Syntax-based Translation

Part 2: Synchronous Grammars

Machine Translation
Lecture 12

Instructor: Chris Callison-Burch
TAs: Mitchell Stern, Justin Chiu

Website: mt-class.org/penn

Goals

- Revisit why people thought syntax would not help machine translation
- Learn about Synchronous Context Free Grammars
- Introduce notation, and basic algorithm
- Understand how we learn SCFGs from bitexts
- Get a sense of the different flavors of SCFGs
  - Hiero
  - SAMT

Every time I fire a linguist my performance goes up

- Longstanding debate about whet
  information can help statistical tra
- Two camps

Syntax will improve translation

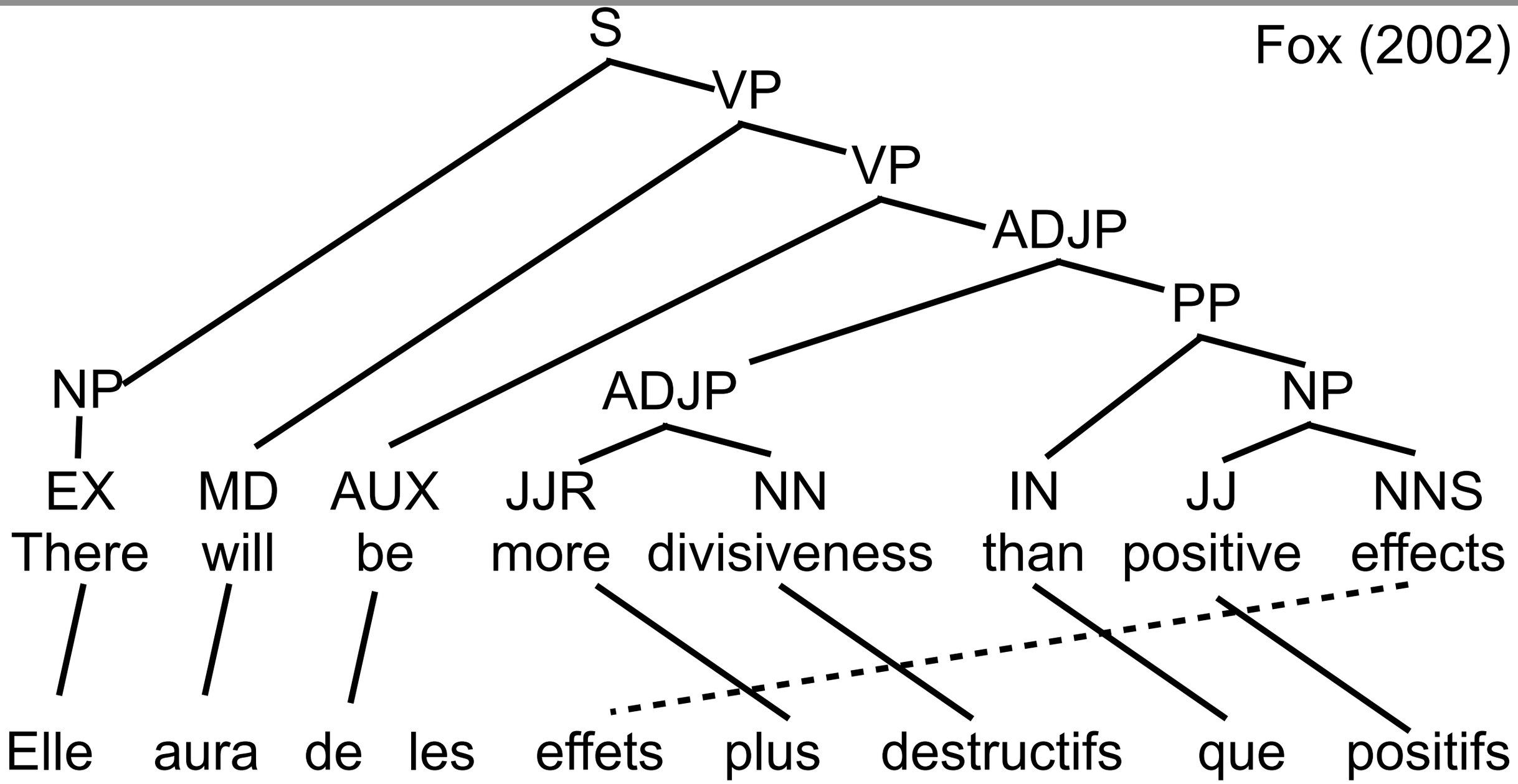
Simpler data-driven models will always win

Syntax is bad for translation

- The IBM Models were the dominant approach to SMT from the `90s until mid 2000s
  - Eschewed linguistic information
- A number of studies cast doubt on whether linguistic info could help SMT
  - Fox (2002) showed that "phrasal cohesion" was less common than assumed across even related languages
  - Koehn et al (2003) empirically demonstrated that syntactically motivated phrases made PBMT worse

Phrases aren't coherent in bitexts

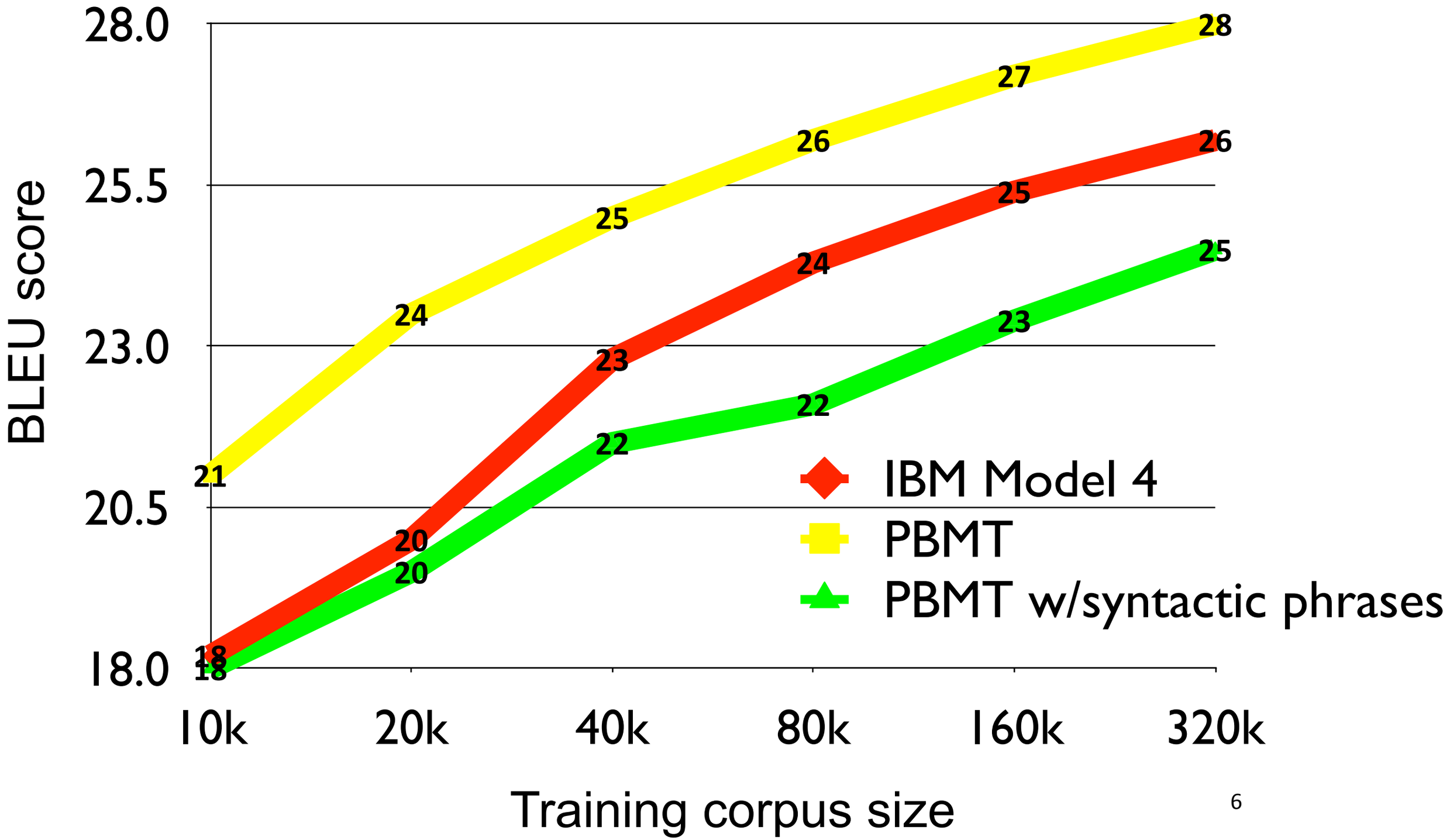
Fox (2002)



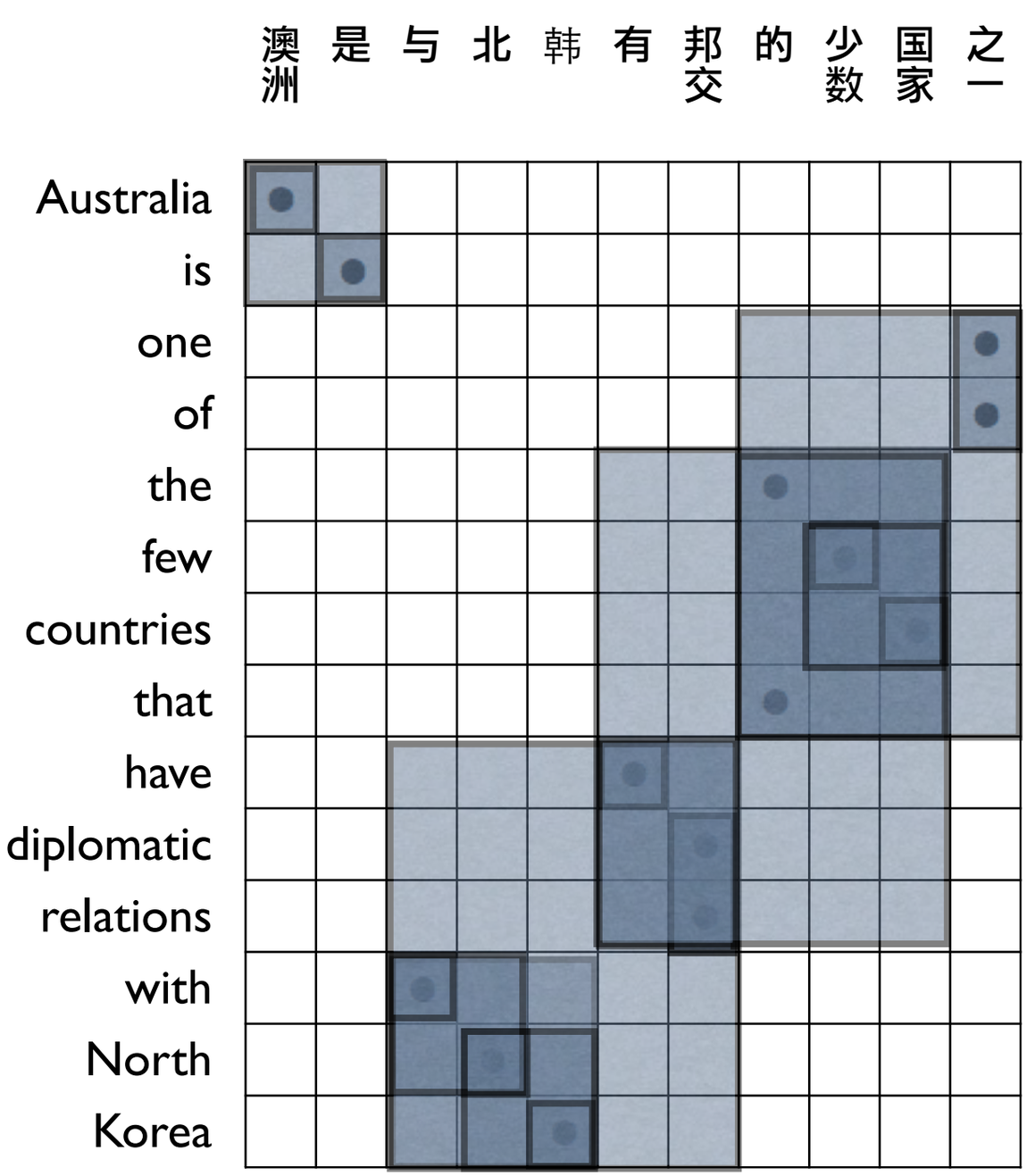
Gloss: *It will have effects more destructive than positive*

Ouch! Syntax hurts!

Koehn et al (2003)



Extracting phrase pairs



澳洲, Australia
是, is
之一, one of
少数, few
国家, countries
有, have
邦交, diplomatic relations
与, with
北, North
韩, Korea

澳洲是, Australia is
少数 国家, few countries
有邦交, have diplomatic relations
与北, with North
北韩, North Korea

的少数 国家, the few countries that
与北韩, with North Korea

之一的少数 国家, one of the the few countries that
与北韩 有邦交, have diplomatic relations with North Korea

有邦交 的少数 国家, the few countries that have diplomatic relations

Why does it hurt to limit to constituents?

- Massively reduces the inventory of phrases that can be used as translation units
- Eliminates non-constituent phrases, many of which are quite useful
  - *there are*
  - *note that*
  - *according to*

So, what should we do?

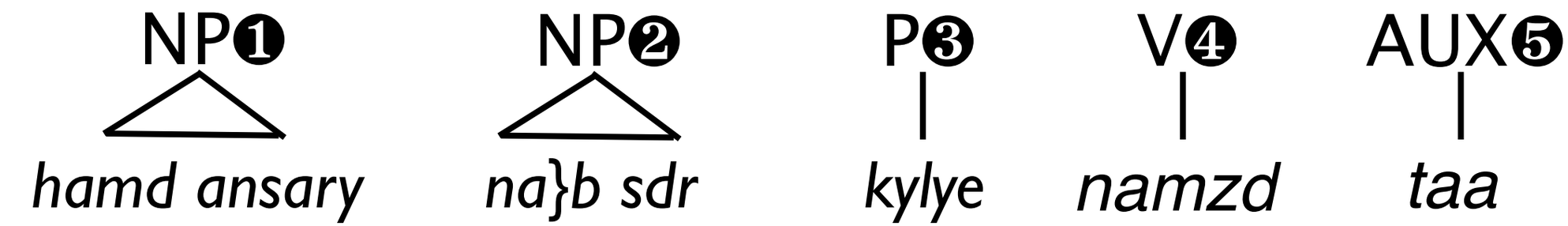
- Drop syntax from statistical machine translation, since syntax is a bad fit for the data
- Abandon conventional English syntax and move towards more robust grammars that adapt to the parallel training corpus
- Maintain English syntax but design different syntactic models

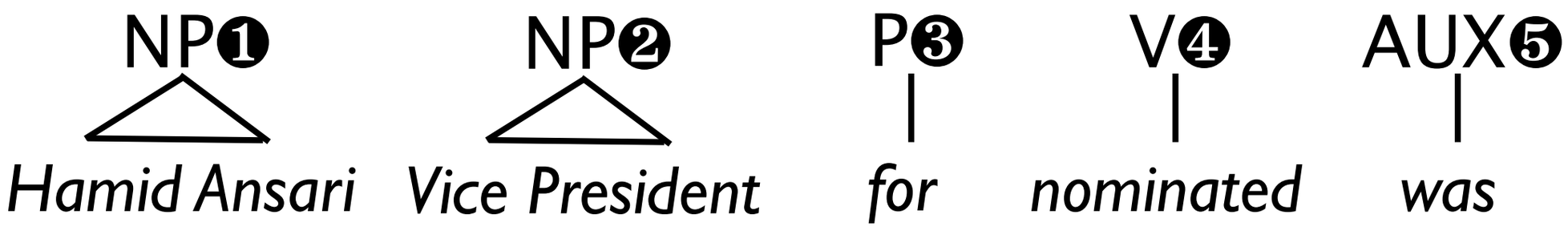
Synchronous Context Free Grammars

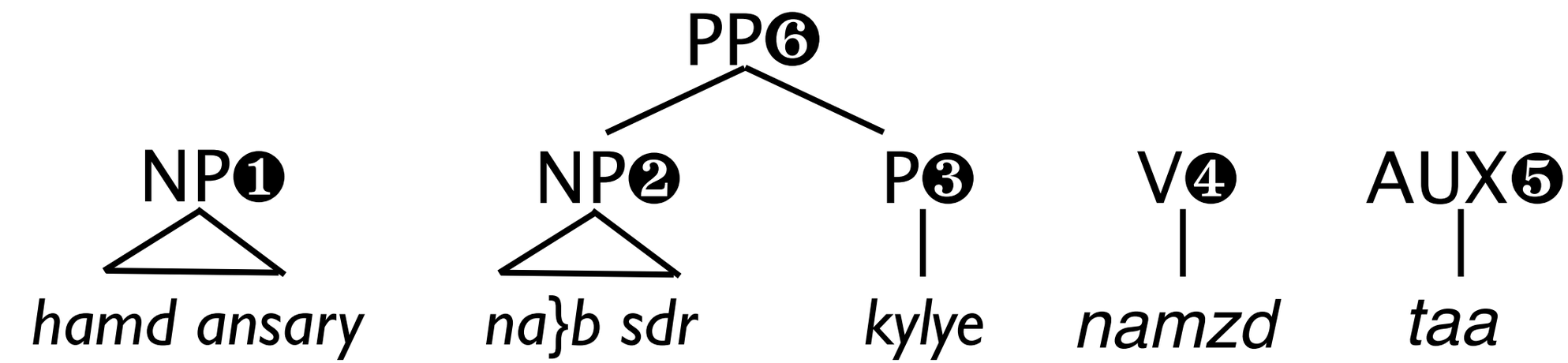
- A common way of representing syntax in NLP is through context free grammars
- Synchronous context free grammars generate pairs of corresponding strings
- Can be used to describe translation and re-ordering between languages
- SCFGs translate sentences by parsing them

