Improvements to Urdu-English

SCALE Summer Workshop Results

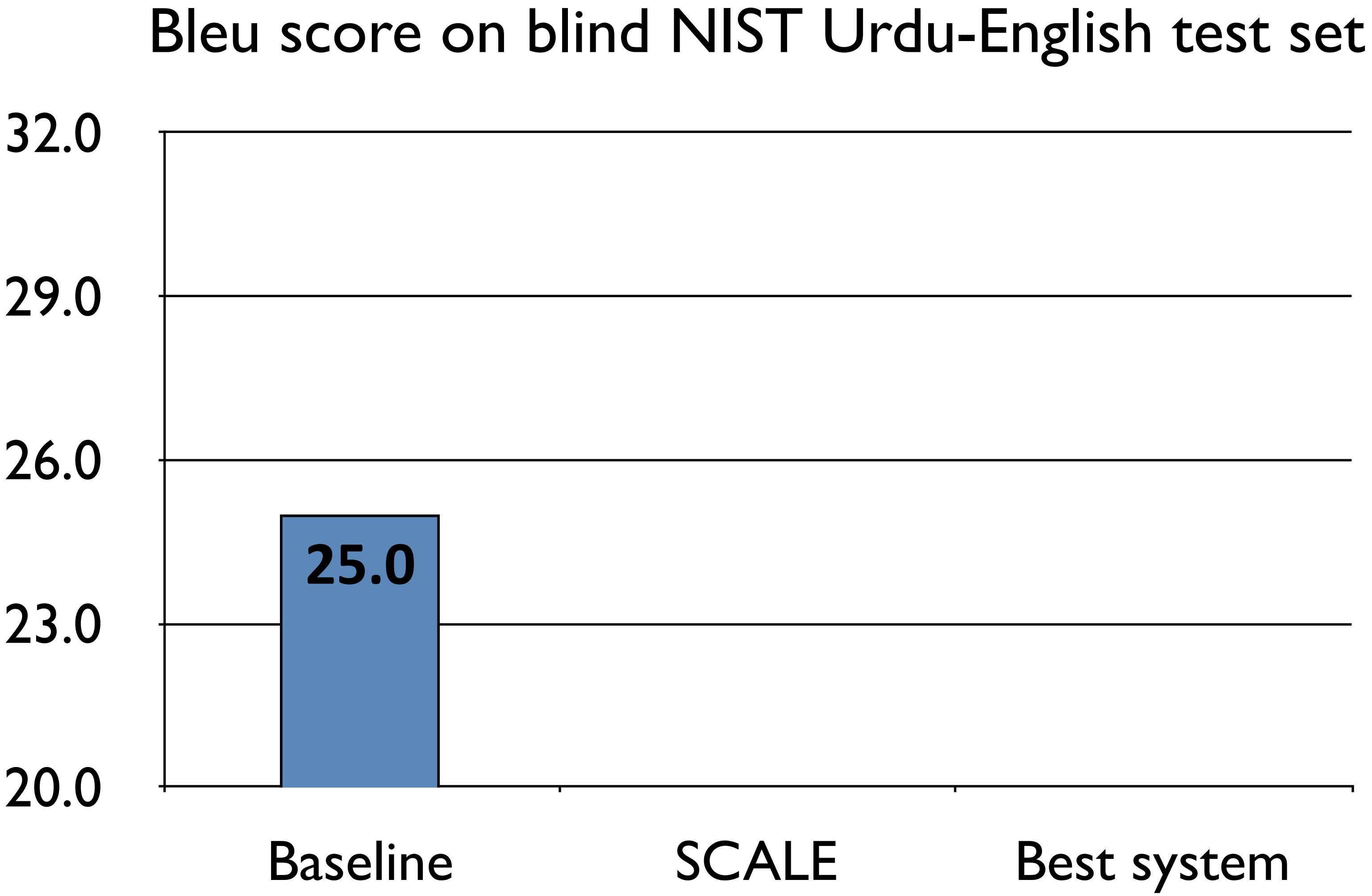
Chris Callison-Burch

Kathy Baker, Steven Bethard, Michael Bloodgood, Ralf Brown, Glen Coppersmith, Bonnie Dorr, Wes Filardo, Kendall Giles, Ann Irvine, Mike Kayser, Lori Levin, Justin Martineau, Jim Mayfield, Scott Miller, Aaron Phillips, Andrew Philpot, Christine Piatko, Lane Schwartz, and David Zajic

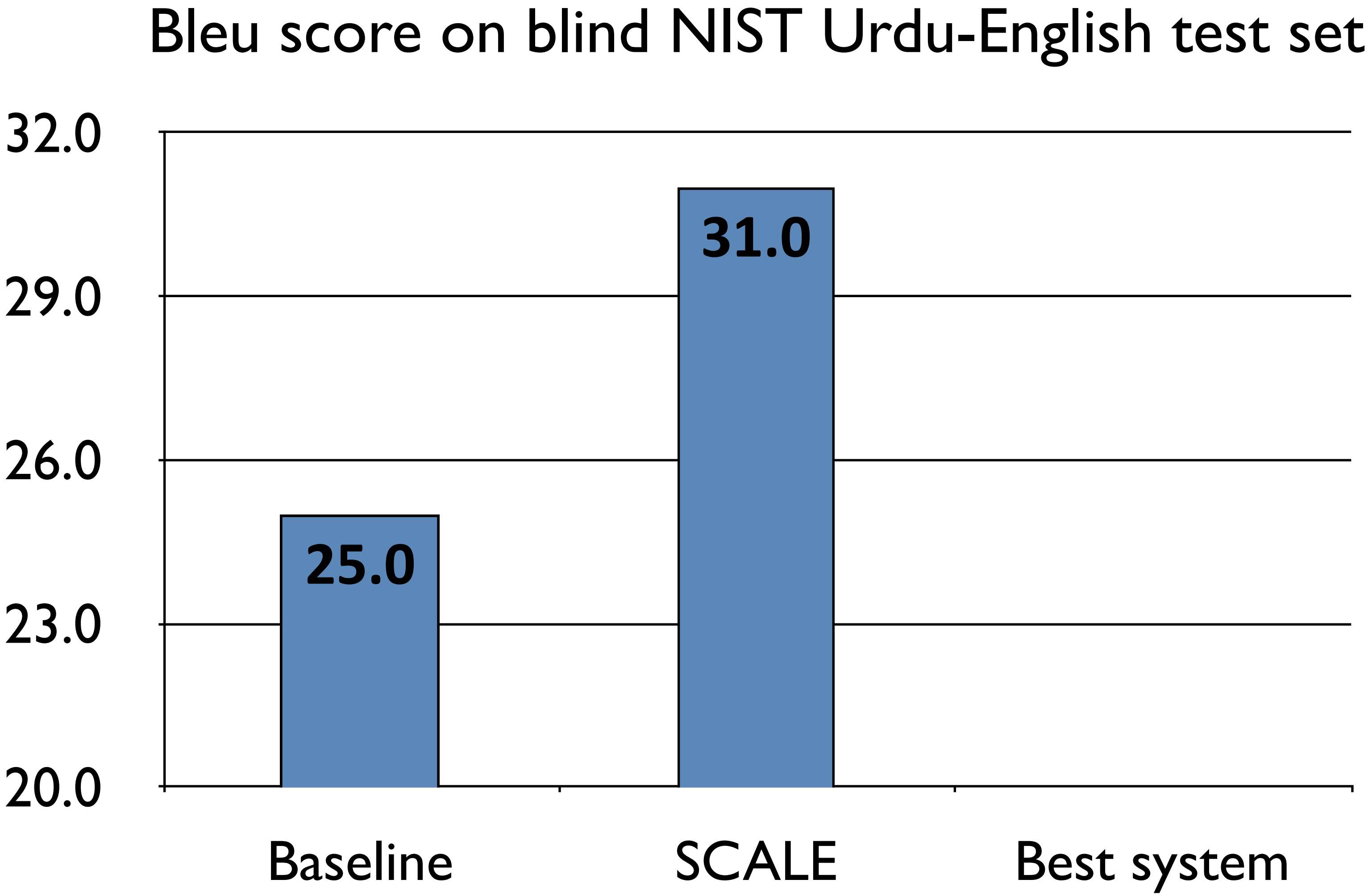
human language technology
center of excellence

The Punch Line

The Punch Line

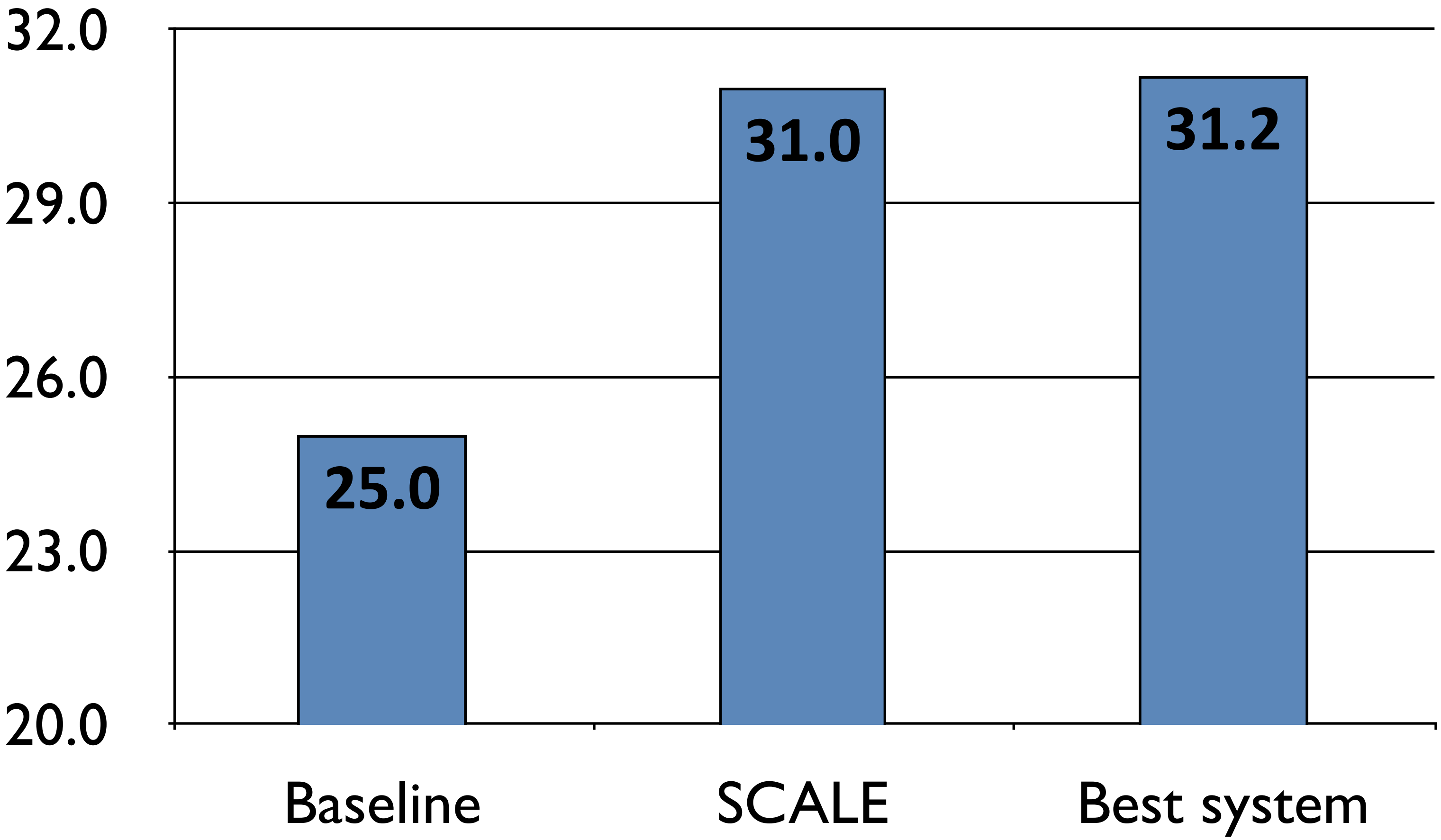


The Punch Line

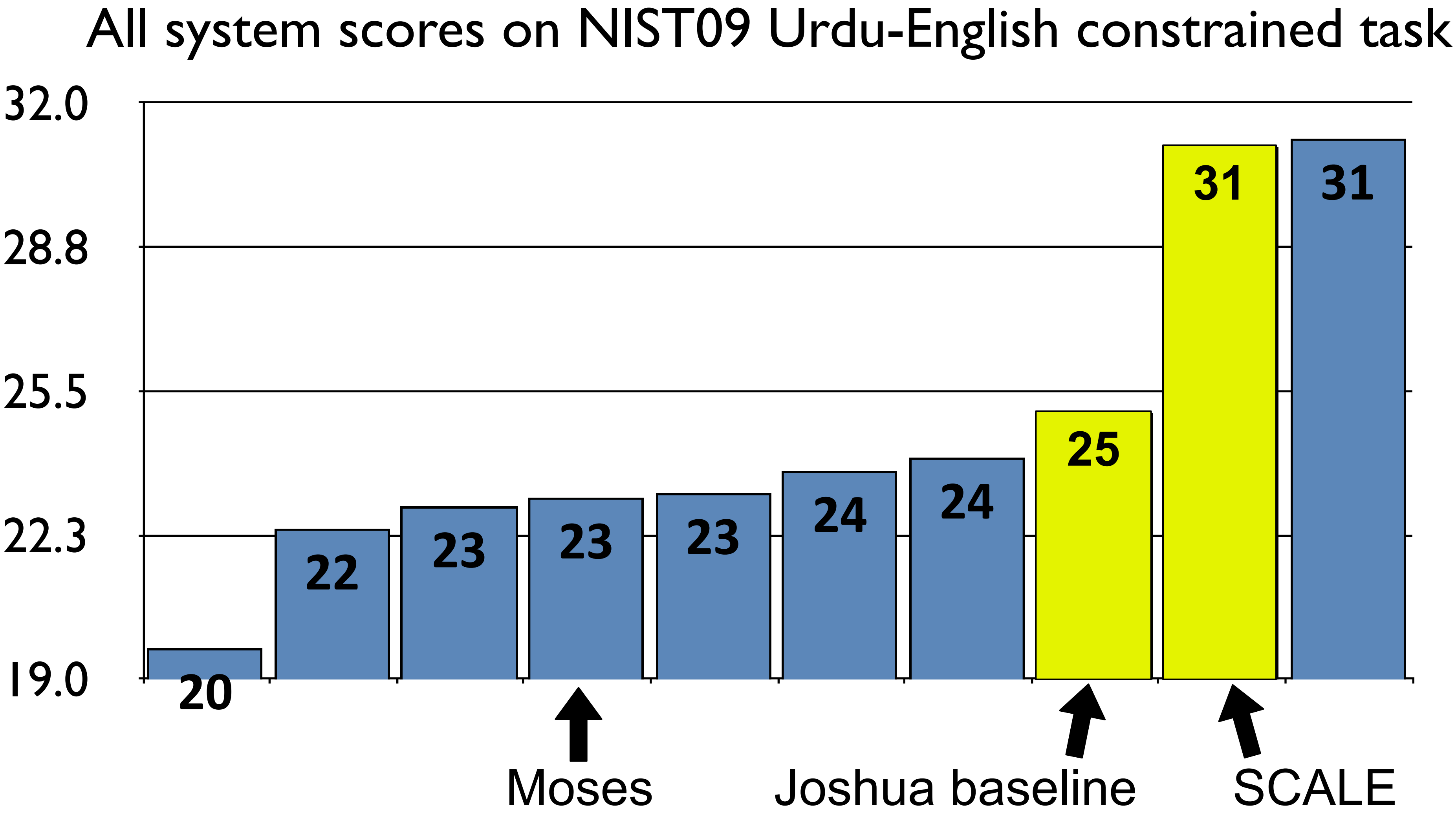


The Punch Line

Bleu score on blind NIST Urdu-English test set



The field for the Urdu task



How we got the gains

- Improvements over baseline decoder
  - Syntax-based model replaced hierarchal phrase-based model
  - Semantic elements (named entities, modalities) are grafted onto syntax trees
  - Unknown Urdu words are transliterated instead of left untranslated
  - Larger number of feature fns score translations
  - Word alignments improved by incorporating multiple aligners including a syntax-based aligner

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 or that of any other nuclear چرائیے 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.

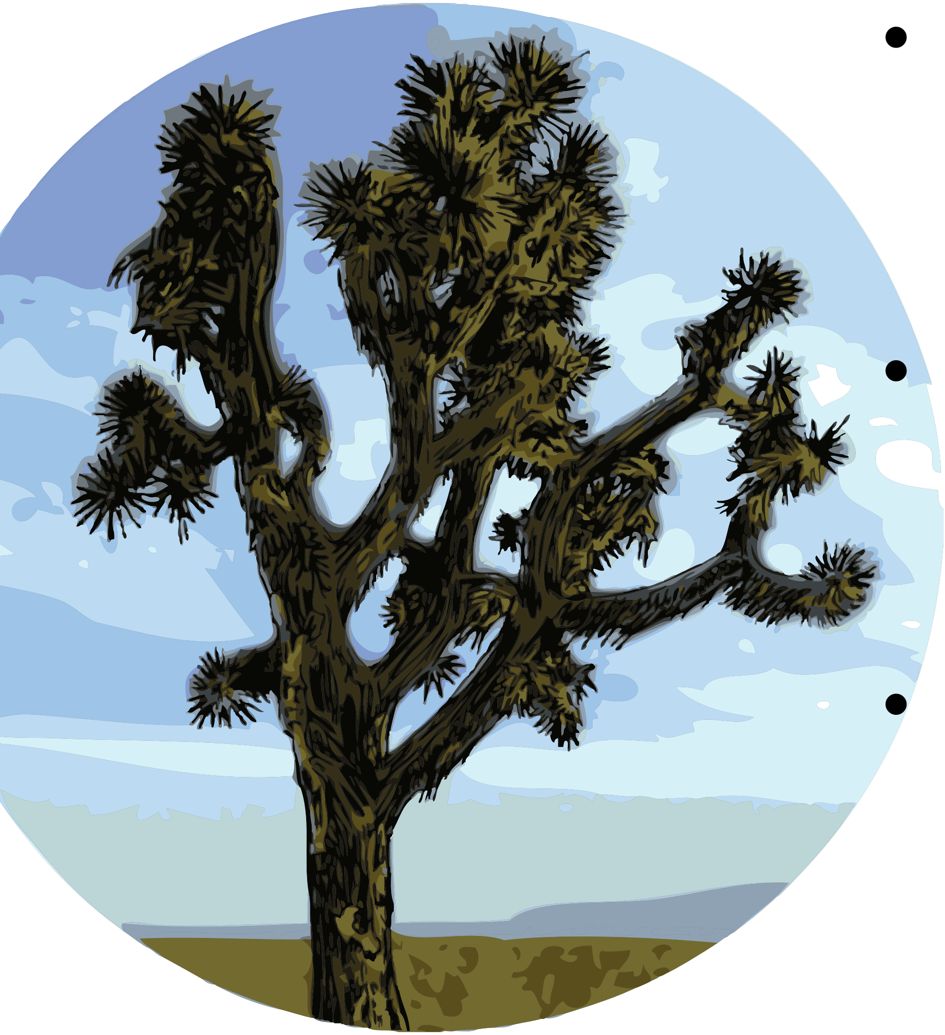
Thomas was red when this question why China has provided the nuclear technology to Pakistan, In response, He said as China and India was joint enemy of Pakistan.

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.

Joshua Decoder

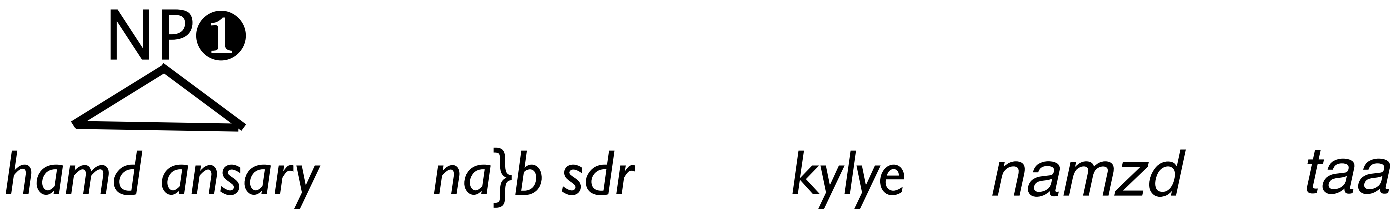


- Synchronous context free grammars generate pairs of corresponding strings
- Can be used to describe translation and re-ordering between languages
- Because Joshua uses SCFGs, it translates sentences by parsing them

Example SCFG for Urdu

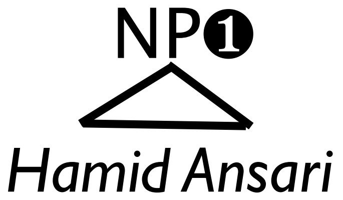
	Urdu	English
S →	NP① VP②	NP① VP②
VP→	PP① VP②	VP② PP①
VP→	V① AUX②	AUX② V①
PP →	NP① P②	P② NP①
NP →	*hamd ansary*	*Hamid Ansari*
NP →	*na}b sdr*	*Vice President*
V →	*namzd*	*nominated*
P →	*kylye*	*for*
AUX →	*taa*	*was*

