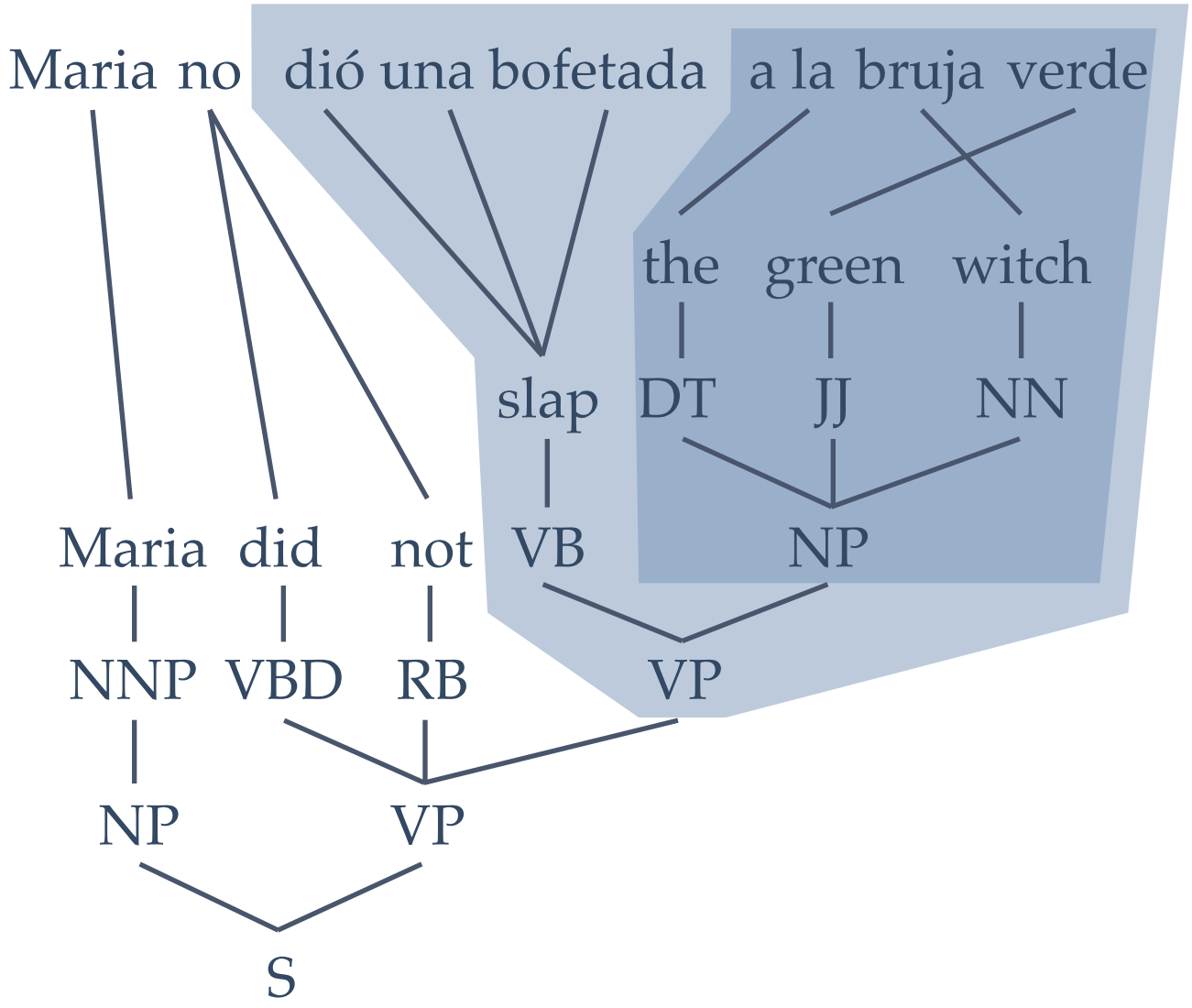
Rule extraction

1. Phrases
   - respect word alignments
   - are syntactic constituents
2. Subtract phrases to form rules



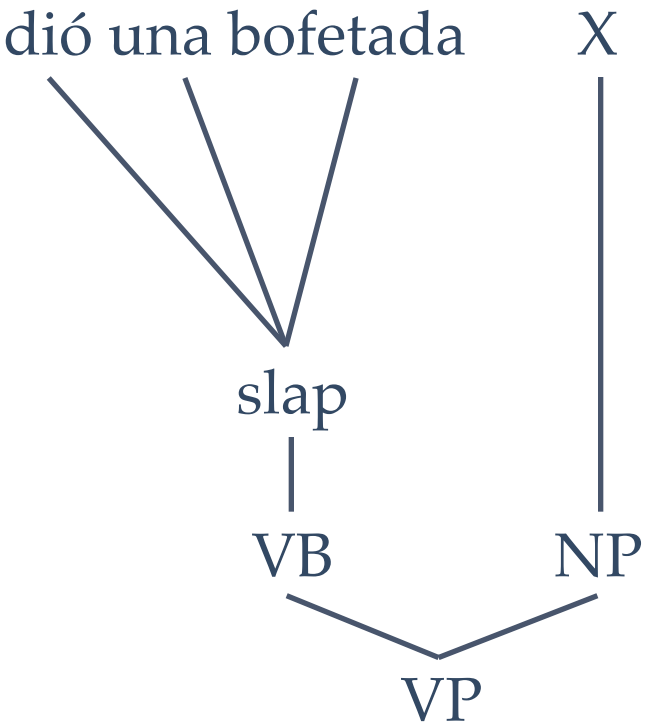
Rule extraction

1. Phrases
   - respect word alignments
   - are syntactic constituents
2. Subtract phrases to form rules

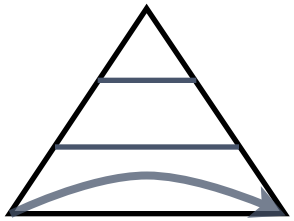
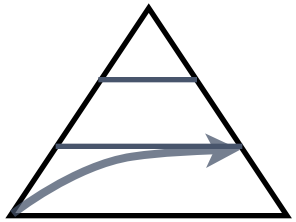
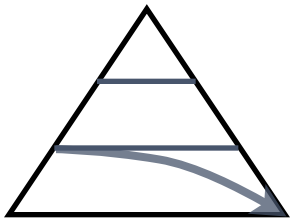
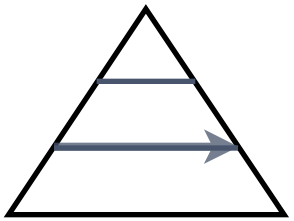