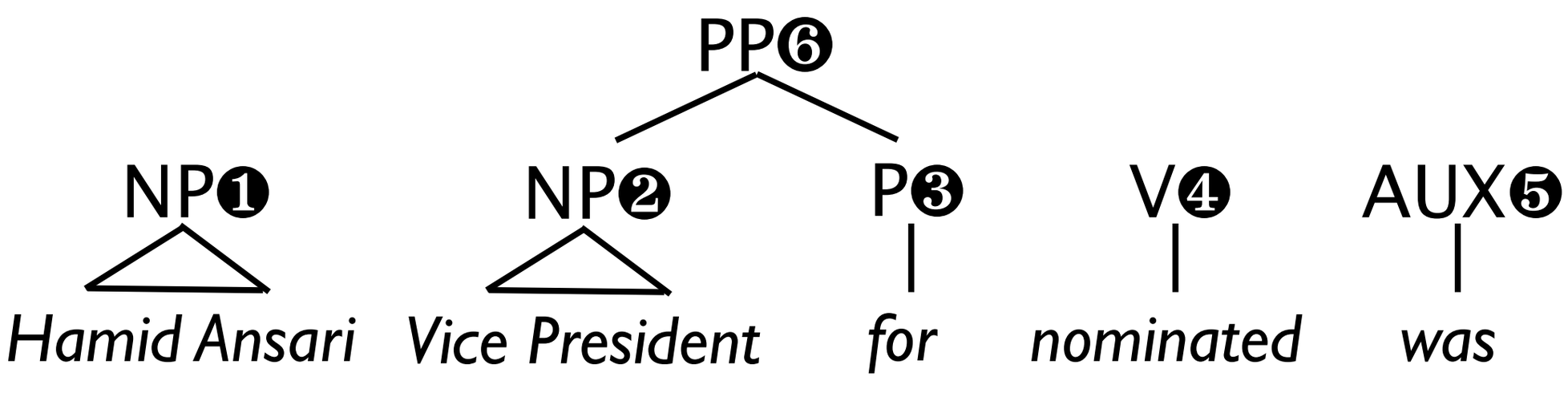
Example SCFG for Urdu

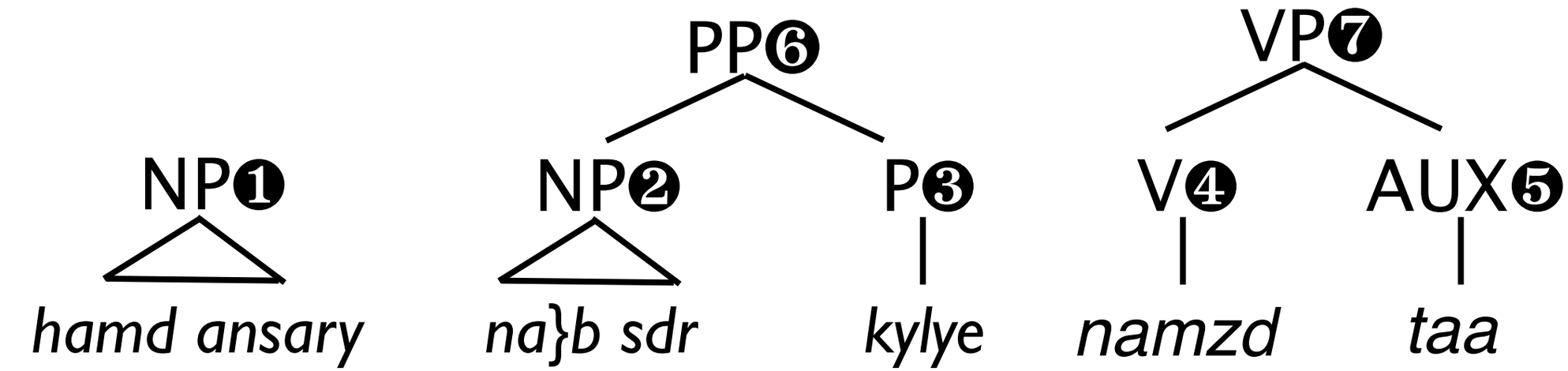
	Urdu	English
S →	NP① VP②	NP① VP②
VP→	PP① VP②	VP② PP①
VP→	V① AUX②	AUX② V①
PP →	NP① P②	P② NP①
NP →	*hamd ansary*	*Hamid Ansari*
NP →	*na}b sdr*	*Vice President*
V →	*namzd*	*nominated*
P →	*kylye*	*for*
AUX →	*taa*	*was*