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



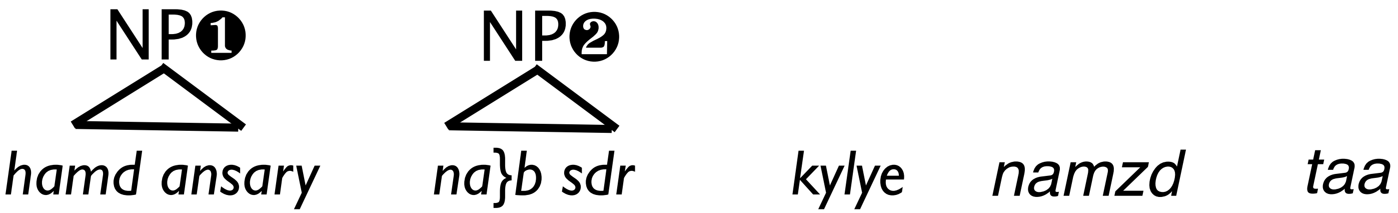
NP❶

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



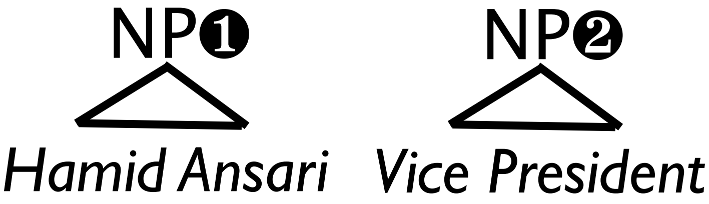
NP❶

Hamid Ansari



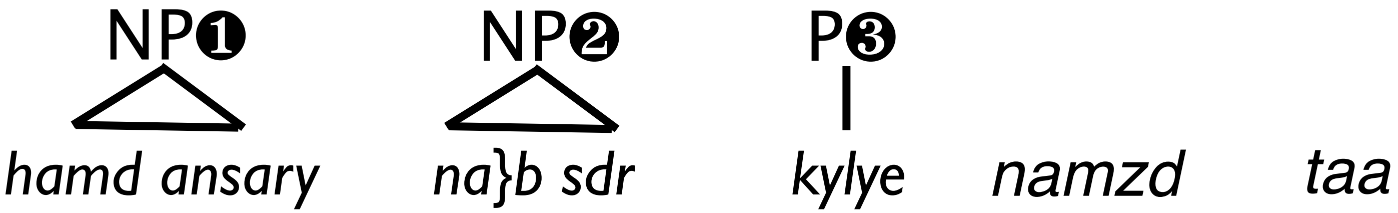
NP❶ NP❷

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



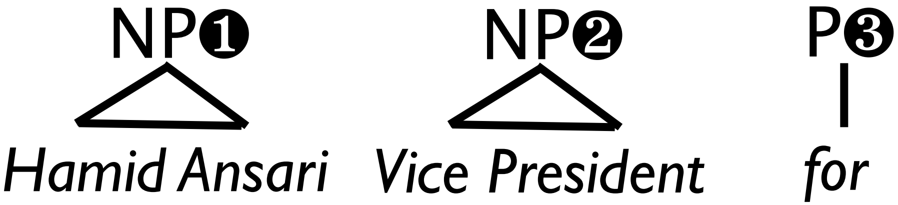
NP❶ NP❷

Hamid Ansari *Vice President*



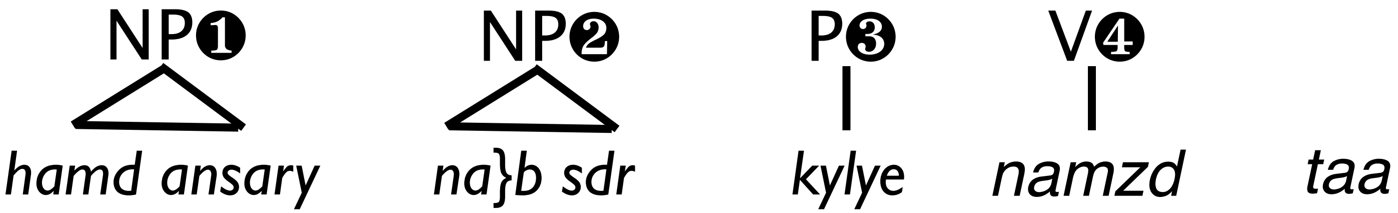
NP❶ NP❷ P❸

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



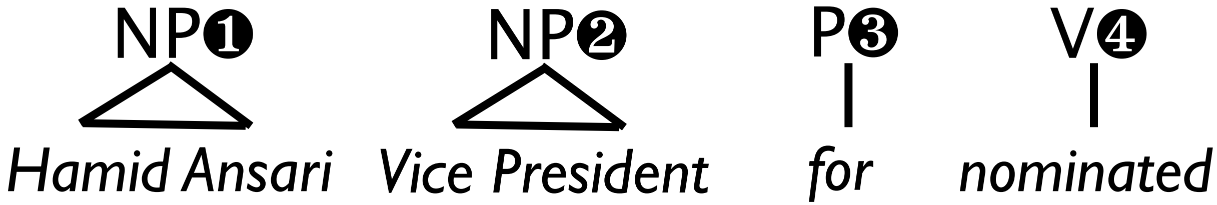
NP❶ NP❷ P❸

Hamid Ansari *Vice President* *for*



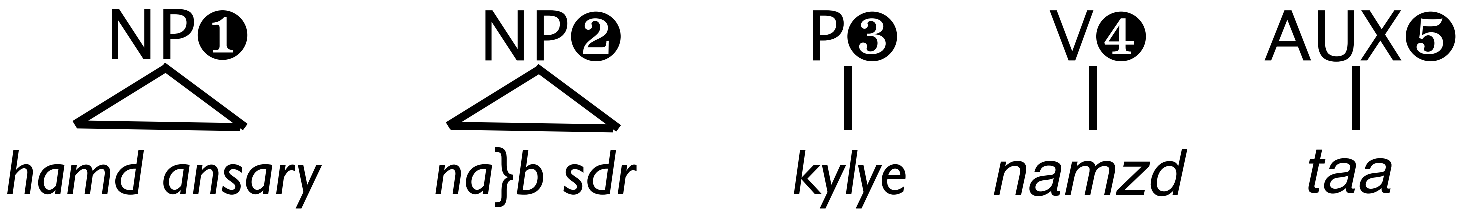
NP❶ NP❷ P❸ V❹

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



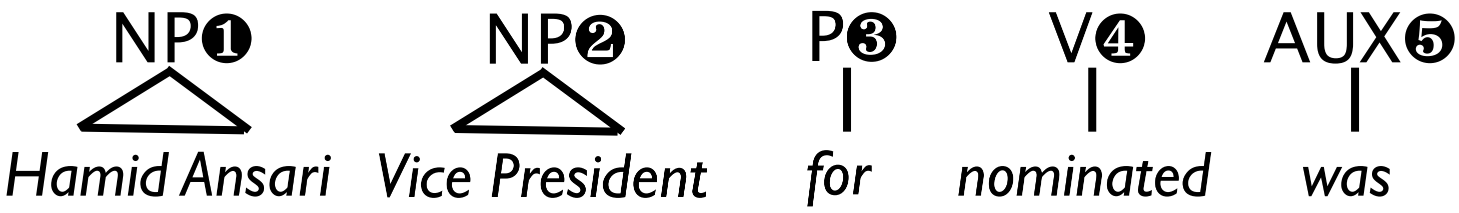
NP❶ NP❷ P❸ V❹

Hamid Ansari *Vice President* *for* *nominated*



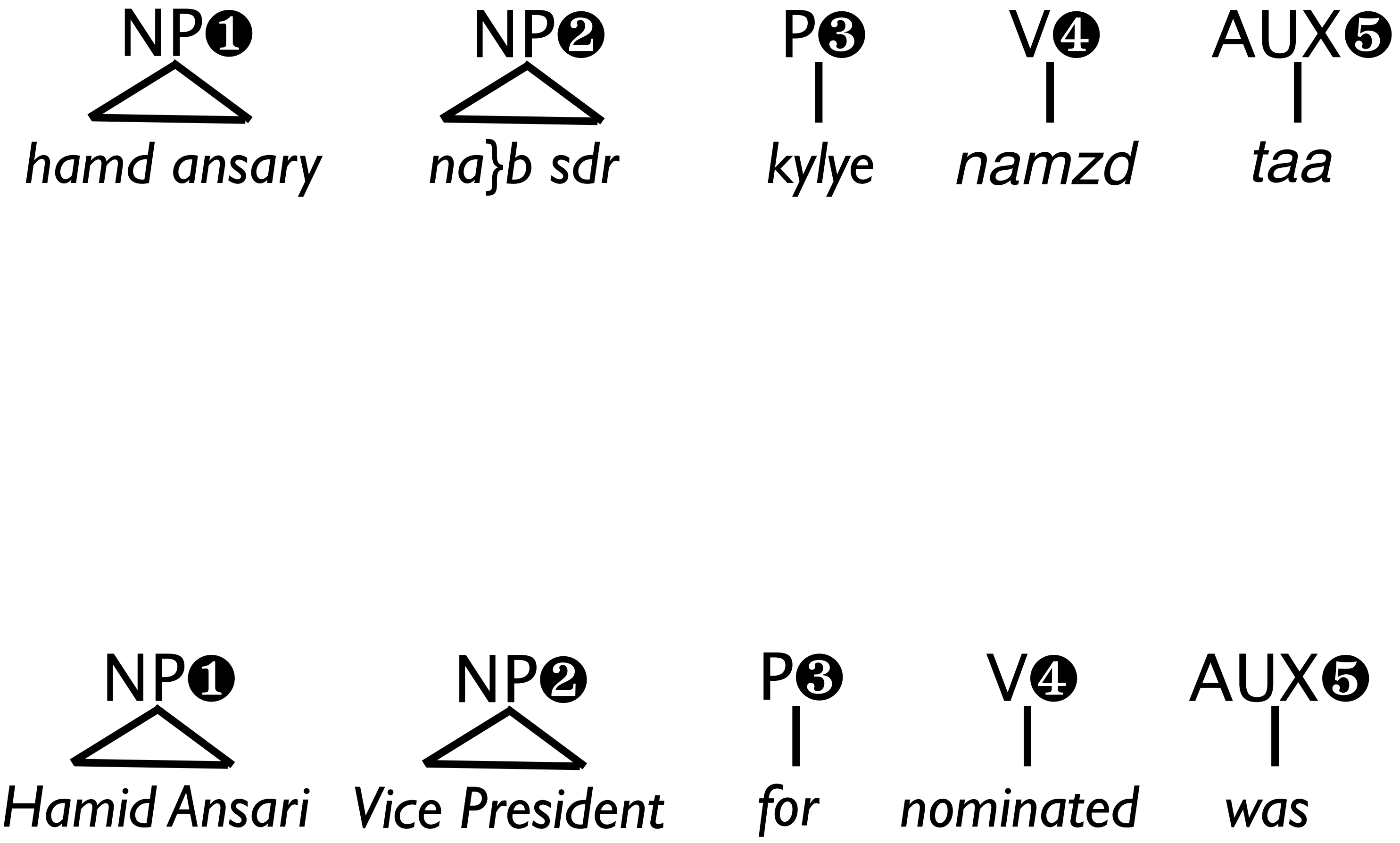
NP❶ NP❷ P❸ V❹ AUX❺

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



NP❶ NP❷ P❸ V❹ AUX❺

Hamid Ansari *Vice President* *for* *nominated* *was*

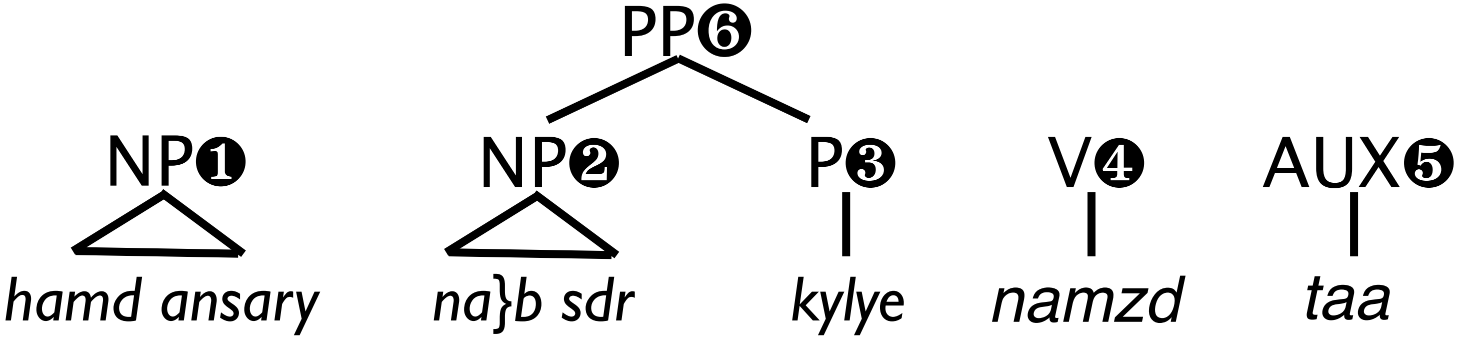


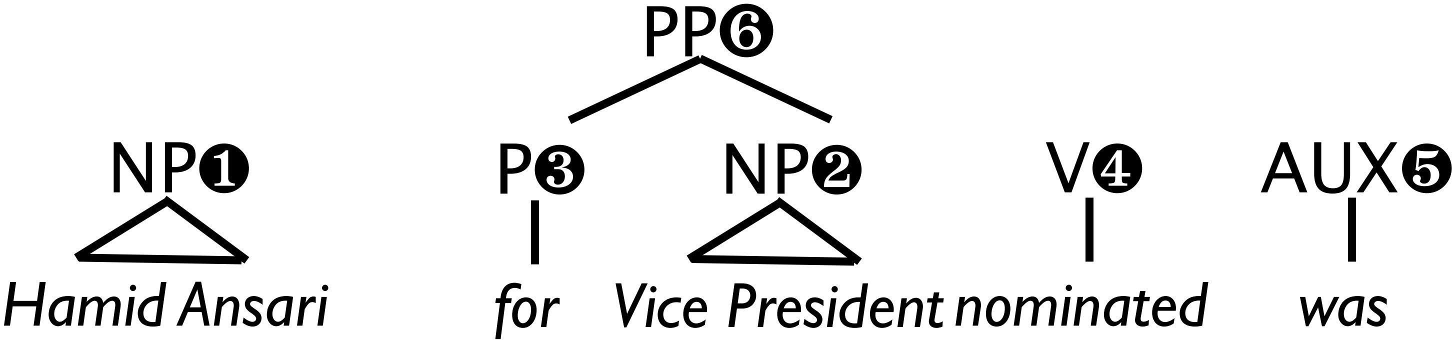
NP❶ NP❷ P❸ V❹ AUX❺

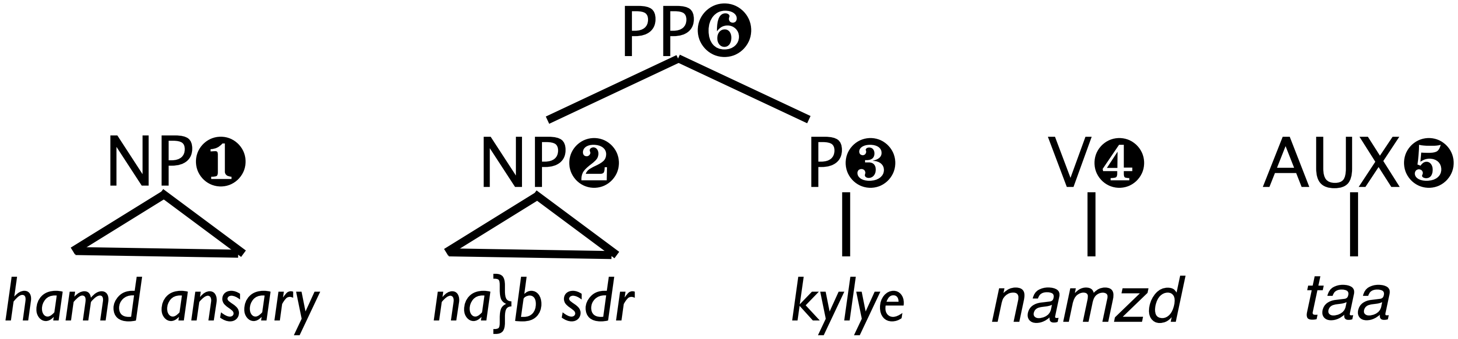
hamd ansary *na}b sdr* *kylye* *namzd* *taa*

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

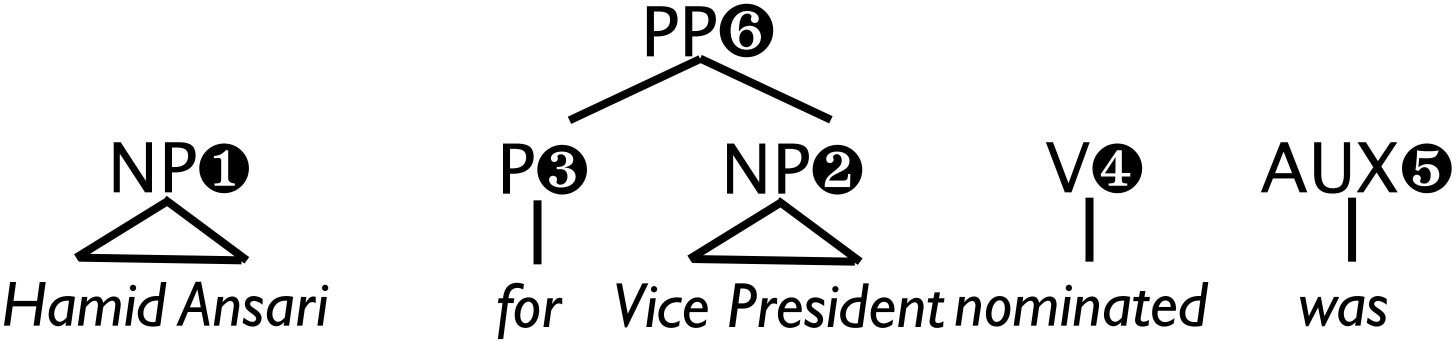

PP❻
NP❶
P❸
NP❷
V❹
AUX❺
Hamid Ansari
for
Vice President
nominated
was



PP❻

NP❶ NP❷ P❸ V❹ AUX❺

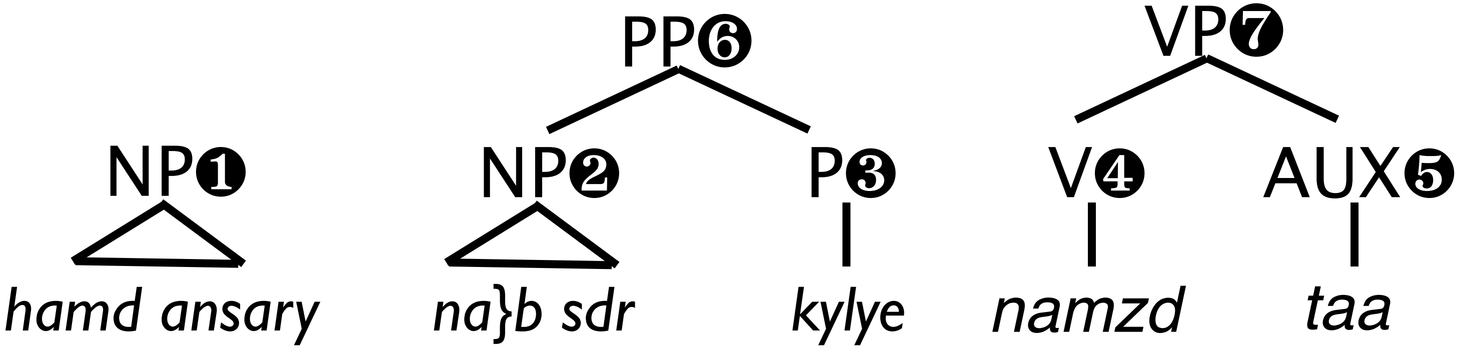
hamd ansary *na}b sdr* *kylye* *namzd* *taa*

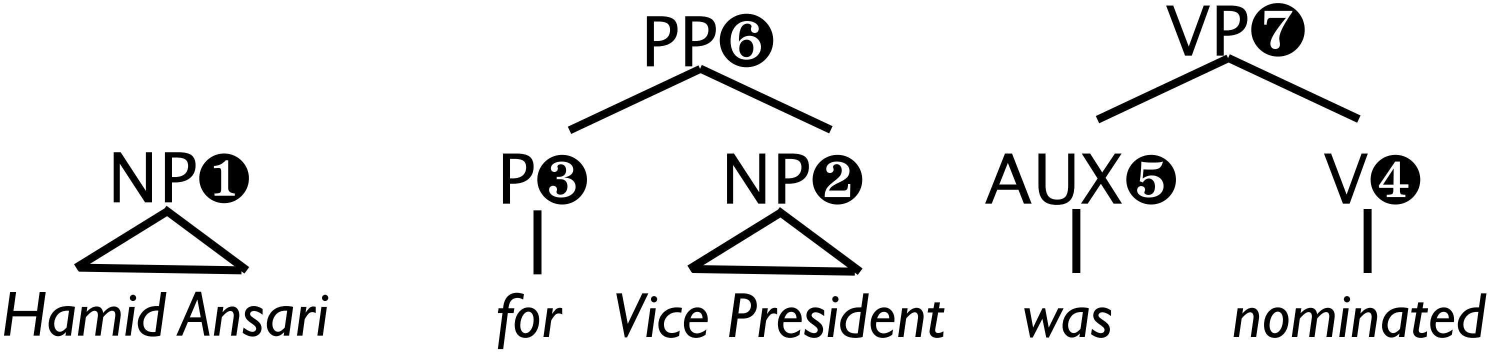


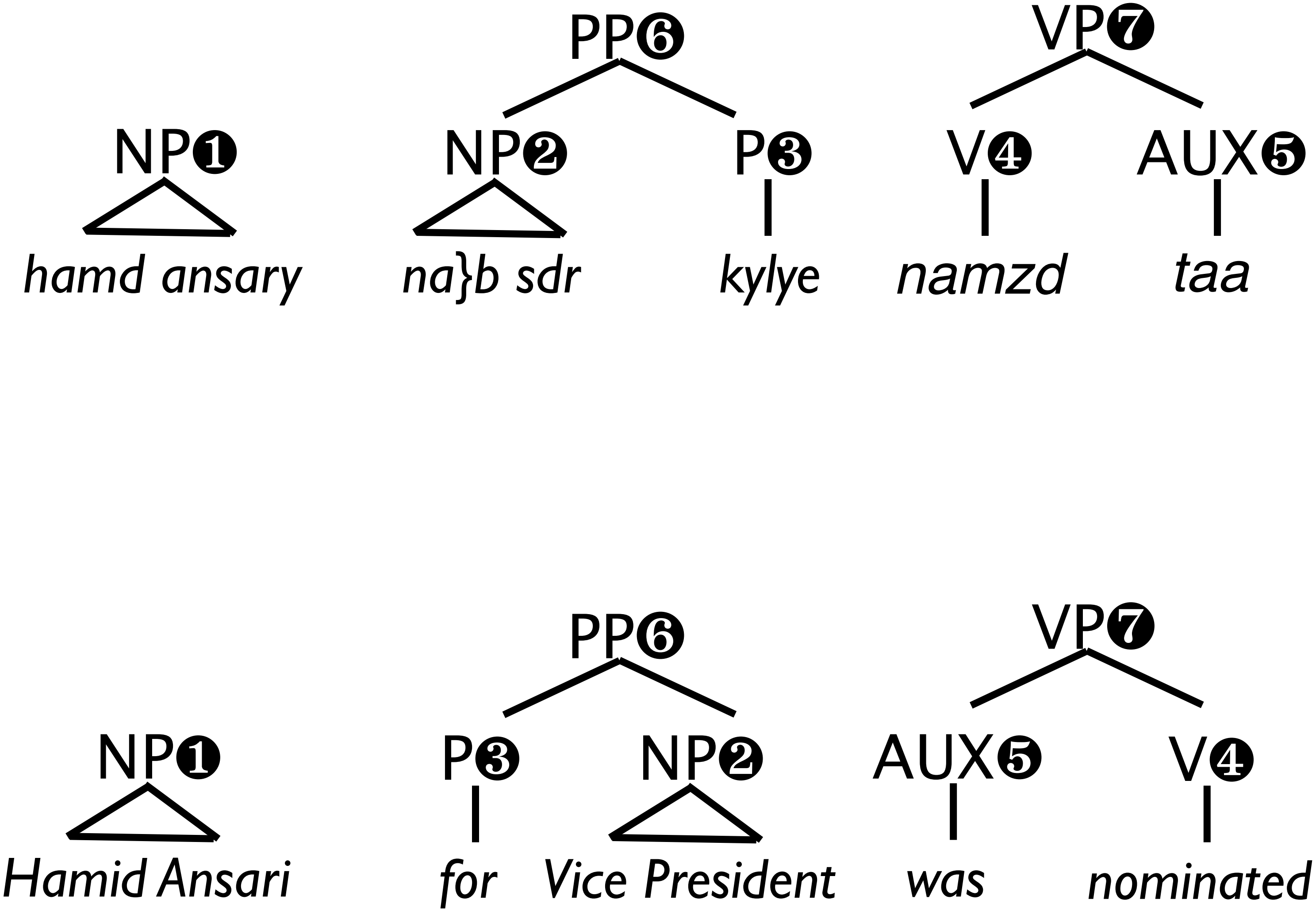
PP❻

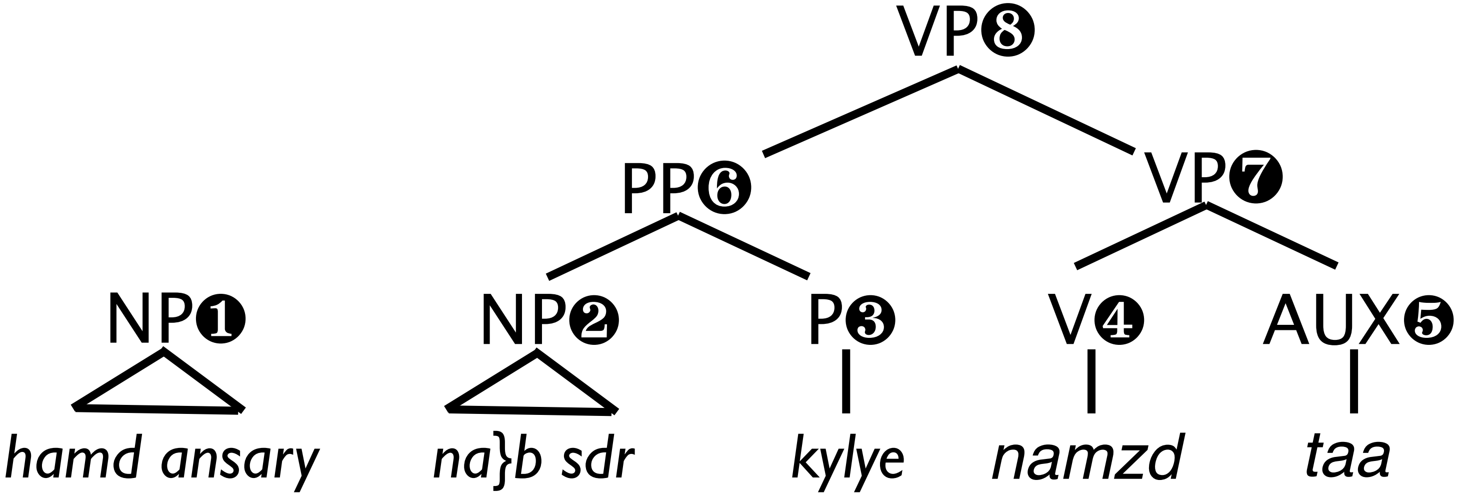
NP❶ P❸ NP❷ V❹ AUX❺

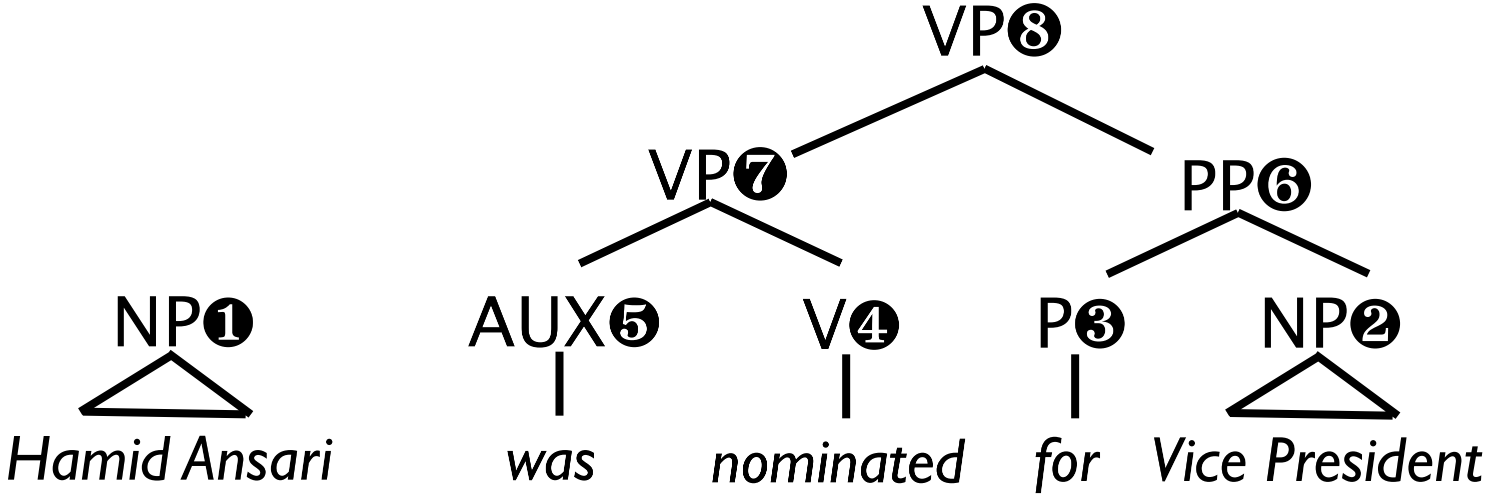
Hamid Ansari *for* *Vice President* *nominated* *was*