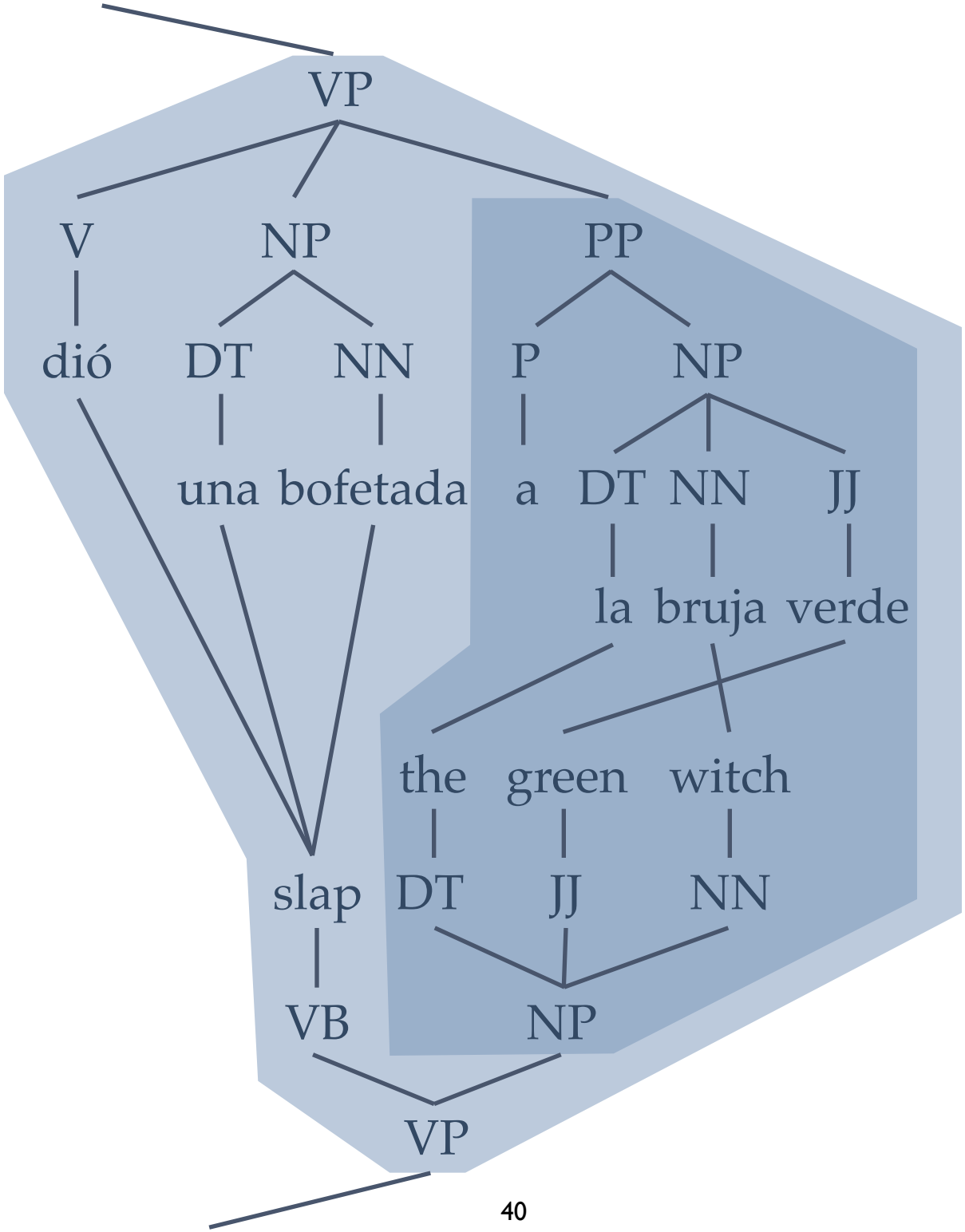
Rule extraction

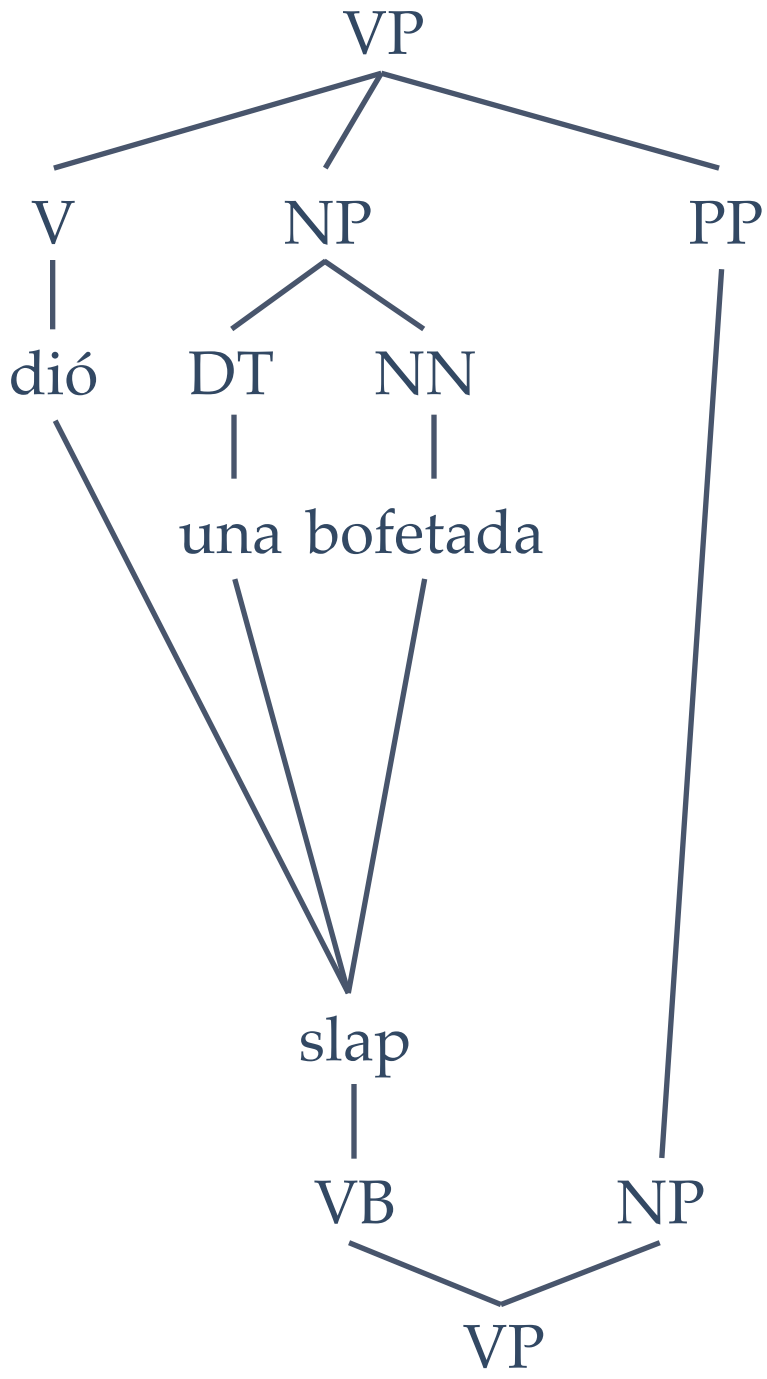
1. Phrases
   - respect word alignments
   - are syntactic constituents
2. Subtract phrases to form rules