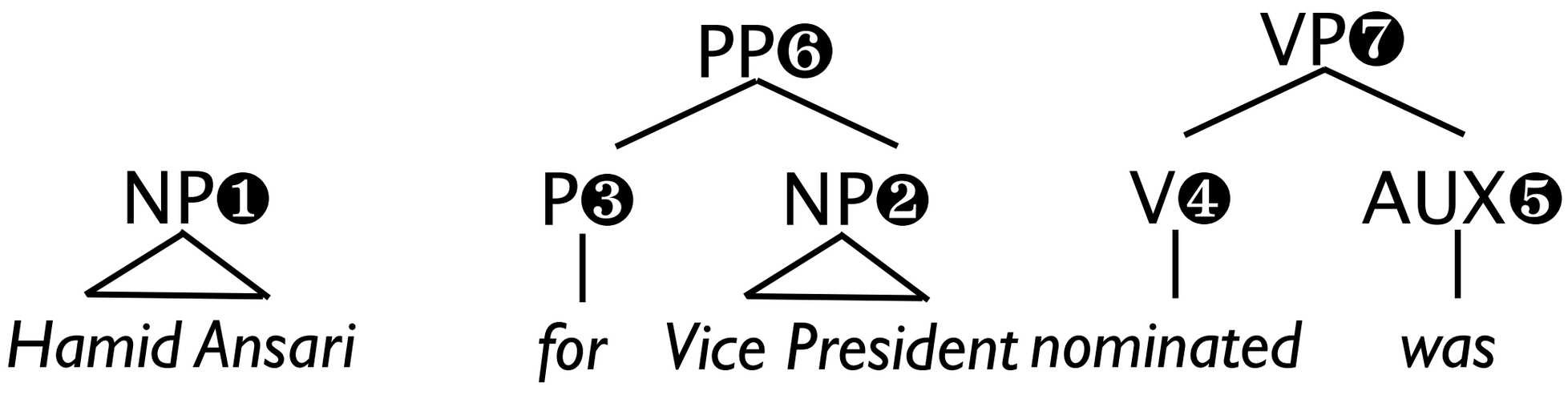

NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa

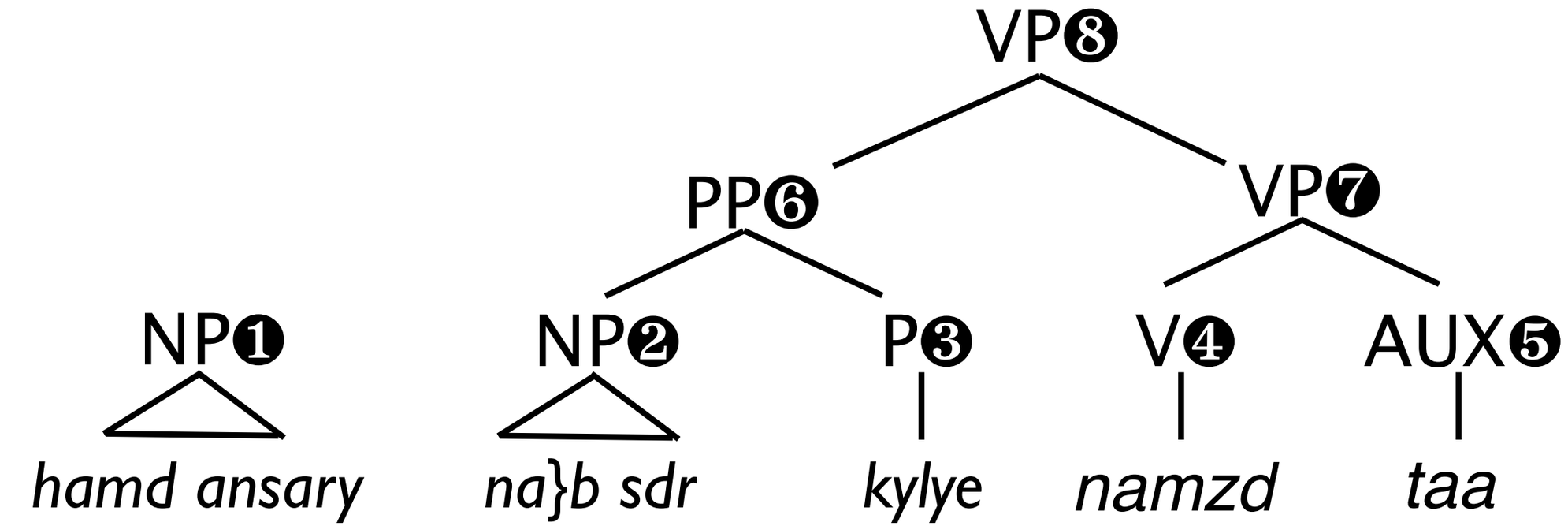

NP❶
NP❷
P❸
V❹
AUX❺
Hamid Ansari
Vice President
for
nominated
was

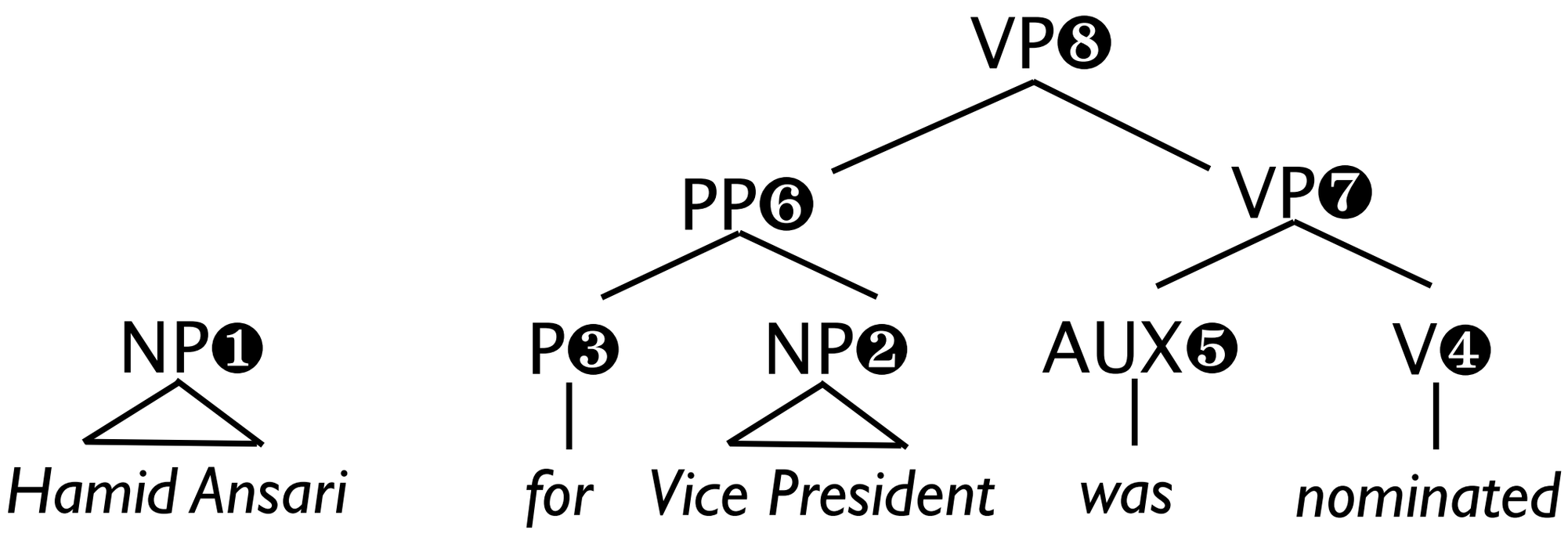

PP❻
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa

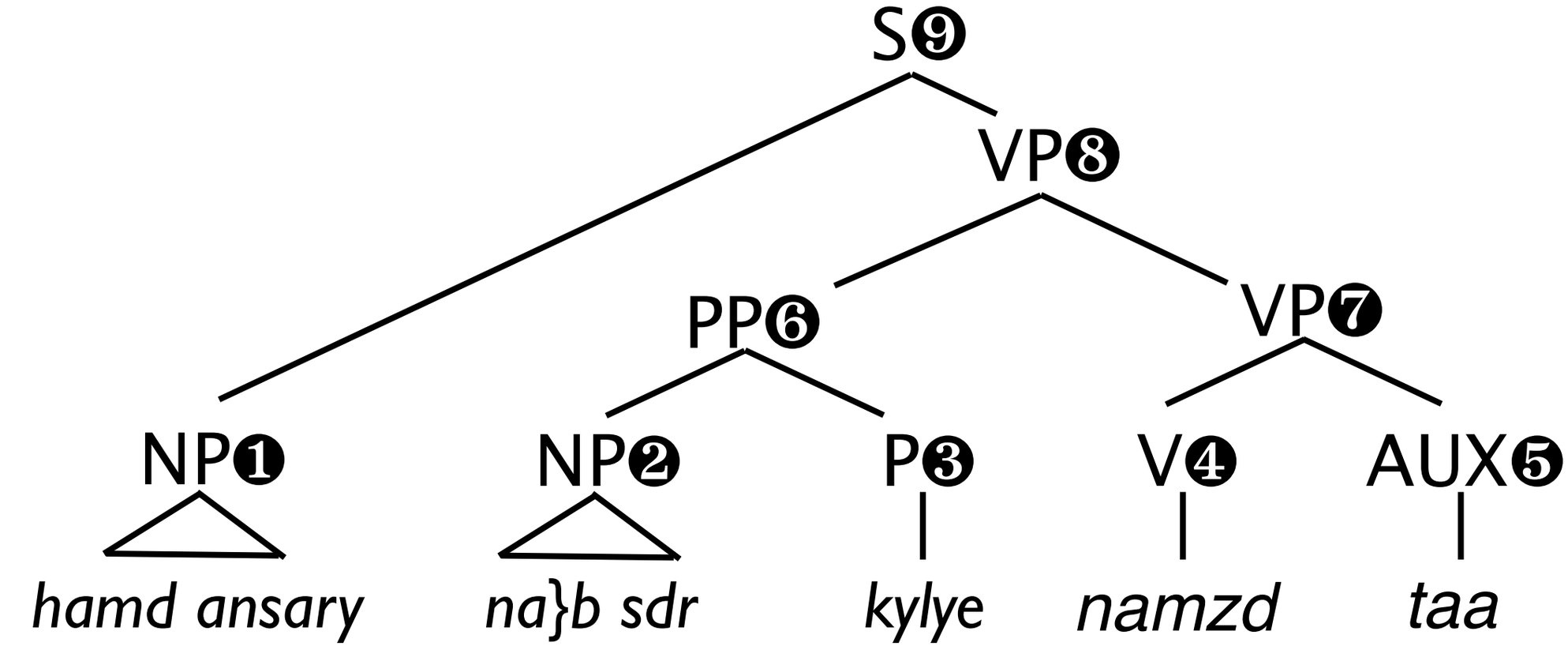

PP❻
NP❶
NP❷
P❸
V❹
AUX❺
Hamid Ansari
Vice President
for
nominated
was

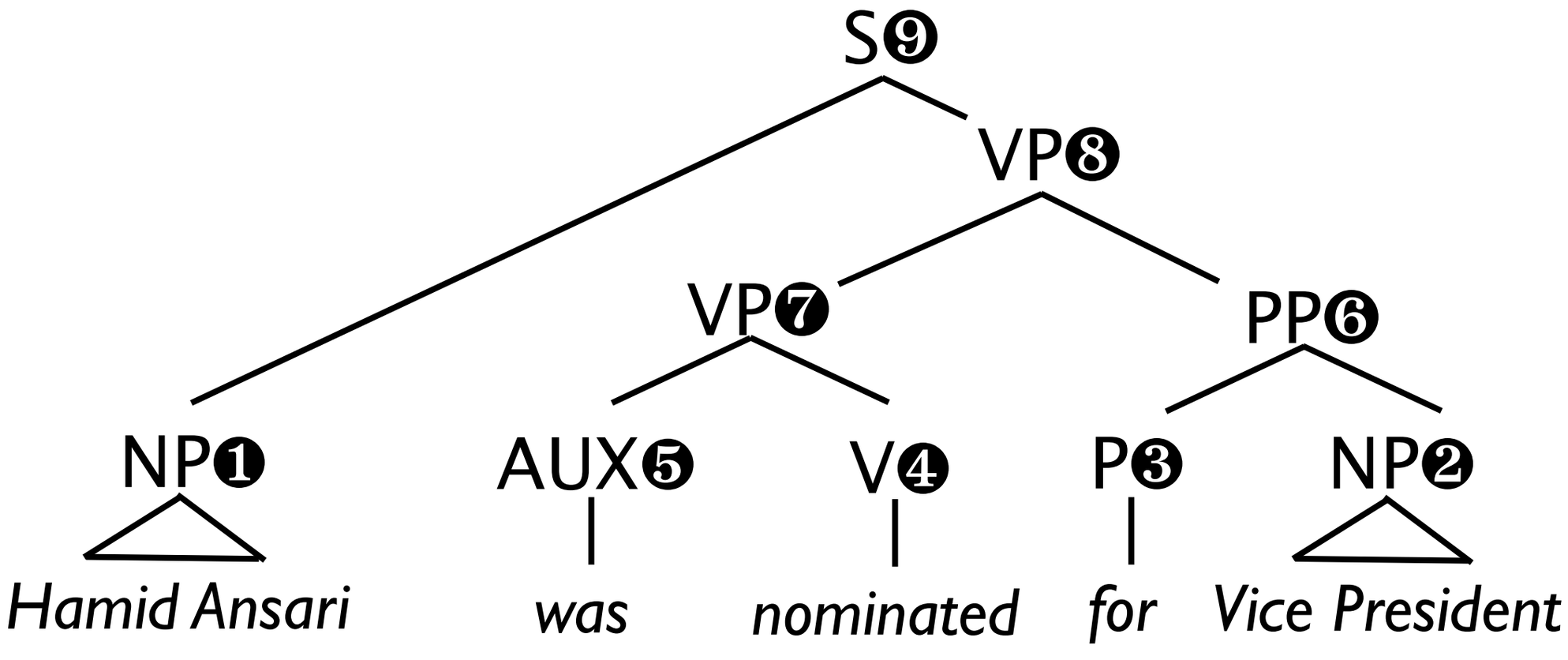

PP❻
VP❼
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa


PP❻
VP❼
NP❶
P❸
NP❷
V❹
AUX❺
Hamid Ansari
for
Vice President
nominated
was


VP❽
PP❻
VP❼
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa


VP❽
PP❻
VP❼
NP❶
P❸
NP❷
AUX❺
V❹
Hamid Ansari
for
Vice President
was
nominated


S❾
VP❽
PP❻
VP❼
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa

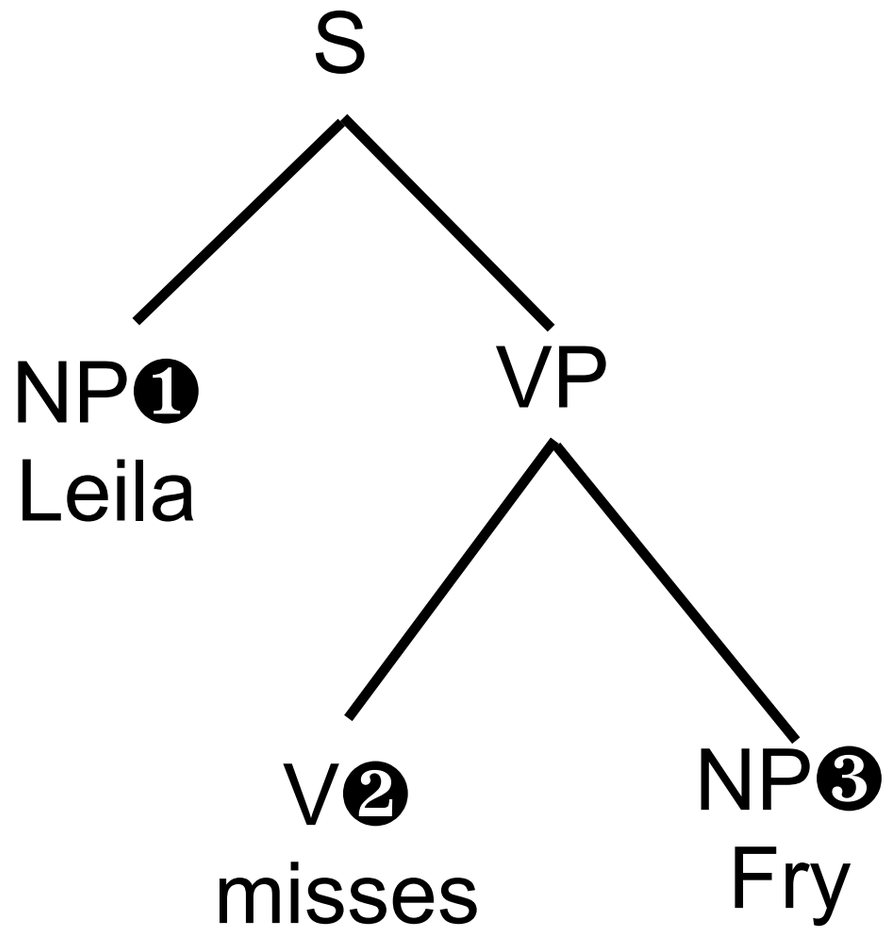

S❾
VP❽
VP❼
PP❻
NP❶
AUX❺
V❹
P❸
NP❷
Hamid Ansari
was
nominated
for
Vice President

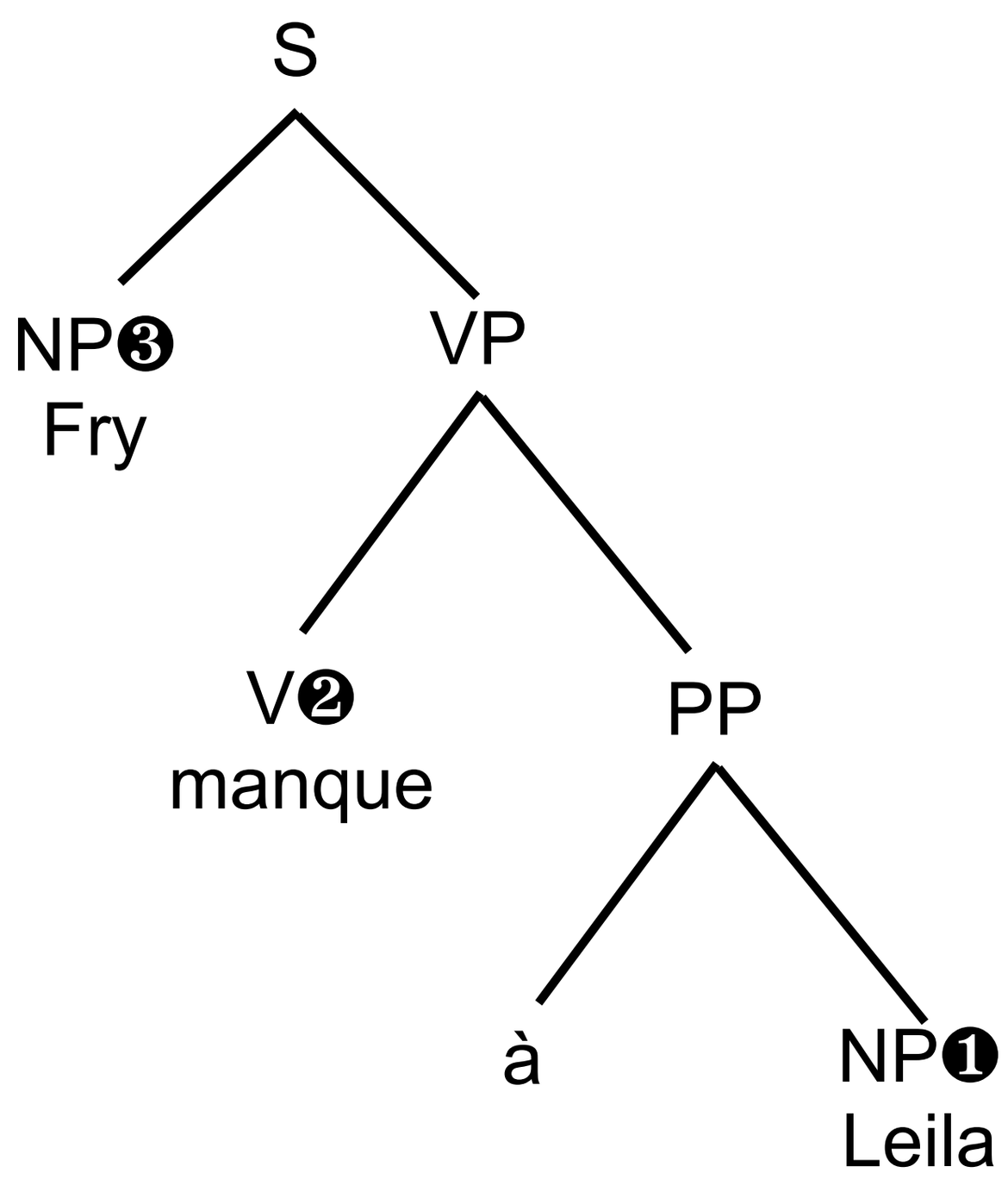
Discussion: Do you like SCFG?

- In what ways are SCFGs better for describing reordering than what we saw before?
- Is this a good model of how languages relate?
- What do you think of the synchronous requirement?

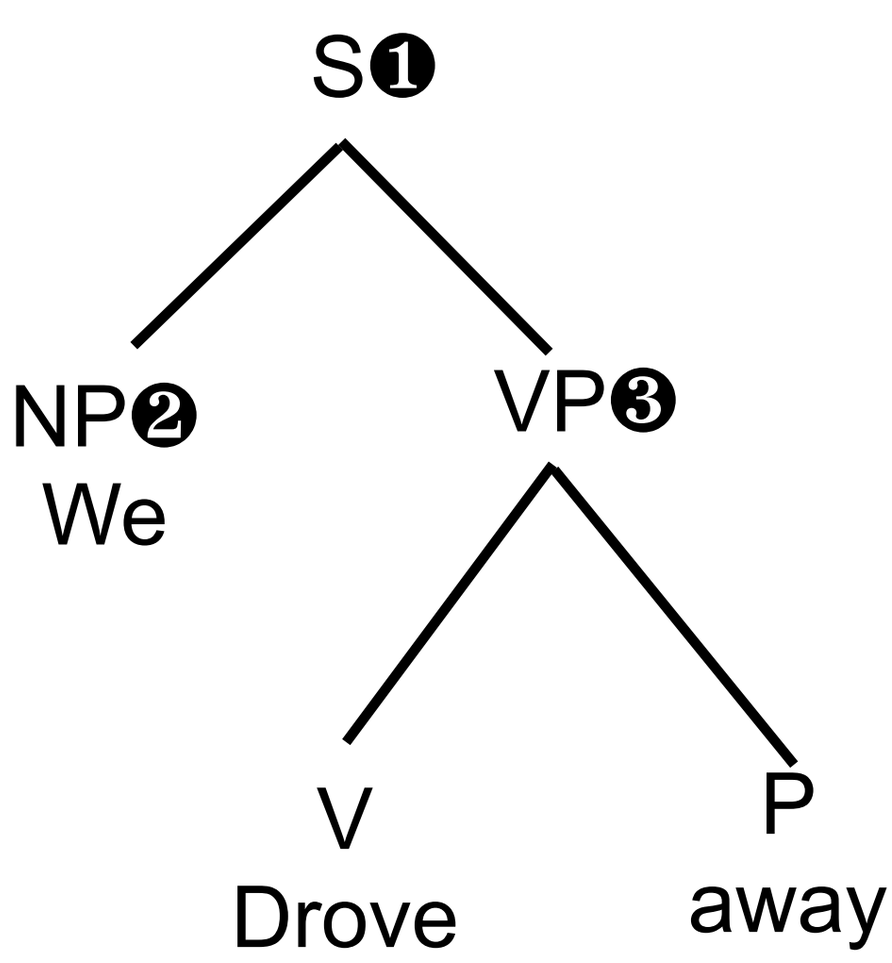
  (Discuss with your neighbor)