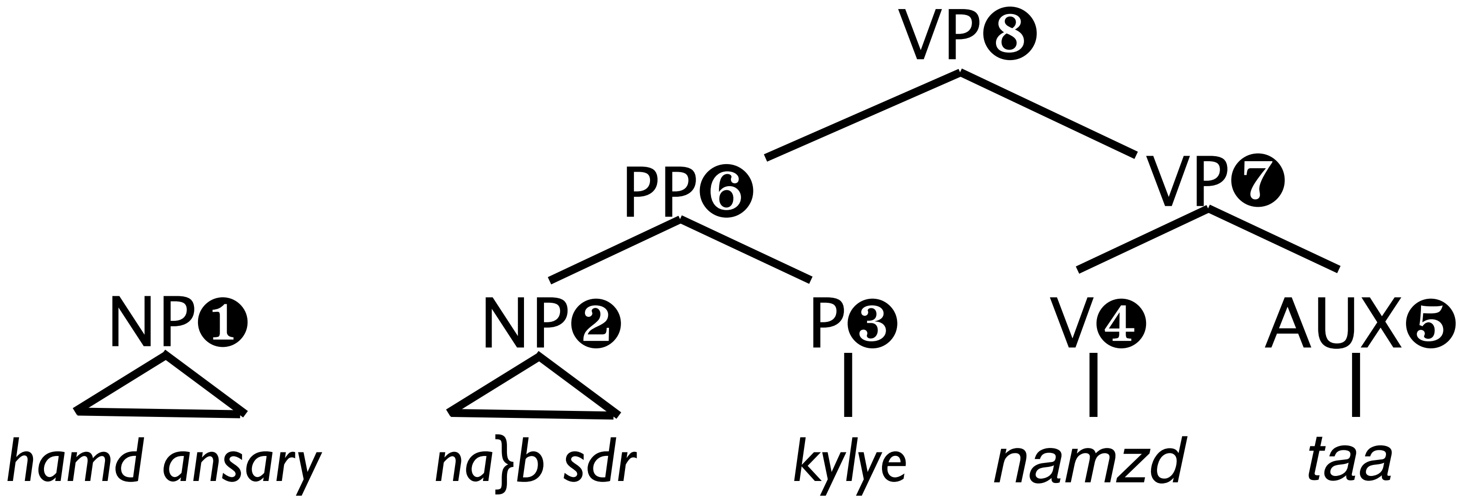

PP❻
VP❼
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa

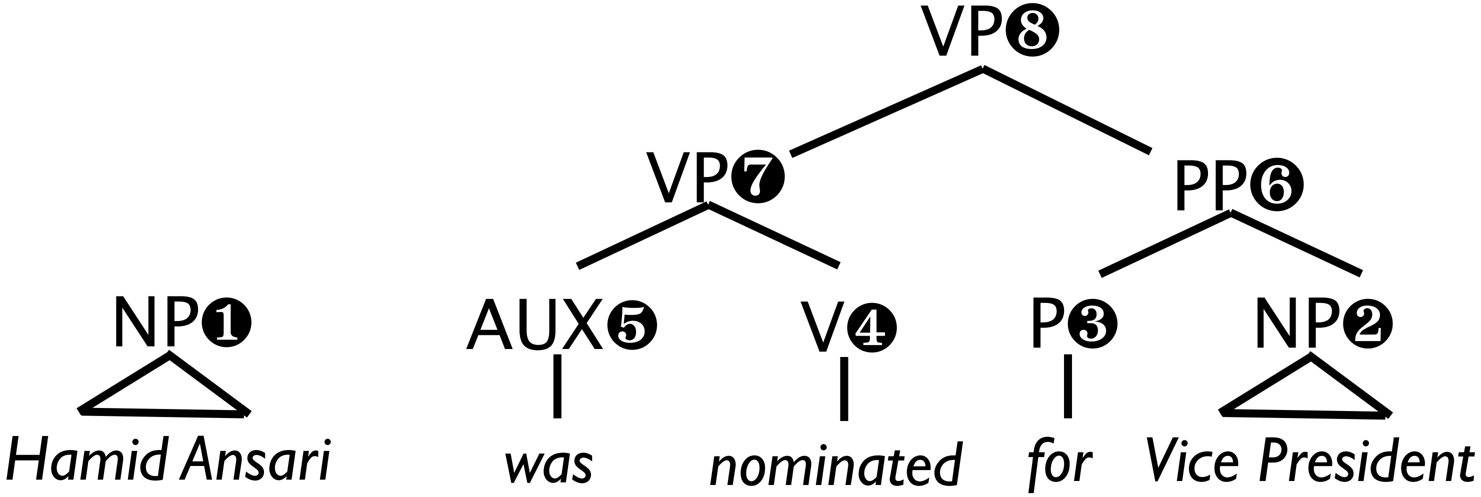

PP❻
VP❼
NP❶
P❸
NP❷
AUX❺
V❹
Hamid Ansari
for
Vice President
was
nominated

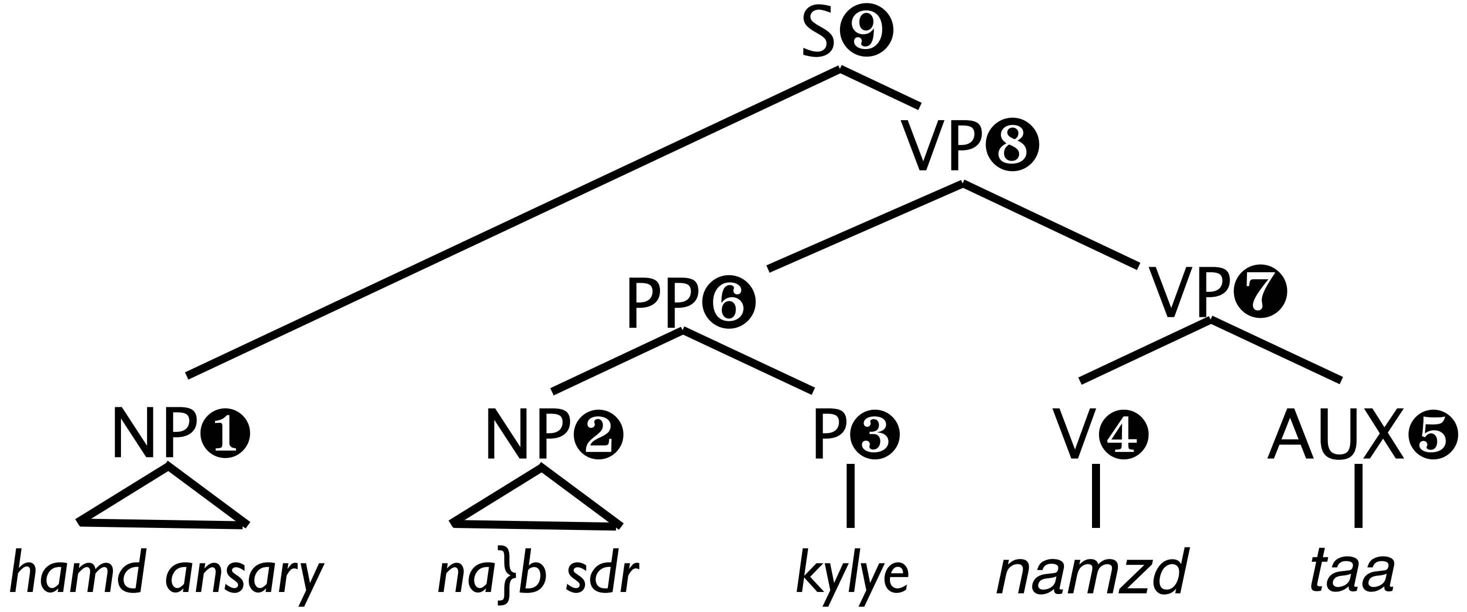

PP❻
VP❼
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa
PP❻
VP❼
NP❶
P❸
NP❷
AUX❺
V❹
Hamid Ansari
for
Vice President
was
nominated

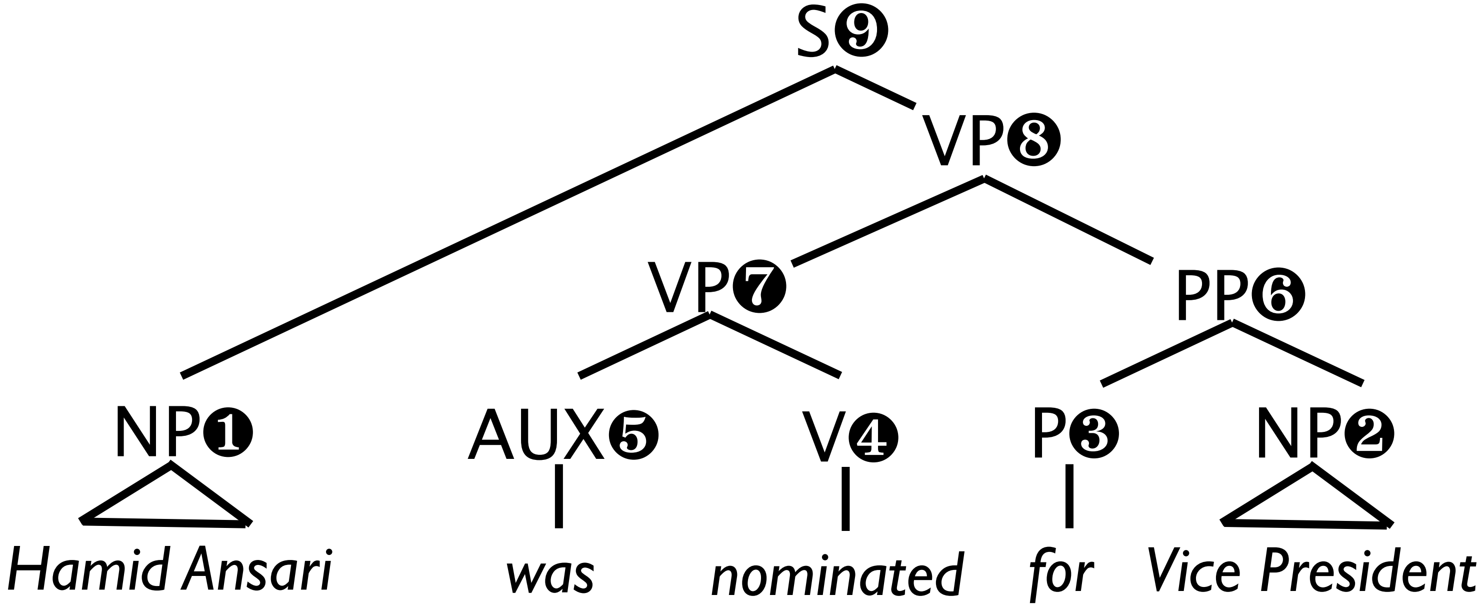

VP❽
PP❻
VP❼
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa


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


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

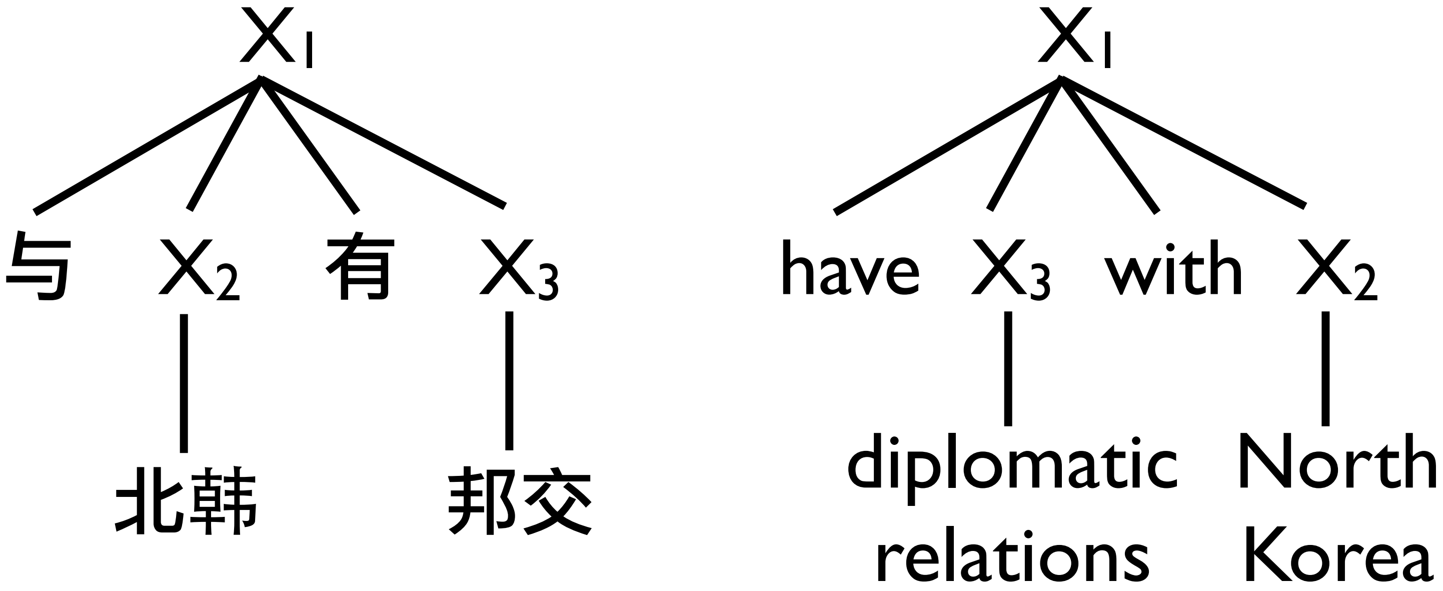

VP❽
VP❼
PP❻
NP❶
AUX❺
V❹
P❸
NP❷
Hamid Ansari
was
nominated
for
Vice President


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

Hiero-style SCFG rules

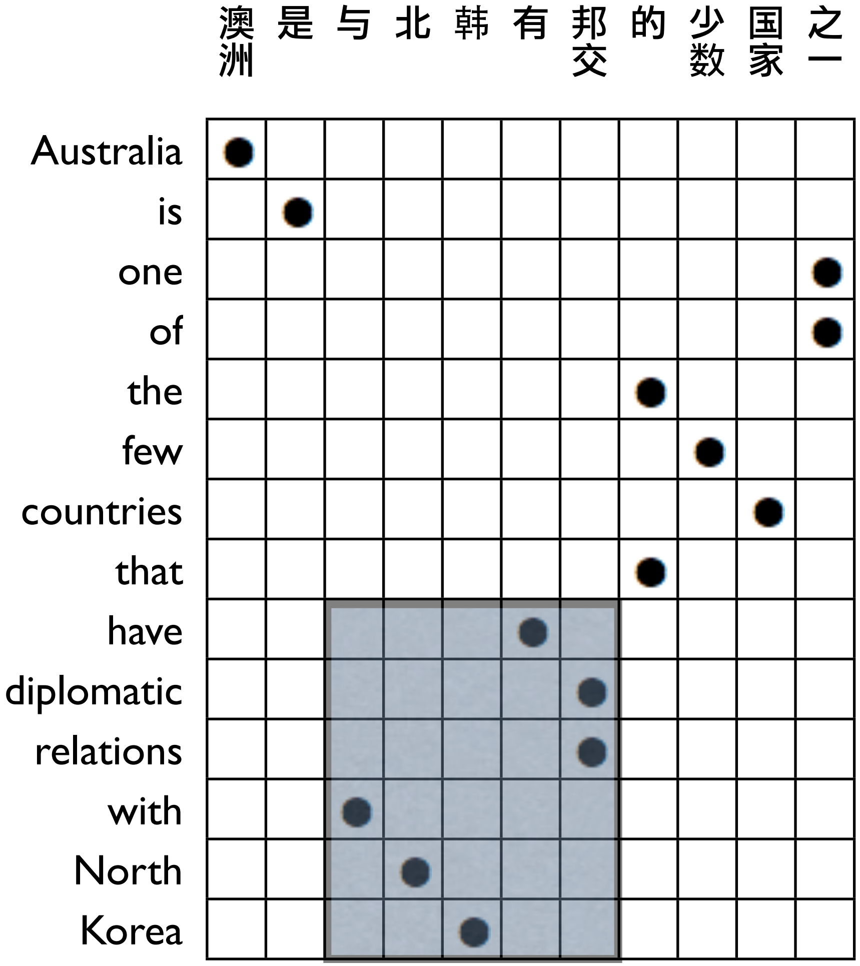
- Baseline only supported Hiero-style SCFG rules which have only one non-terminal symbol
- Not as nice as linguistically motivated rules, do not capture the reordering in Urdu



Extracting Hiero rules

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

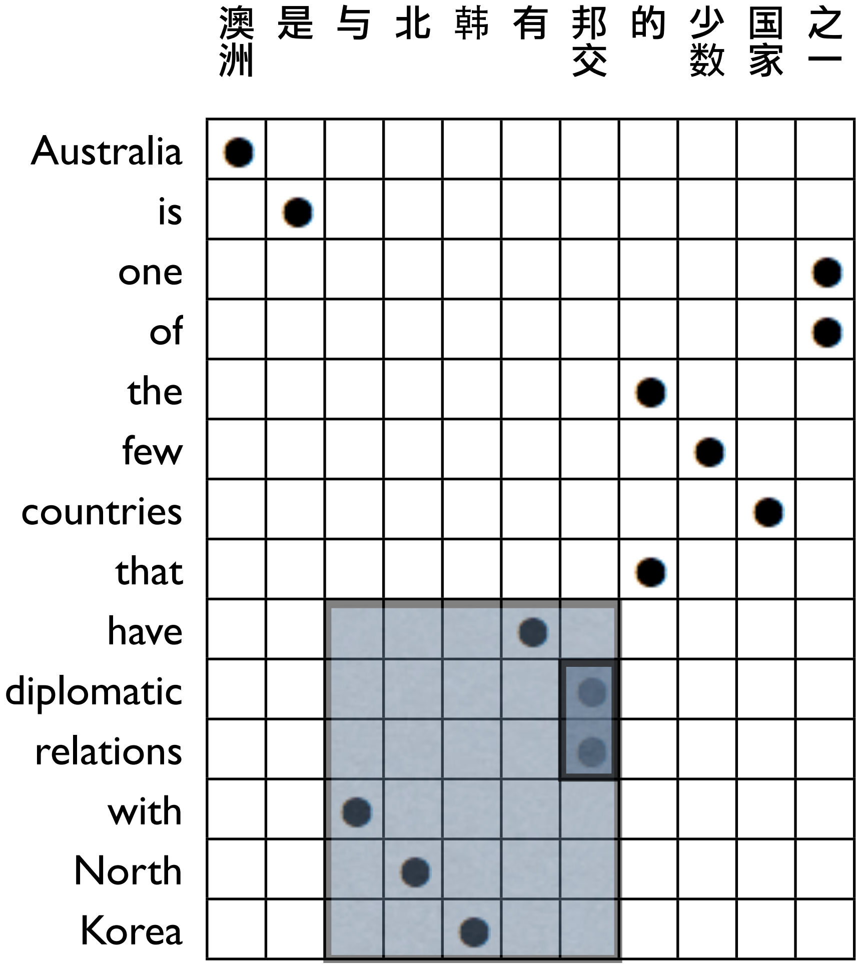
Extracting Hiero rules



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

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

Extracting Hiero rules

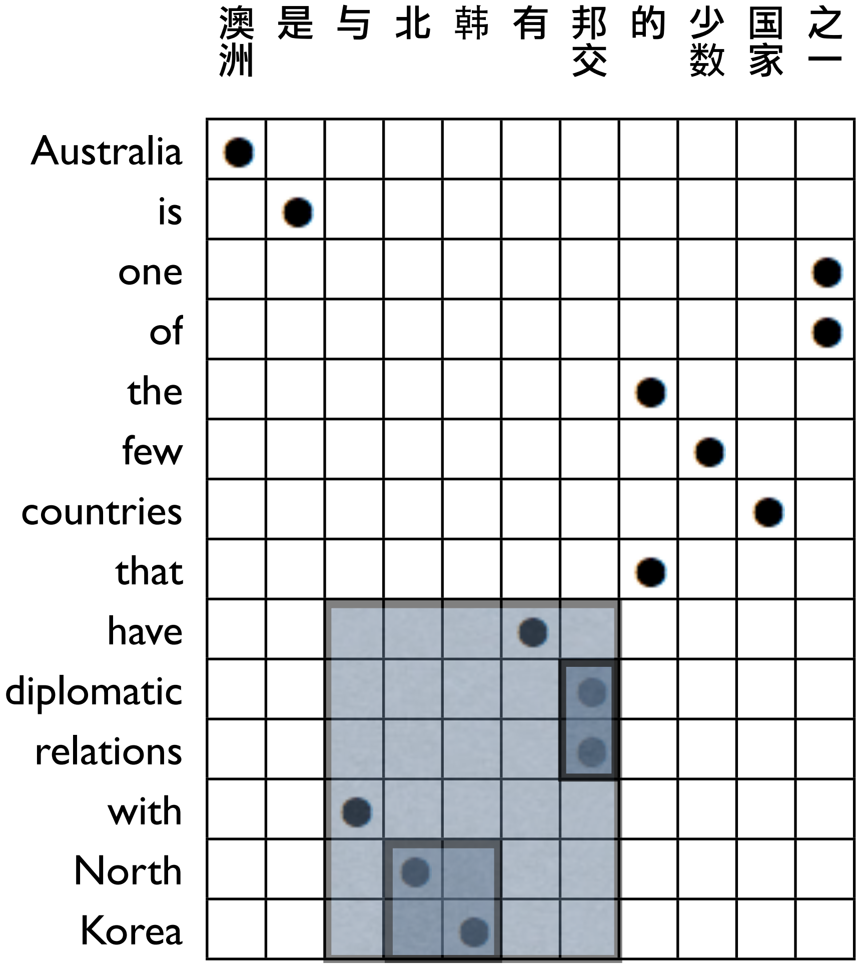


	澳洲	是	与	北	韩	有	邦交	的	少数	国家	之一
Australia	●										
is		●									
one											●
of											●
the								●			
few									●		
countries										●	
that								●			
have						●					
diplomatic							●				
relations							●				
with			●								
North				●							
Korea					●						

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

X → 邦交,
diplomatic relations

Extracting Hiero rules

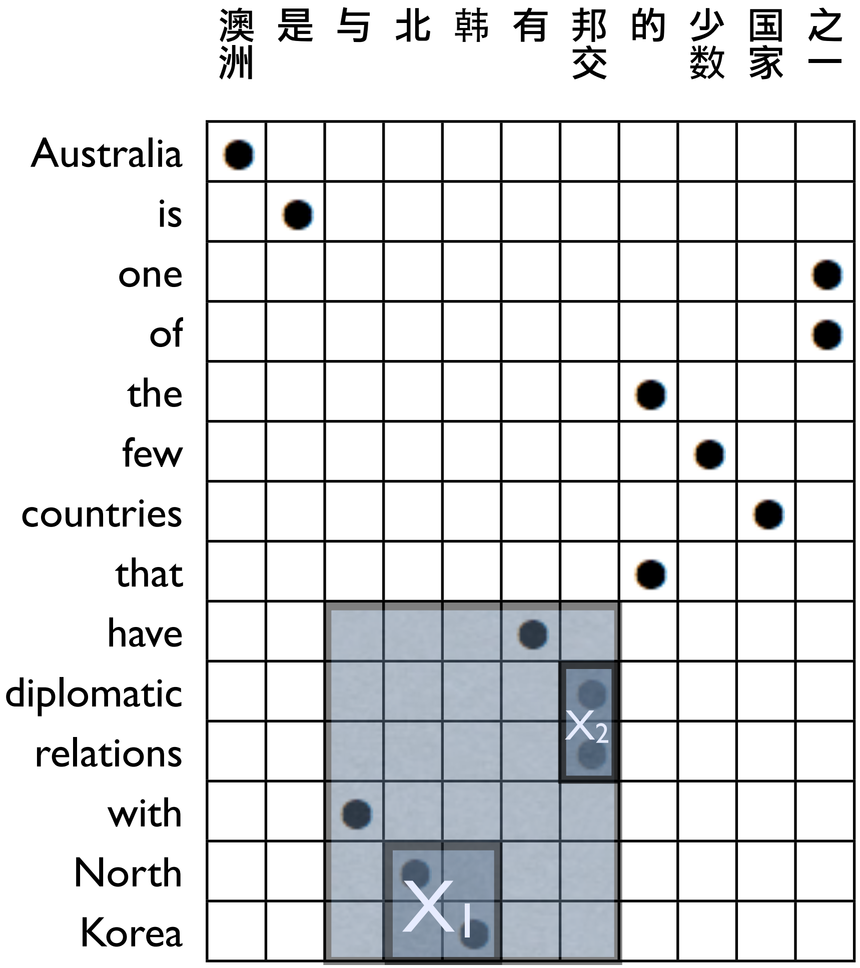


X → 与 北 韩 有 邦交,
have diplomatic relations
with North Korea

X → 邦交,
diplomatic relations

X → 北 韩,
North Korea

Extracting Hiero rules

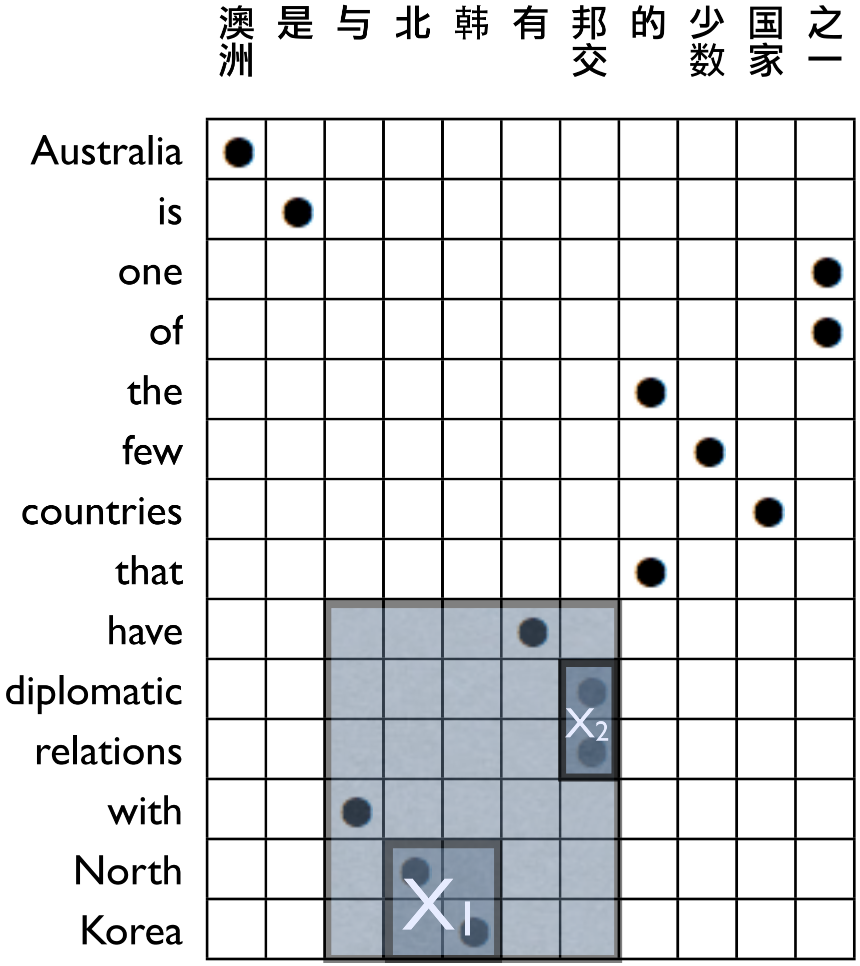


X → 与 北 韩 有 邦交,
have diplomatic relations
with North Korea

X → 邦交,
diplomatic relations

X → 北 韩,
North Korea

Extracting Hiero rules



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

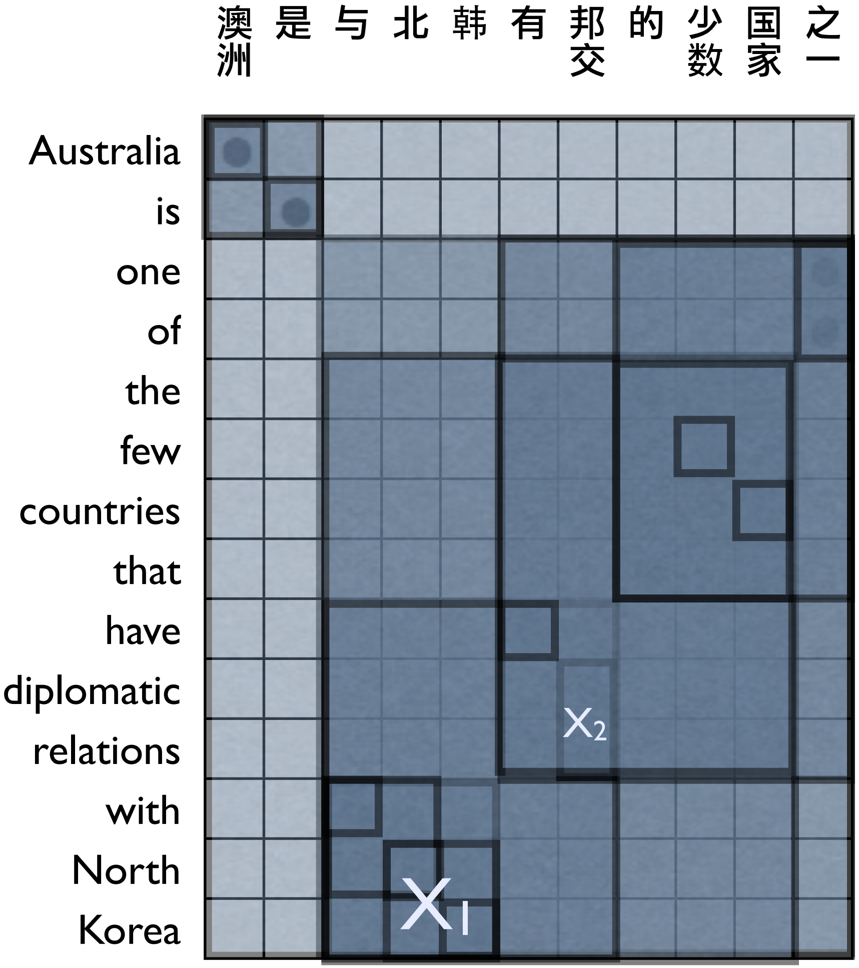
X → 与 北 韩 有 邦交,
have diplomatic relations with North Korea