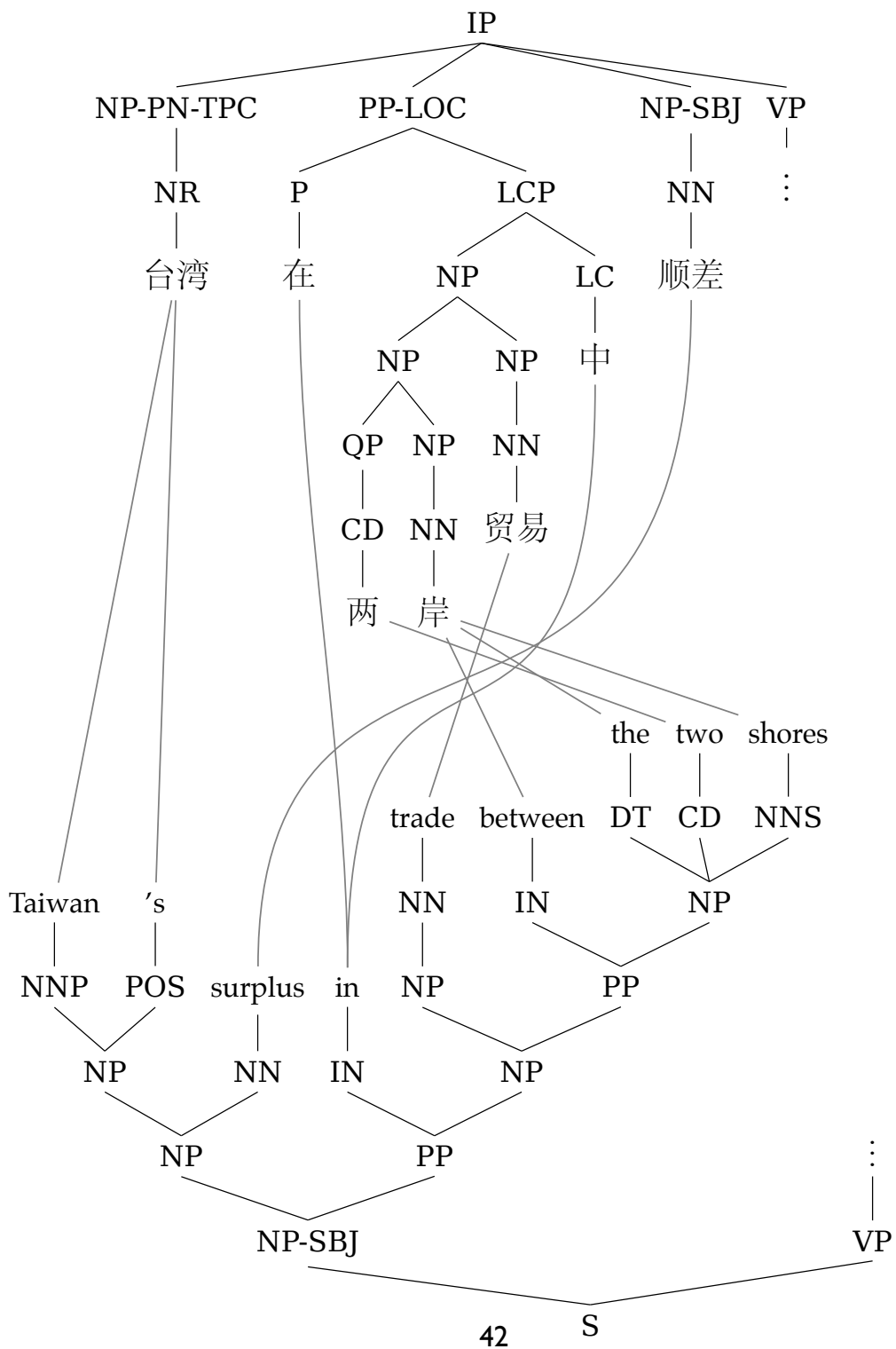


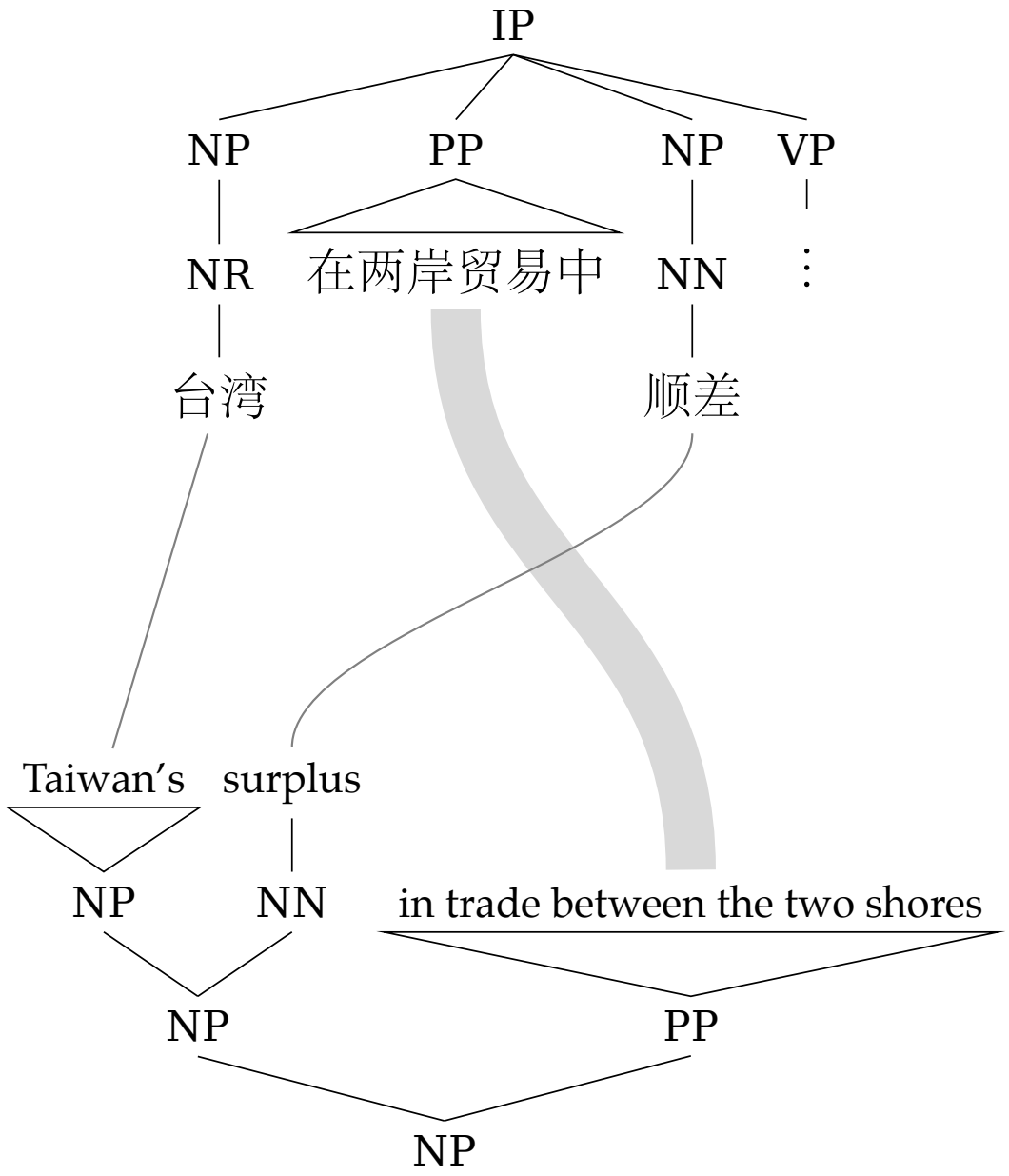
Varieties of rule extraction

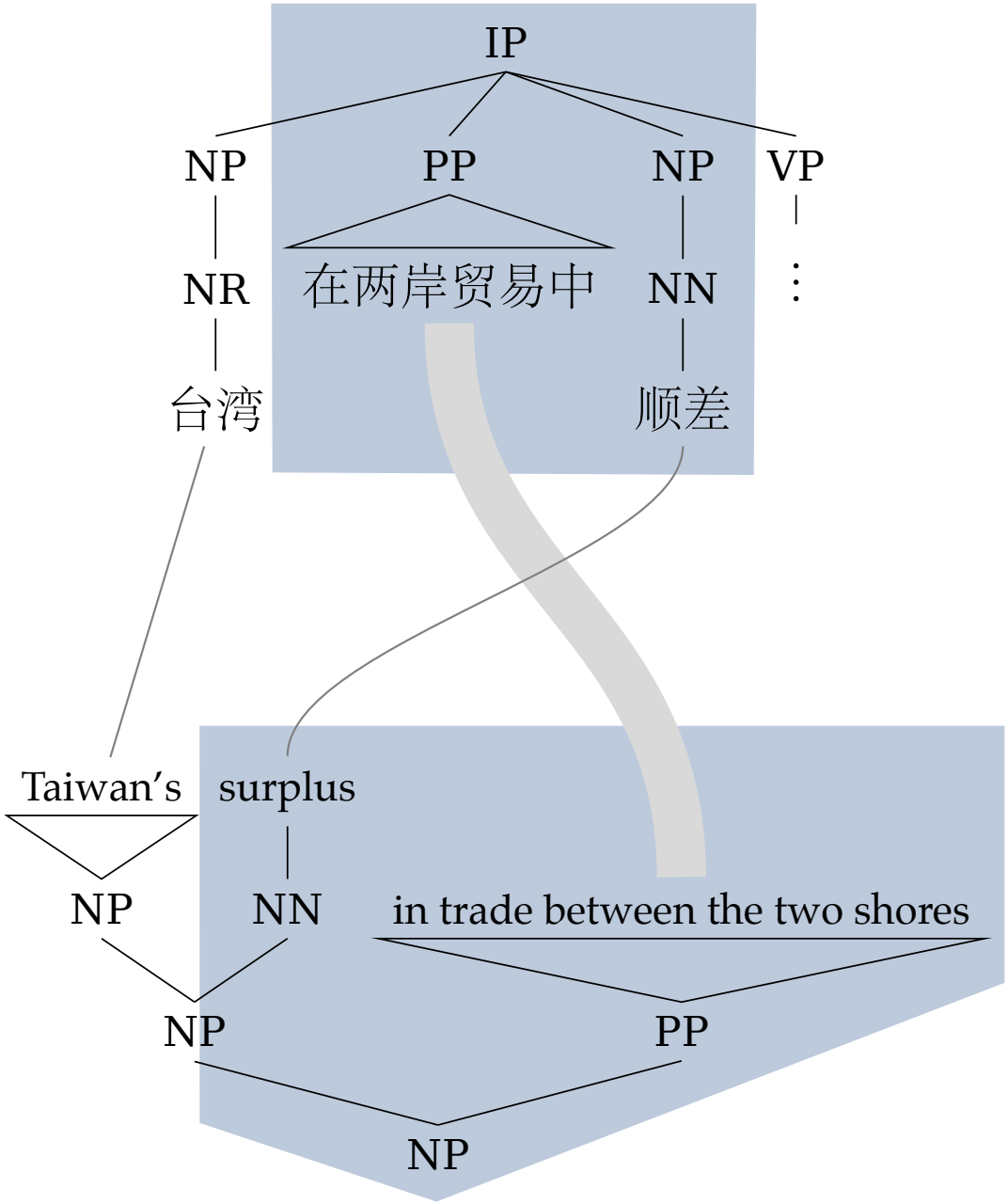
	string-to-string	ITG (Wu, 1997)	Hiero (Chiang, 2005)
	string-to-tree	Yamada & Knight, 2001	Galley et al, 2004/2006
	tree-to-string		Huang et al, 2006 Liu et al, 2006
	tree-to-tree	DOT (Poutsma, 2000) Eisner, 2003	Zhang et al., 2008 Liu et al., 2009

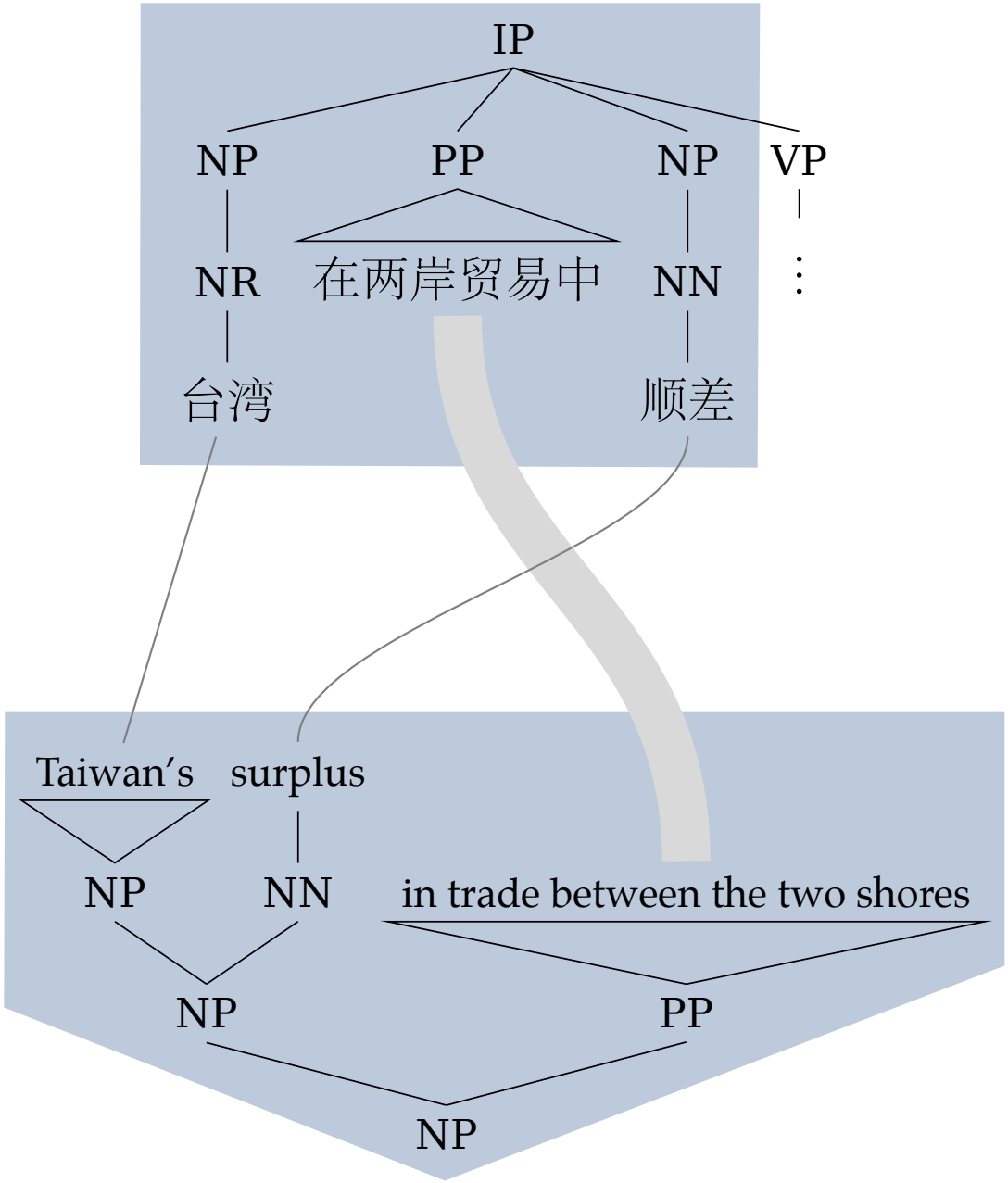

VP
V
NP
PP
dió
DT
NN
P
NP
una
bofetada
a
DT
NN
JJ
la
bruja
verde
the
green
witch
slap
DT
JJ
NN
VB
NP
VP