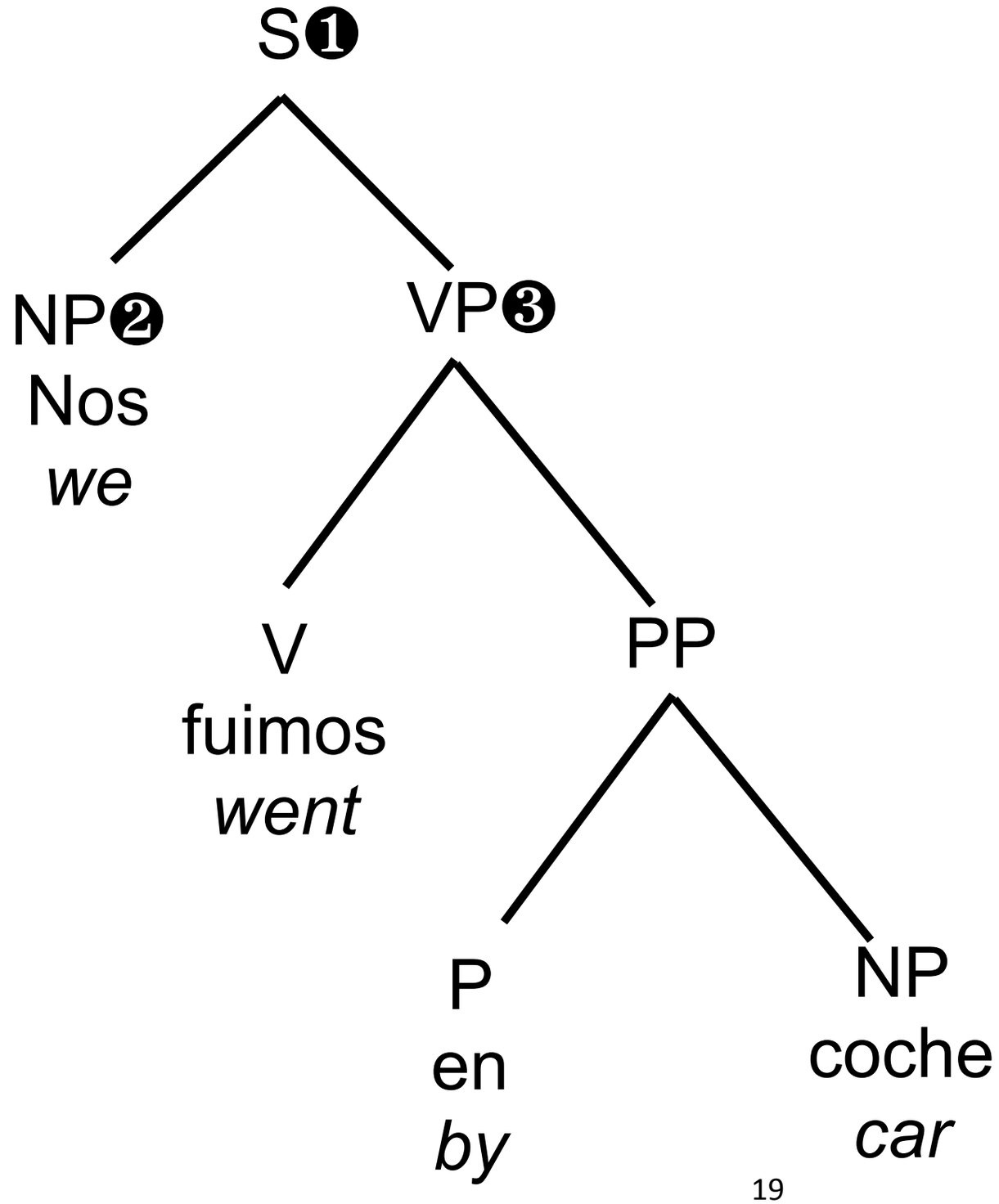
Sometimes languages are mismatched



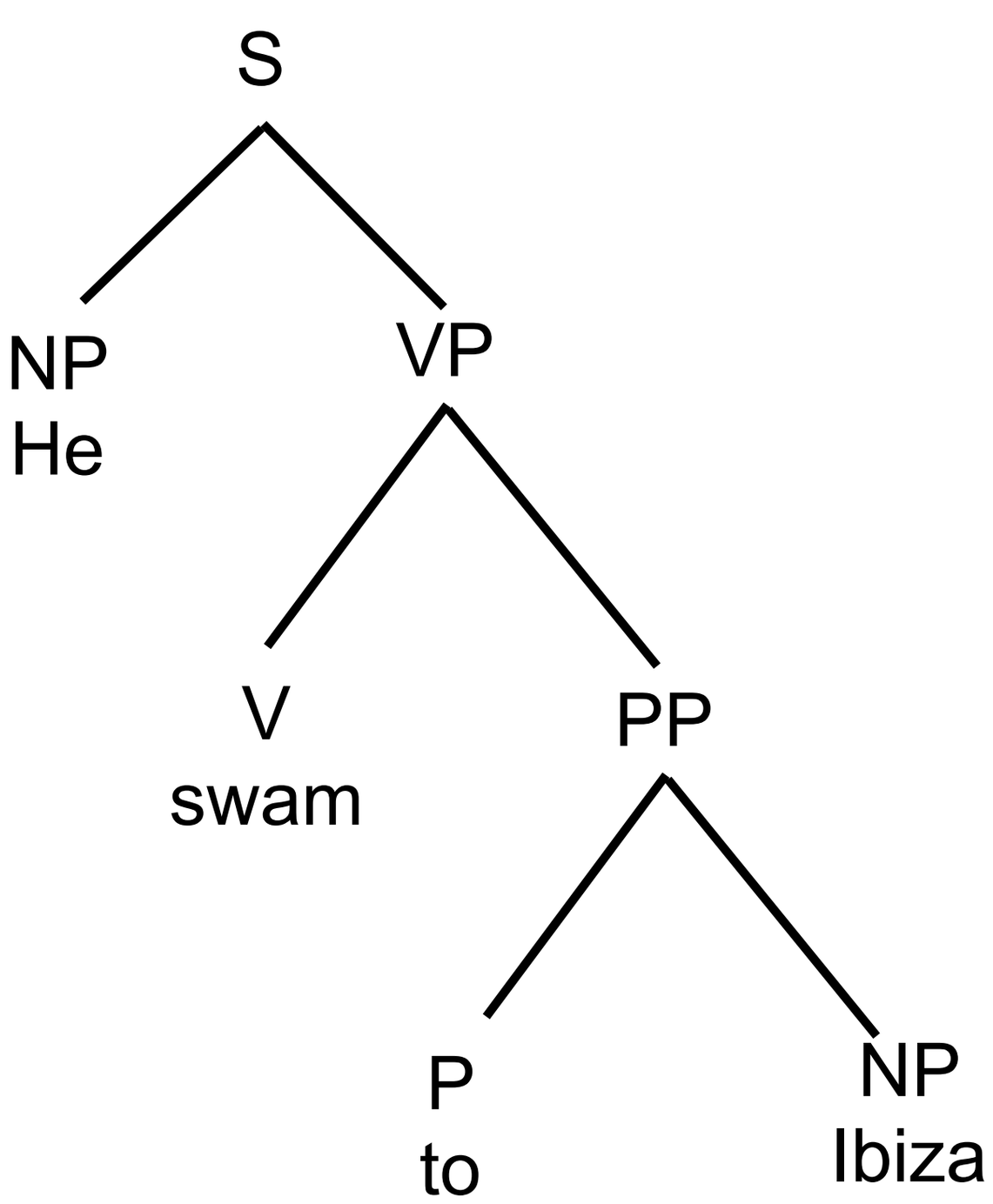


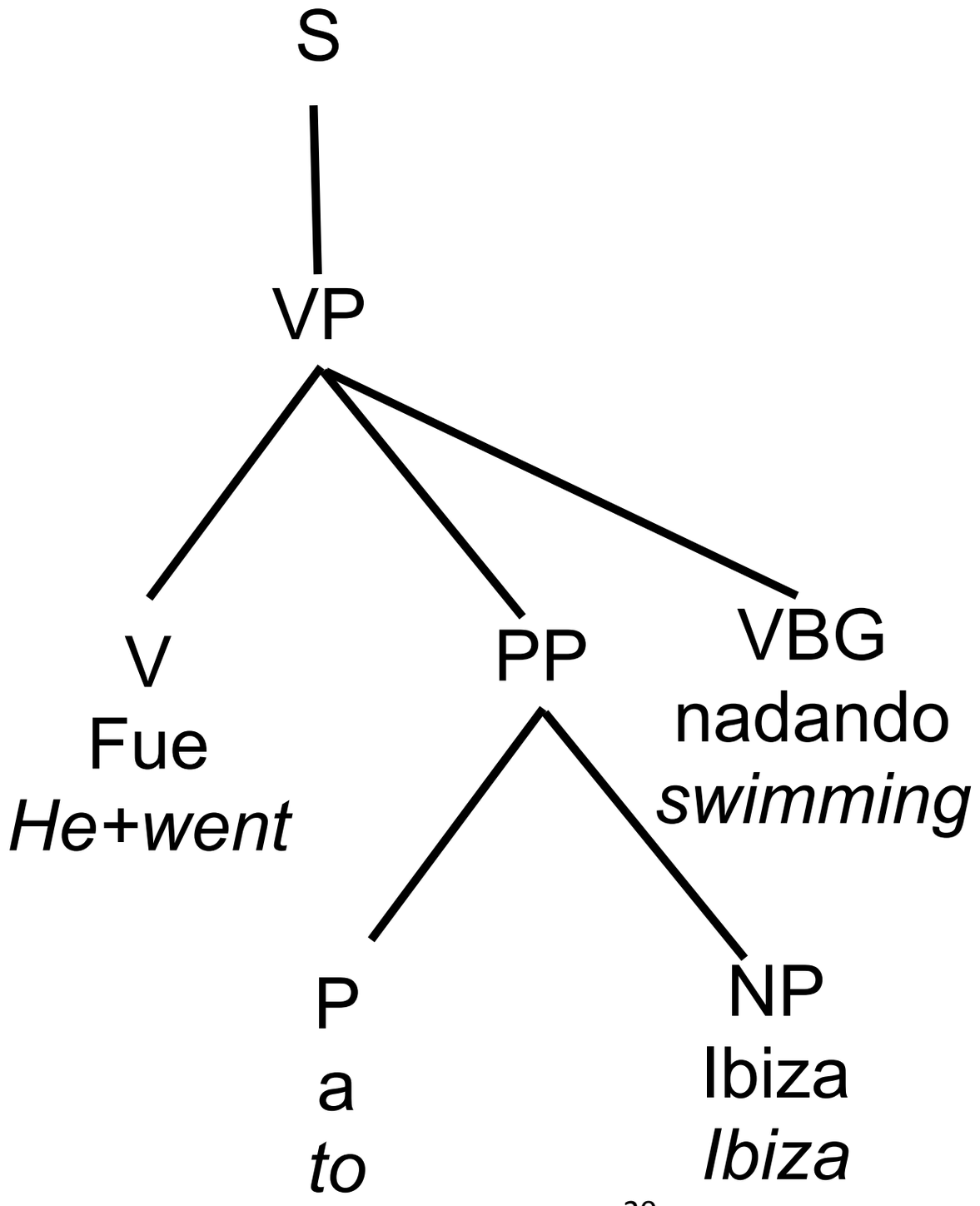
Spanish motion verb





Spanish motion verb, pro-drop

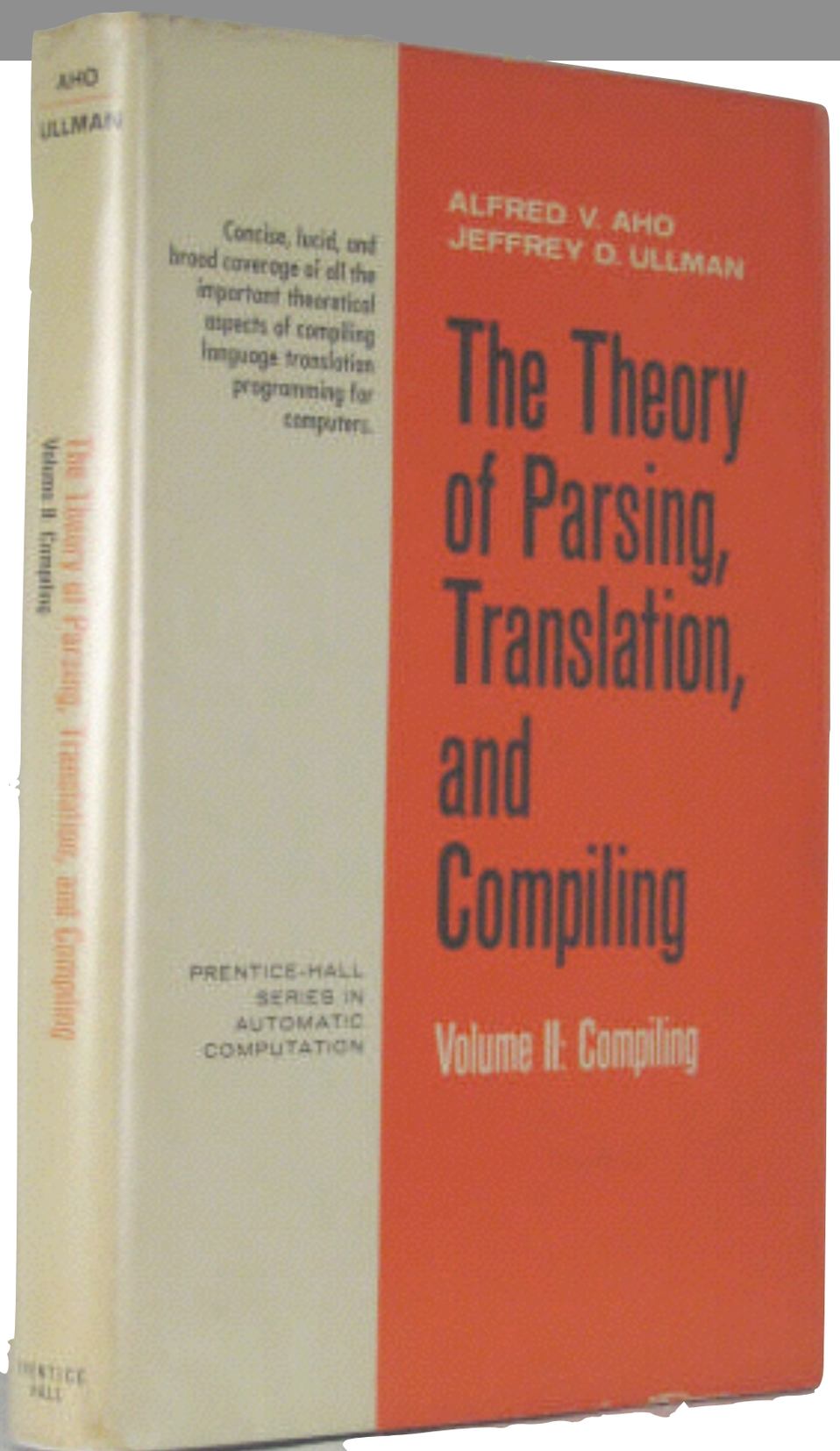




We are going to use them anyway

- SCFGs are mismatched with some linguistic phenomena
- But they have nice formal properties and well-defined algorithms

Formal definition of SCFGs



- Aho and Ullman worked all of this out in the `60s and `70s
- Compiler theory

Formal definition of SCFGs

hamd ansary, na}b sdr, namzd, kylye, taa

for, Hamid Ansari, nominated, Vice President, was

A synchronous context free grammar is defined by a tuple

S, NP, VP, PP, P, V, AUX

S

$$G = \langle N, T_S, T_T, R, S \rangle$$

- Where
  - N is a shared set of non-terminal symbols
  - T_S is the set of source language terminals
  - T_T is the set of target language terminals
  - R is a set of production rules
  - $S \in N$, designated as the goal state

Formal definition of SCFGs

- Each production rule has the form

$$X \rightarrow \langle \alpha, \beta, \sim, w \rangle$$

- Where
  - $X \in N$
  - $\alpha \in (N \cup T_S)^*$
  - $\beta \in (N \cup T_T)^*$
  - ~ is a one-to-one correspondence between the non terminals in α and β
  - w is a weight assigned to the rule

Algorithms for SCFGs

- Translation with SCFGs is done via parsing
- How do we write an algorithm for parsing?
- One way to do it is as a deductive proof system

The CKY Parsing Algorithm

Axioms	$\dfrac{}{A \rightarrow \alpha}$	for all $(A \rightarrow \alpha) \in R$
Inference rules	$\dfrac{A \rightarrow w_{i+1}}{[A, i, i+1]}$ $\dfrac{[B, i, j] \quad [C, j, k] \quad A \rightarrow BC}{[A, i, k]}$	
Goal	$[S, 0, n]$	

Axioms

S →	NP VP	
VP→	PP VP	
VP→	V AUX	
PP →	NP P	
NP →	*hamd ansary*	
NP →	*na}b sdr*	
V →	*namzd*	
P →	*kylye*	
AUX →	*taa*	

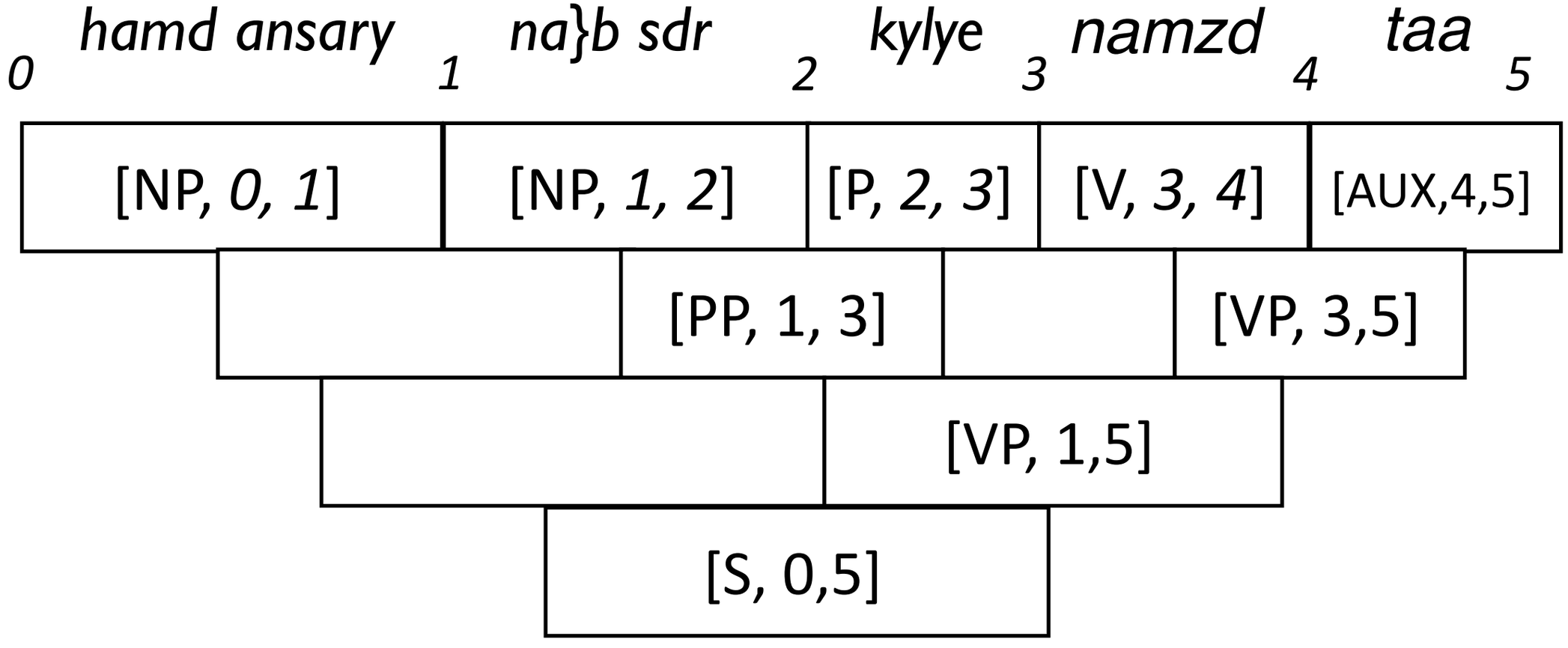
Inference rule used

[illegible] [illegible]

[illegible]

Goal

[S, 0, 5]



The CKY Parsing Algorithm

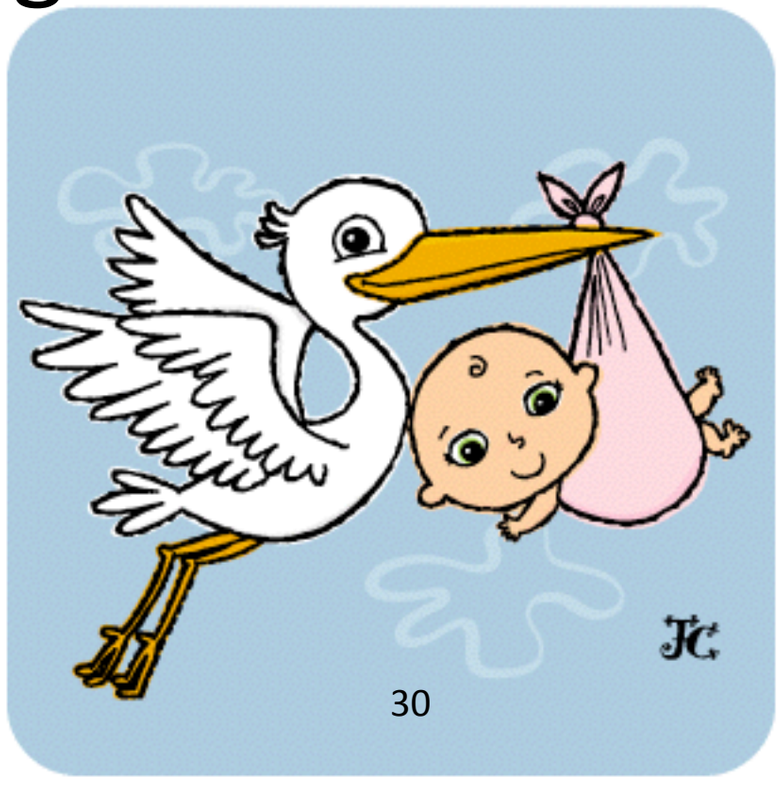
Axioms	$\dfrac{}{A \rightarrow \alpha}$	for all $(A \rightarrow \alpha) \in R$
Inference rules	$\dfrac{A \rightarrow w_{i+1}}{[A, i, i+1]}$ $\dfrac{[B, i, j] \quad [C, j, k] \quad A \rightarrow BC}{[A, i, k]}$	
Goal	$[S, 0, n]$	

The CKY Translation Algorithm

Axioms	$\frac{}{\mathbf{A \rightarrow \alpha,\ \beta}}$	for all $(\mathbf{A \rightarrow \alpha,\ \beta}) \in R$
Inference rules	$\frac{A \rightarrow w_{i+1}}{[A, i, i+1]}$ $\frac{[B, i, j] \quad [C, j, k] \quad A \rightarrow BC}{[A, i, k]}$	
Goal	$[S, 0, n]$	

Where do grammars come from?

- Great! We now have
  - a formalism for describing the relationship between two languages,
  - an algorithm for producing translations
- All we need now is a synchronous grammar
- Where do grammars come from?
- Well, when two languages love each other very much...



Data-driven grammar extraction

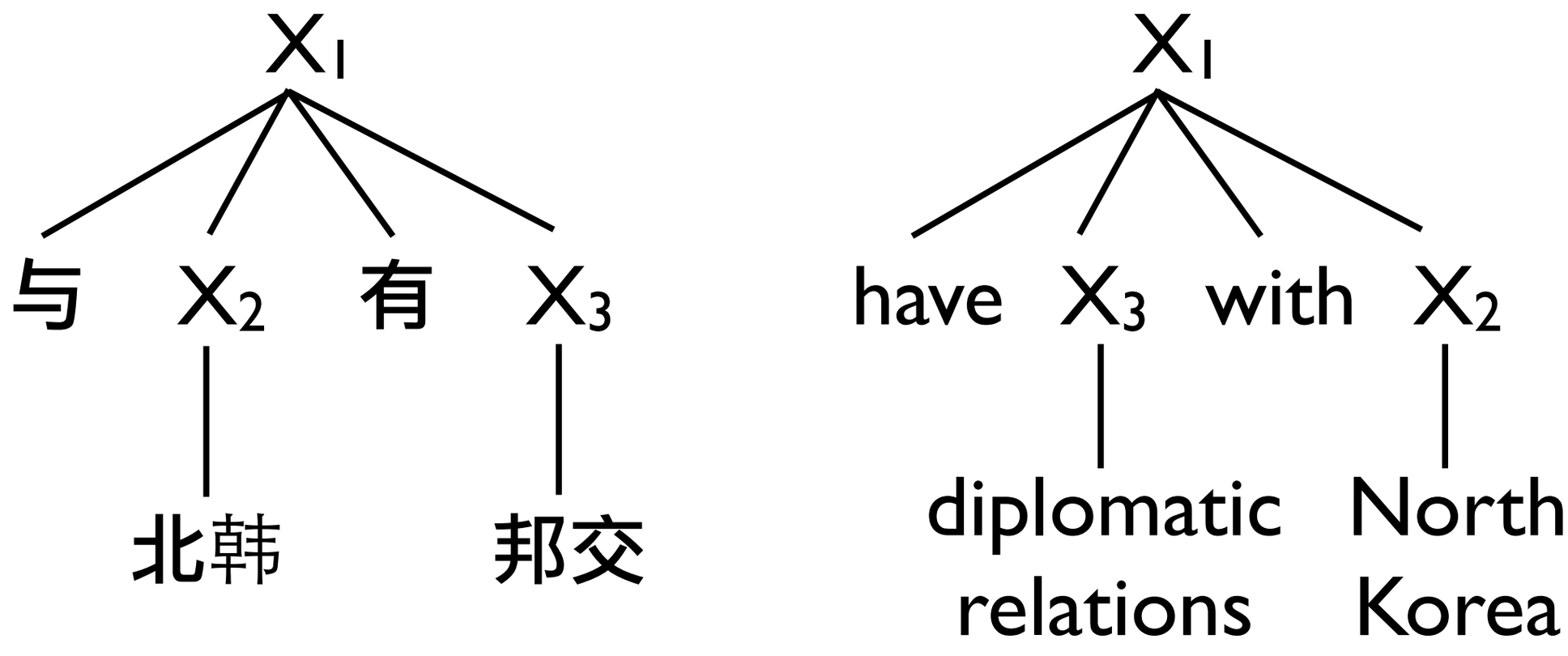
- Grammar rules are not written by hand, they are extracted from bilingual parallel corpora

Arabic	English
فالتعذيب لا يزال يمارس على نطاق واسع	Torture is still being practised on a wide scale.
وتتم عمليات الاعتقال والاحتجاز دون سبب بصورة روتينية	Arrest and detention without cause take place routinely.
وحان وقت التحلى بالبصيرة والشجاعة السياسية .	This is a time for vision and political courage
. . .	. . .