X → 邦交,
diplomatic relations

X → 北 韩,
North Korea

X → 与 X_1 有 X_2,
have X_2 with X_1

Extracting Hiero rules



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

X → 邦交,
diplomatic relations

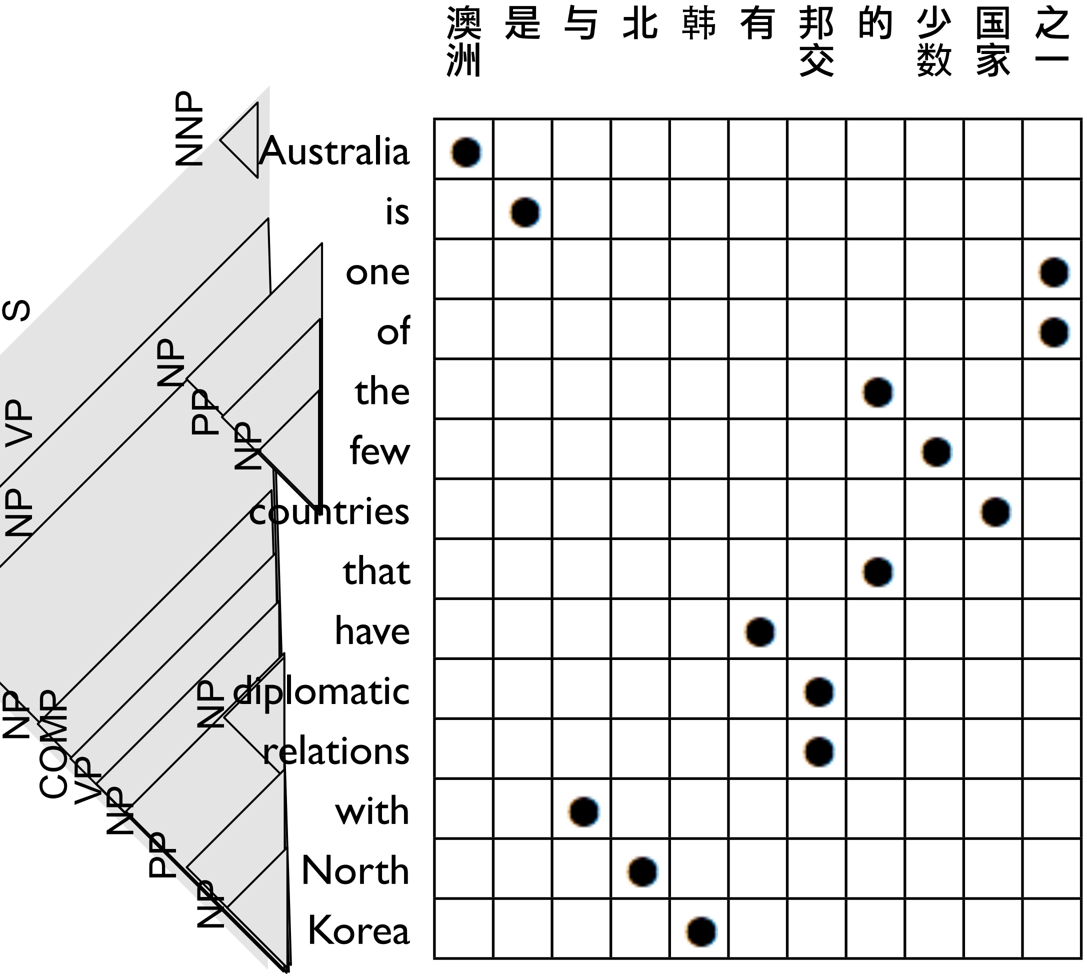
X → 北 韩,
North Korea

X → 与 X_1 有 X_2,
have X_2 with X_1

Extracting Syntactic Rules

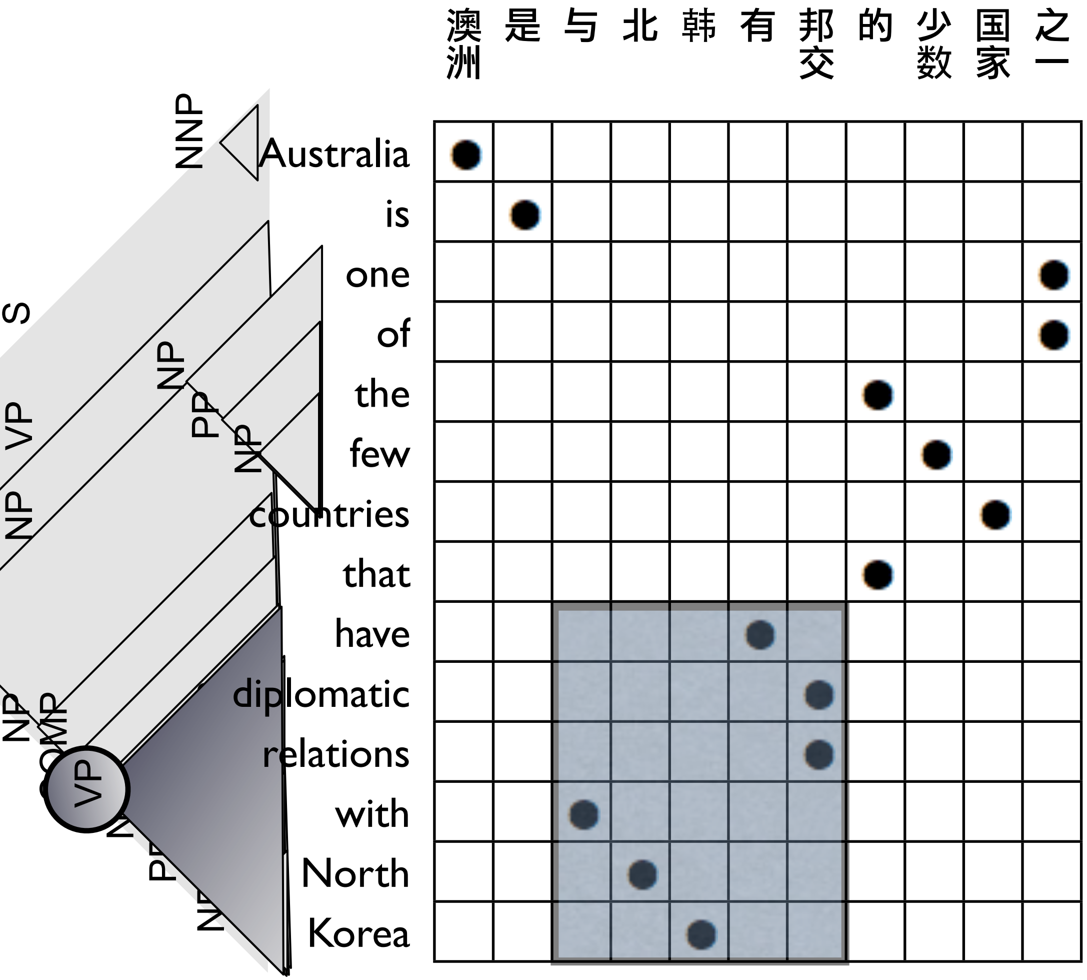
	澳洲	是	与	北	韩	有	邦交	的	少数	国家	之一
Australia	●										
is		●									
one											●
of											●
the								●			
few									●		
countries										●	
that								●			
have						●					
diplomatic							●				
relations							●				
with			●								
North				●							
Korea					●						

Extracting Syntactic Rules



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

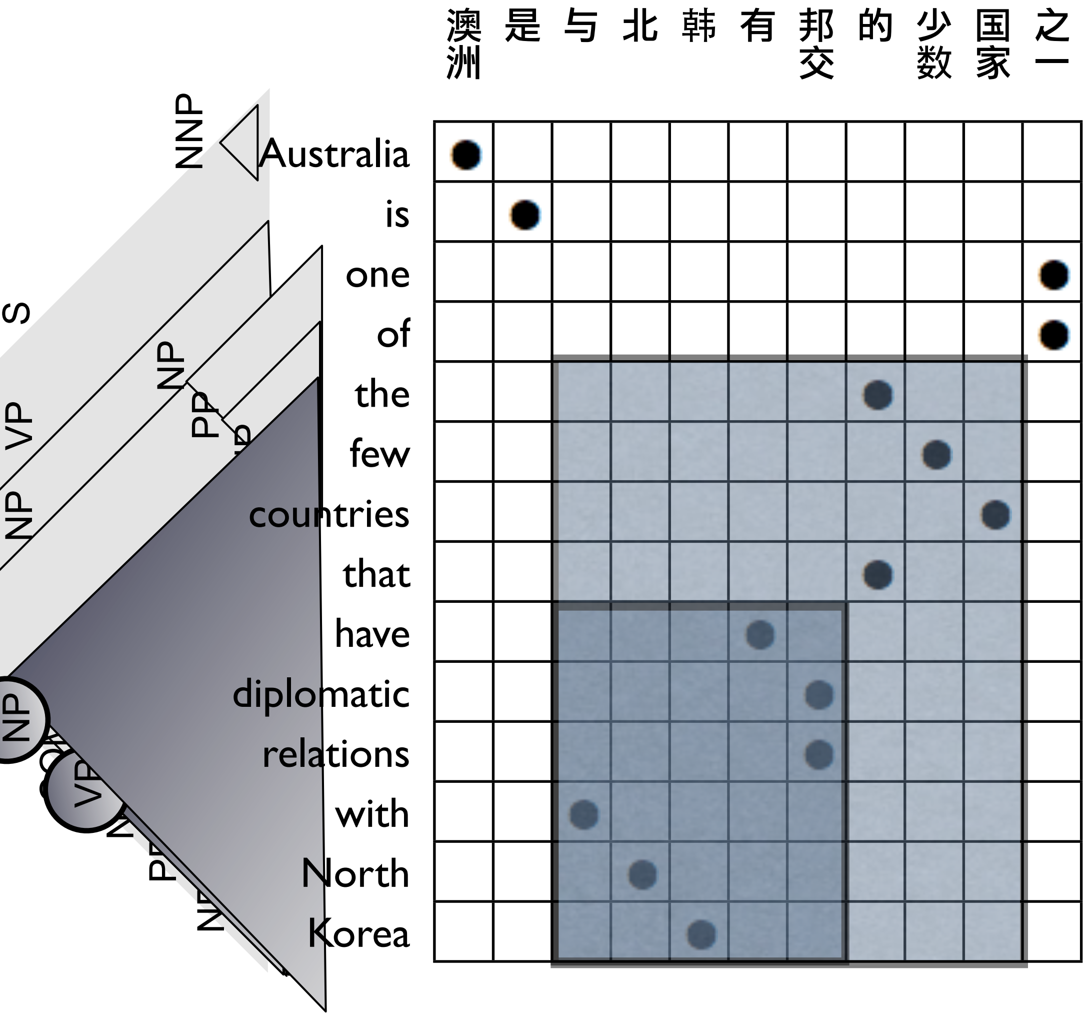
Extracting Syntactic Rules



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

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

Extracting Syntactic Rules

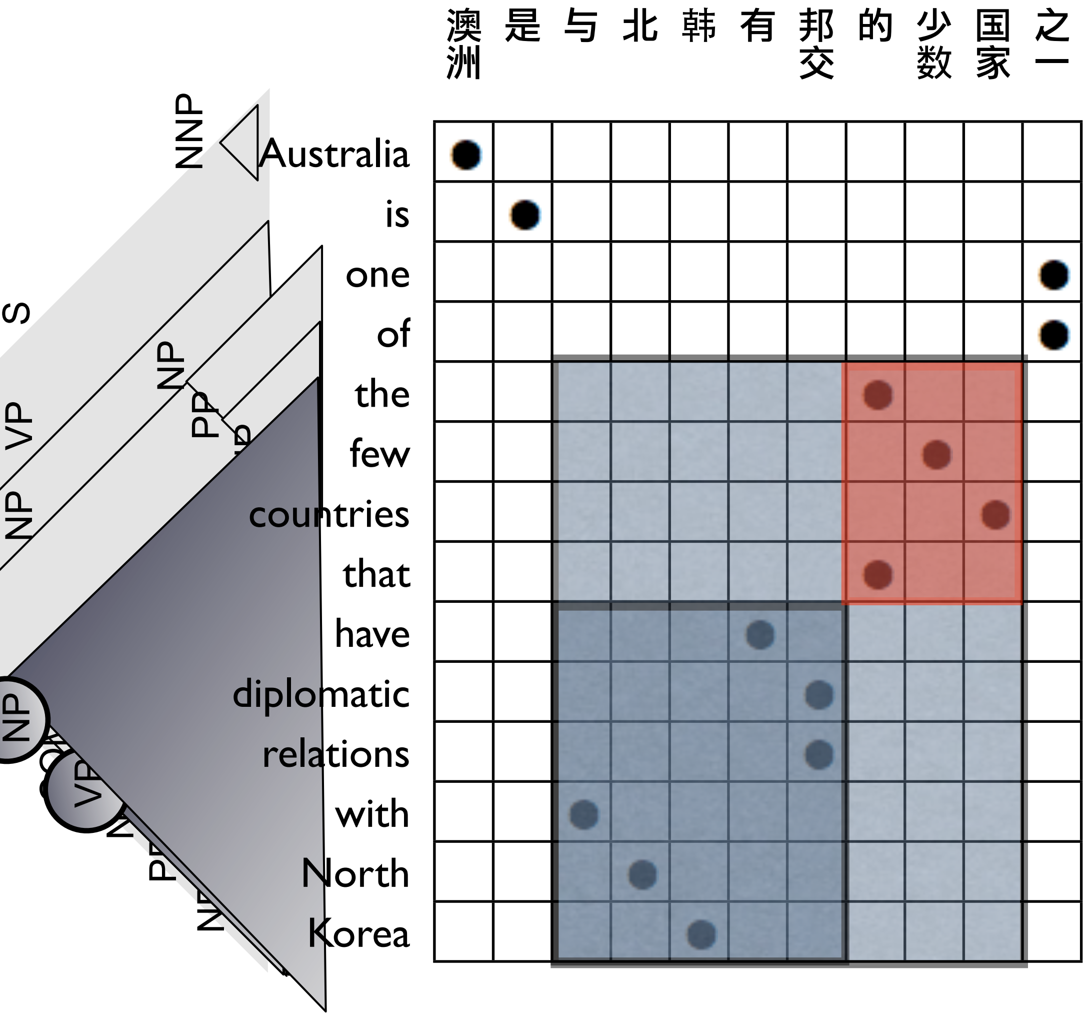


	澳洲	是	与	北	韩	有	邦交	的	少数	国家	之一
Australia	●										
is		●									
one											●
of											●
the								●			
few									●		
countries										●	
that								●			
have						●					
diplomatic							●				
relations							●				
with			●								
North				●							
Korea					●						

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

Extracting Syntactic Rules

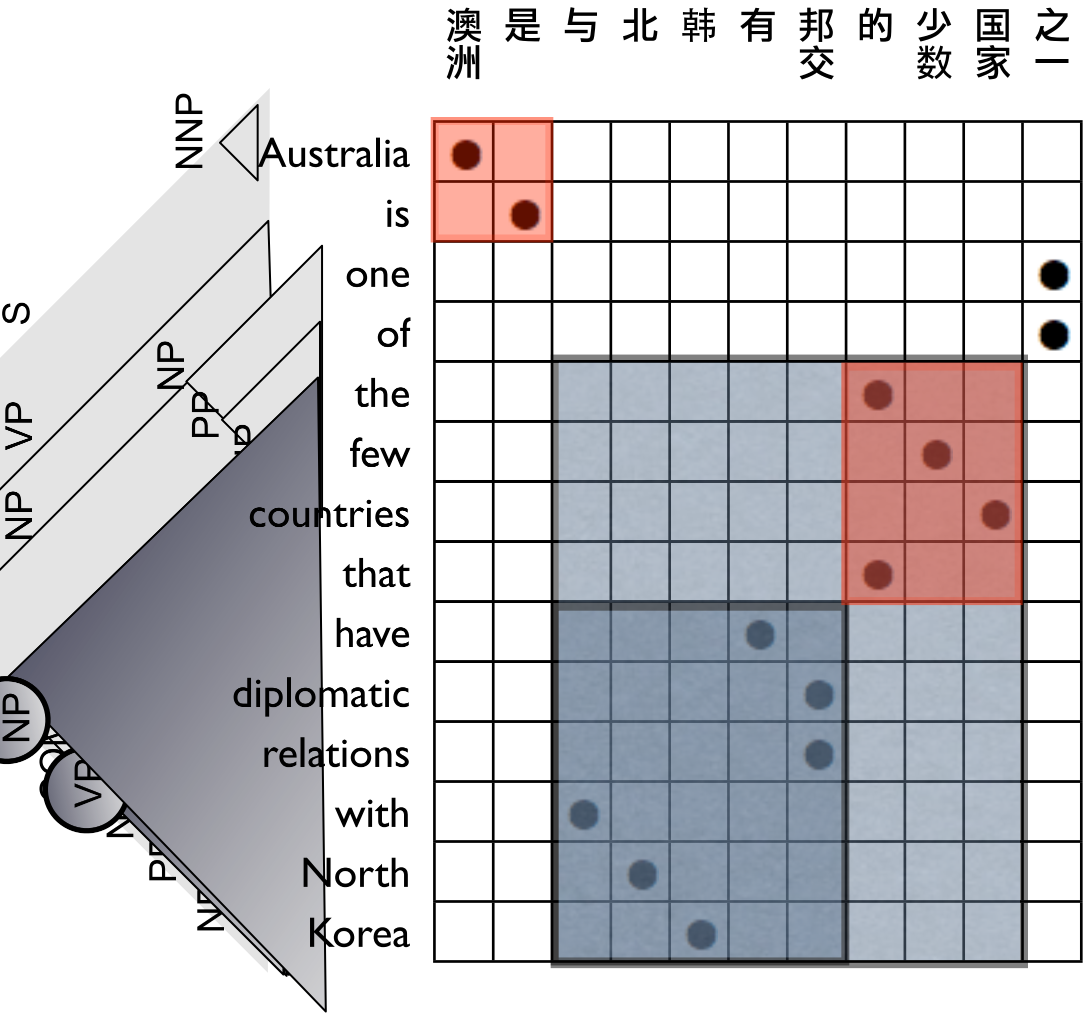


VP → 与 北 韩 有 邦交, have diplomatic relations with North Korea

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

Extracting Syntactic Rules



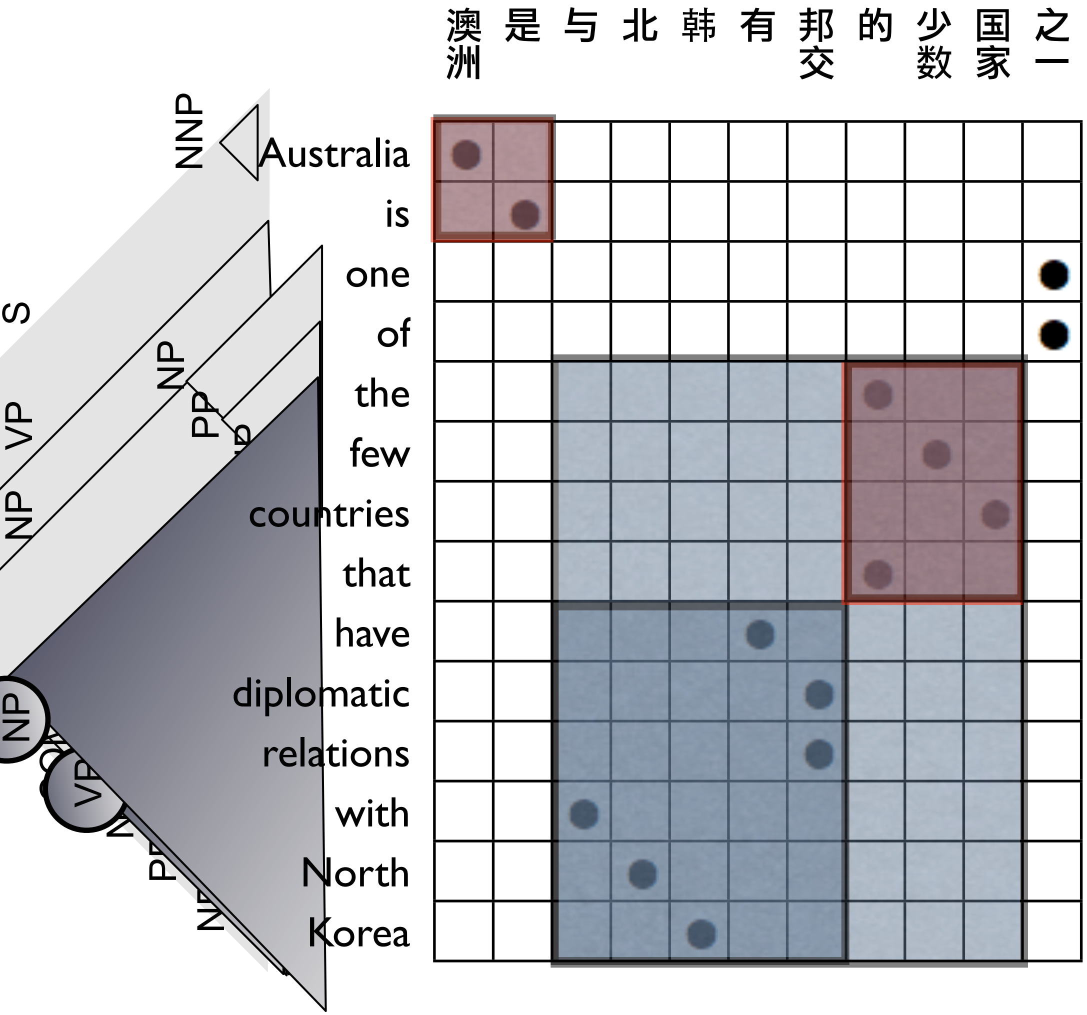
VP → 与 北 韩 有 邦交, have diplomatic relations with North Korea

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



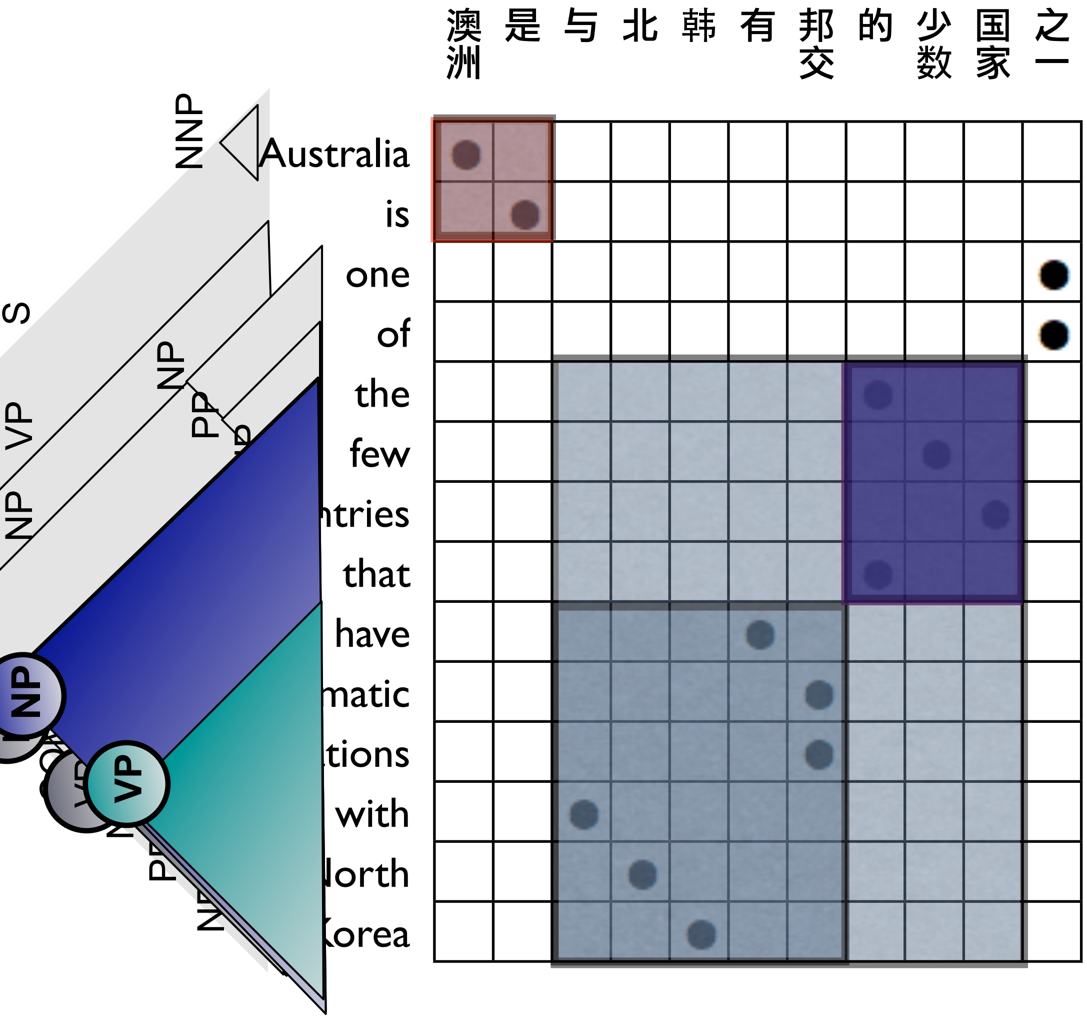
VP → 与 北 韩 有 邦交, have diplomatic relations with North Korea