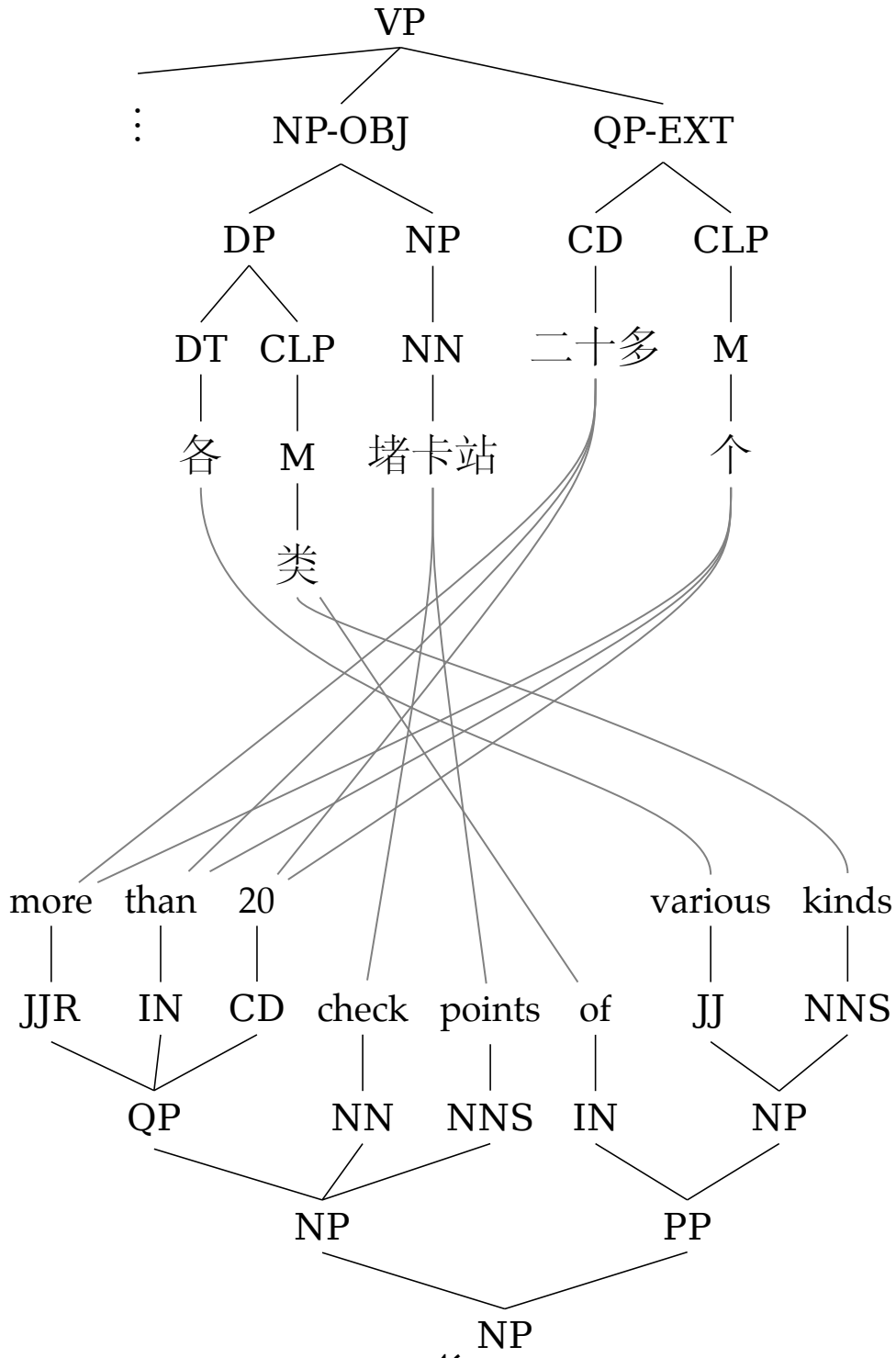

VP
V
NP
PP
dió
DT
NN
una
bofetada
slap
VB
NP
VP

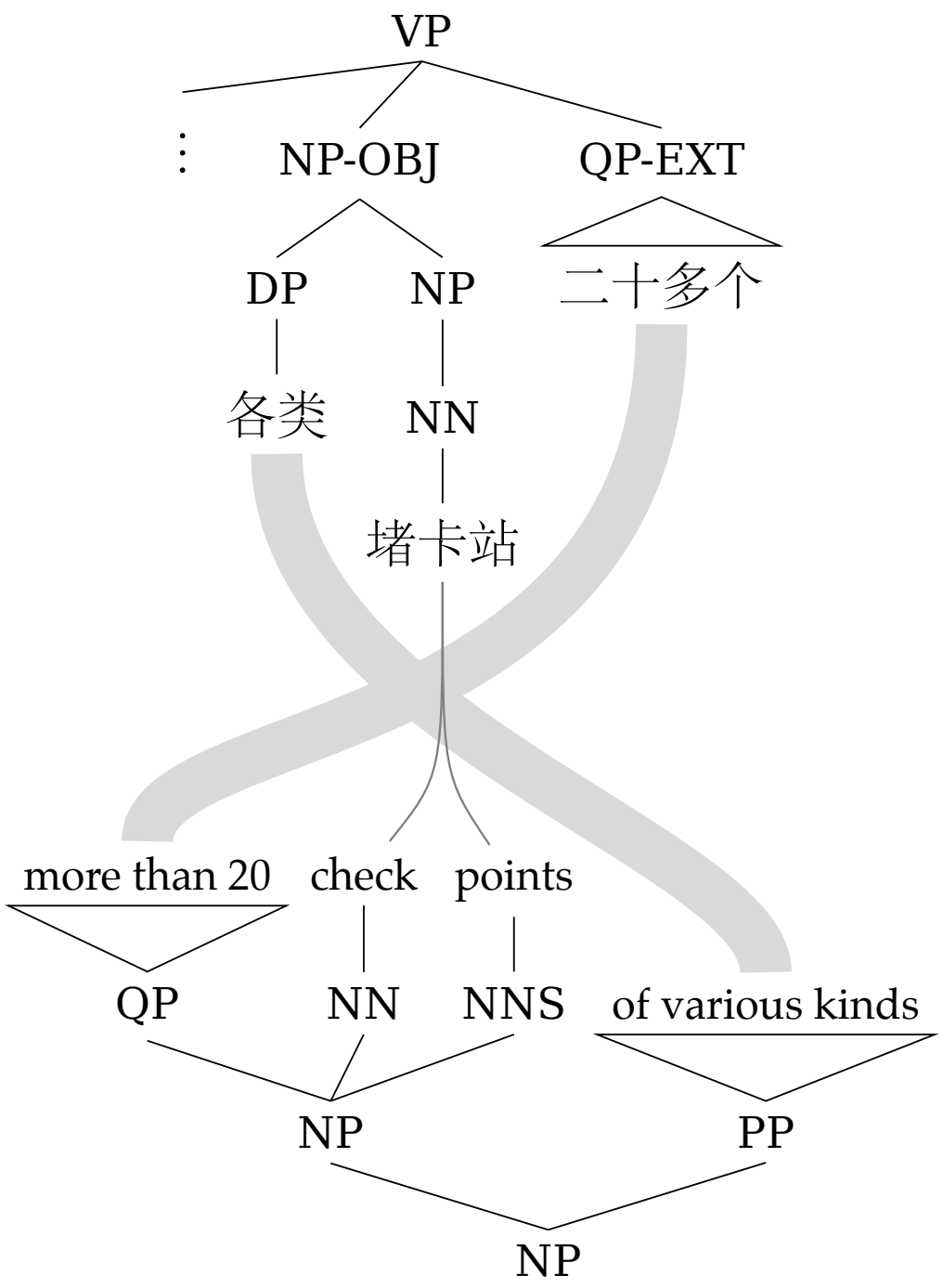

IP
NP-PN-TPC
PP-LOC
NP-SBJ
VP
NR
P
LCP
NN
台湾
在
NP
LC
顺差
NP
NP
中
QP
NP
NN
CD
NN
贸易
两
岸
the
two
shores
trade
between
DT
CD
NNS
Taiwan
's
NN
IN
NP
NNP
POS
surplus
in
NP
PP
NP
NN
IN
NP
NP
PP
NP-SBJ
VP
S

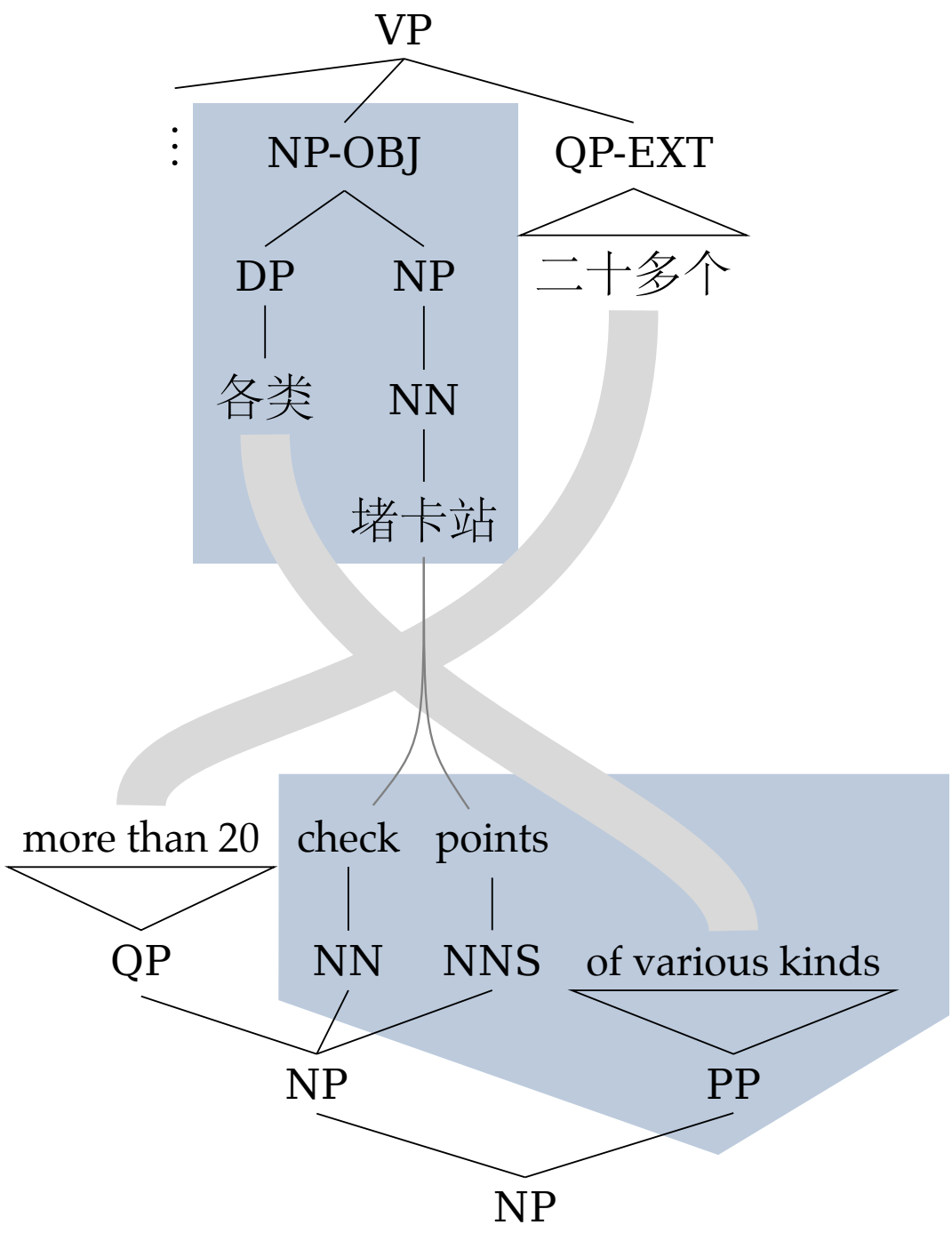

IP
NP
PP
NP
VP
NR
在两岸贸易中
NN
台湾
顺差
Taiwan's
surplus
NP
NN
in trade between the two shores
NP
PP
NP