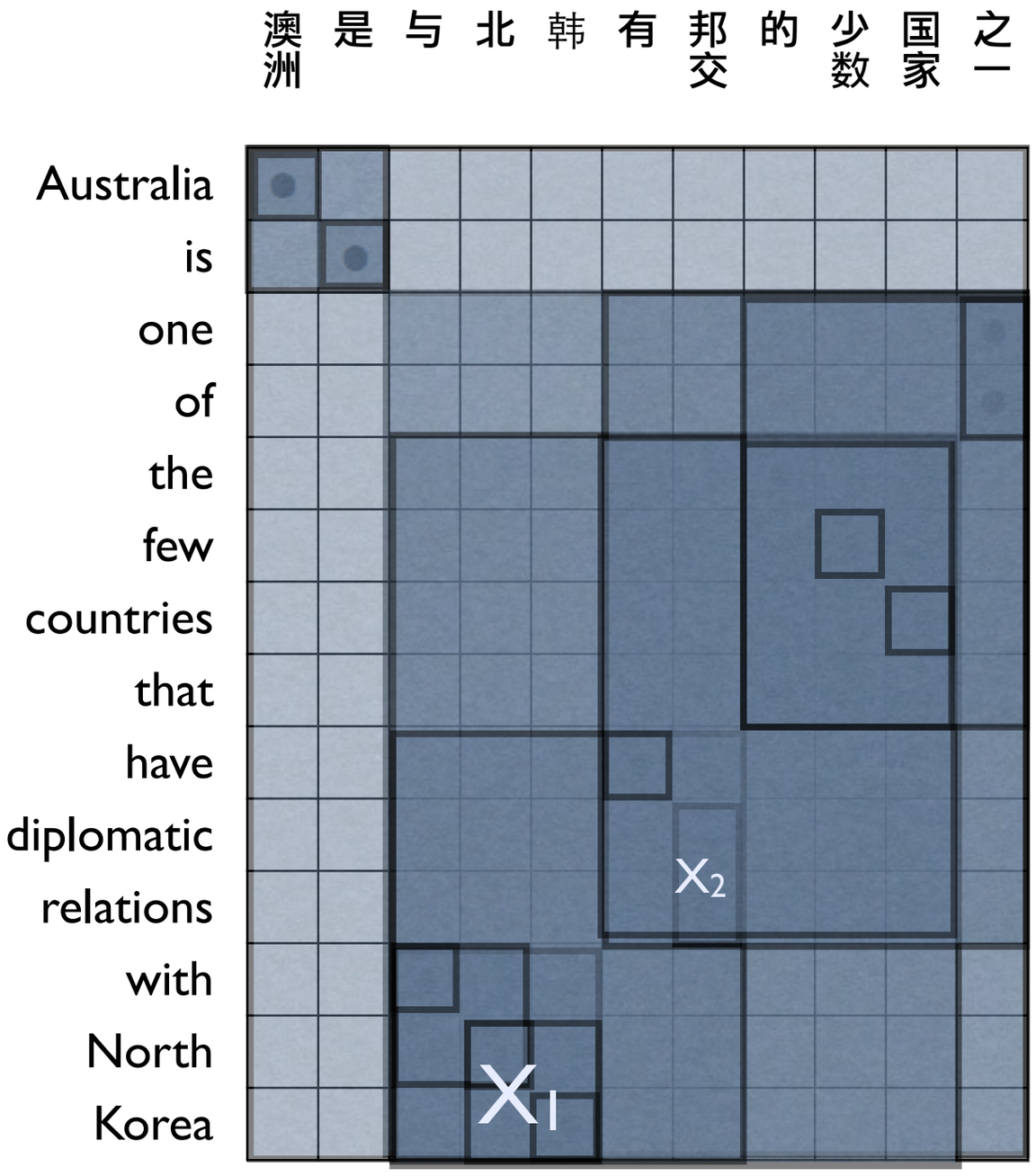
Chinese	English
我国 能源 原材料 工业 生产 大幅度 增长 .	China's energy and raw materials production up.
非国大 要求 阻止 更 多 被 拘留 人员 死亡 .	ANC calls for steps to prevent deaths in police custody .
. . .	. . .

Hiero-style SCFG rules

- Most common type of SCFG in SMT is Hiero which has rules w/one non-terminal symbol
- Not as nice as linguistically motivated rules, does not capture the reordering in Urdu



Extracting Hiero rules



X → 与 北 韩 有 邦交,
have diplomatic relations with North Korea

X → 邦交,
diplomatic relations

X → 北 韩,
North Korea

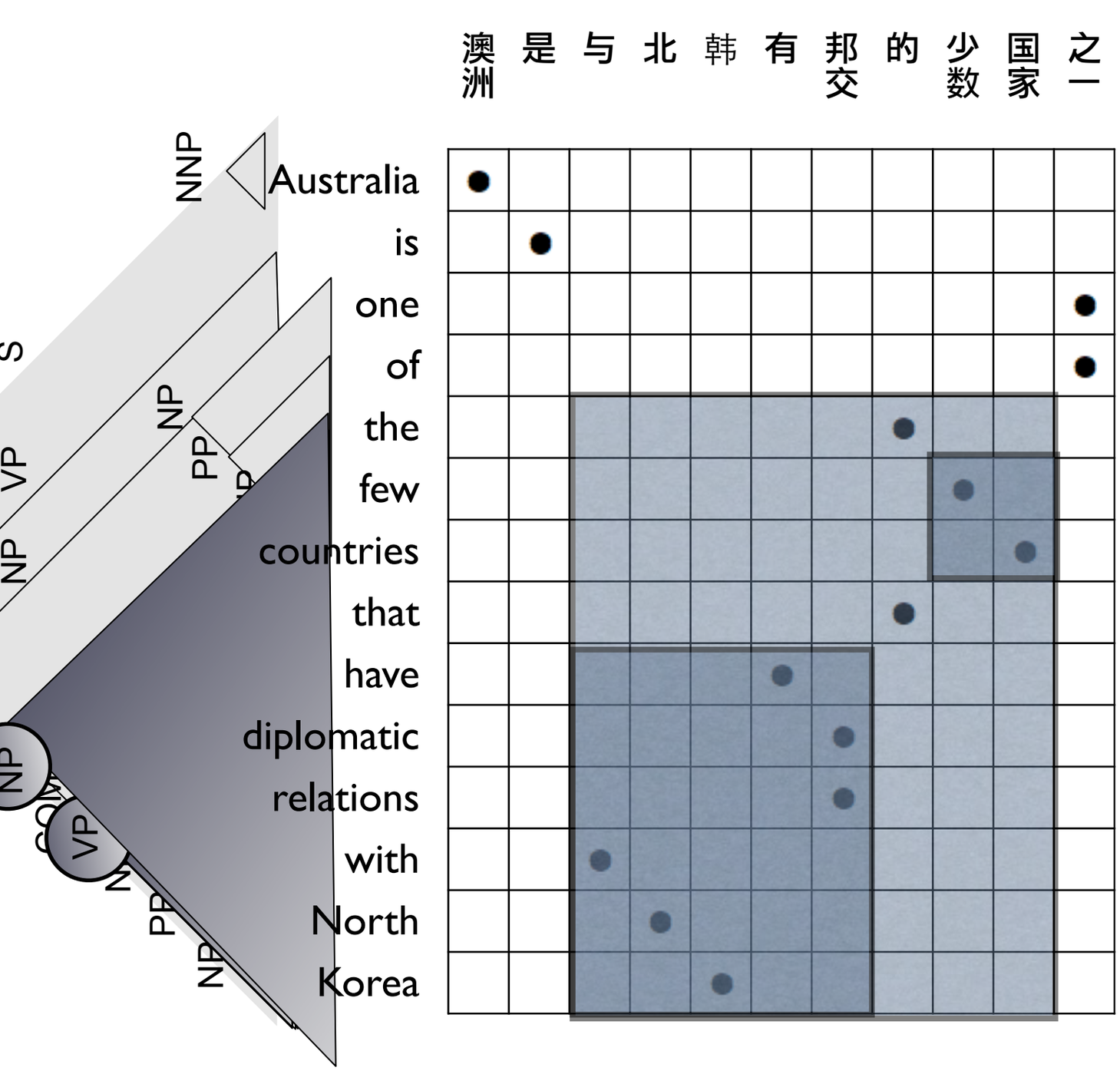
X → 与 X_1 有 X_2,
have X_2 with X_1

Discussion: what do you think of Hiero?

- So, we now have a way of extracting SCFGs from bitexts. Great! So what?
- Is this any better than the phrase based model?
- How?
- Do you feel that it is lacking anything?

(Discuss with your neighbor)

Extracting Syntactic Rules



VP → 与 北 韩 有 邦交, have diplomatic relations with North Korea

NP → 与 北 韩 有 邦交 的 少数 国家, the few countries that have diplomatic relations with North Korea

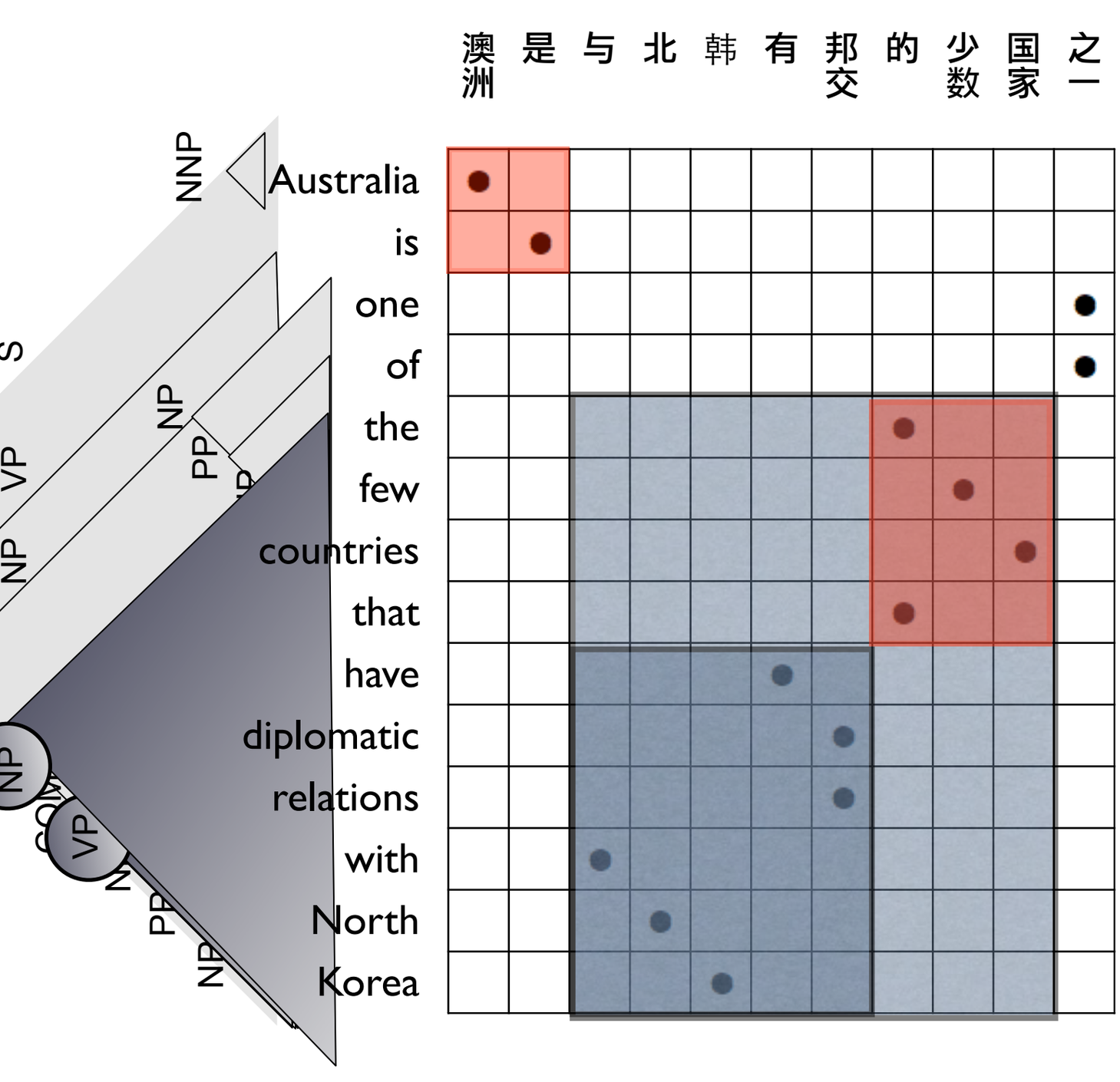
NP → VP 的 少数 国家, the few countries that VP

NP → VP 的 NP, the NP that VP

Wait a minute...

- Didn’t we see this earlier in Koehn’s paper?
- Aren’t we giving up a ton of rules that you said were valuable?
- Something about a reduced inventory because we got rid of non-constituent phrases?

Extracting Syntactic Rules



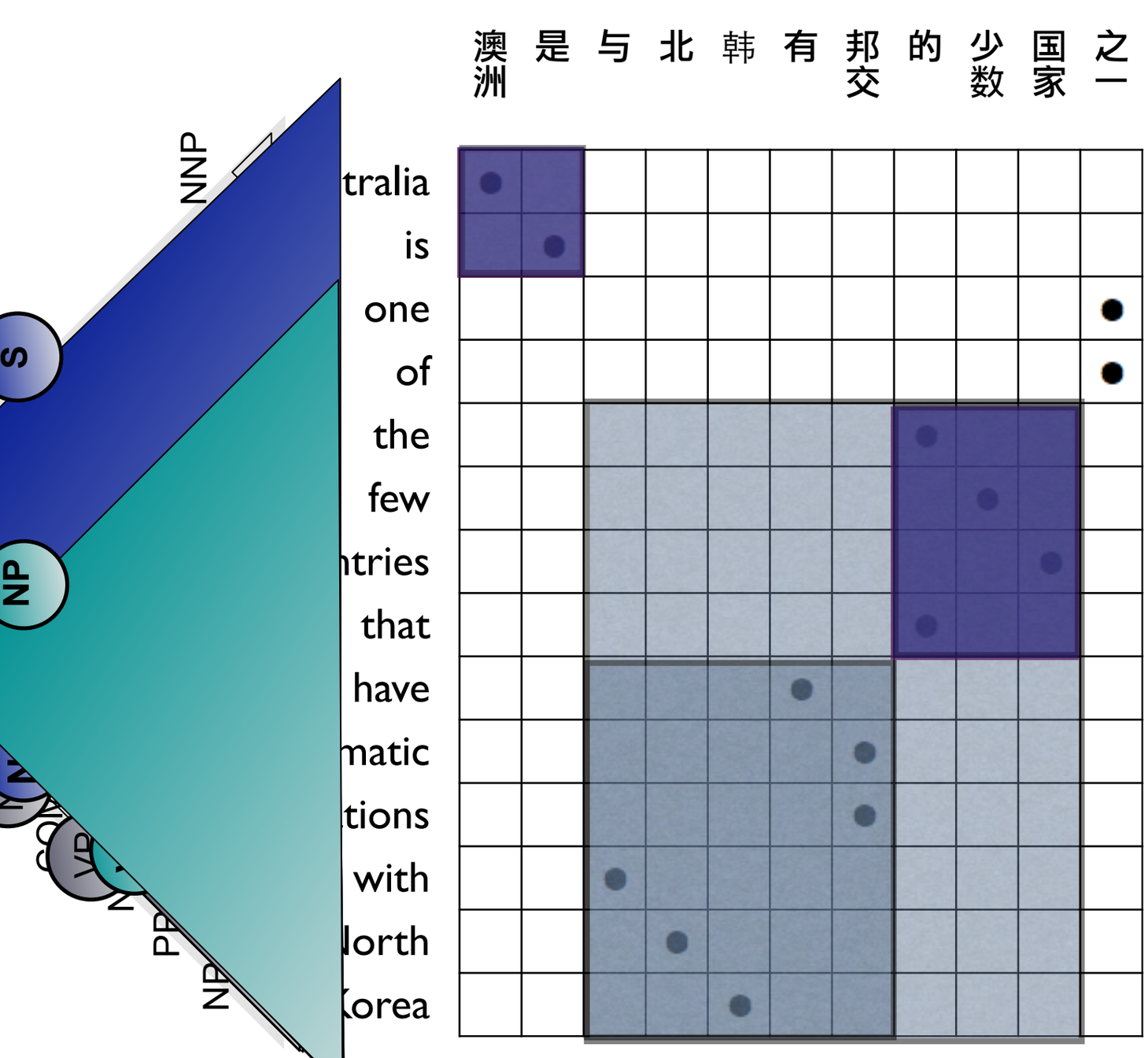
VP → 与 北 韩 有 邦交, have diplomatic relations with North Korea

NP → 与 北 韩 有 邦交 的 少数 国家, the few countries that have diplomatic relations with North Korea