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

??? → 的 少数 国家, the few countries that

??? → 澳洲 是, Australia is

Extracting Syntactic Rules



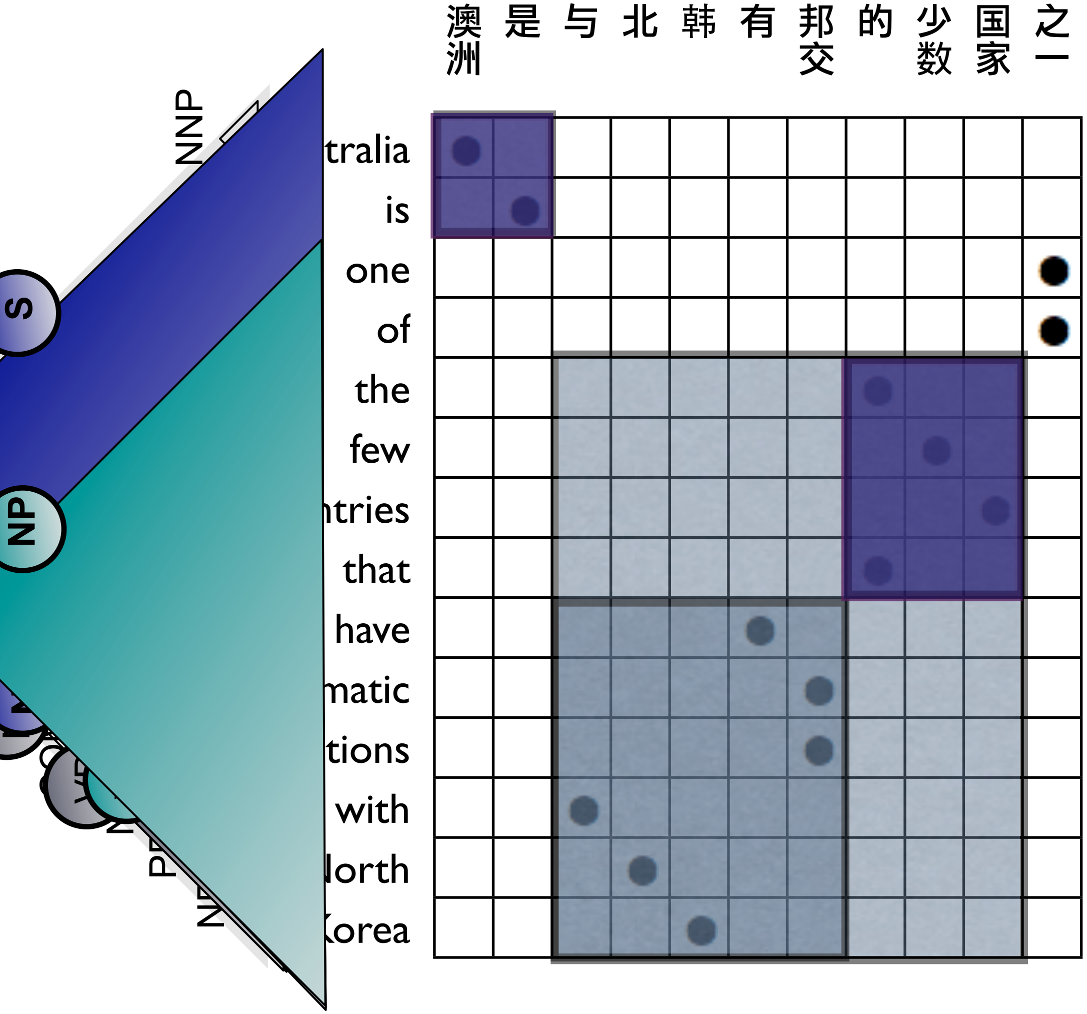
VP → 与 北 韩 有 邦交, have diplomatic relations with North Korea

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

??? → 澳洲 是, Australia is

Extracting Syntactic Rules



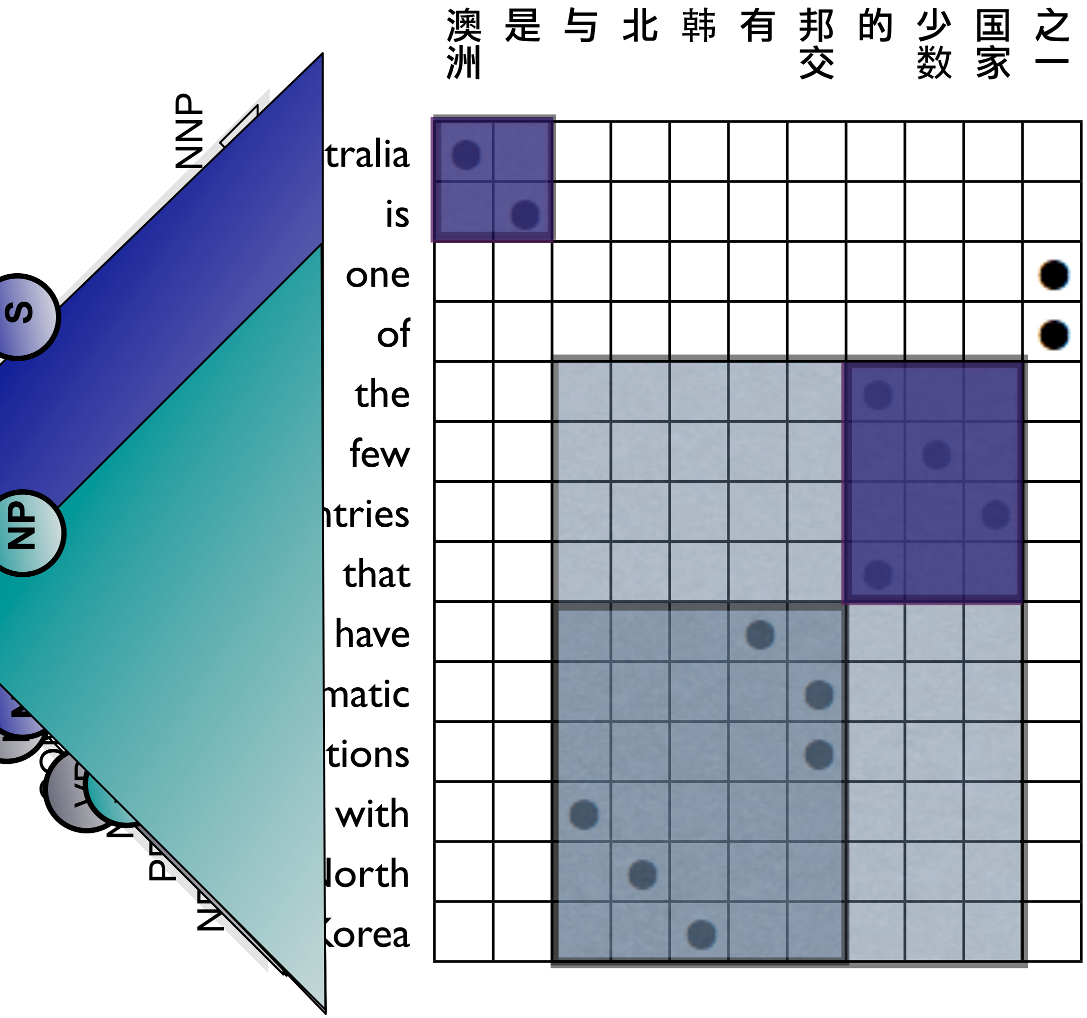
VP → 与 北 韩 有 邦交, have diplomatic relations with North Korea

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



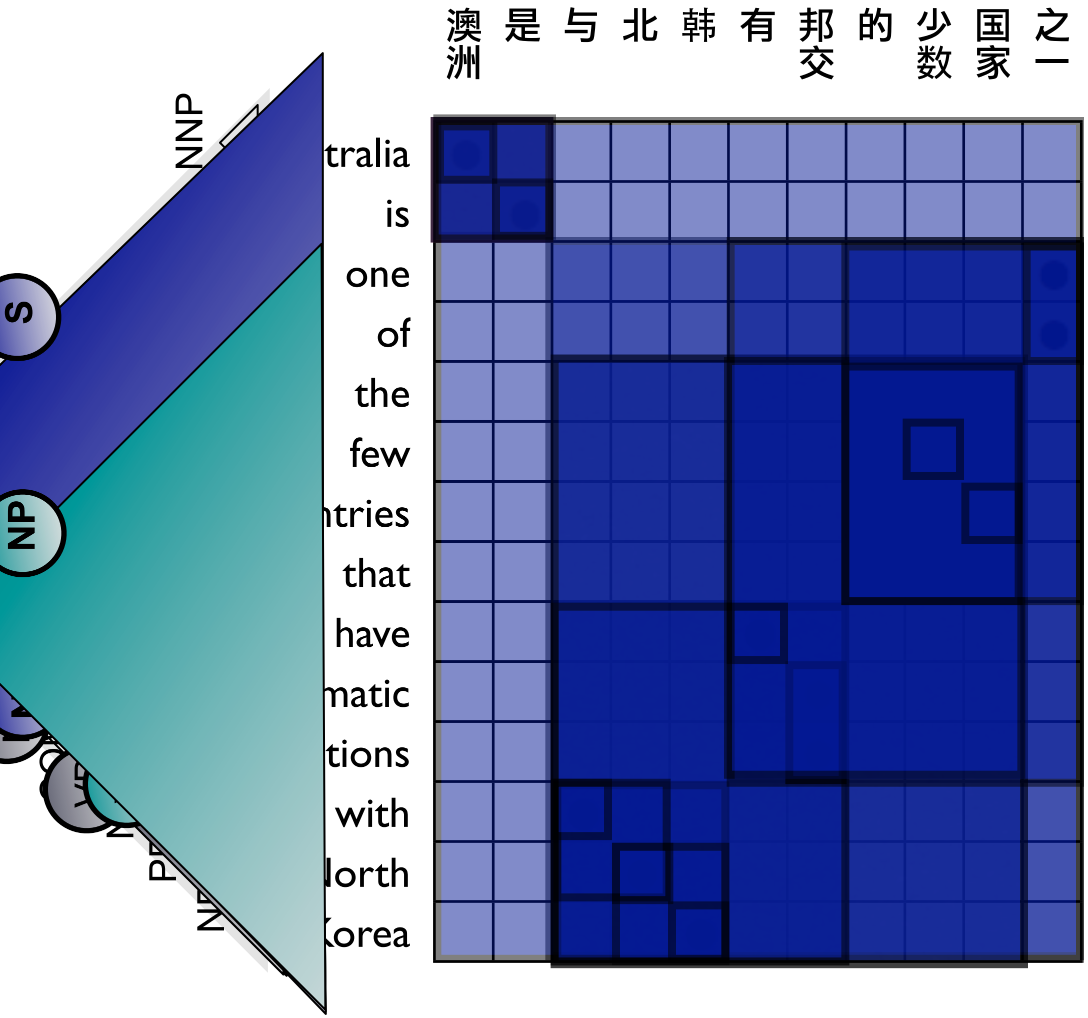
VP → 与 北 韩 有 邦交, have diplomatic relations with North Korea

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

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

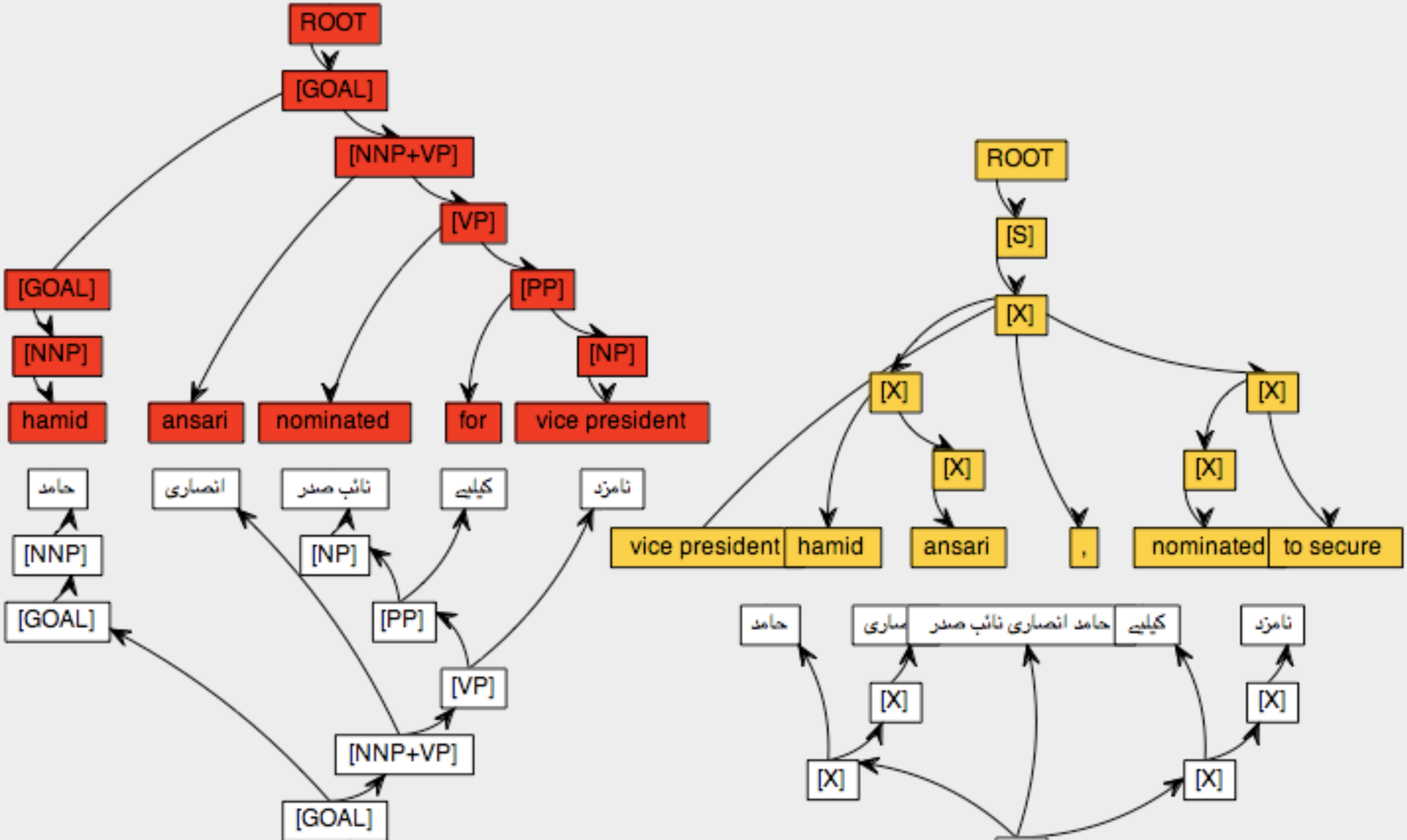
NP/VP → 的 少数 国家, the few countries that

S/NP → 澳洲 是, Australia is

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

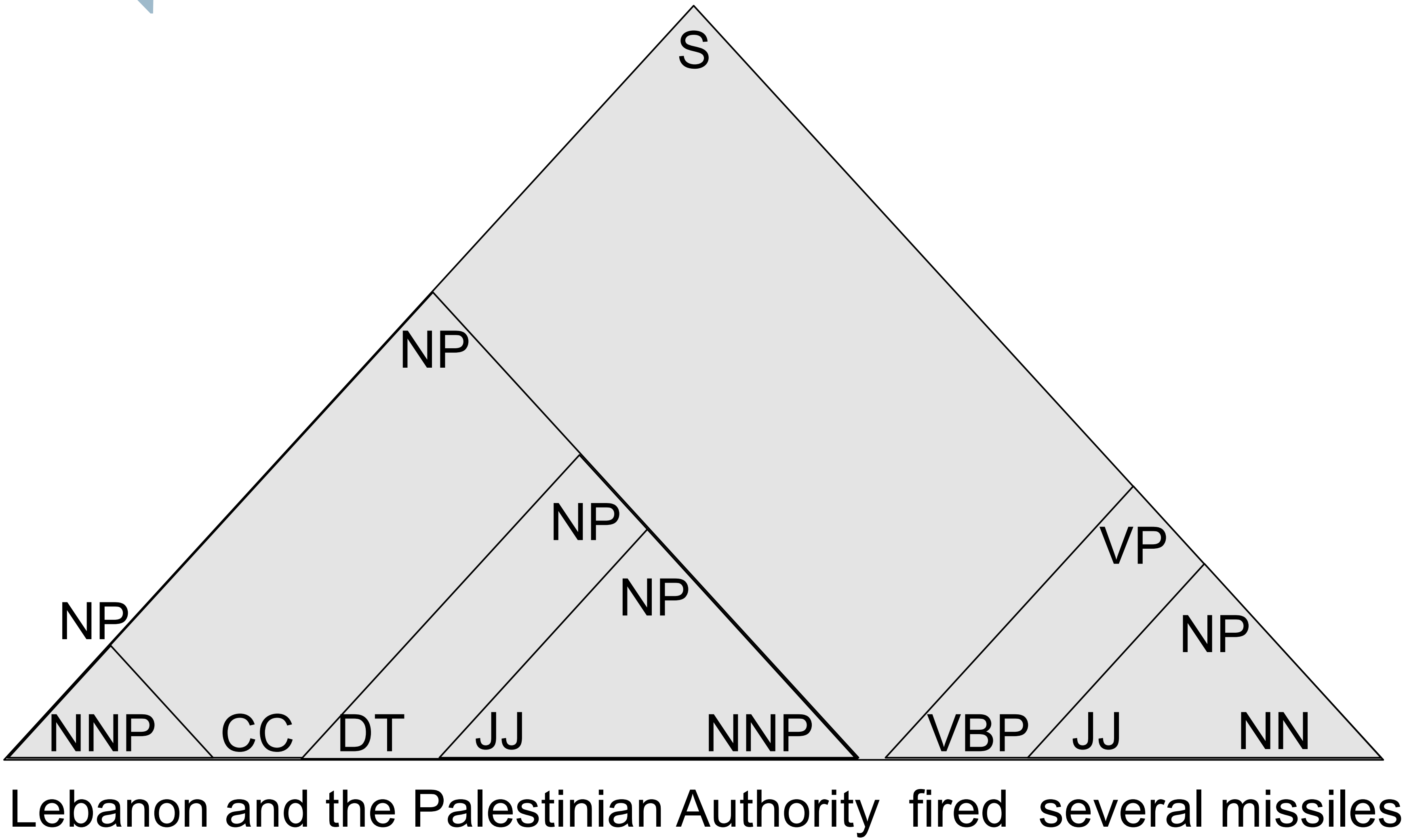
Syntax captures Urdu reordering



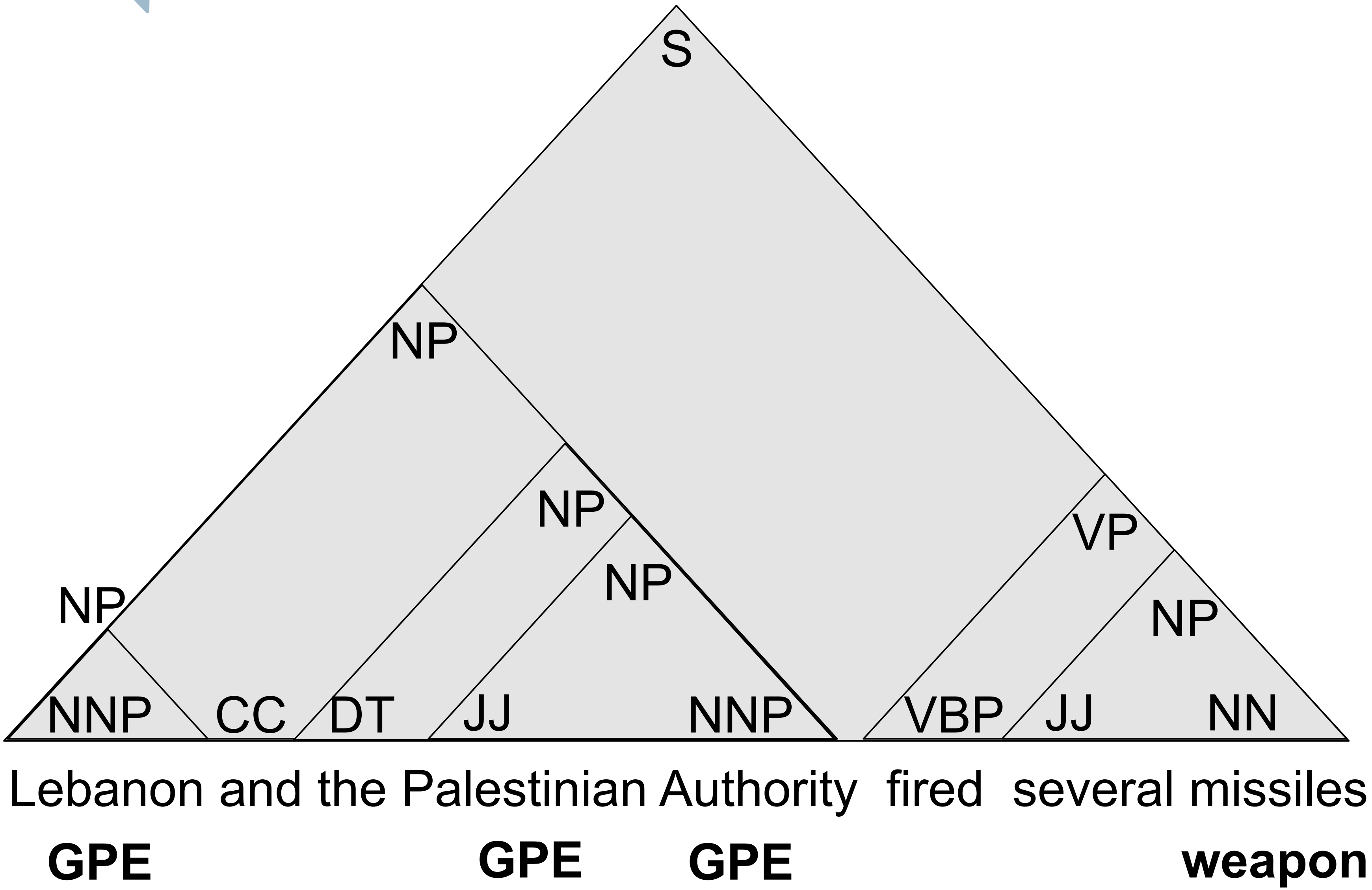
Integrating HIVEs into SCFGs

- A primary goal of the workshop was to include semantic entities in the translation process
- Semantic entities included high information value elements such as
  - Named entities
  - Modality
- These elements are included in SCFGs by grafting onto parse trees

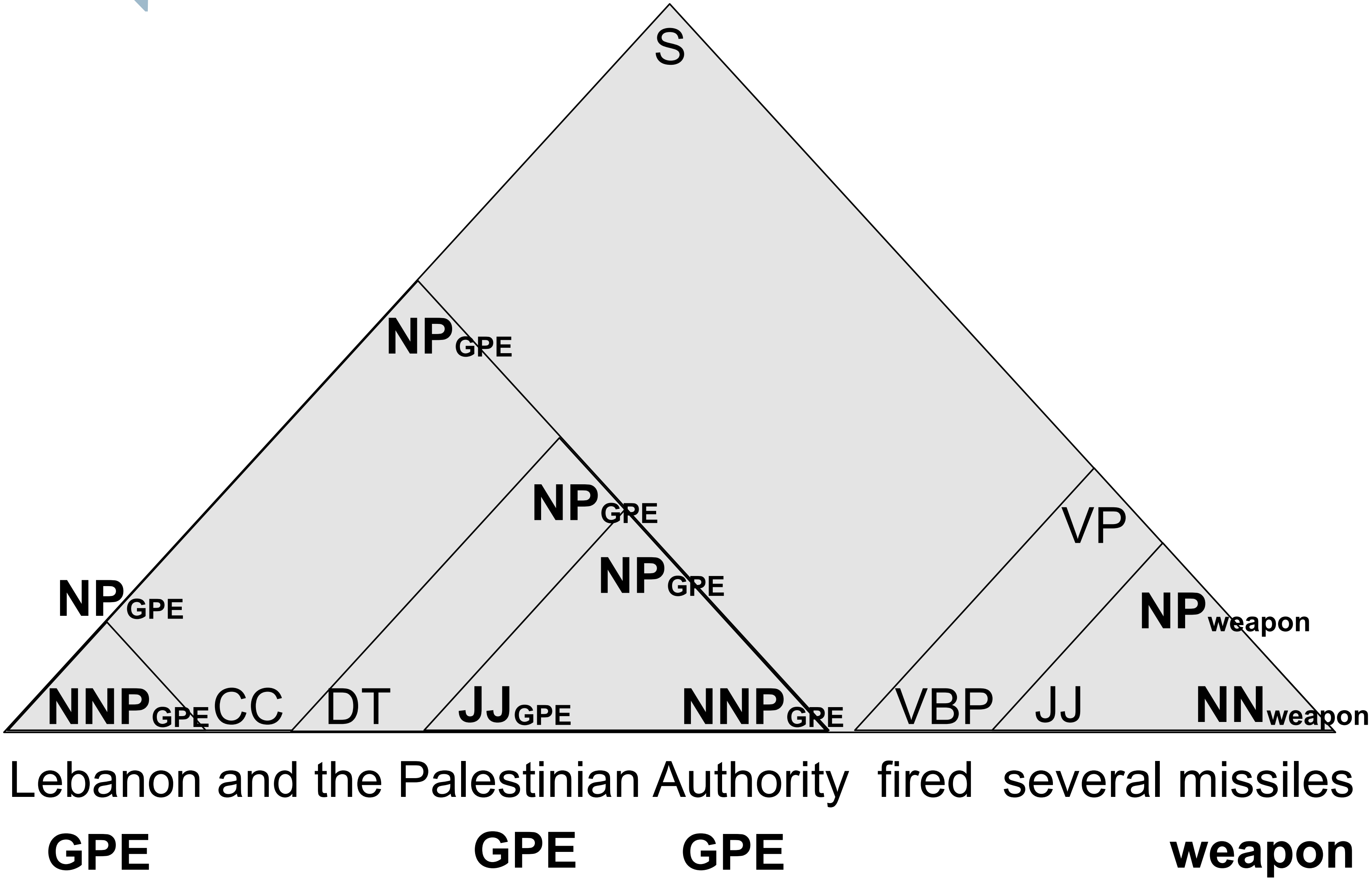
Example tree graft


S
NP
NP
NP
NP
VP
NP
NNP
CC
DT
JJ
NNP
VBP
JJ
NN
Lebanon and the Palestinian Authority fired several missiles