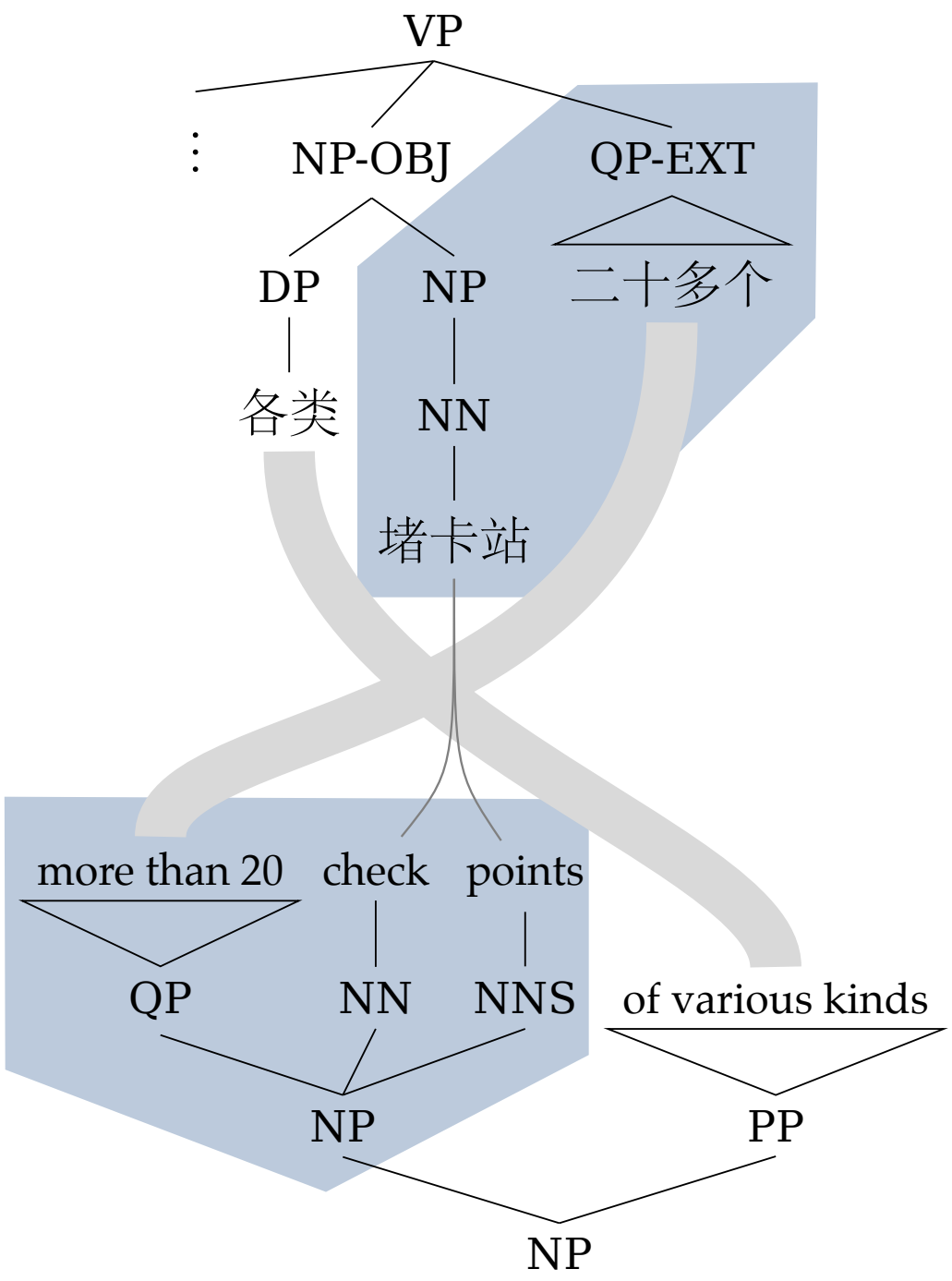

IP
NP
PP
NP
VP
NR
在两岸贸易中
NN
台湾
顺差
Taiwan's
surplus
NP
NN
in trade between the two shores
NP
PP
NP


IP
NP
PP
NP
VP
NR
在两岸贸易中
NN
台湾
顺差
Taiwan's
surplus
NP
NN
in trade between the two shores
NP
PP
NP


VP
⋮
NP-OBJ
QP-EXT
DP
NP
CD
CLP
DT
CLP
NN
二十多
M
各
M
堵卡站
个
类
more
than
20
various
kinds
JJR
IN
CD
check
points
of
JJ
NNS
QP
NN
NNS
IN
NP
NP
PP
NP


VP
⋮
NP-OBJ
QP-EXT
DP
NP
二十多个
各类
NN
堵卡站
more than 20
check
points
QP
NN
NNS
of various kinds
NP
PP
NP


VP
NP-OBJ
QP-EXT
DP
NP
二十多个
各类
NN
堵卡站
more than 20
check
points
QP
NN
NNS
of various kinds
NP
PP
NP


VP
NP-OBJ
QP-EXT
DP
NP
二十多个
各类
NN
堵卡站
more than 20
check
points
QP
NN
NNS
of various kinds
NP
PP
NP

Model

Probability model

- Synchronous CFG gives us possible translations (derivations)
- What makes one translation better than another?
- Log-linear model that combines small number of feature functions

Probability model

Log-linear model (Och and Ney, 2002)

$$P(D) \propto \prod_{r \in D} \prod_{i} v_i(r)^{\lambda_i(r)}$$

D	derivation
r	rules in D
v_i	feature functions
λ_i	feature weights

Weights λ_i learned by maximum-BLEU training (Och 2003)

Model features

- Phrase translation:

  $p(X \rightarrow X_1 \text{之一} \mid X \rightarrow \text{one of } X_1)$

  $p(X \rightarrow \text{one of } X_1 \mid X \rightarrow X_1 \text{之一})$

  - Problem: rules not actually observed?

Model features

- Lexical weighting (Koehn):

  $\frac{1}{2}[p(\text{之一} \mid \text{one}) + p(\text{之一} \mid \text{of})]$

  $p(\text{one} \mid \text{之一}) \times p(\text{of} \mid \text{之一})$

Model features

- Binary features: extracted rules; glue rule S → SX; green rules
- Number of English words
- (Word) *n*-gram language model:

$p(\text{Australia} \mid \textit{START}) \times p(\text{is} \mid \textit{START}\ \text{Australia}) \cdots$