??? → 的 少数 国家, the few countries that

??? → 澳洲 是, Australia is

Extracting Syntactic Rules



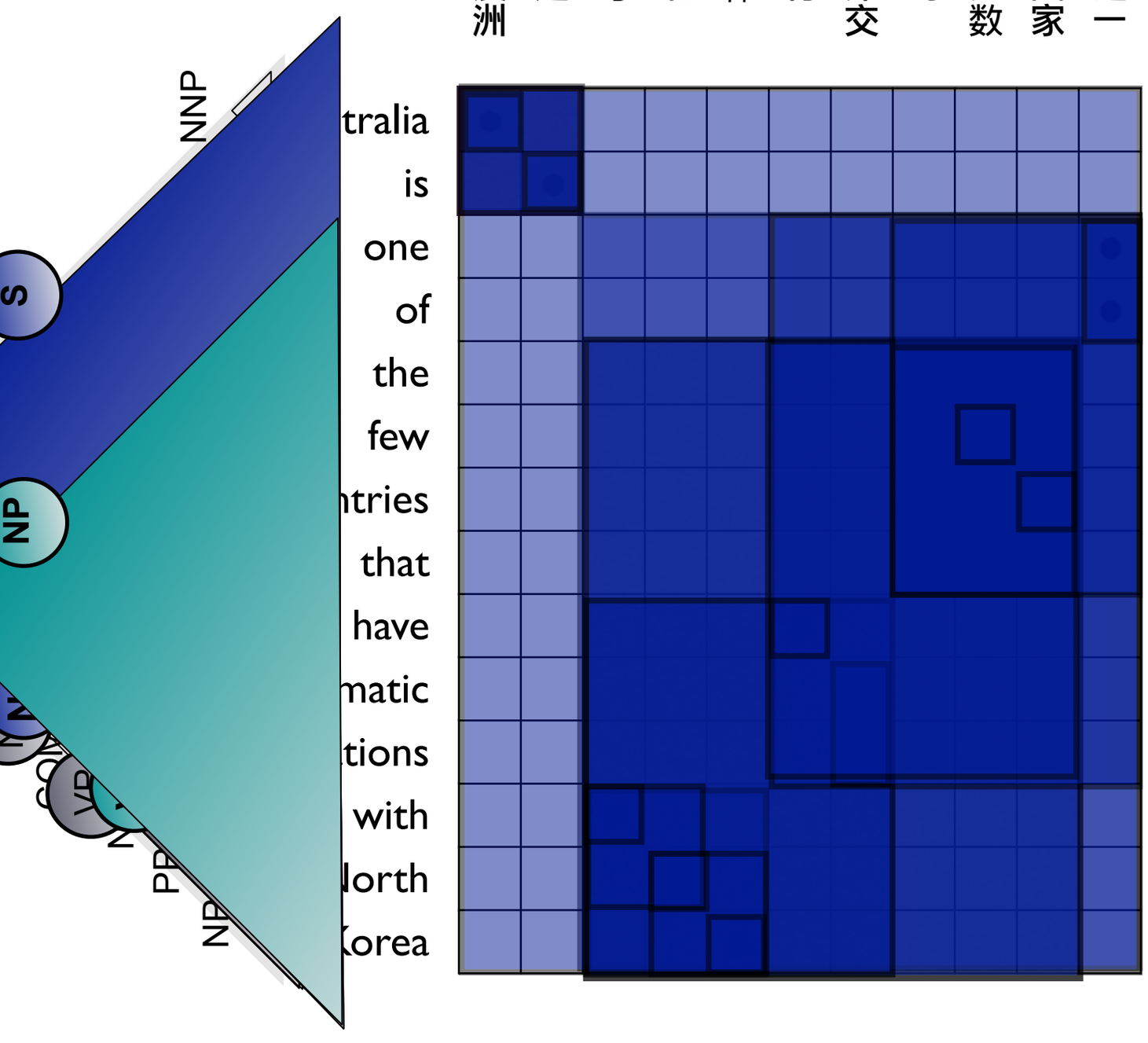
VP → 与 北 韩 有 邦交, have diplomatic relations with North Korea

NP → 与 北 韩 有 邦交 的 少数 国家, the few countries that have diplomatic relations with North Korea

NP/ VP → 的 少数 国家, the few countries that

S/ NP → 澳洲 是, Australia is

Extracting Syntactic Rules



VP → 与 北 韩 有 邦交, have diplomatic relations with North Korea

NP → 与 北 韩 有 邦交 的 少数 国家, the few countries that have diplomatic relations with North Korea

NP/ VP → 的 少数 国家, the few countries that

S/ NP → 澳洲 是, Australia is

Discussion: Is this better?

- What do you think of this flavor of SCFGs?
- What are its limitations?
- Do you think that it is better or worse than Hiero?
- How would you prove it?

(Discuss with your neighbors)

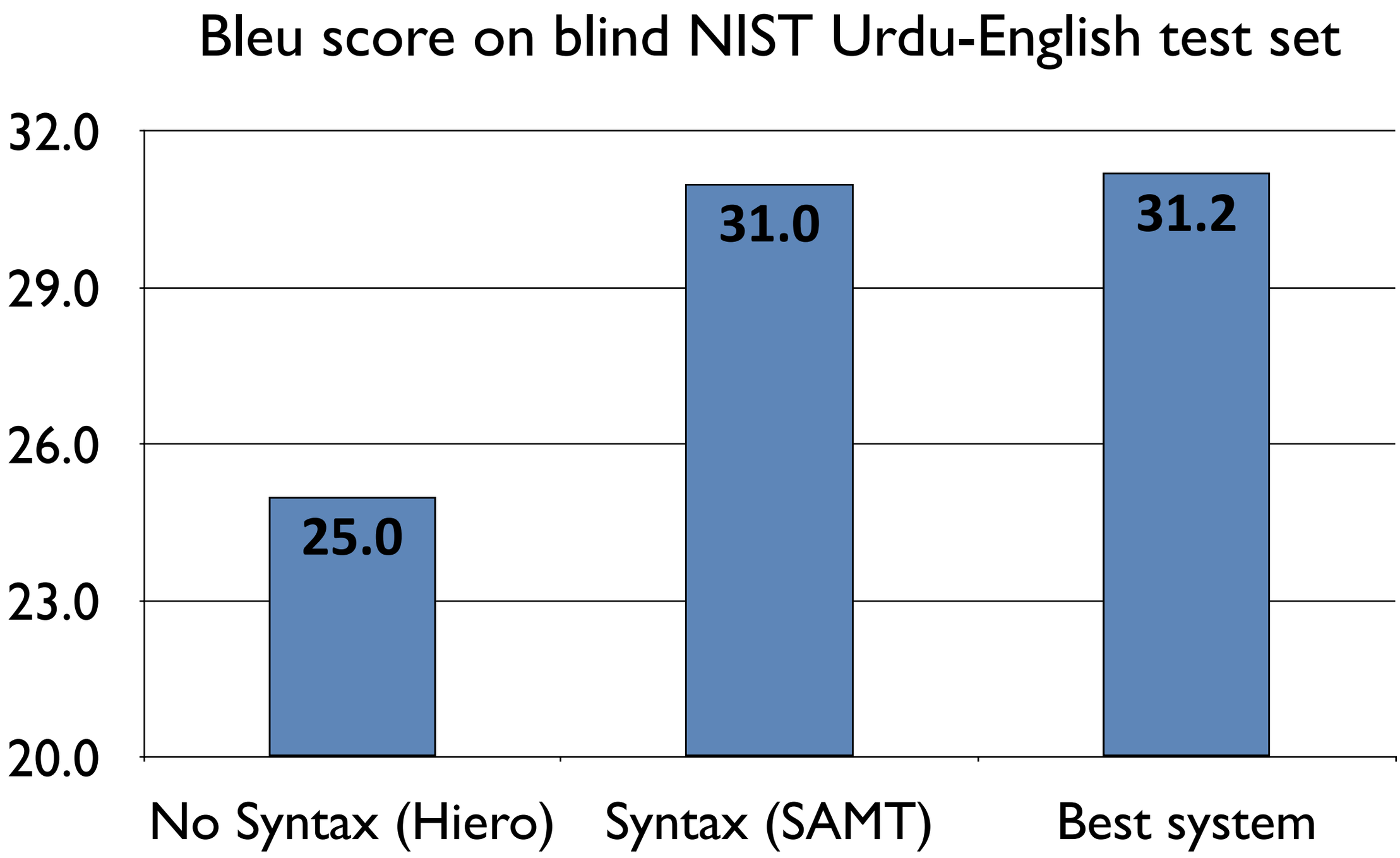
New training paradigm

- Training data: word-aligned bilingual parallel corpus, with parse trees
  - –No need to parse the Urdu, just parse the English
  - –Method is therefore transferable to other resource poor languages
- Extract SCFG rules with syntactic nonterminals
- For non-constituent phrases use CCG-style nonterminals
- Same coverage as Hiero model

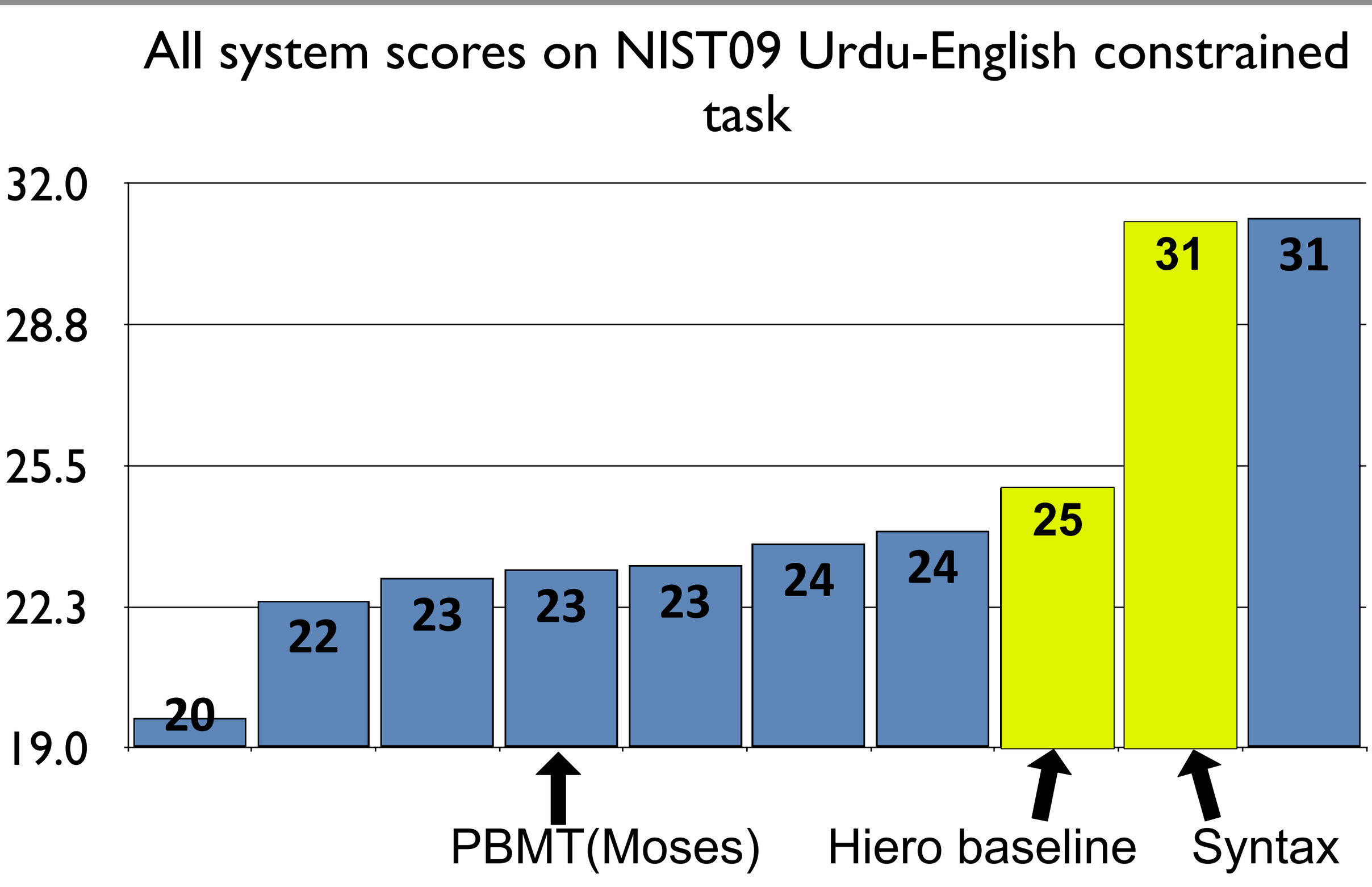
Does it work?

- Tested for Urdu-English MT
- 1.5 Million word parallel corpus
- Two contrastive systems, with different grammar extraction mechanism
  - Hiero
  - Syntax-augmented grammars
- Used same decoder in both cases
- Tested results in a blind test set administered by the National Institute for Standards in Technology

Syntax v. no Syntax



State of the Art Urdu Results



Translation improvements

'first nuclear experiment in 1990 was'
Thomas red Unilever National Laboratory of the United States in ویپن designer, are already working on the book of Los ایلموس National Laboratory ڈینی, former director of the technical انٹیلجنس written with the cooperation of سٹلمین.

This book 'nuclear express: political history and the expansion of bomb' has been written, and the two writers have also claimed that the country has made nuclear bomb is he or any other country's nuclear secrets to چرائیے or that of any other nuclear power cooperation is achieved.

The First Nuclear Test Was in 1990.

Thomas red of the United States, the National Laboratory in designer are already working on the book of Los Alamos National Laboratory, former director of the technical intelligence, with the cooperation of Diana steelman wrote.

This book under the title of the spread of nuclear expressway: the political history of the bomb and this has been written and the two writers have claimed that the country also has made nuclear bomb or any other country, Korea nuclear secrets, or any of the other nuclear power cooperation.

Who did what to whom?

Baseline

He said that China, North Korea, Iran, Syria, Pakistan, through Egypt, Libya and Yemen is to provide nuclear technology.

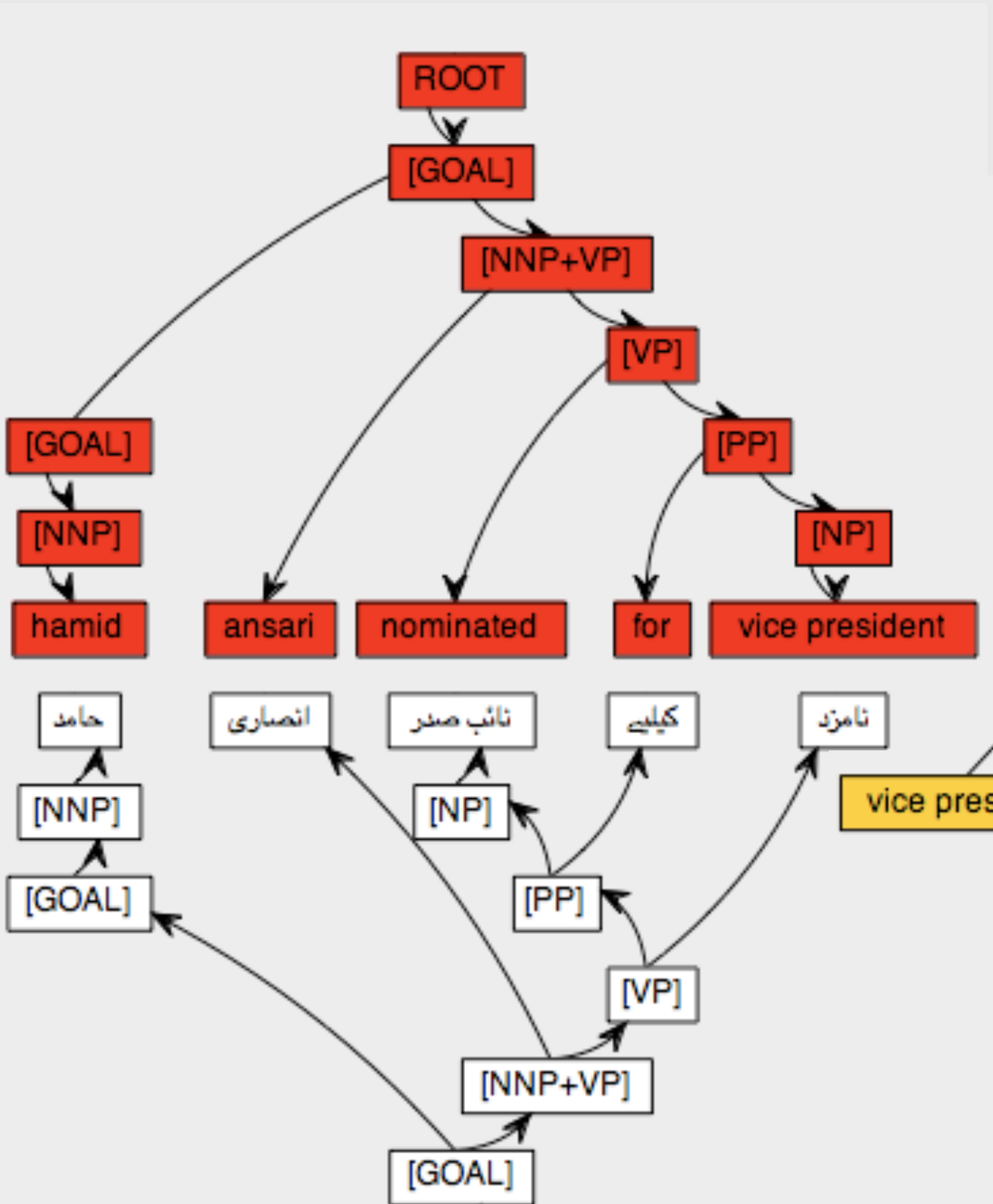
Thomas was red when this question why China has provided the nuclear technology to Pakistan, In response, He said as China and India was joint enemy of Pakistan.