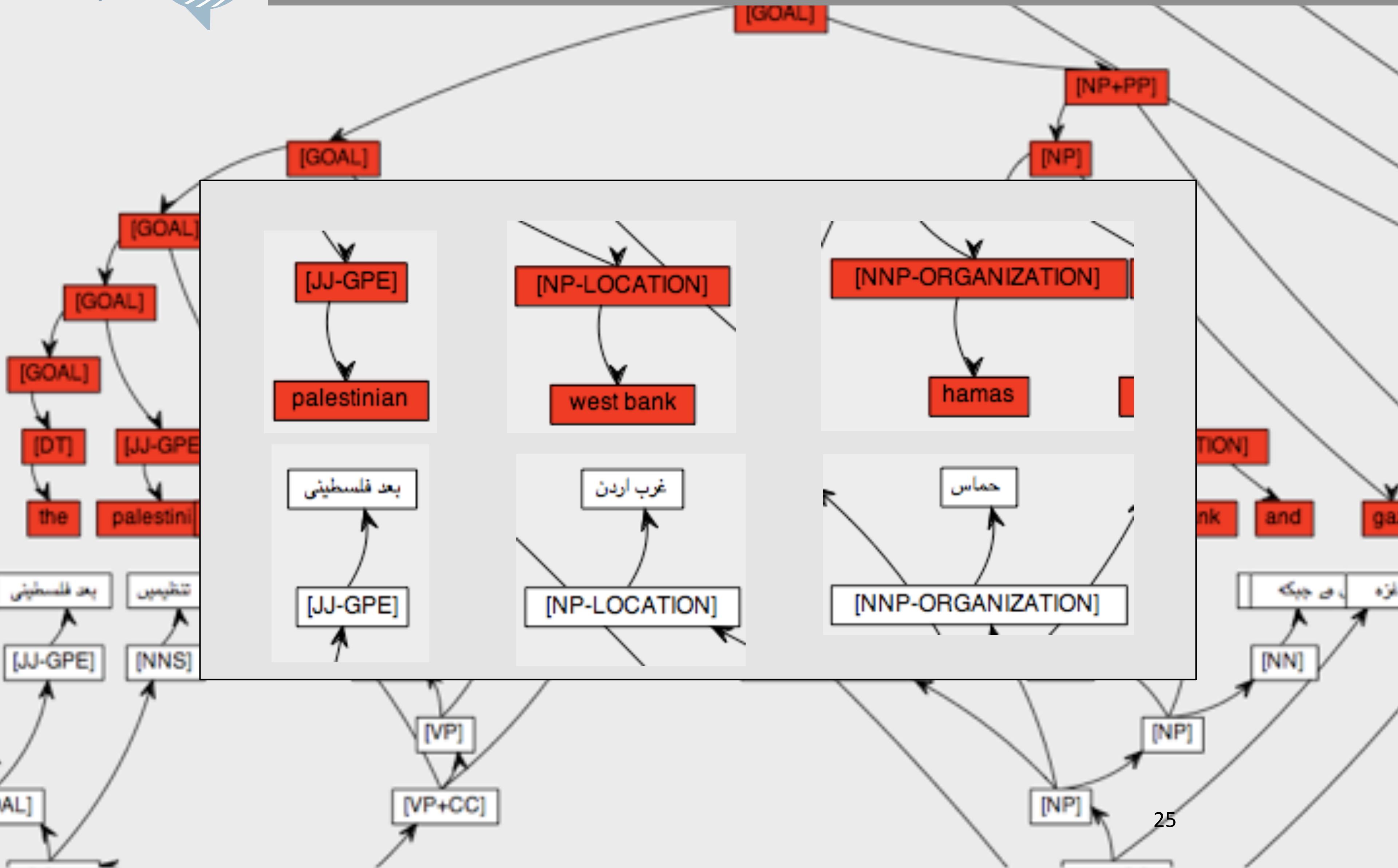
Example tree graft


S
NP
NP
NP
NP
VP
NP
NNP
CC
DT
JJ
NNP
VBP
JJ
NN
Lebanon and the Palestinian Authority fired several missiles
GPE
GPE
GPE
weapon

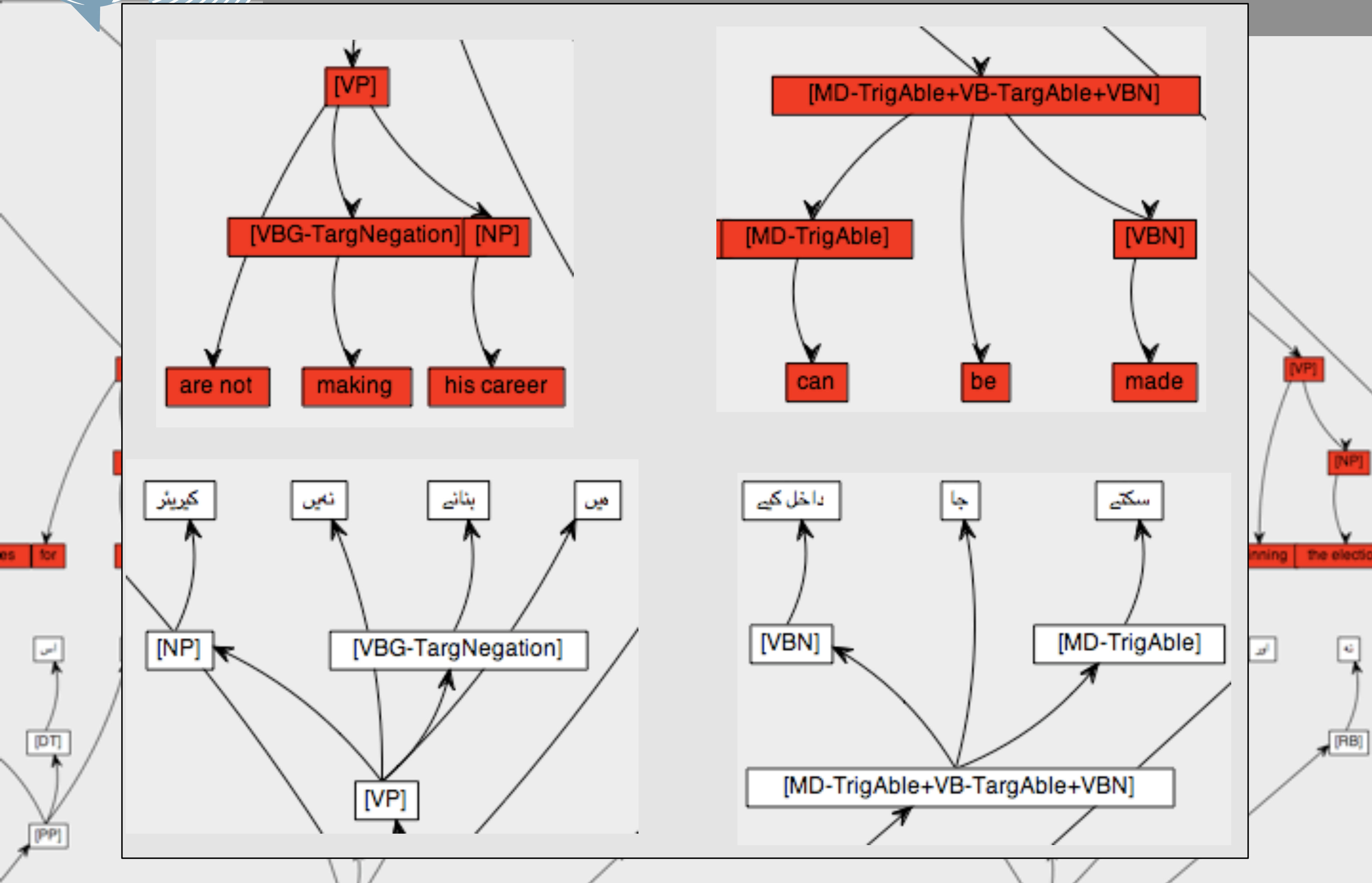
Example tree graft



Simple Urdu NE tagger

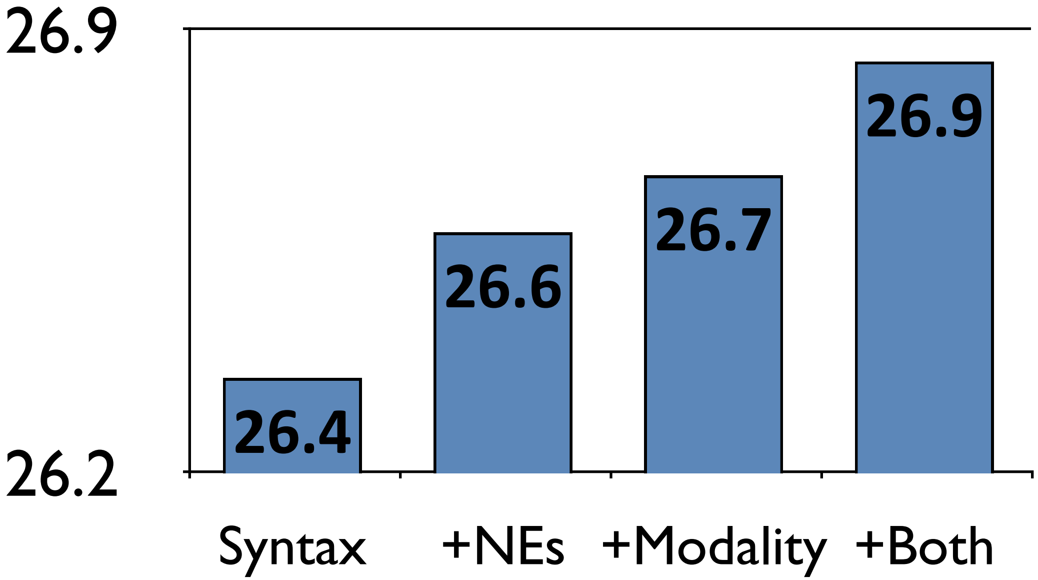


Simple Urdu modality tagger



HIVE improvements

- HIVEs improve quality over using syntax alone by 0.5 Blue points
- NEs better capture distributional properties of people, places, organizations (over generic "NP"). Modality better structures VPs.



Conclusions

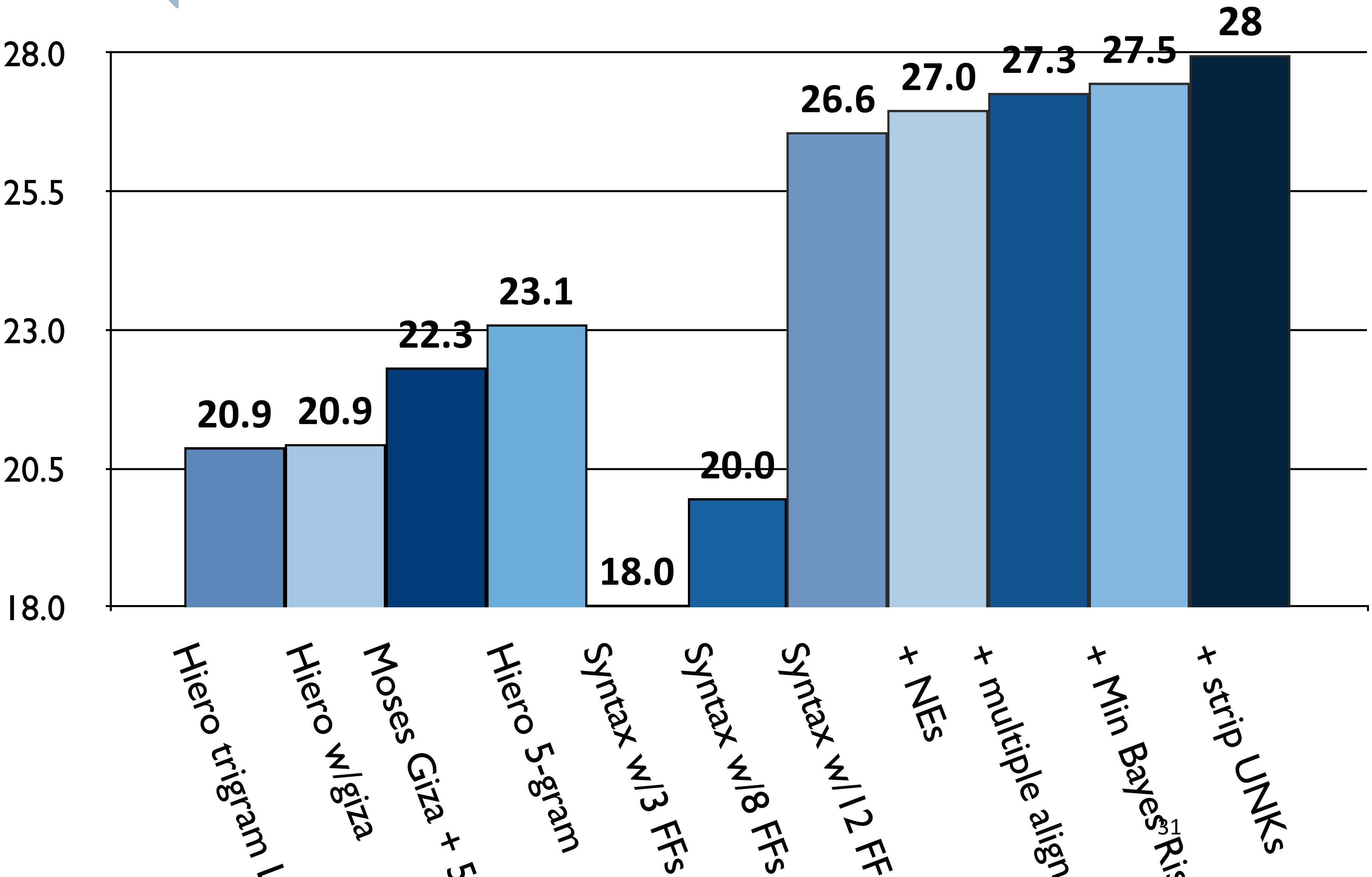
- Using syntax-based translation models resulted in huge improvements in quality
- Urdu is an ideal language to show off the advantages of syntax
  - Very small amount of training data
  - Very different word order than English
- No need for any specialized Urdu tools
  - Therefore applicable to other languages
  - Likely to find similar gains for low-resource languages that have different word order

Free software!

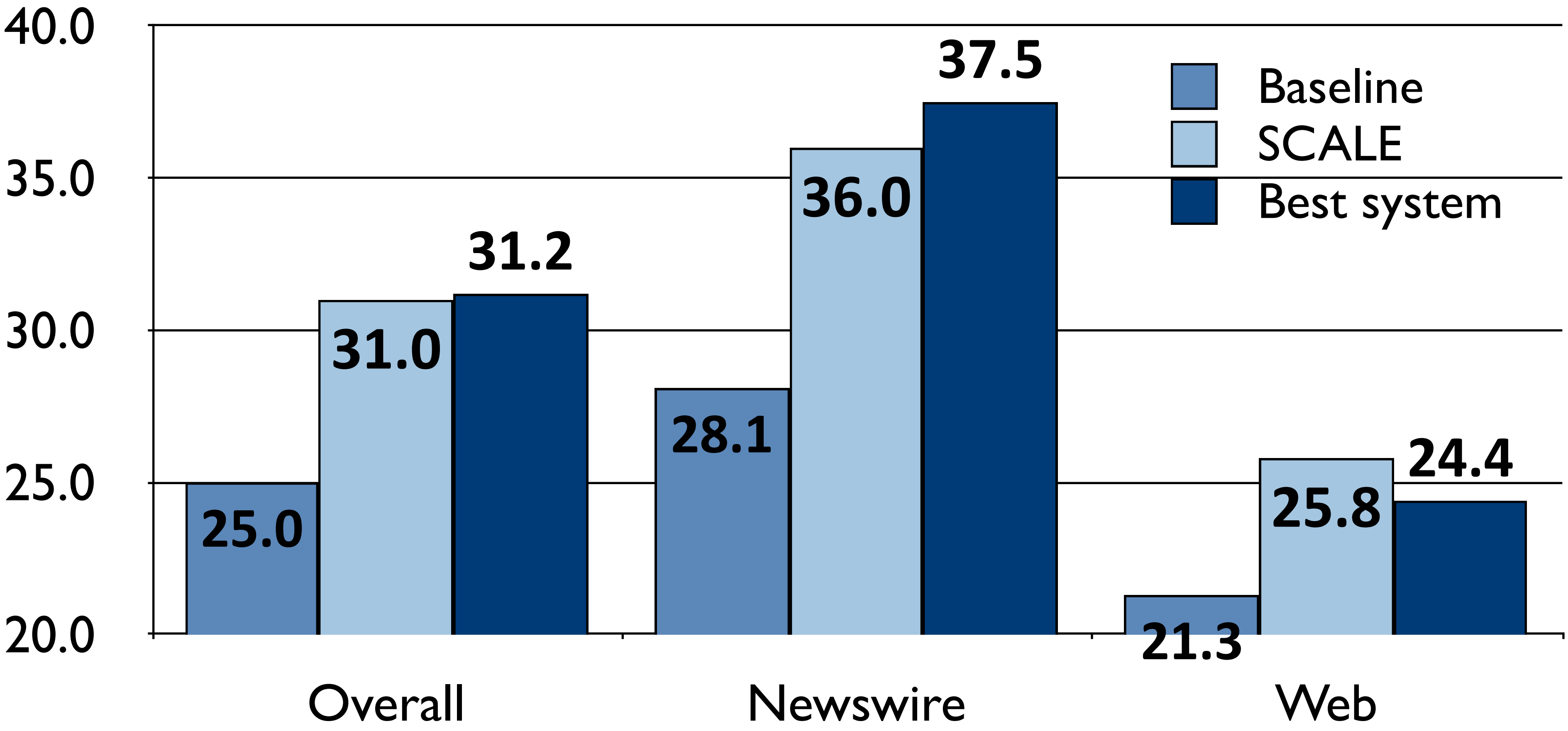
- The Joshua Decoder is open source -- try it out today! http://cs.jhu.edu/~ccb/joshua/
- Zhifei Li, one of our lead developers, is graduating -- hire him!

Thanks!

Breakdown of improvements



NIST Eval Results



Extracting Hiero rules

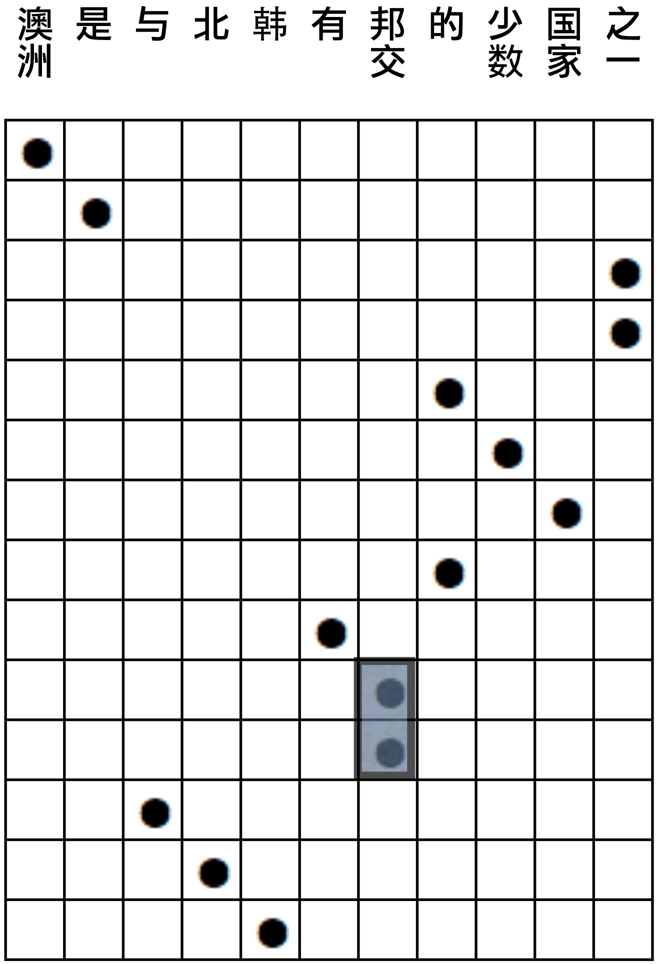
	澳洲	是	与	北	韩	有	邦交	的	少数	国家	之一
Australia	●										
is		●									
one											●
of											●
the								●			
few									●		
countries										●	
that								●			
have						●					
diplomatic							●				
relations							●				
with			●								
North				●							
Korea					●						

Extracting Hiero rules

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

X → 与 北 韩 有 邦交, have diplomati
with North Ko

Extracting Hiero rules

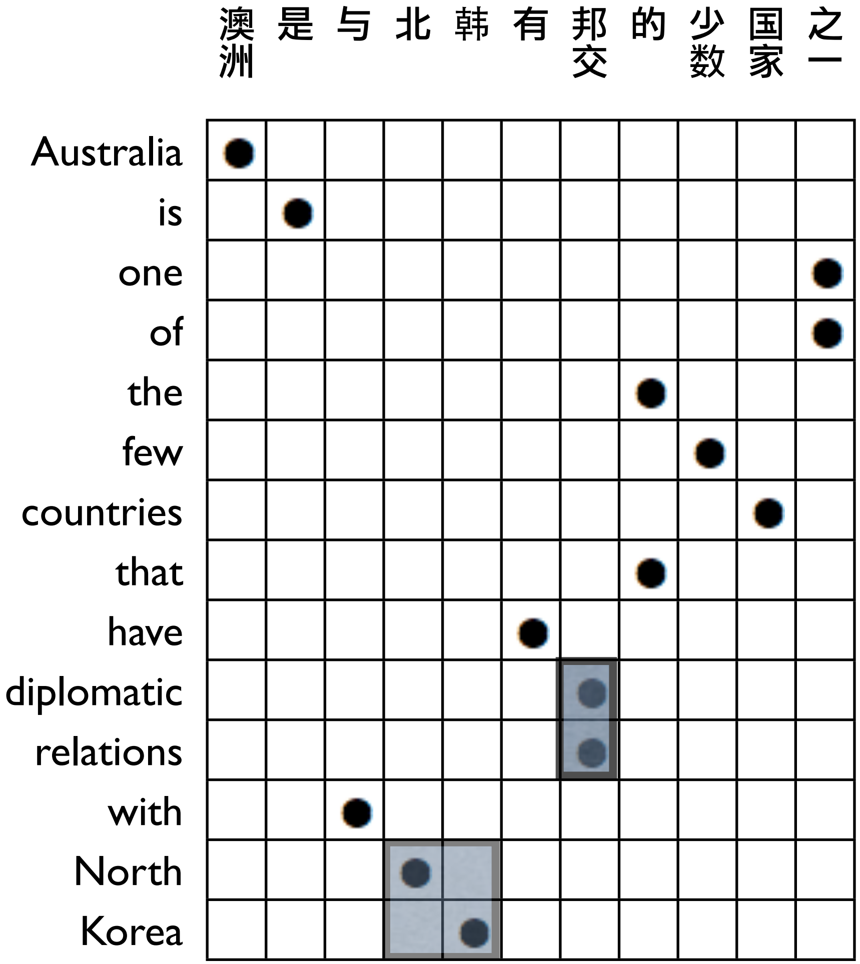


	澳洲	是	与	北	韩	有	邦交	的	少数	国家	之一
Australia	●										
is		●									
one											●
of											●
the								●			
few									●		
countries										●	
that								●			
have						●					
diplomatic							●				
relations							●				
with			●								
North				●							
Korea					●						

X → 与 北 韩 有 邦交, have diplomati
with North Ko

X → 邦交, diplomatic relations

Extracting Hiero rules



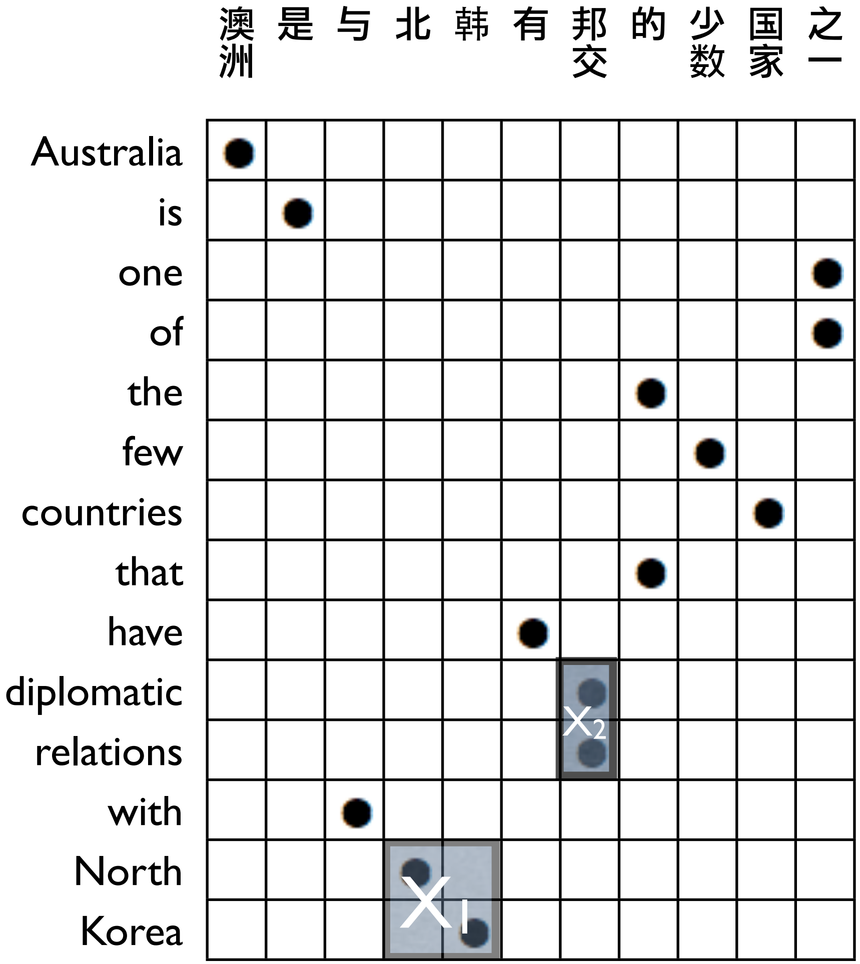
	澳洲	是	与	北	韩	有	邦交	的	少数	国家	之一
Australia	●										
is		●									
one											●
of											●
the								●			
few									●		
countries										●	
that								●			
have						●					
diplomatic							●				
relations							●				
with			●								
North				●							
Korea					●						