Decoding

Review: CKY

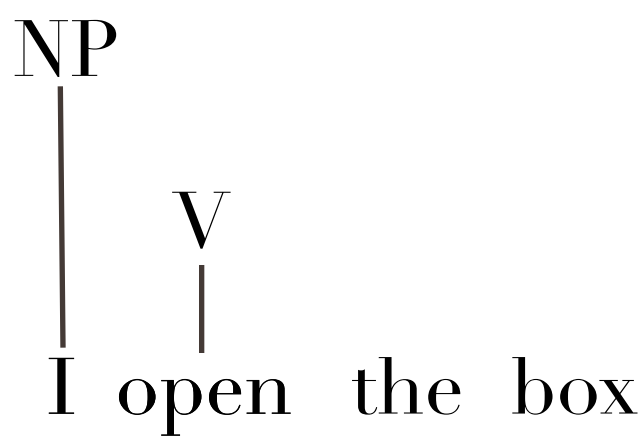
S → NP VP

NP → I

NP → the box

VP → V NP

V → open



Review: CKY

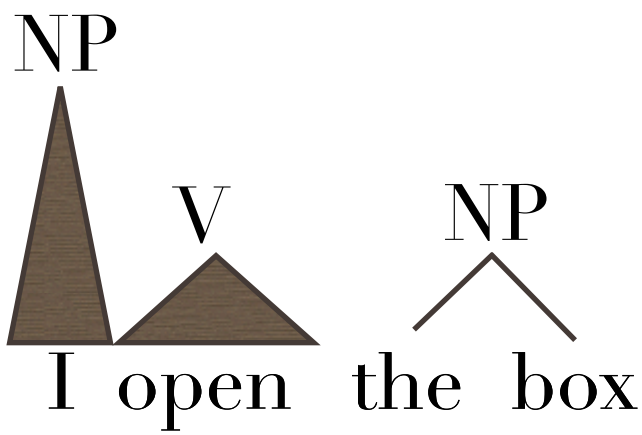
S → NP VP

NP → I

NP → the box

VP → V NP

V → open



Review: CKY

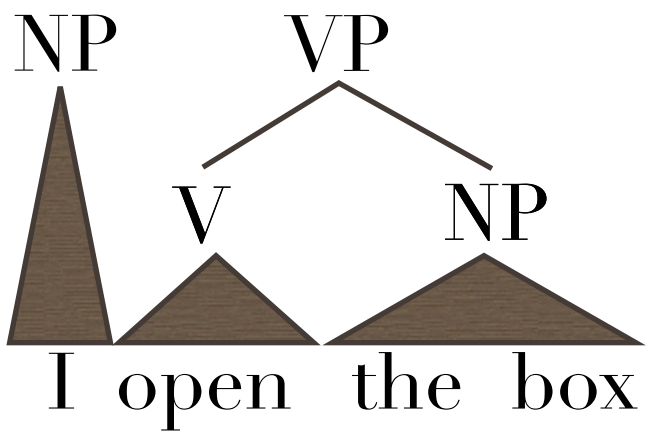
S → NP VP

NP → I

NP → the box

VP → V NP

V → open



Review: CKY

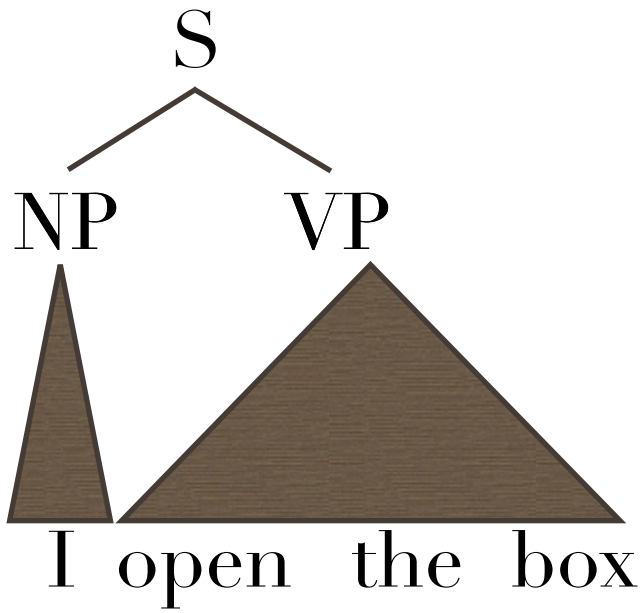
S → NP VP

NP → I

NP → the box

VP → V NP

V → open



Review: CKY

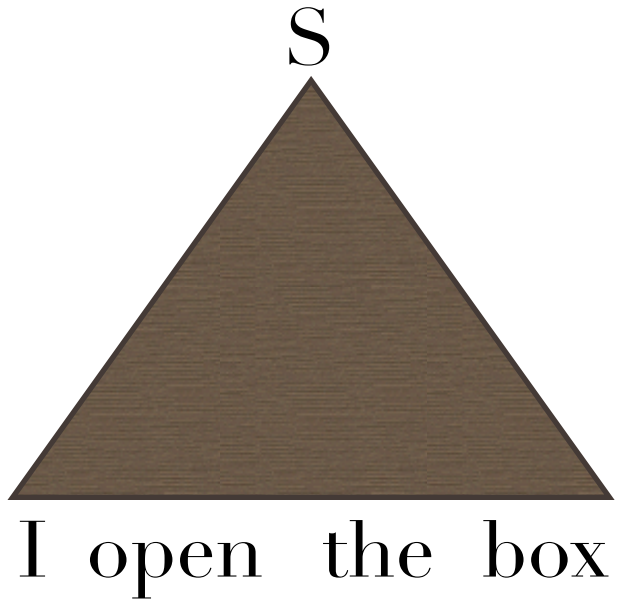