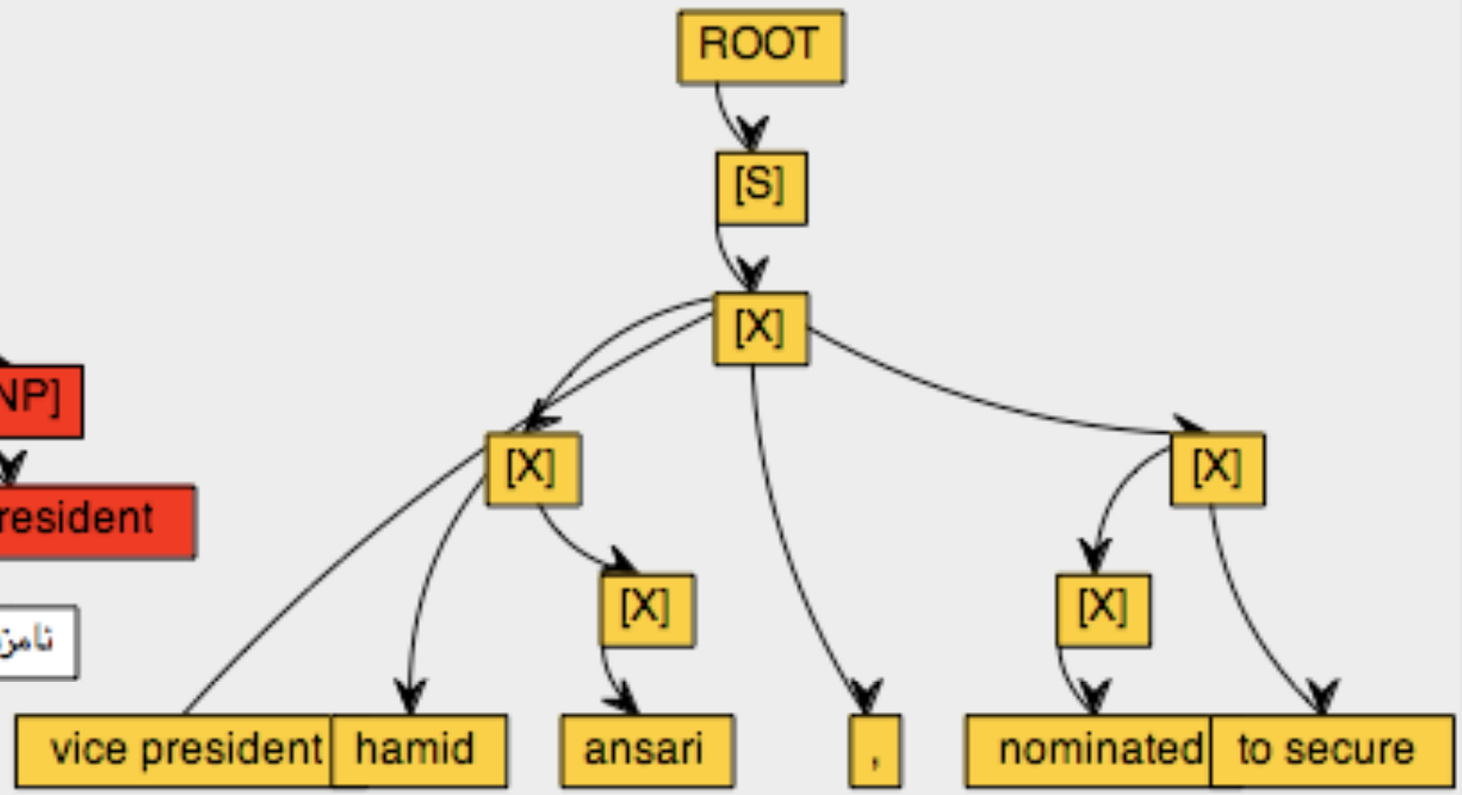
SCALE final system

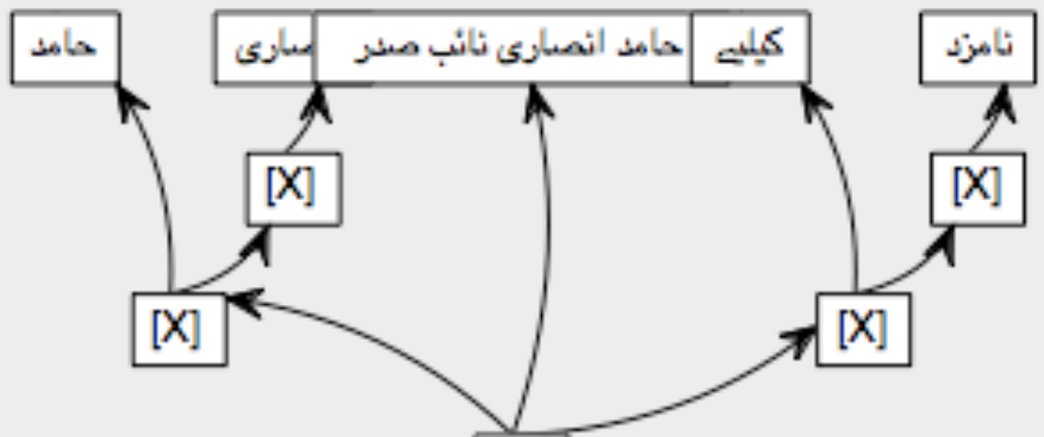
He said that China would provide nuclear technology to North Korea, Iran, Syria, Pakistan, Egypt, Libya and Yemen.

Thomas red when was this question why China has provided to Pakistan nuclear technology, he said in response to China, Pakistan and India as a common enemy.

Syntax captures Urdu reordering



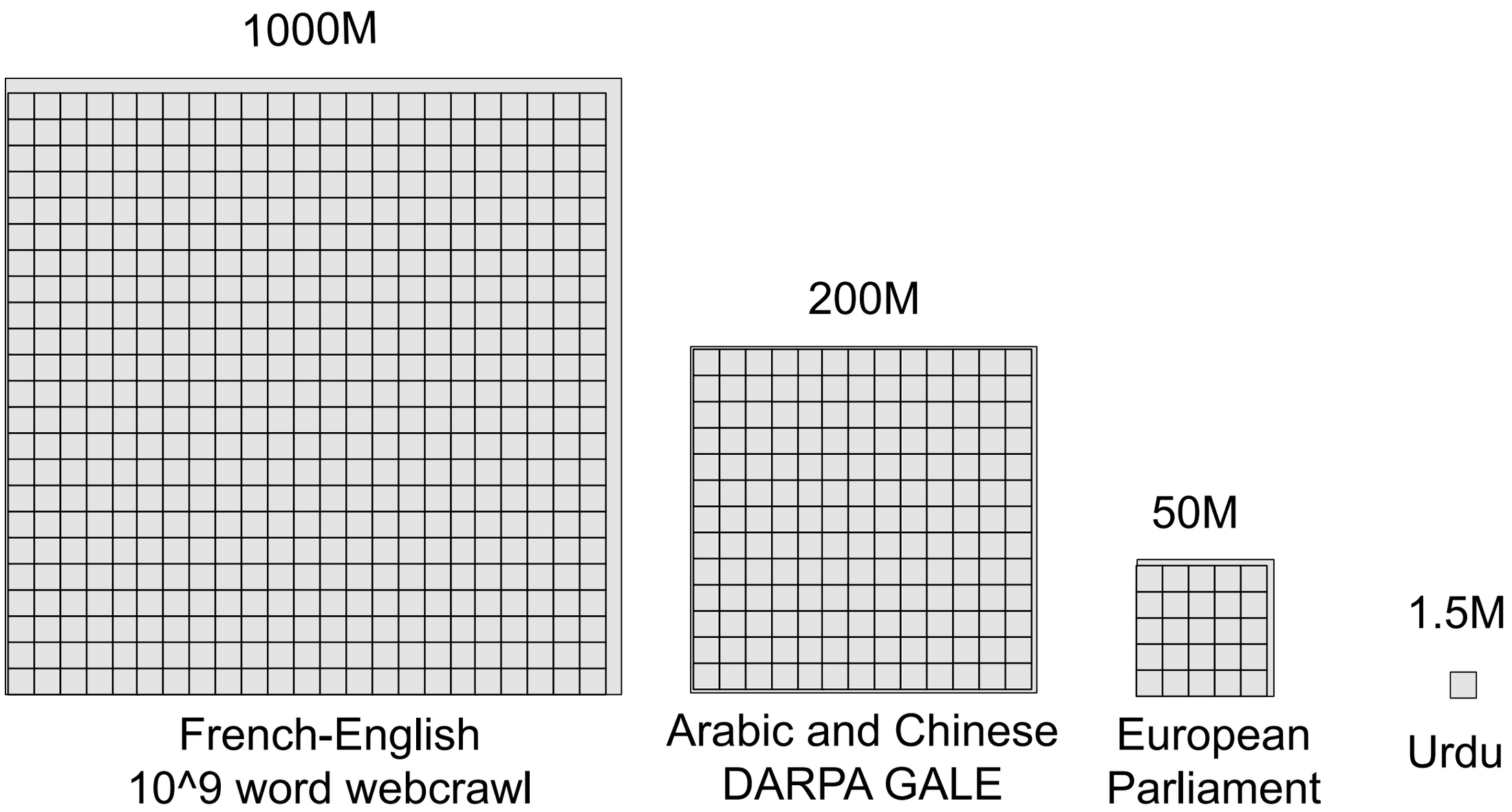




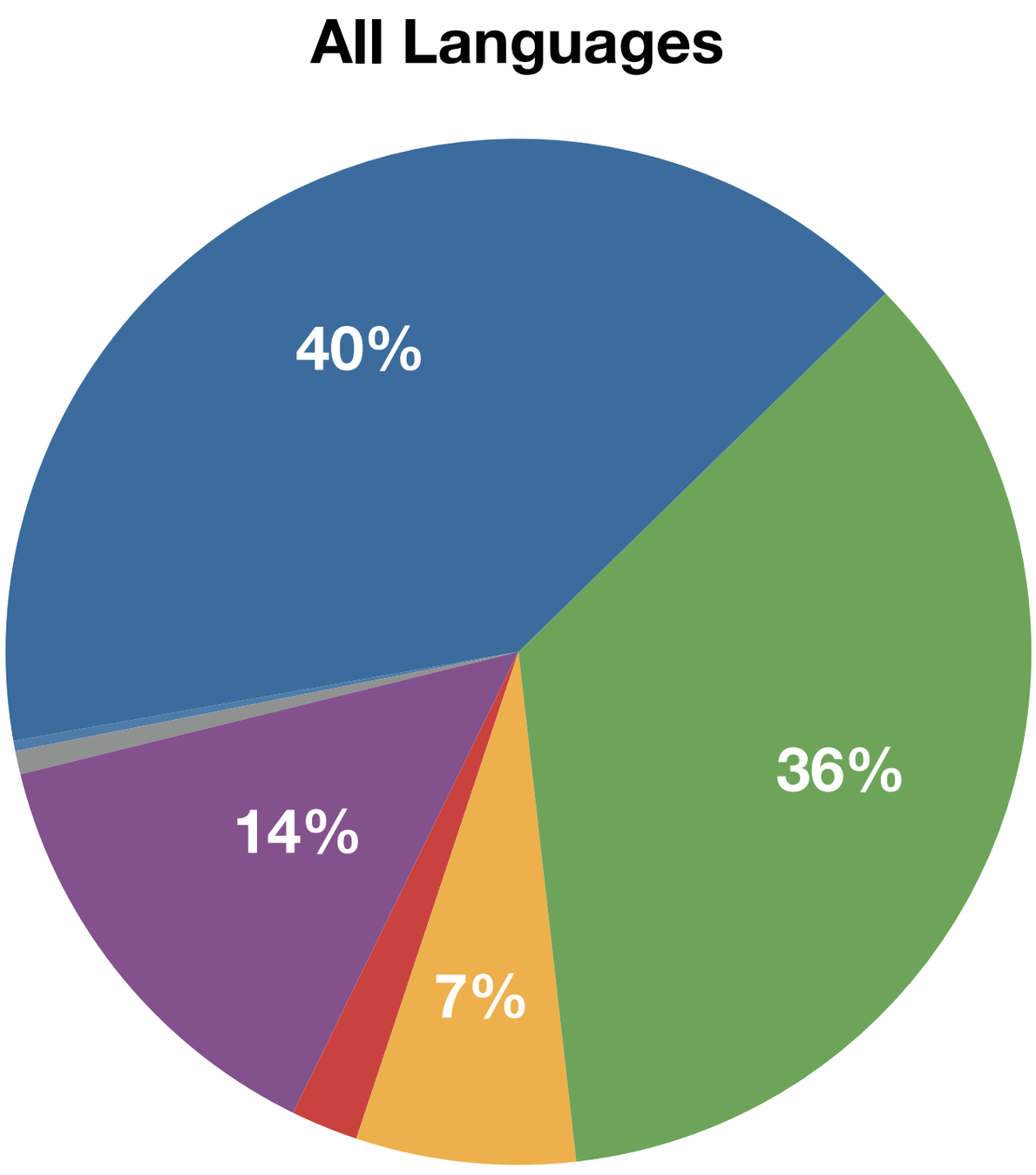
Why did this work?

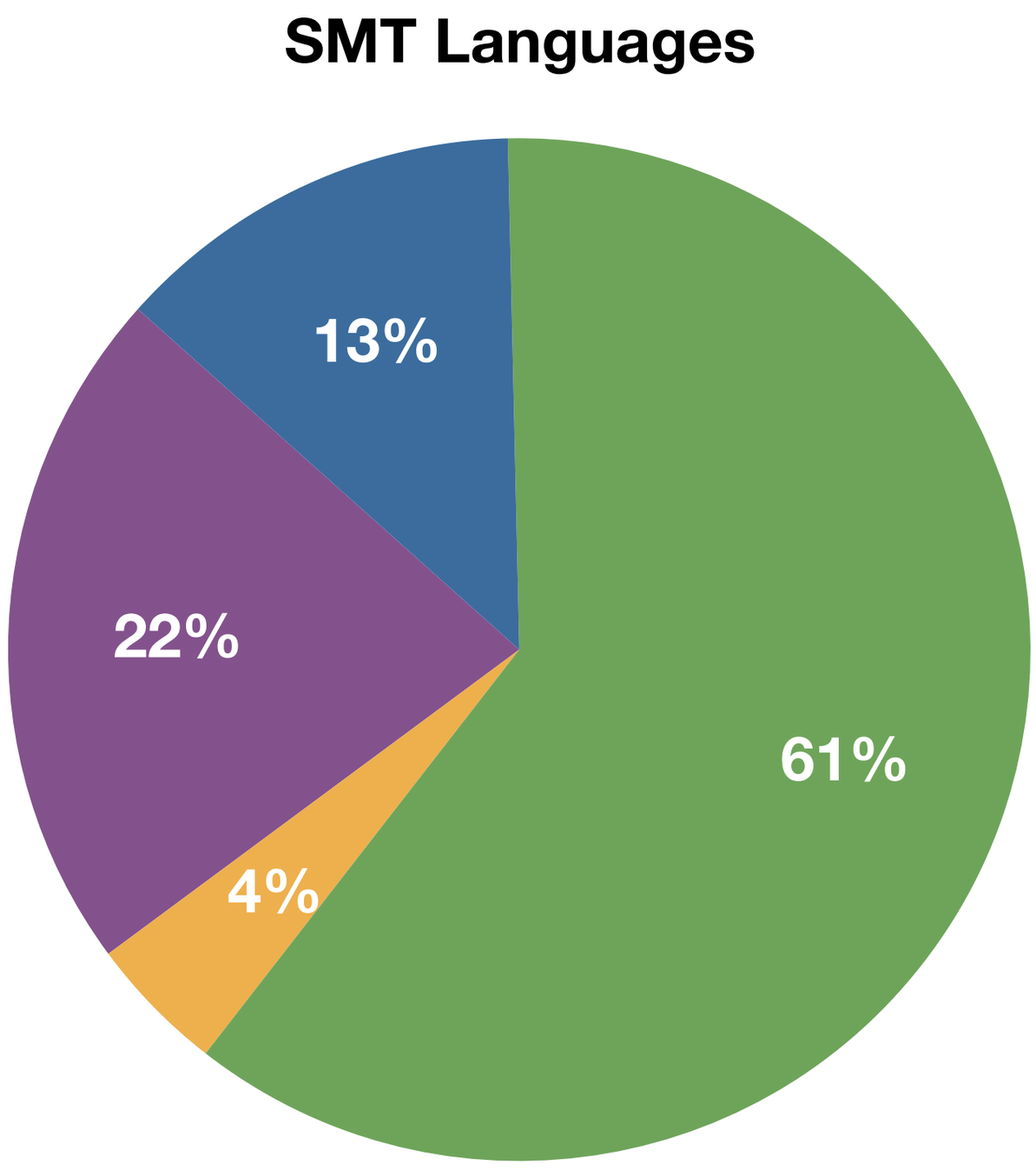
- Using syntax-based translation models resulted in huge improvements in quality
- Previous work on syntax did not shown significant gains, so why did it work here?
- Urdu is an ideal language to show off the advantages of syntax
  - –Very small amount of training data
  - –Very different word order than English
- Can't simply memorize translations of phrases
- Must generalize