X → 与 北 韩 有 邦交, have diplomati
with North Ko

X → 邦交, diplomatic relations

X → 北 韩, North Korea

Extracting Hiero rules



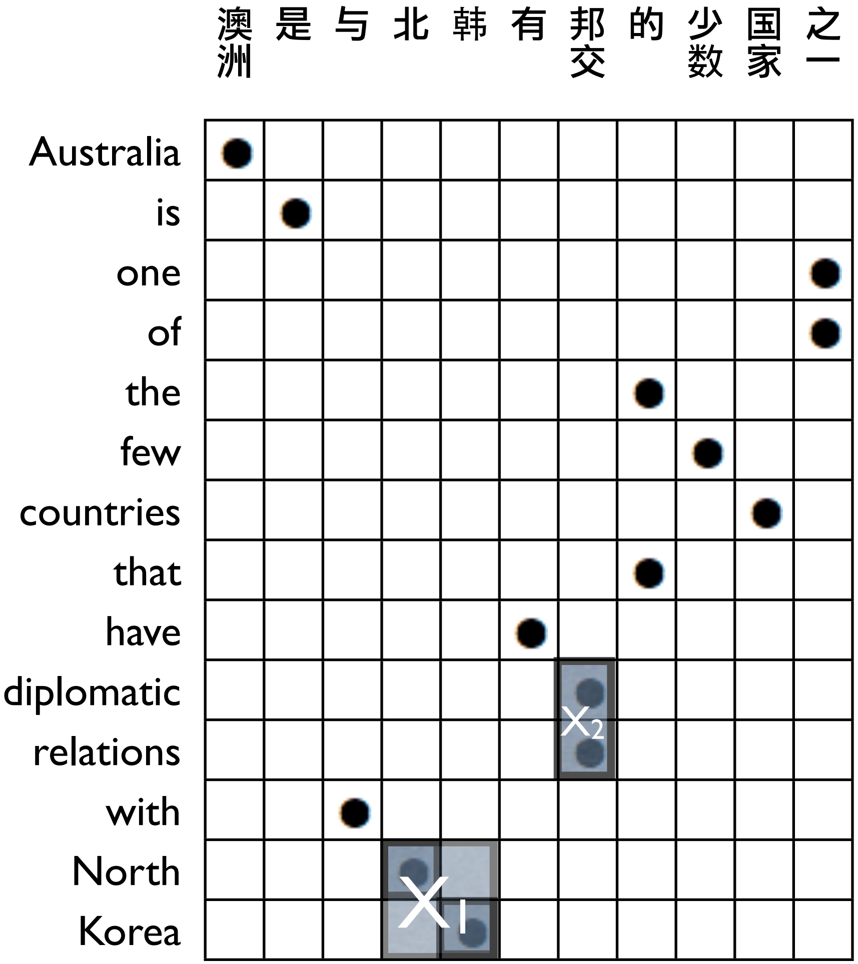
	澳洲	是	与	北	韩	有	邦交	的	少数	国家	之一
Australia	●										
is		●									
one											●
of											●
the								●			
few									●		
countries										●	
that								●			
have						●					
diplomatic							●				
relations							●				
with			●								
North				●							
Korea					●						

X → 与 北 韩 有 邦交, have diplomati
with North Ko

X → 邦交, diplomatic relations

X → 北 韩, North Korea

Extracting Hiero rules



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

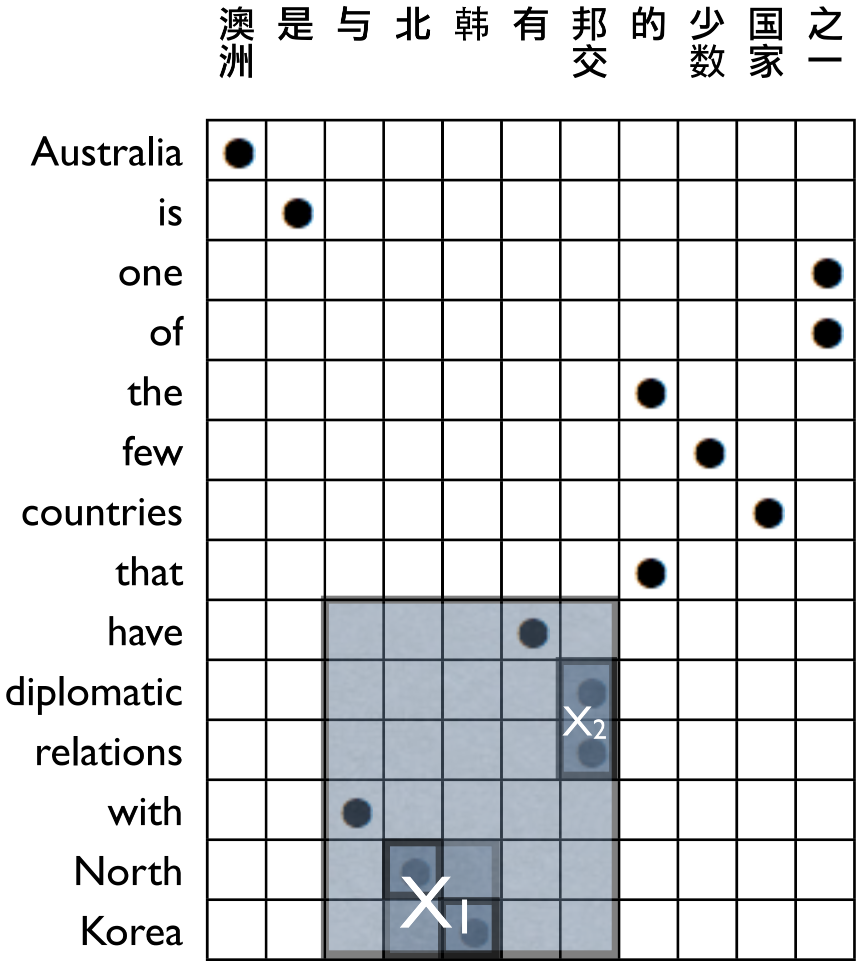
X → 与 北 韩 有 邦交, have diplomati
with North Ko

X → 邦交, diplomatic relations

X → 北 韩, North Korea

X → 与 X_1 有 X_2, have X_2 with X_1

Extracting Hiero rules



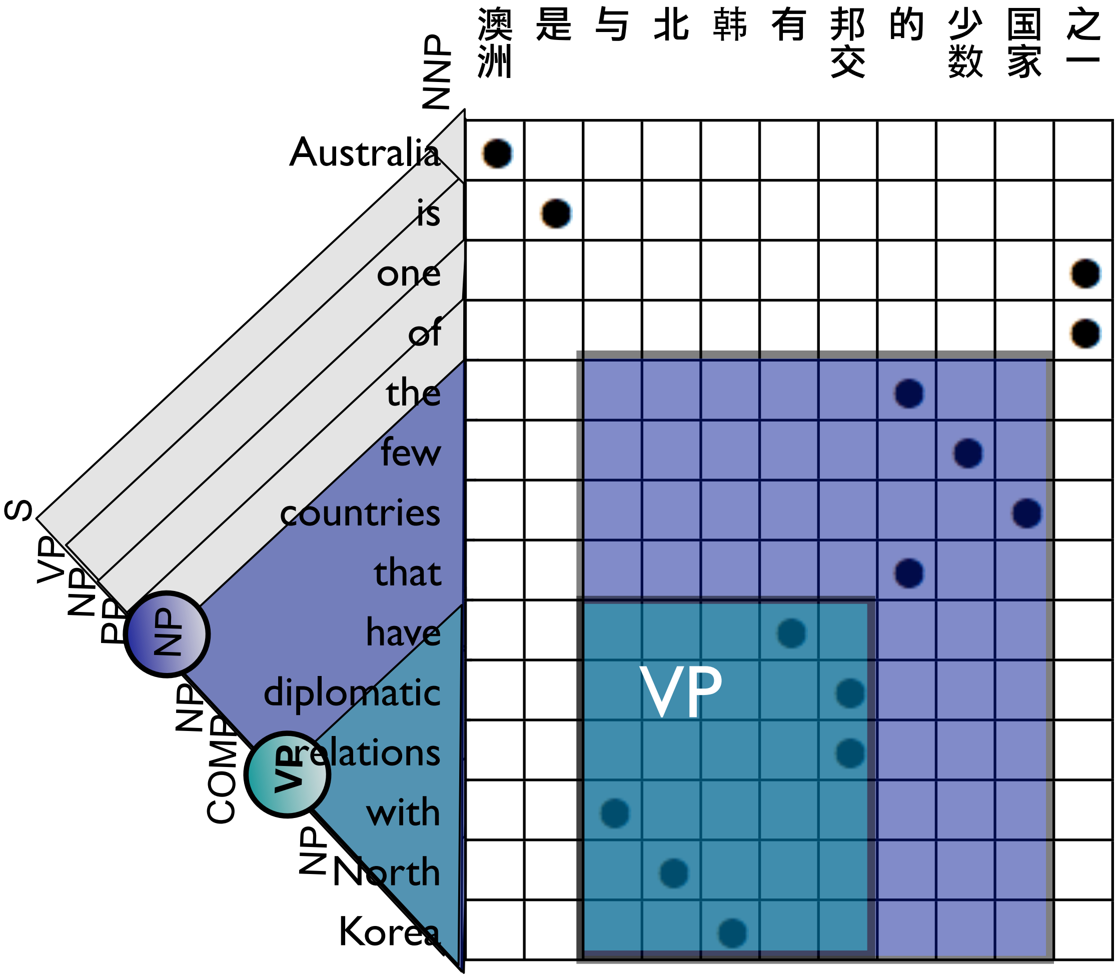
	澳洲	是	与	北	韩	有	邦交	的	少数	国家	之一
Australia	●										
is		●									
one											●
of											●
the								●			
few									●		
countries										●	
that								●			
have						●					
diplomatic							●				
relations							●				
with			●								
North				●							
Korea					●						

X → 与 北 韩 有 邦交, have diplomati
with North Ko

X → 邦交, diplomatic relations

X → 北 韩, North Korea

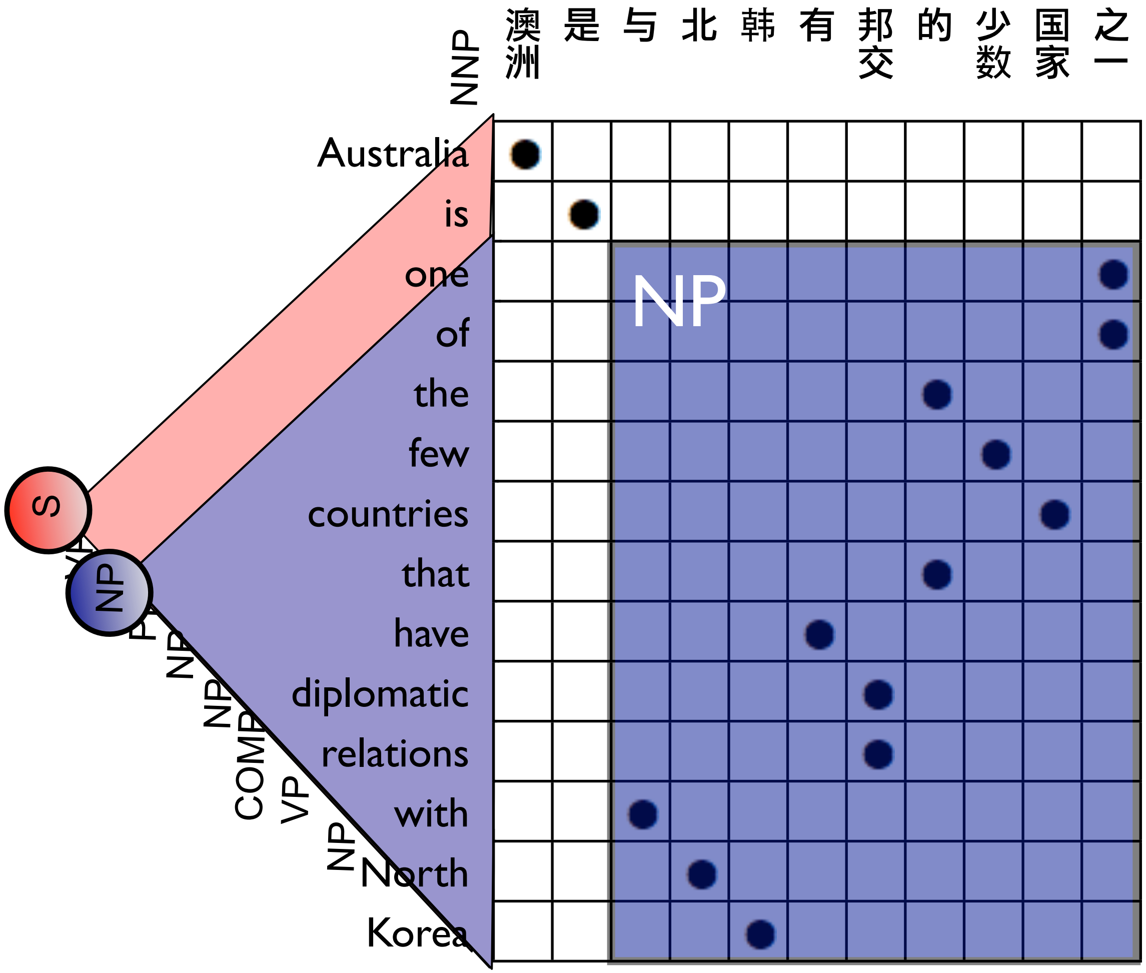
X → 与 X_1 有 X_2, have X_2 with X_1



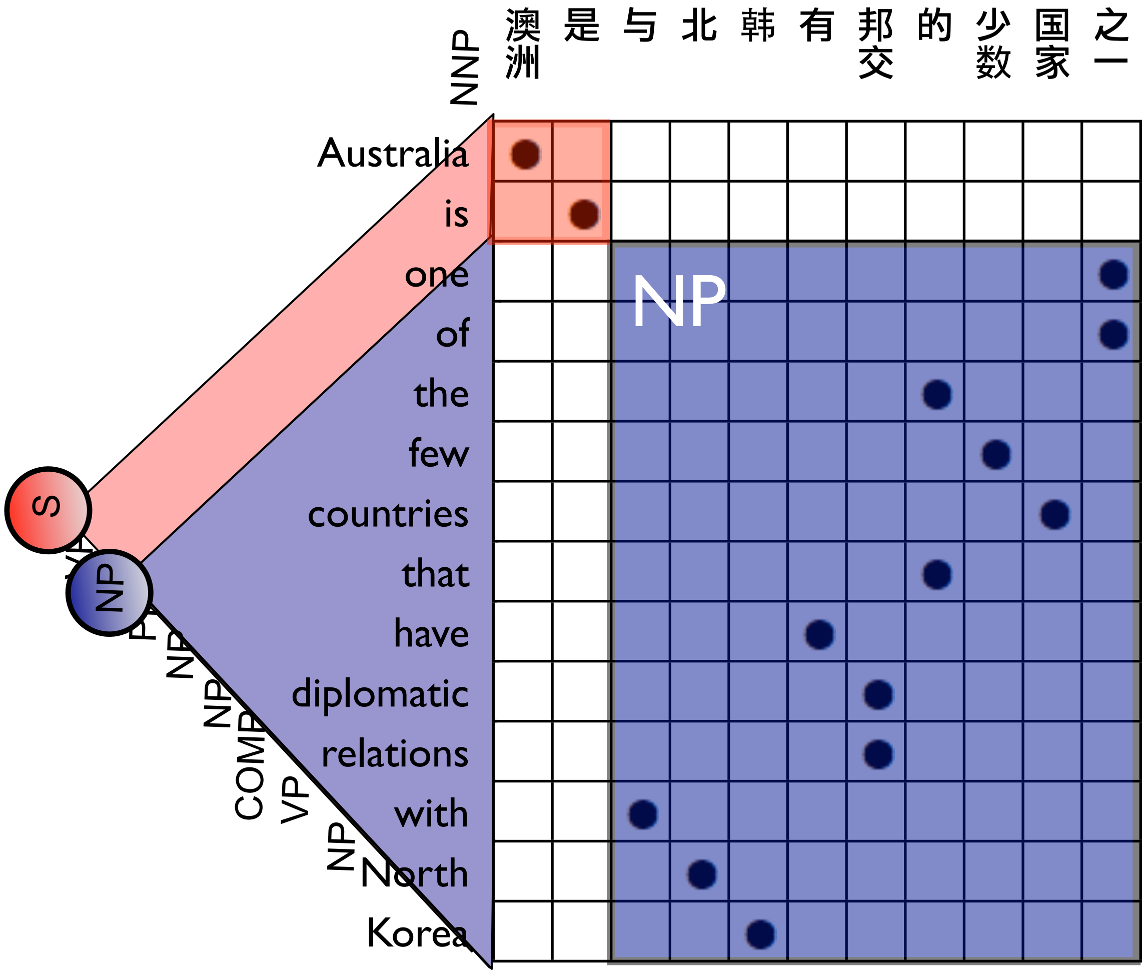
VP → 与 北 韩 有 邦交,
have diplomatic relation

NP → 与 北 韩 有 邦交 的
the few countries that
relations with North K

NP → VP 的 少数 国家,
the few countries that

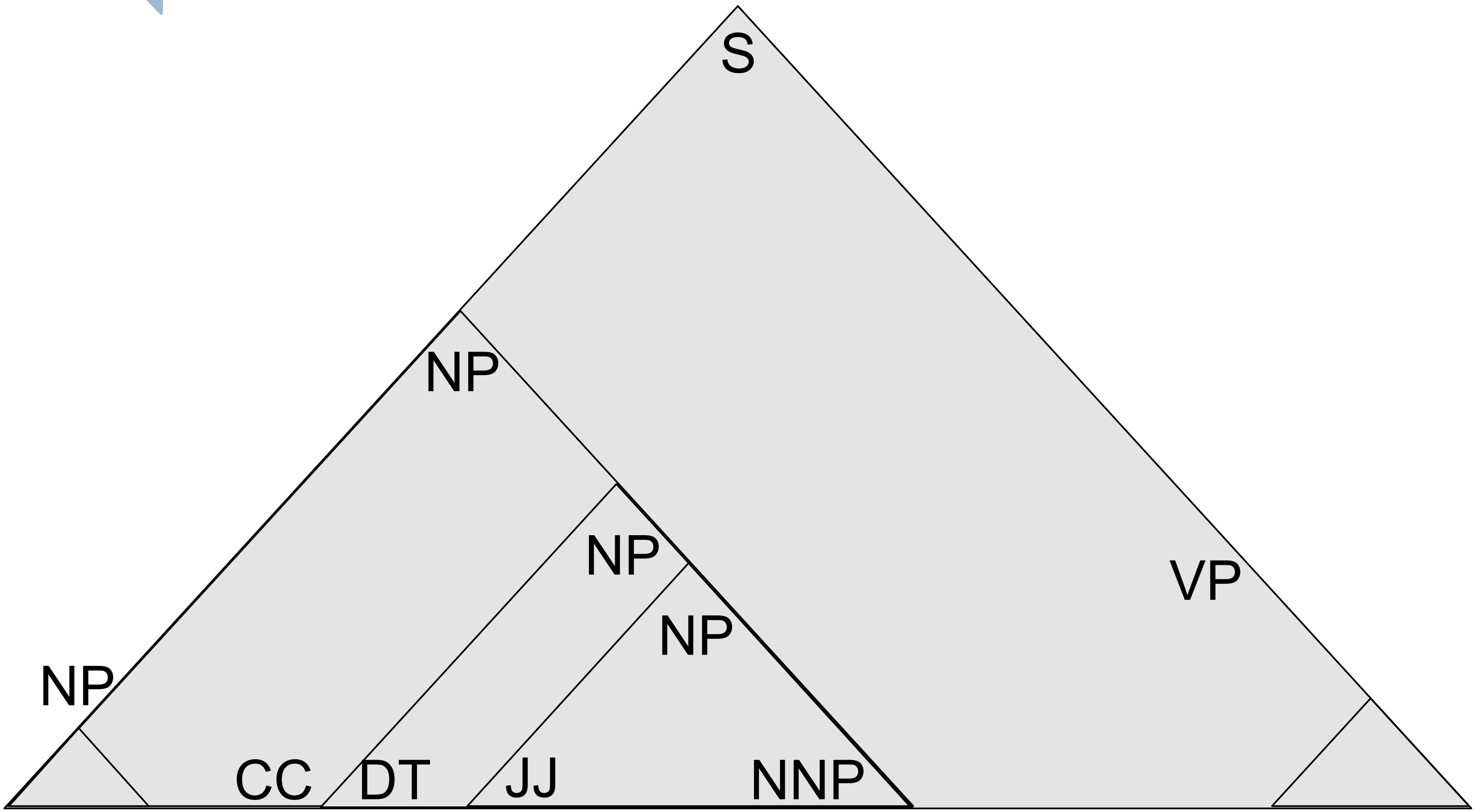


S/NP → 澳洲, Australia is



S/NP → 澳洲, Australia is

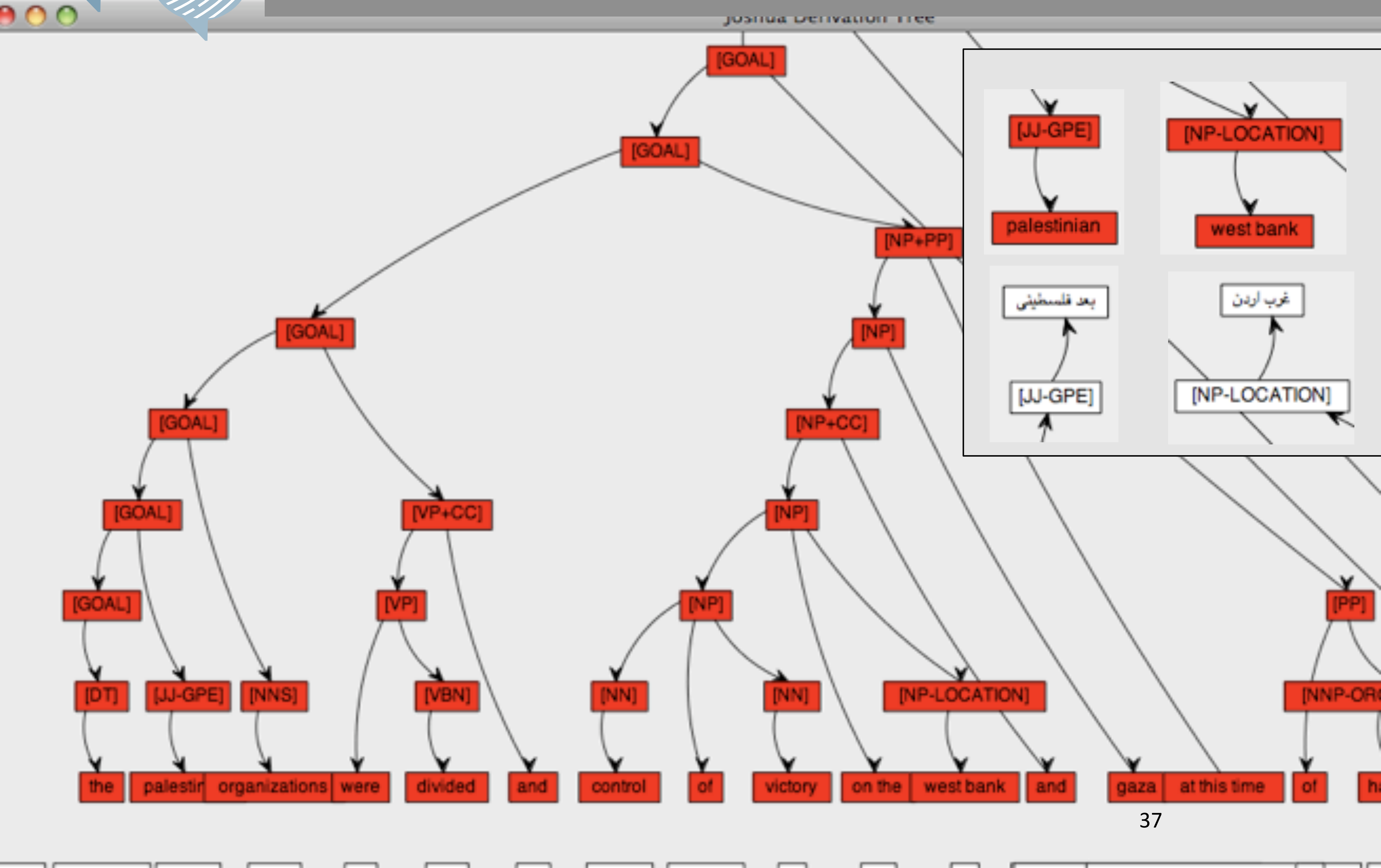
Example tree graft

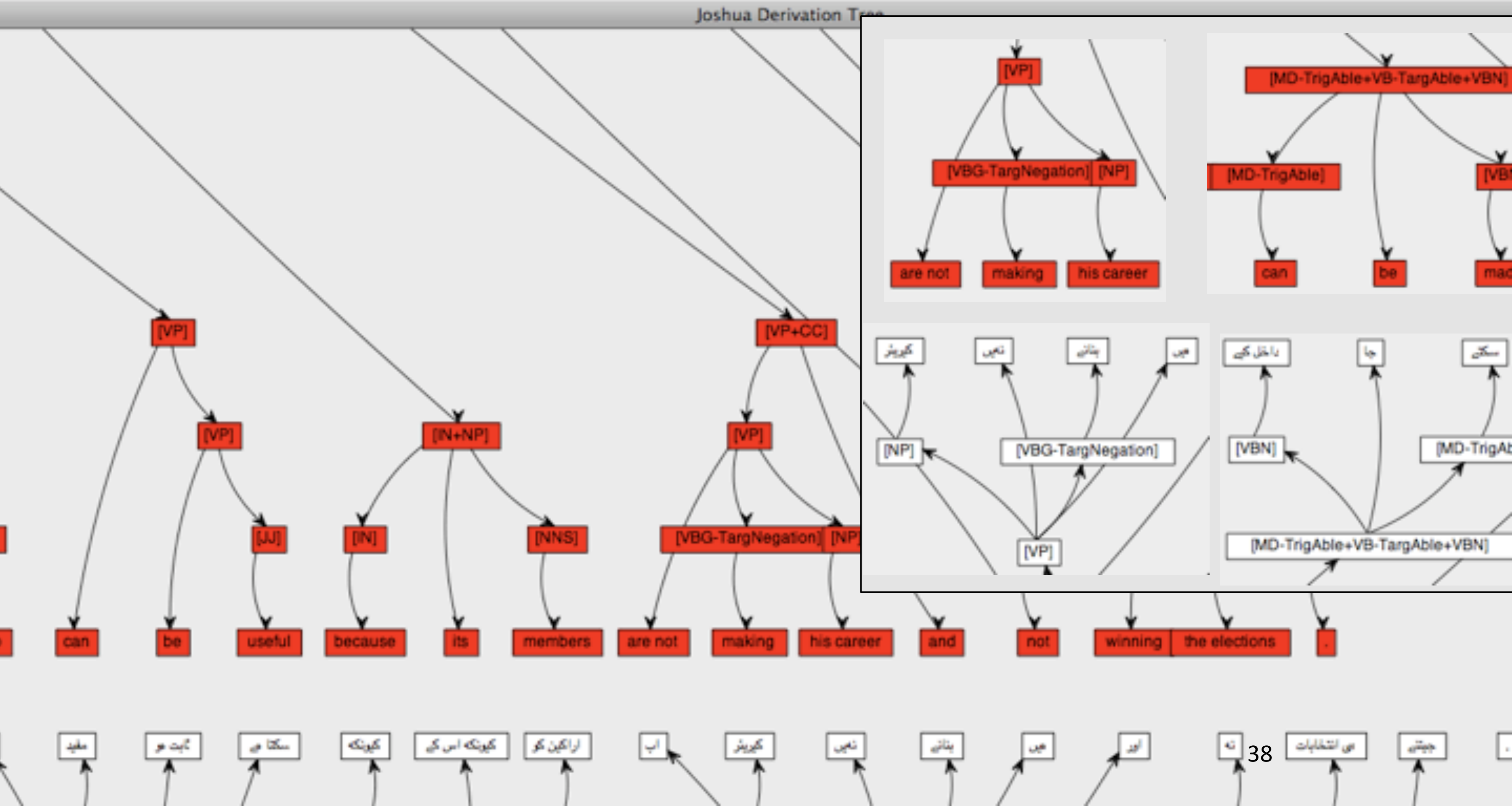


NNP , NP VBZ RB DT VBN NN TO NNP NNP , VBD

Nance , who is also a paid consultant to ABC News , said ....