S → NP VP

NP → I

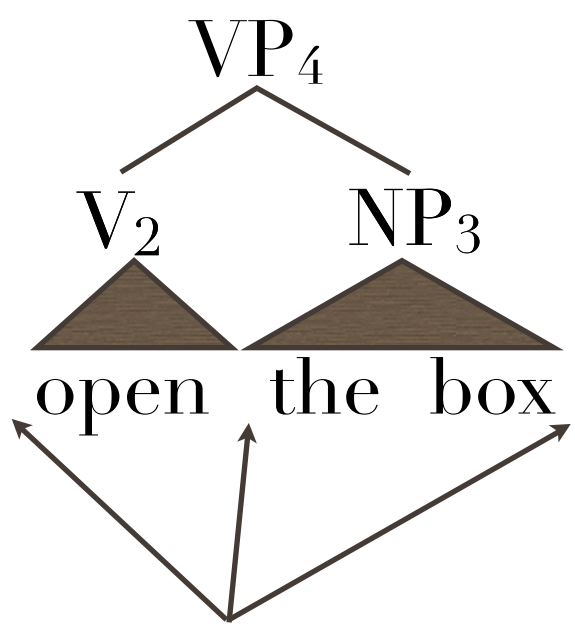
NP → the box

VP → V NP

V → open

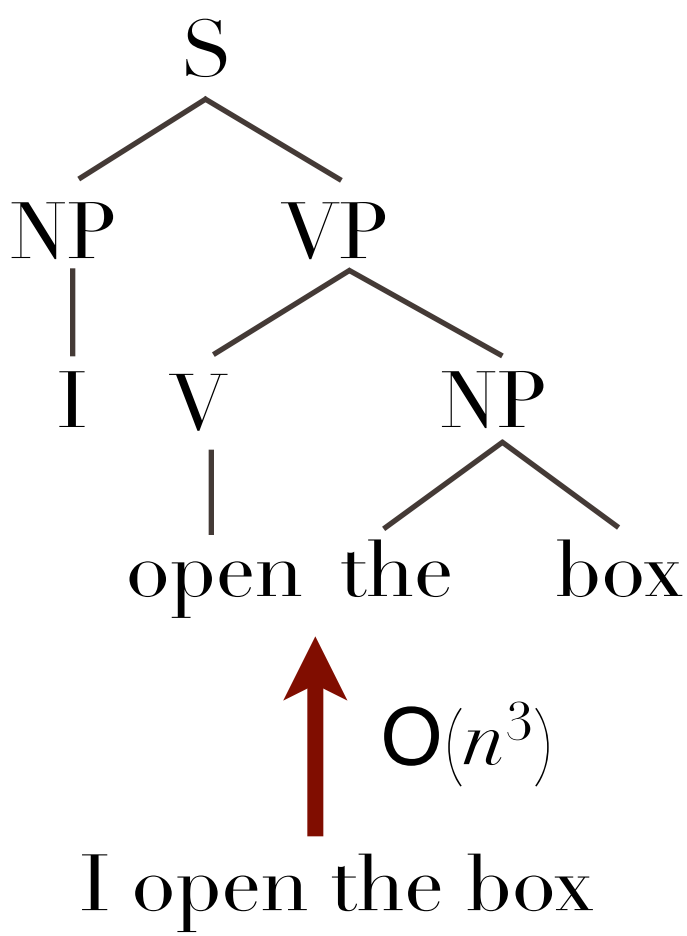


Review: CKY

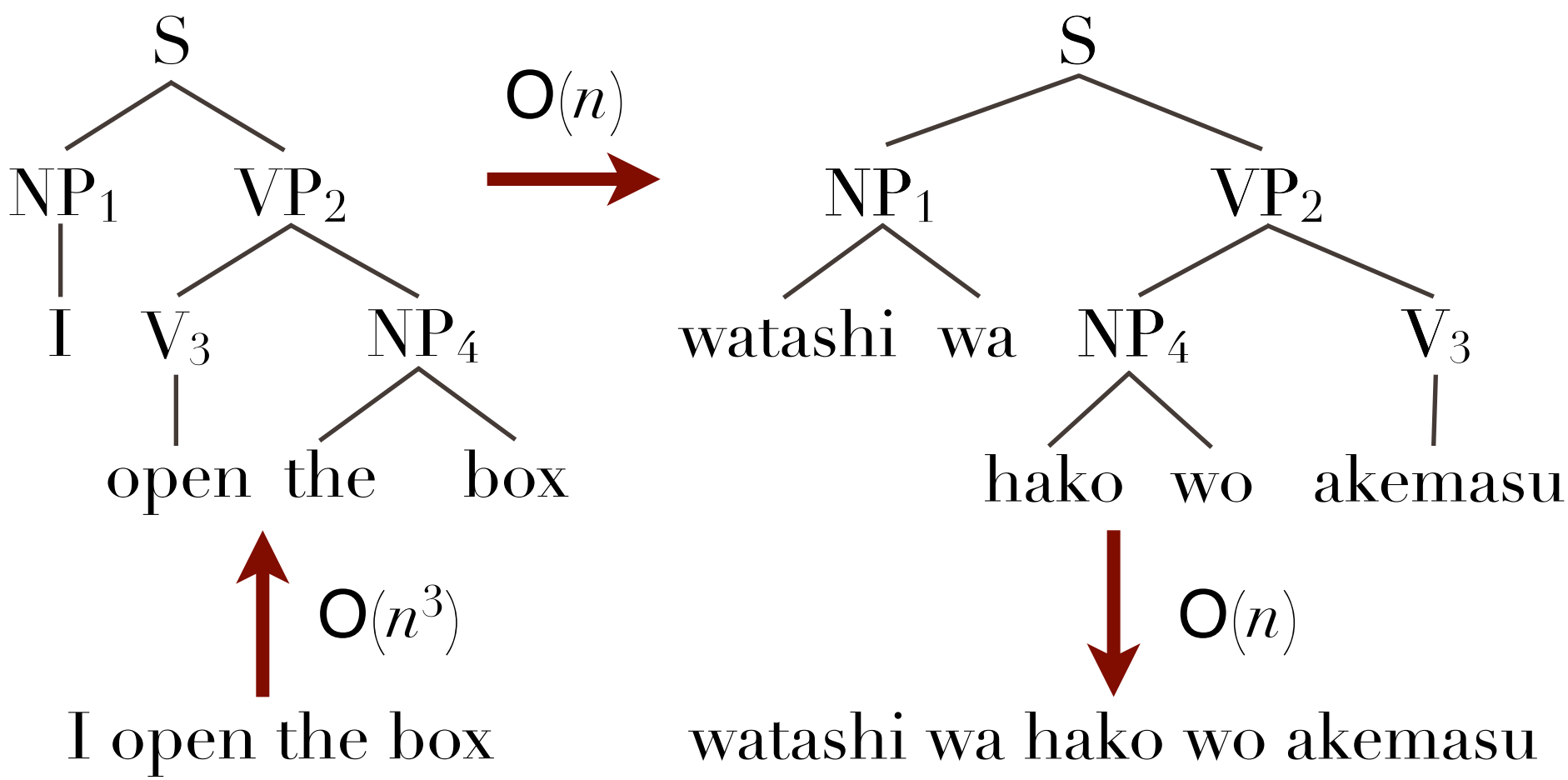


$O(n^3)$ ways of matching

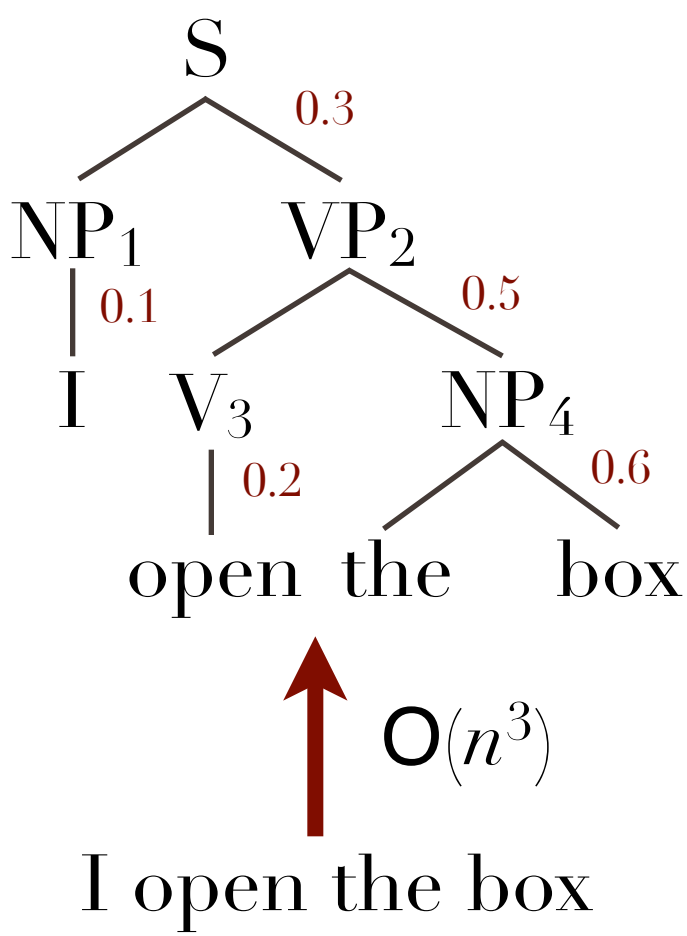
Translation



Translation

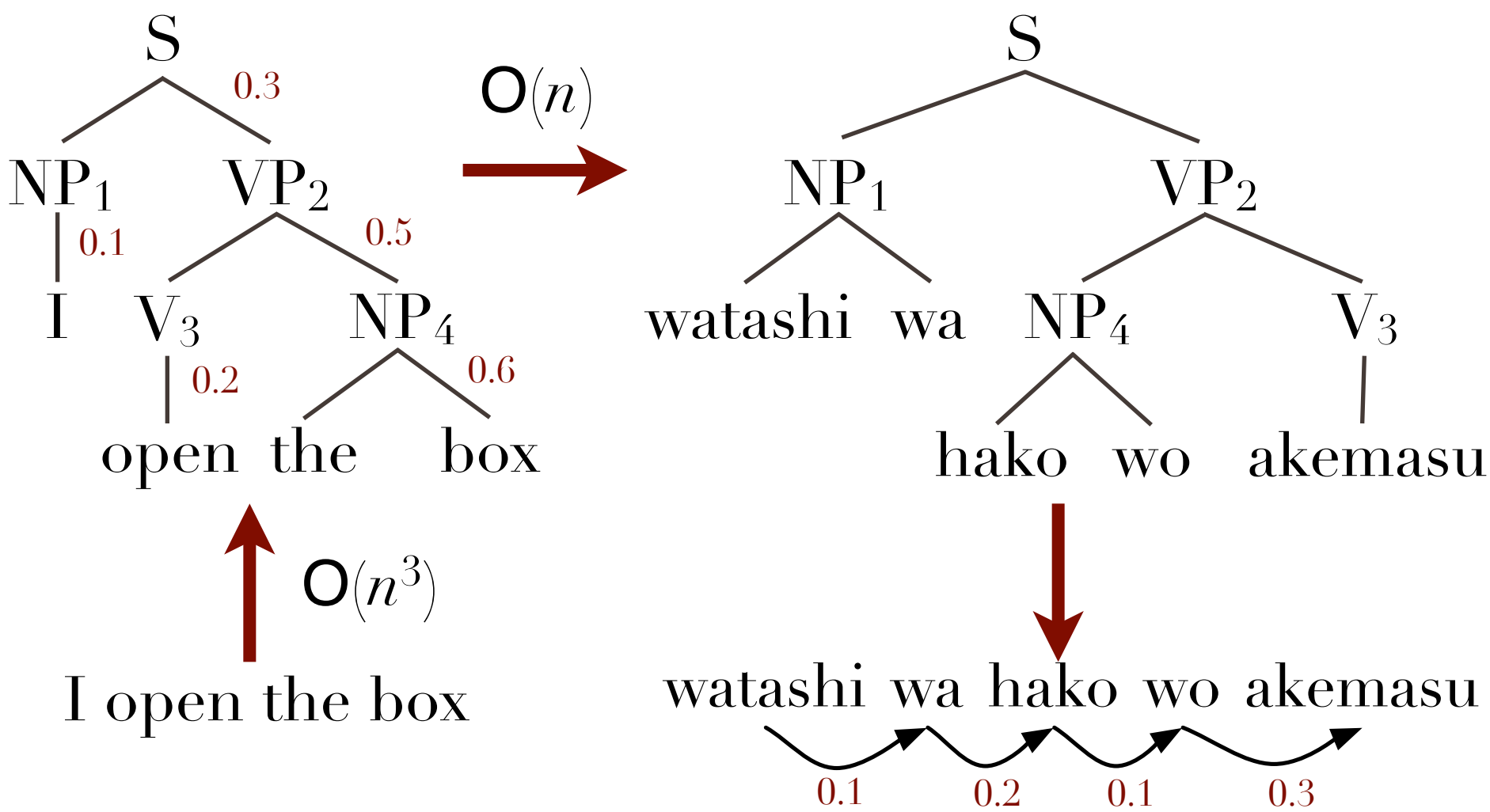


Finding the best translation

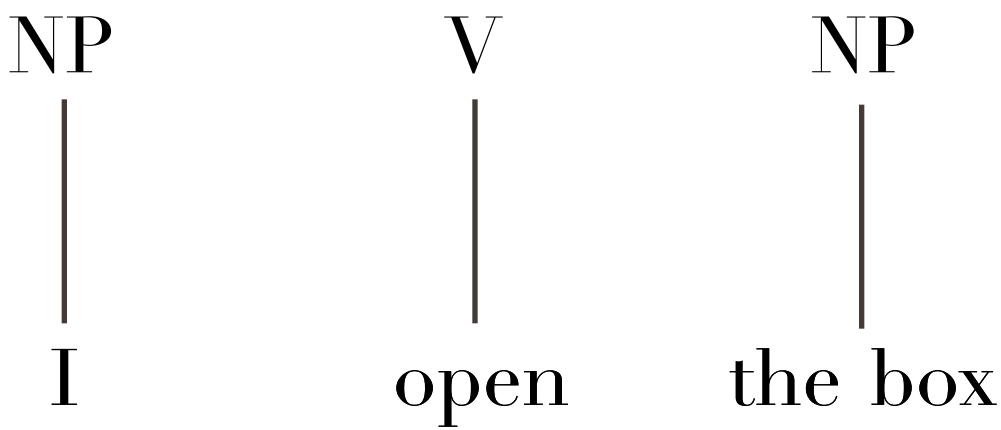


- Just like PCFG parsing
- *n*-best translations is efficient with the right tricks