Training data for MT Research



Distribution of Word Orders





SOV SVO VSO VOS No dominant order

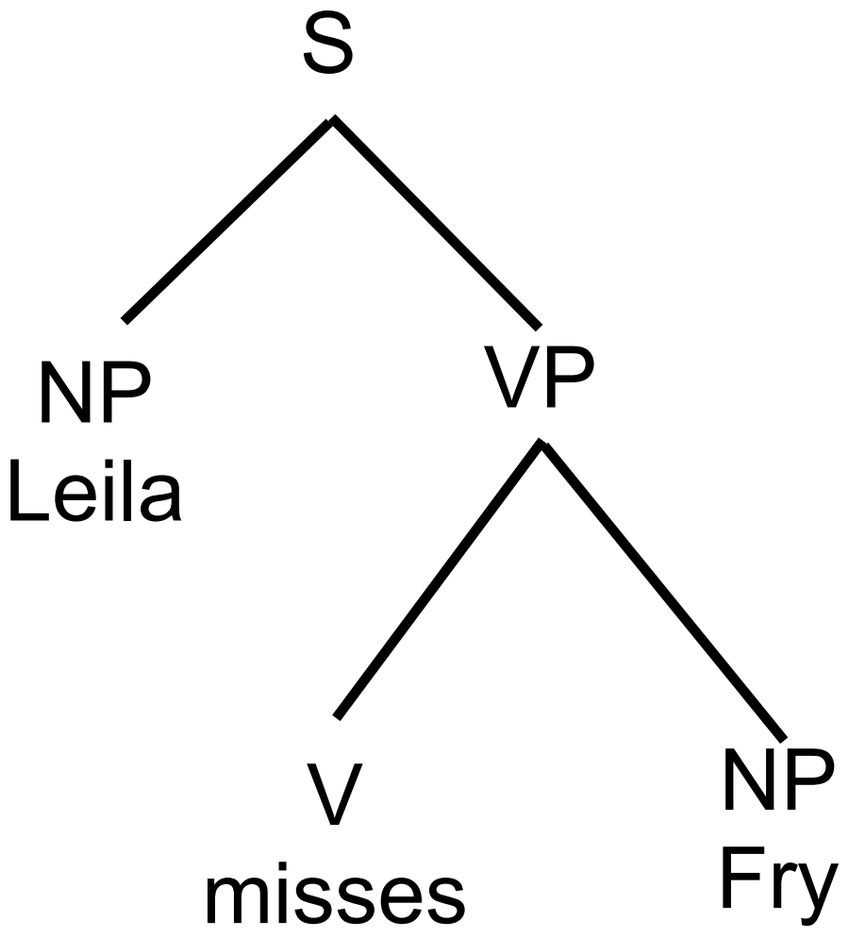
Joshua Decoder

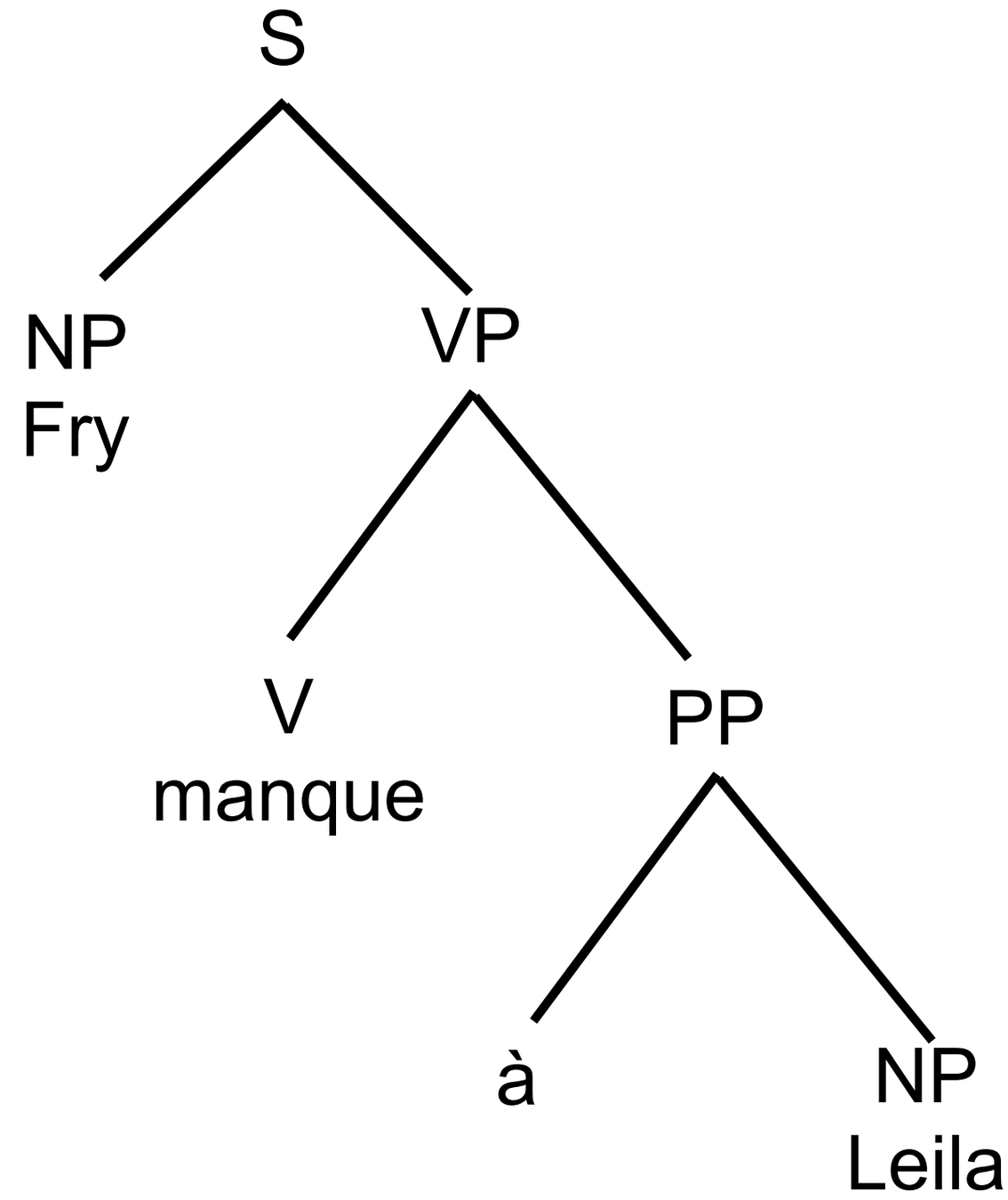
- An open source decoder
- Uses synchronous context free grammars to translate
- Implements all algorithms needed for translating with SCFGs
  - grammar extraction (Thrax!)
  - chart-parsing
  - n-gram language model integration
  - pruning, and k-best extraction

Joshua Decoder

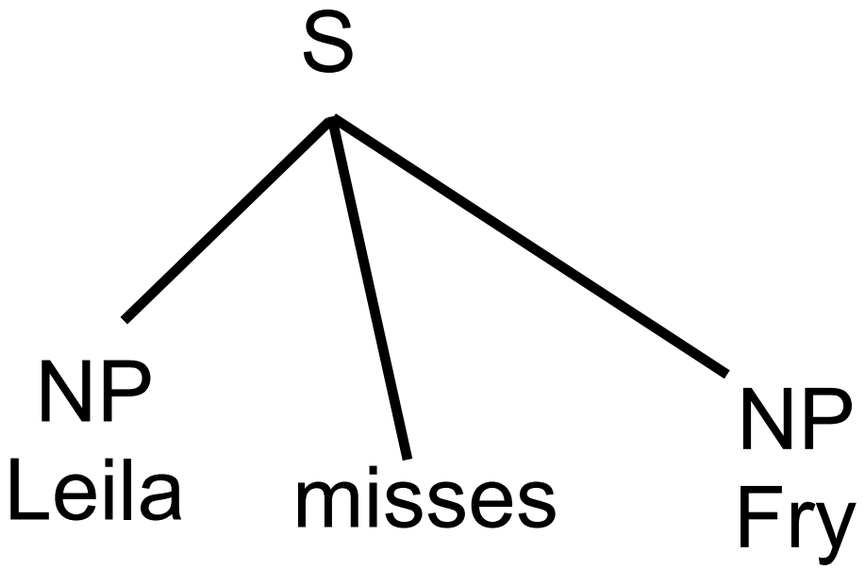
- Download it from
  - http://joshua-decoder.org
- Brownie points if you use it in your final projects
- Use Jonny's Thrax grammar extractor to test different kinds of SFCGs for your problems

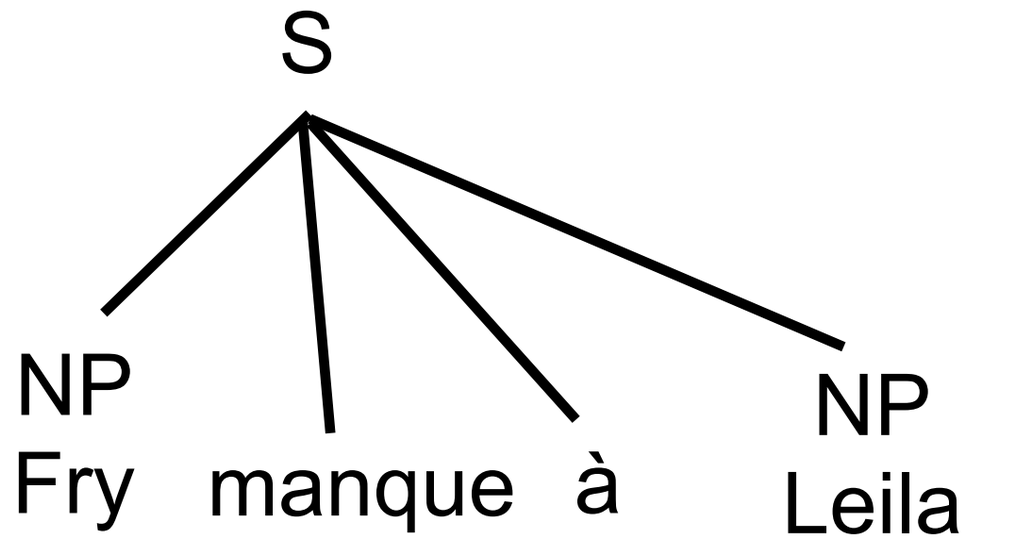
Dealing with language mismatches





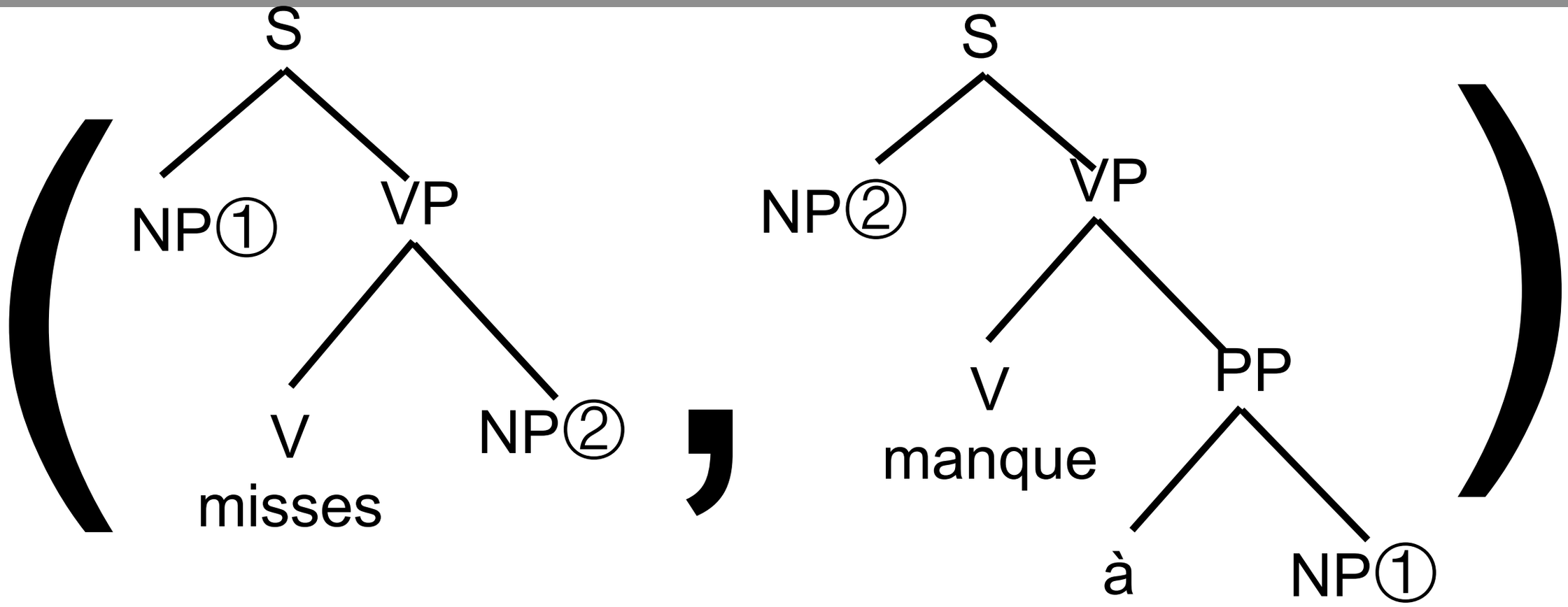
Dealing with language mismatches

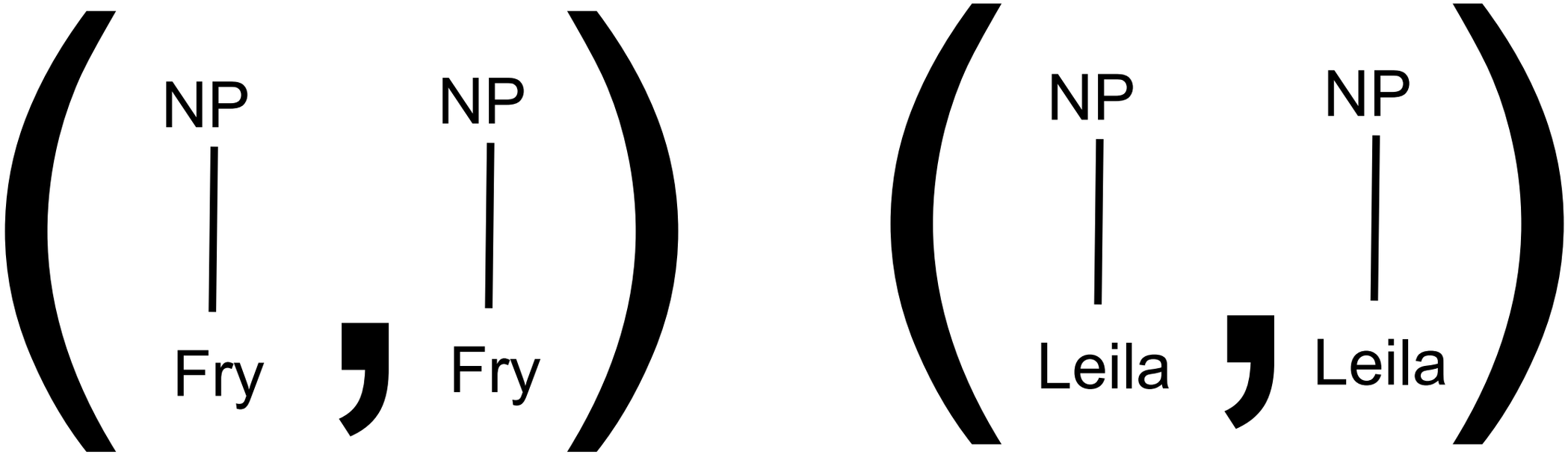




S →	NP①	misses NP②	NP② manque à NP①
NP →		Fry	Fry
NP →		Leila	Leila

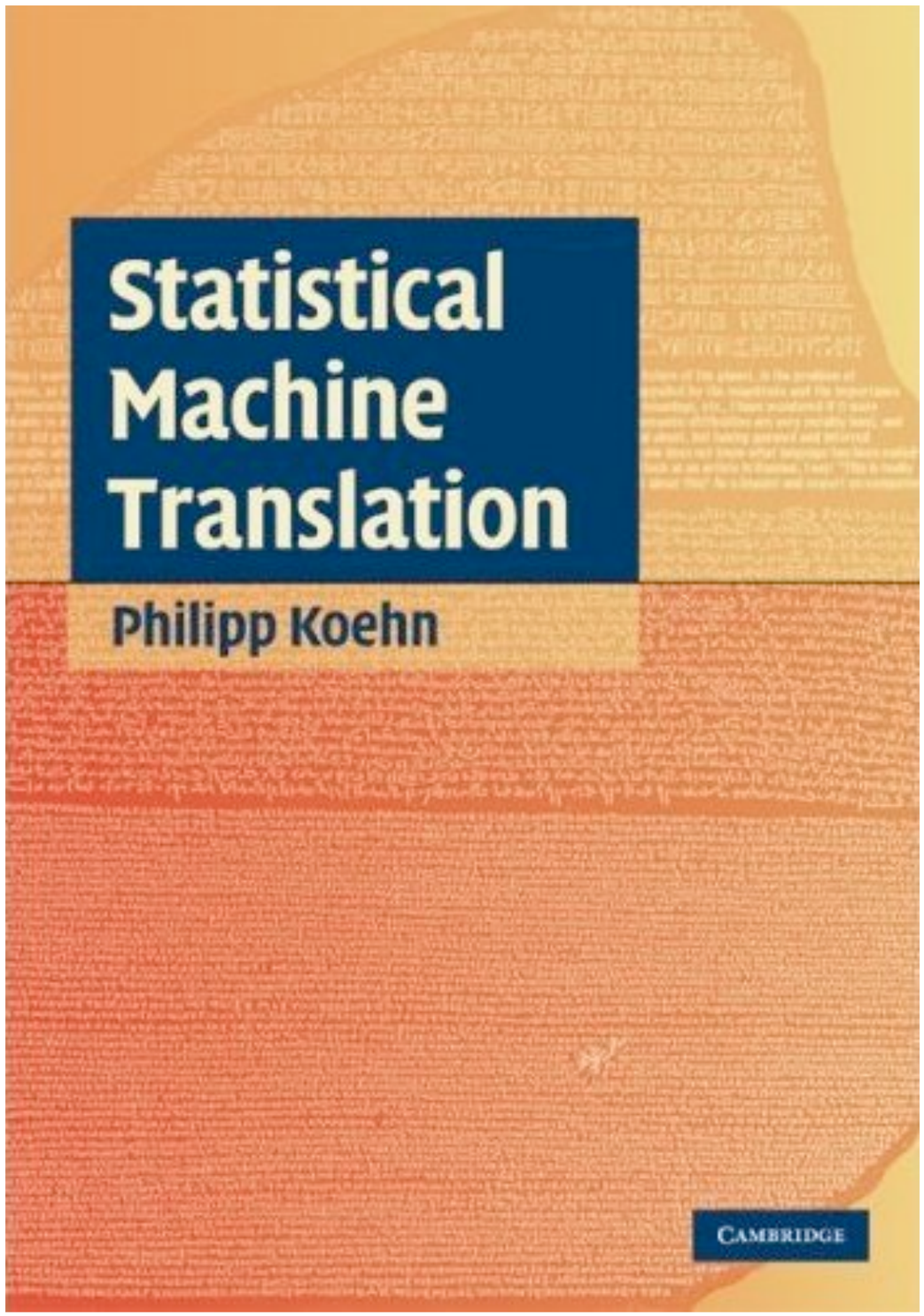
Synchronous Tree Substitution





Reading

- Read Chapter 11 from the textbook



Announcements

- Next week:
  - On Thursday 3/27 you have two items due
  - HW4 is due
    - You must send the training/development/test data that you will use for your term project to Jonny
- Language in 10 minutes:
  - Tuesday: Jessy - Serbian and Rigel - Javanese
  - Thursday: Kai - Italian