Simple Urdu NE tagger




Joshua Derivation Tree
[VP]
[VP+CC]
[VP]
[IN+NP]
[VP]
[JJ]
[IN]
[NNS]
[VBG-TargNegation]
[NP]
can
be
useful
because
its
members
are not
making
his career
and
not
winning
the elections
.
مفید
ثابت ہو
سکتا ہے
کیونکہ
کیونکہ اس کی
اراکین کو
اب
کیریئر
نہیں
بنانے
ہیں
اور
نہ
ی انتخابات
جیتنے
[VP]
[VBG-TargNegation]
[NP]
are not
making
his career
[MD-TrigAble+VB-TargAble+VBN]
[MD-TrigAble]
[VB
can
be
mac
کیریئر
نہیں
بنانے
ہیں
داخل کیے
جا
سکتی
[NP]
[VBG-TargNegation]
[VP]
[VBN]
[MD-TrigAb
[MD-TrigAble+VB-TargAble+VBN]

Translation improvements

Baseline

'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.

Thomas Reid said in a news یوایس interview that in 1990 in the era of Benazir Bhutto China had the experience of Pakistan's first nuclear bomb.

The First Nuclear T

Thomas red of the U
Laboratory in design
the book of Los Alan
former director of the
the cooperation of D

This book under the
expressway: the poli
this has been writter
claimed that the cou
bomb or any other c
secrets, or any of the
cooperation.

Thomas Reid said in
that Benazir Bhutto i

Source

ناگاؤں نے آسام میں آگ لگا دی

بدھ کے روز مشتعل ناگا قبائلیوں نے منی پور کے دس سکولوں
کو بھی نذر آتش کردیا تھا۔

پولیس کے مطابق سینکڑوں کی تعداد میں ناگالینڈ کے مسلح
قبائلیوں نے آسام کے گلیکی اور سیبسا گر کے تین گاؤں میں آگ
لگا دی۔

اس حملہ کے بعد بڑی تعداد میں مقامی باشندوں نے علاقوں کو
خالی کردیا ہے۔

ناگالینڈ دعویٰ کرتا ہے کہ ریاست آسام اس کے بعض خطوں پر
قابض ہے۔

Nagas Set Fire i

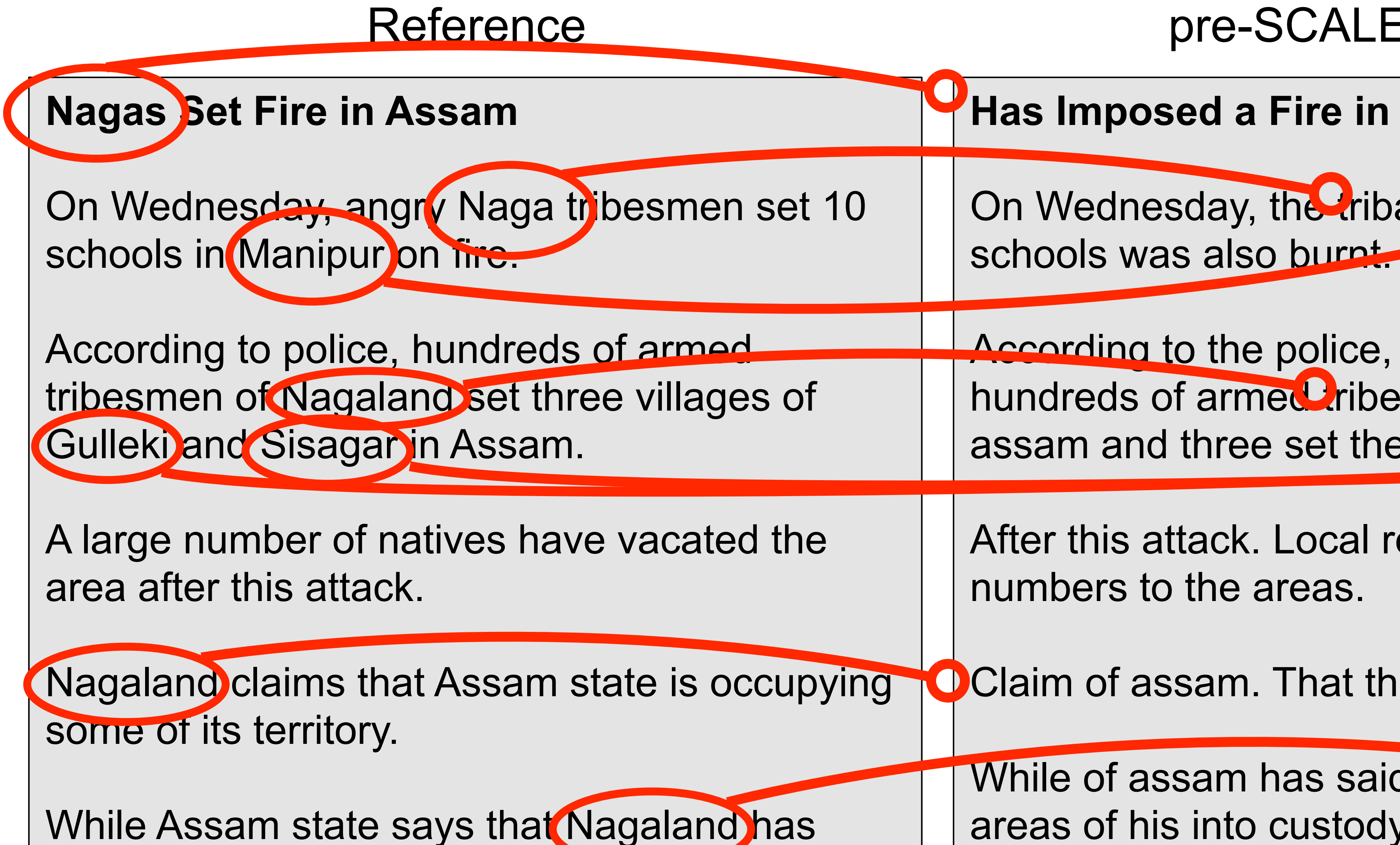
On Wednesday,
schools in Manip

According to poli
tribesmen of Nag
Gulleki and Sisag

A large number c
area after this att

Nagaland claims
some of its territo

While Assam sta



Reference

Nagas Set Fire in Assam

On Wednesday, angry Naga tribesmen set 10 schools in Manipur on fire.

According to police, hundreds of armed tribesmen of Nagaland set three villages of Gulleki and Sisagar in Assam.

A large number of natives have vacated the area after this attack.

Nagaland claims that Assam state is occupying some of its territory.

While Assam state says that Nagaland has

pre-SCALE

Has Imposed a Fire in

On Wednesday, the triba
schools was also burnt.

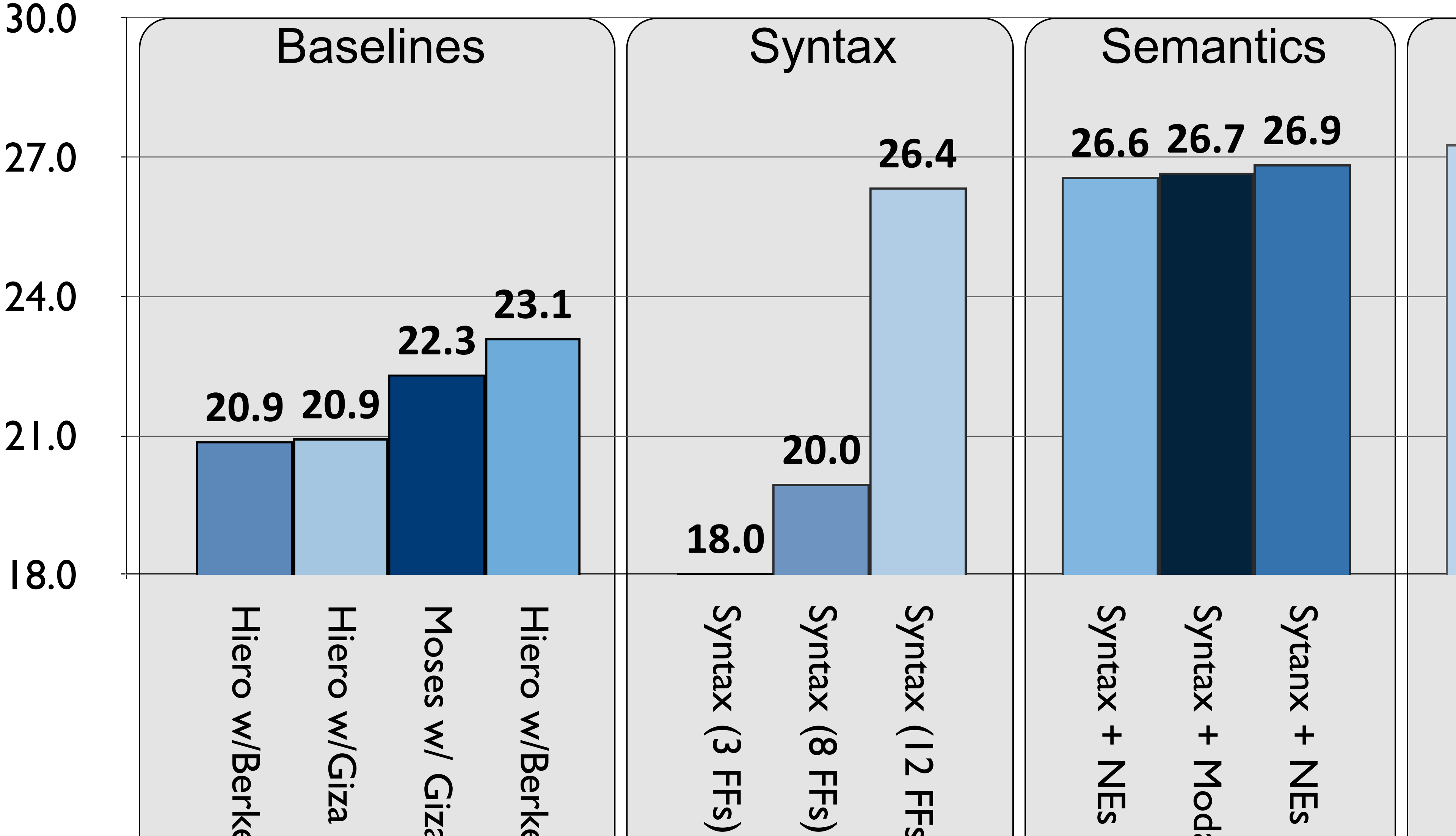
According to the police,
hundreds of armed tribe
assam and three set the

After this attack. Local r
numbers to the areas.

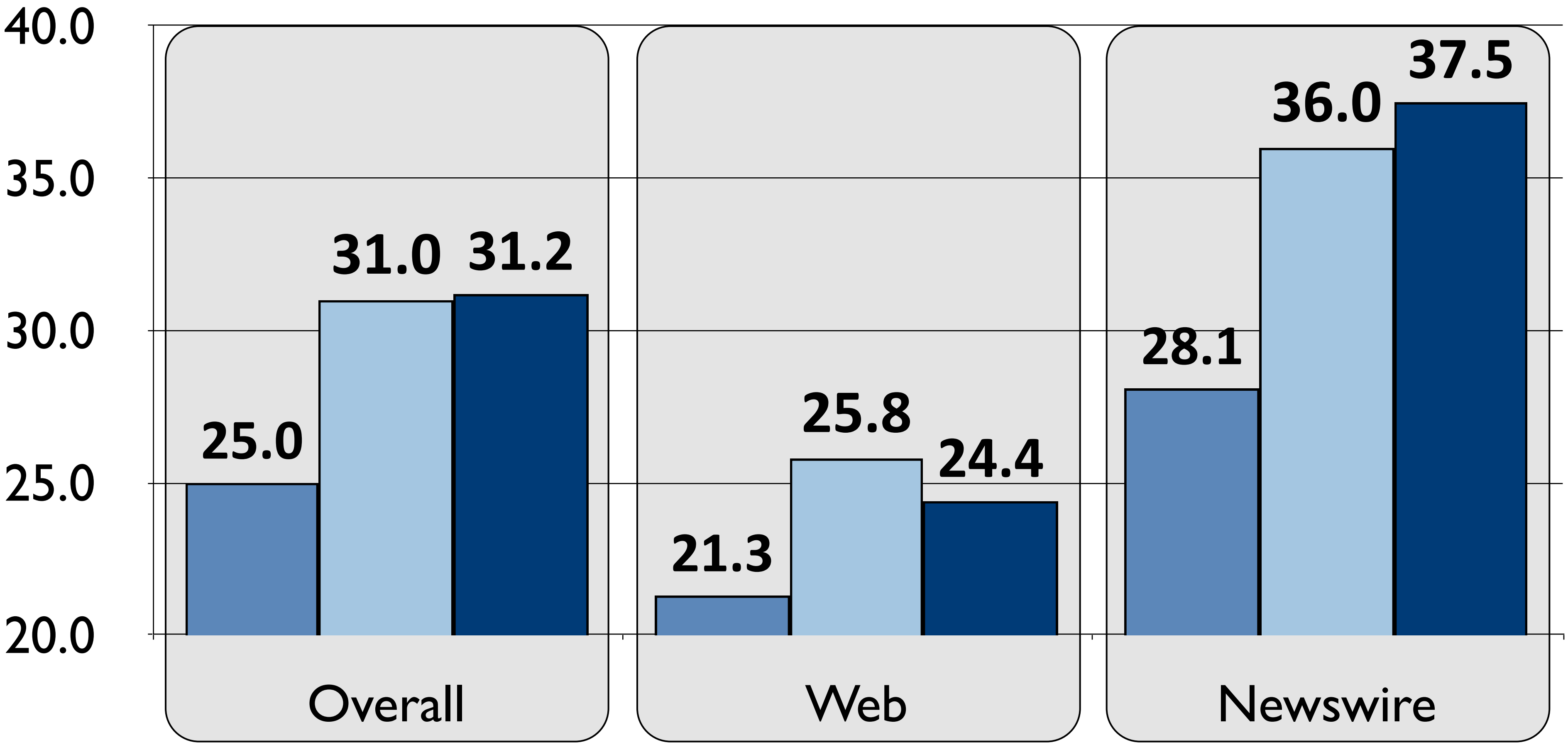
Claim of assam. That th

While of assam has saic
areas of his into custody

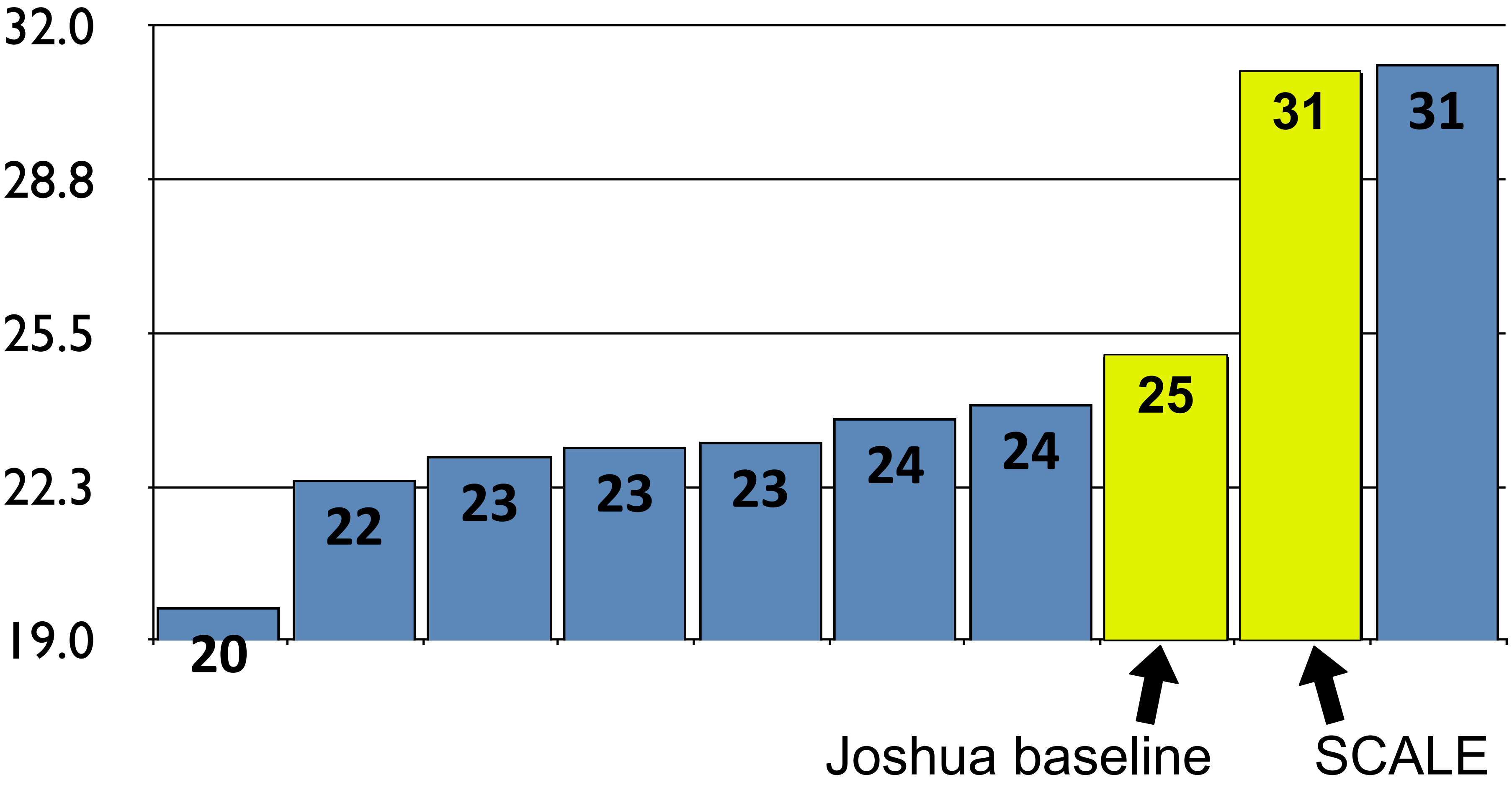
Breakdown of improvements

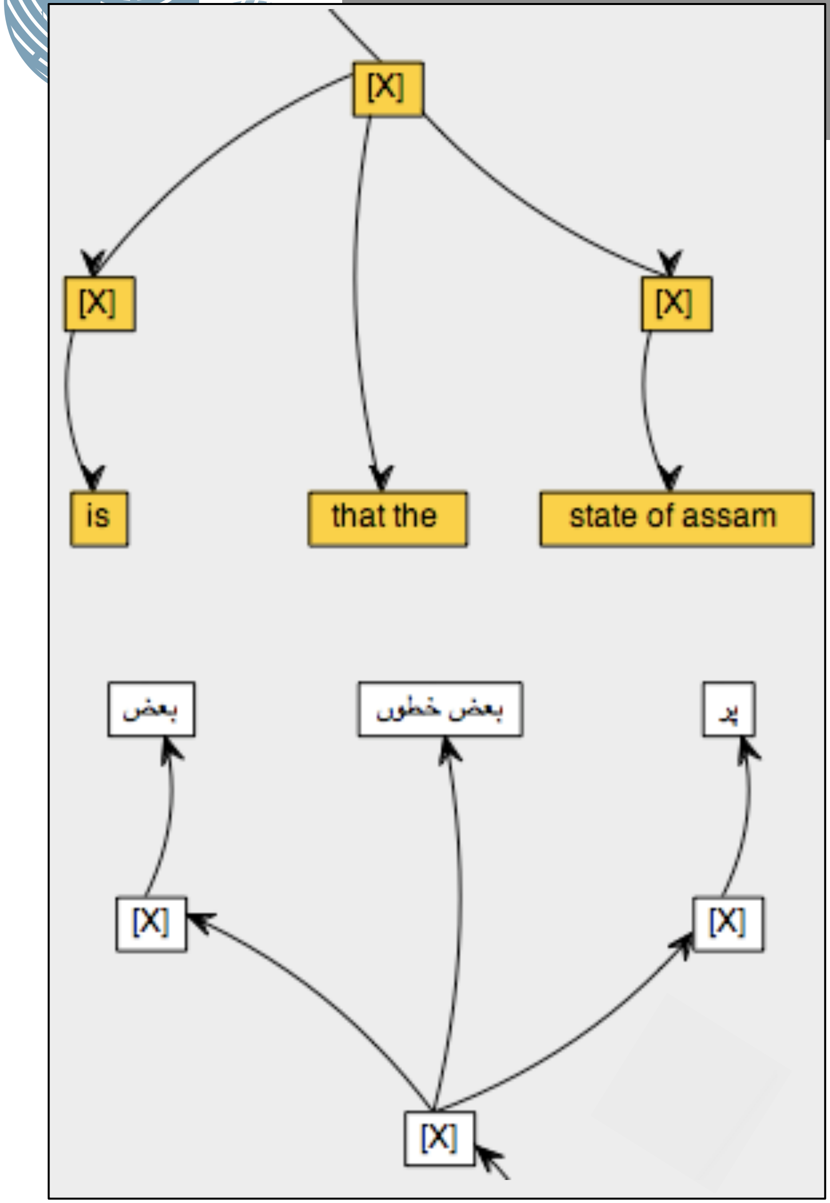


NIST Eval Results

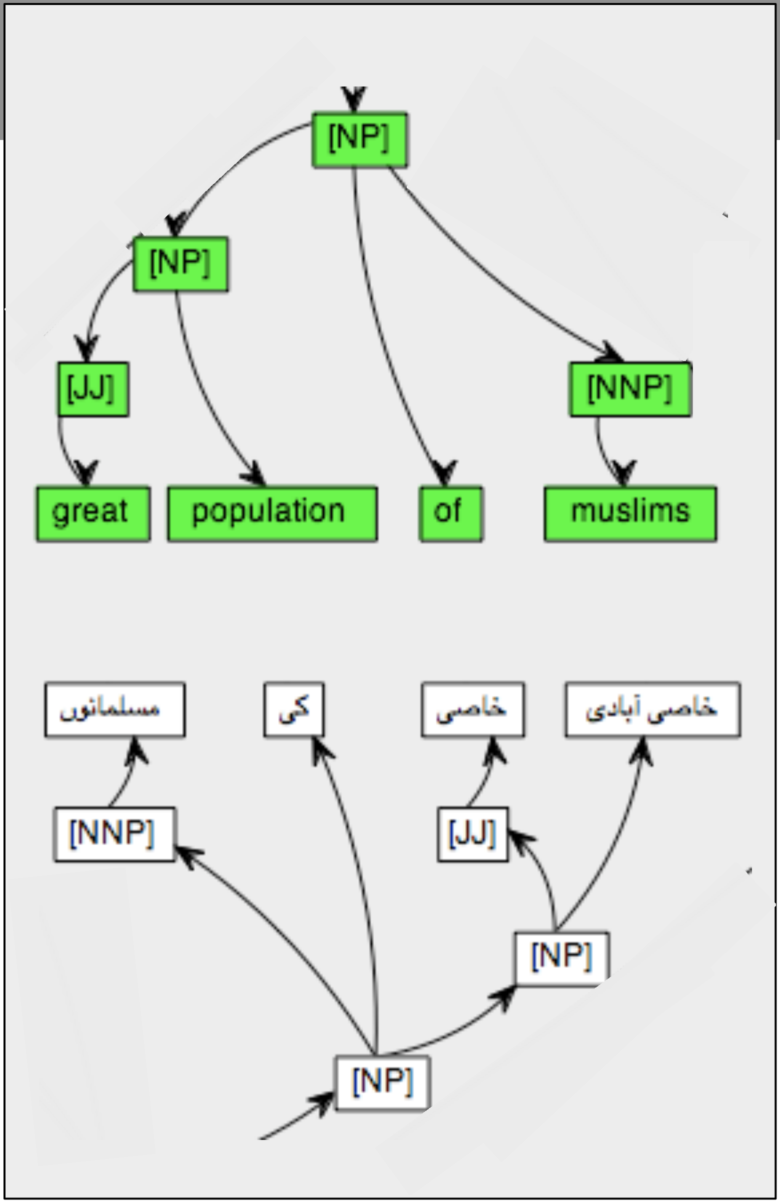


The field for the Urdu task

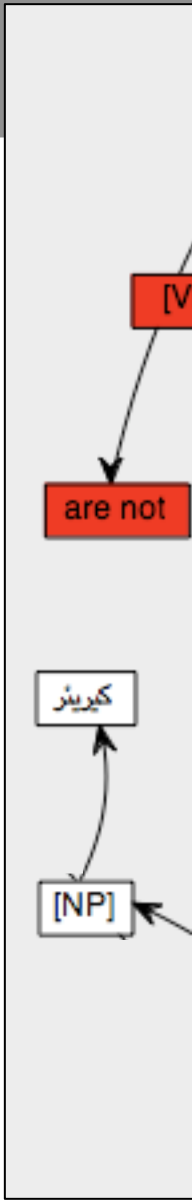




Standard Hierarchical Rules



Syntactic Enhancements



Sema