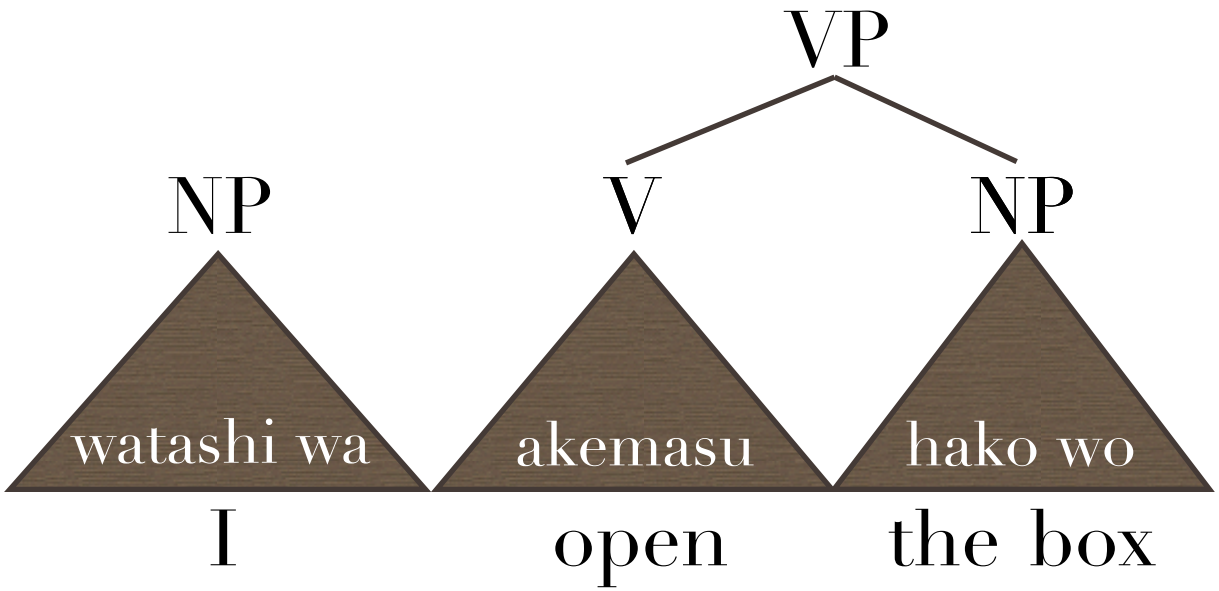
Bigram language model



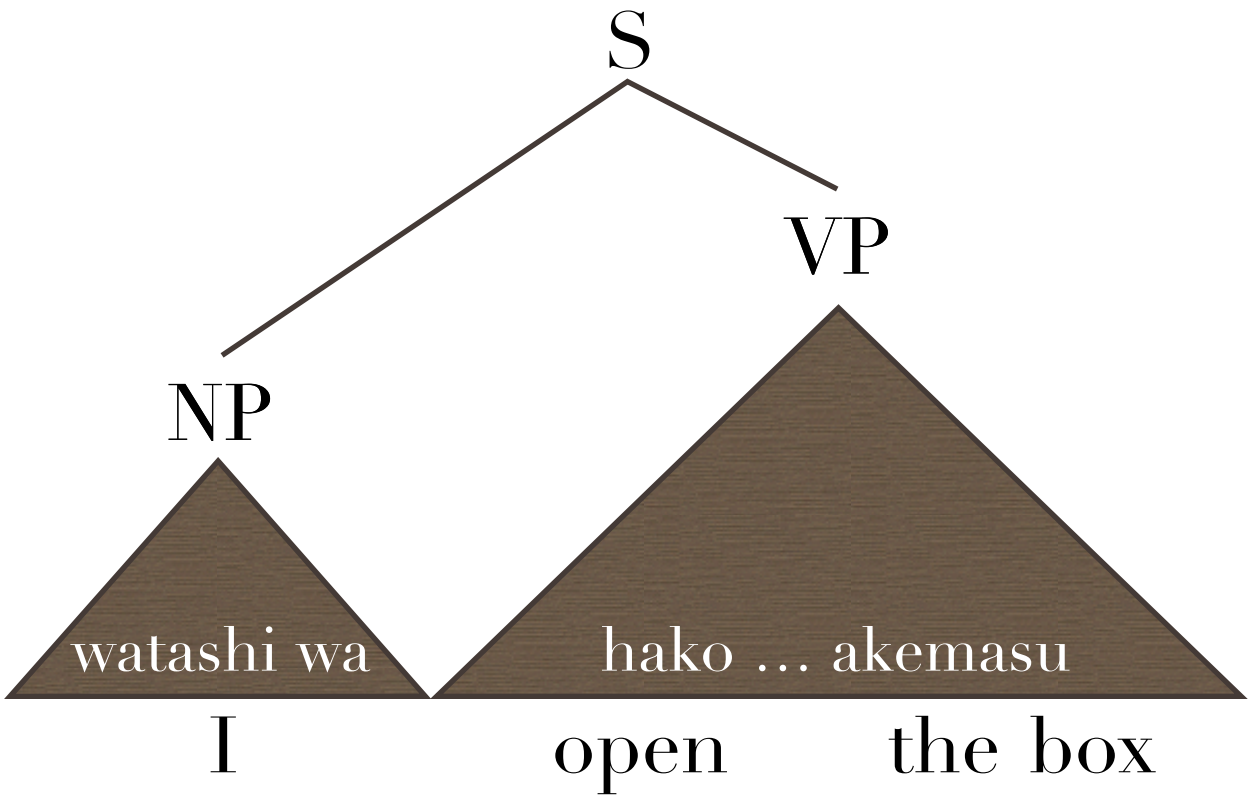
Translation with a LM



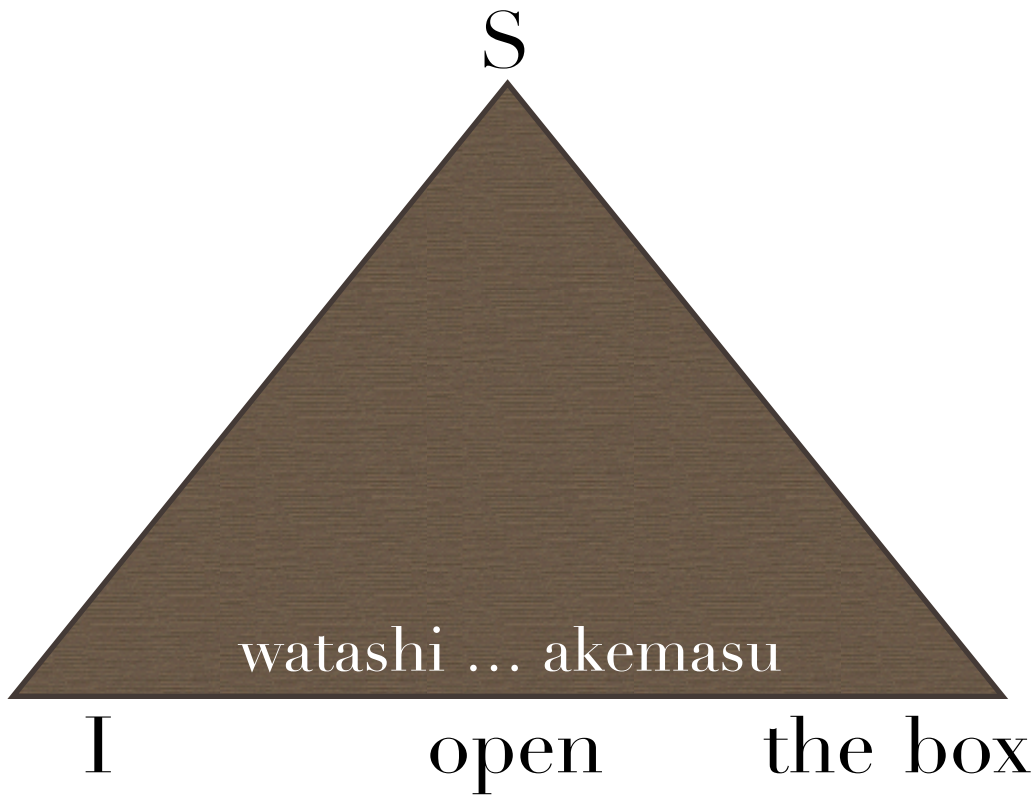
Translation with a LM



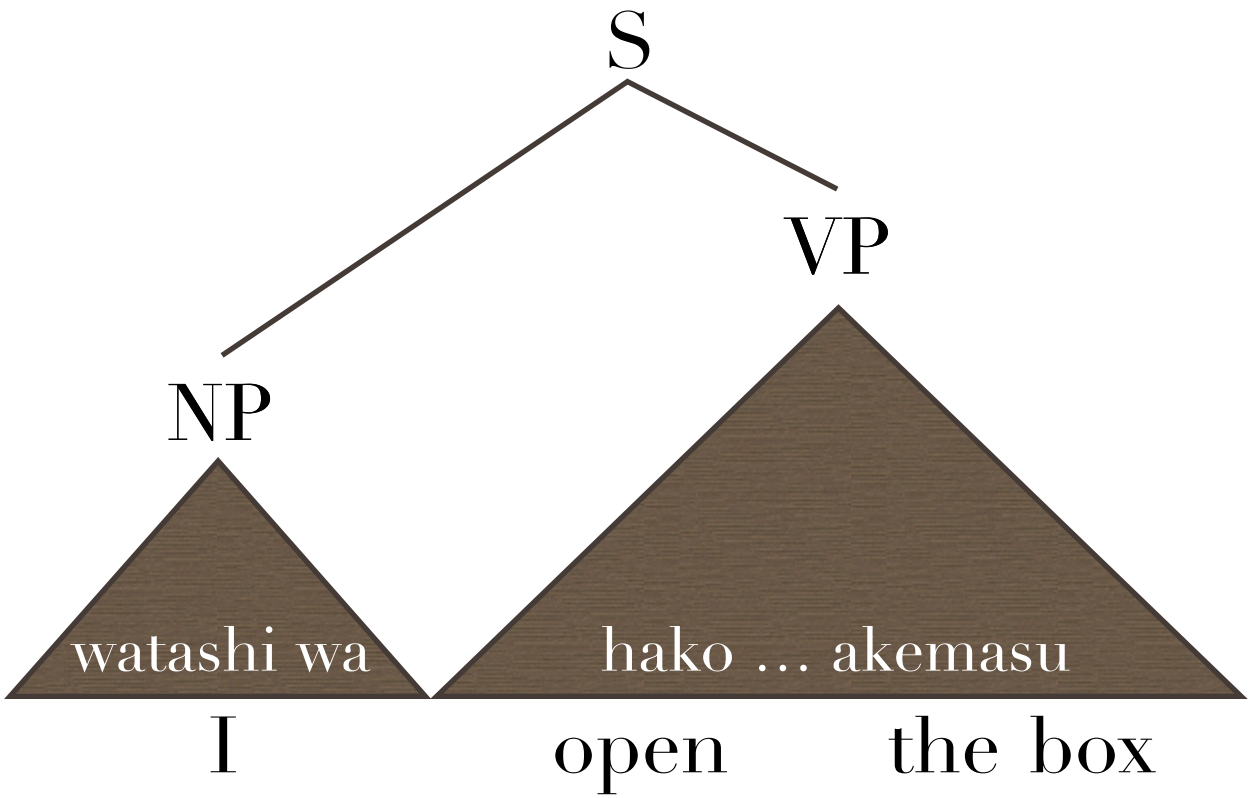
Translation with a LM



Translation with a LM

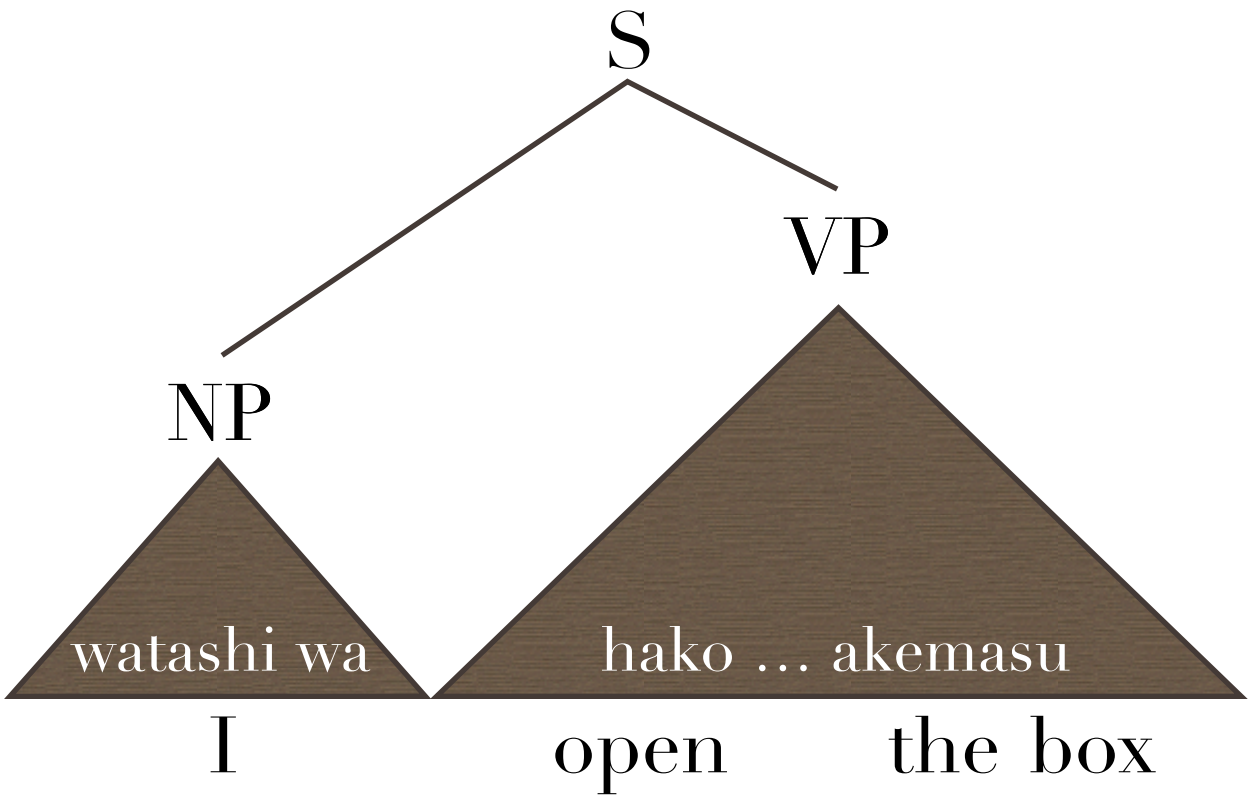


Translation with a LM



$O(n^3 V^{4(g-1)})$ combinations

Translation with a LM



$O(n^{3+4(g-1)})$ combinations

Summary

- Synchronous CFG/TSG allows translation based on hierarchical/syntactic structure
- Binarization not free
- Can be extracted from parallel data
- Tree-to-tree translation rare
- LM adds complexity to decoding