Learning to translate with source and target syntax

David Chiang, USC Information Sciences Institute

14 July 2010

Overview

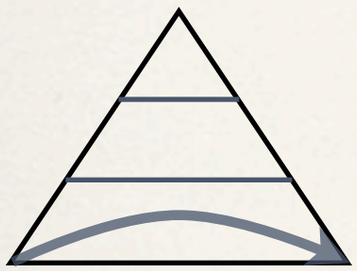
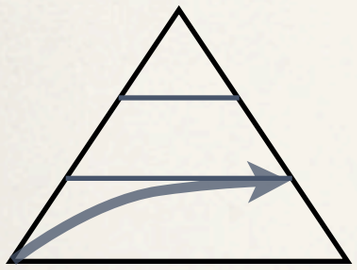
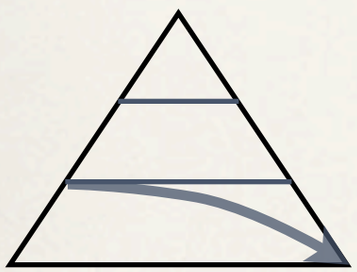
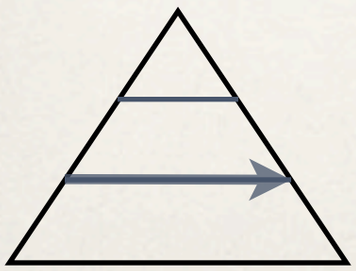
- Using source and target syntax
- Why is it hard?
- How can we make it better?
  - Let the model learn how much syntax to use
  - The model does choose syntax, for improvements of +0.6–0.8 BLEU

reference: An official from Japan ’s science and technology ministry said , " We are highly encouraged by Abraham ’s comment .

Hiero: Officials of the Japanese ministry of education and science , " said Abraham speeches , we are deeply encouraged by .

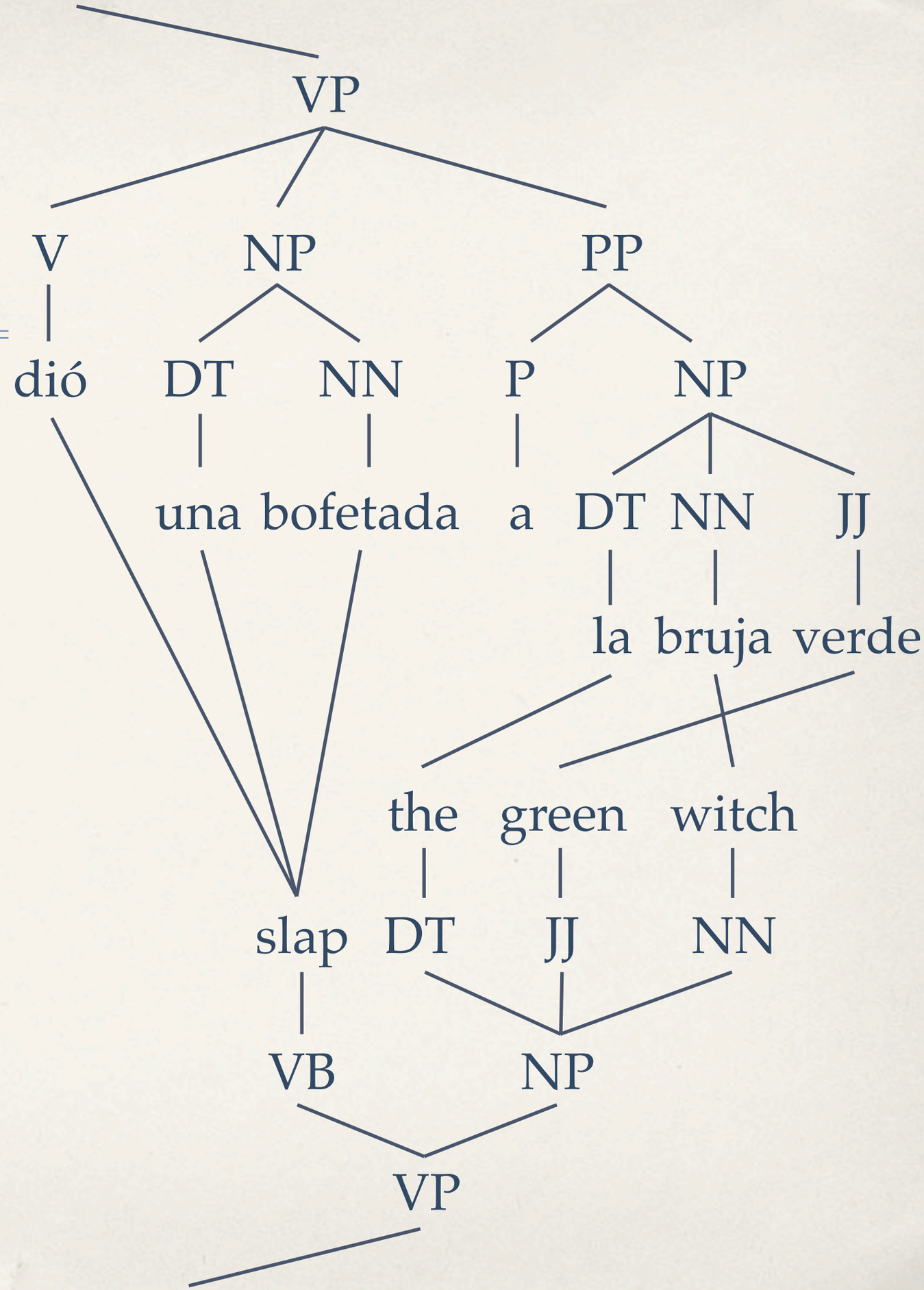
string-to-tree: Japan ’s ministry of education , culture , sports , science and technology , " Abraham ’s statement , which is most encouraging , " the official said .

Previous work

	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 Y Liu et al 2006
	tree-to-tree	DOT (Poutsma 2000) Eisner 2003	Stat-XFER (Lavie et al 2008) M Zhang et al. 2008 Y Liu et al., 2009

STSG extraction

1. Phrases
   - respect word alignments
   - are syntactic constituents on *both* sides
2. Phrase pairs form rules
3. Subtract phrases to form rules



STSG extraction

1. Phrases

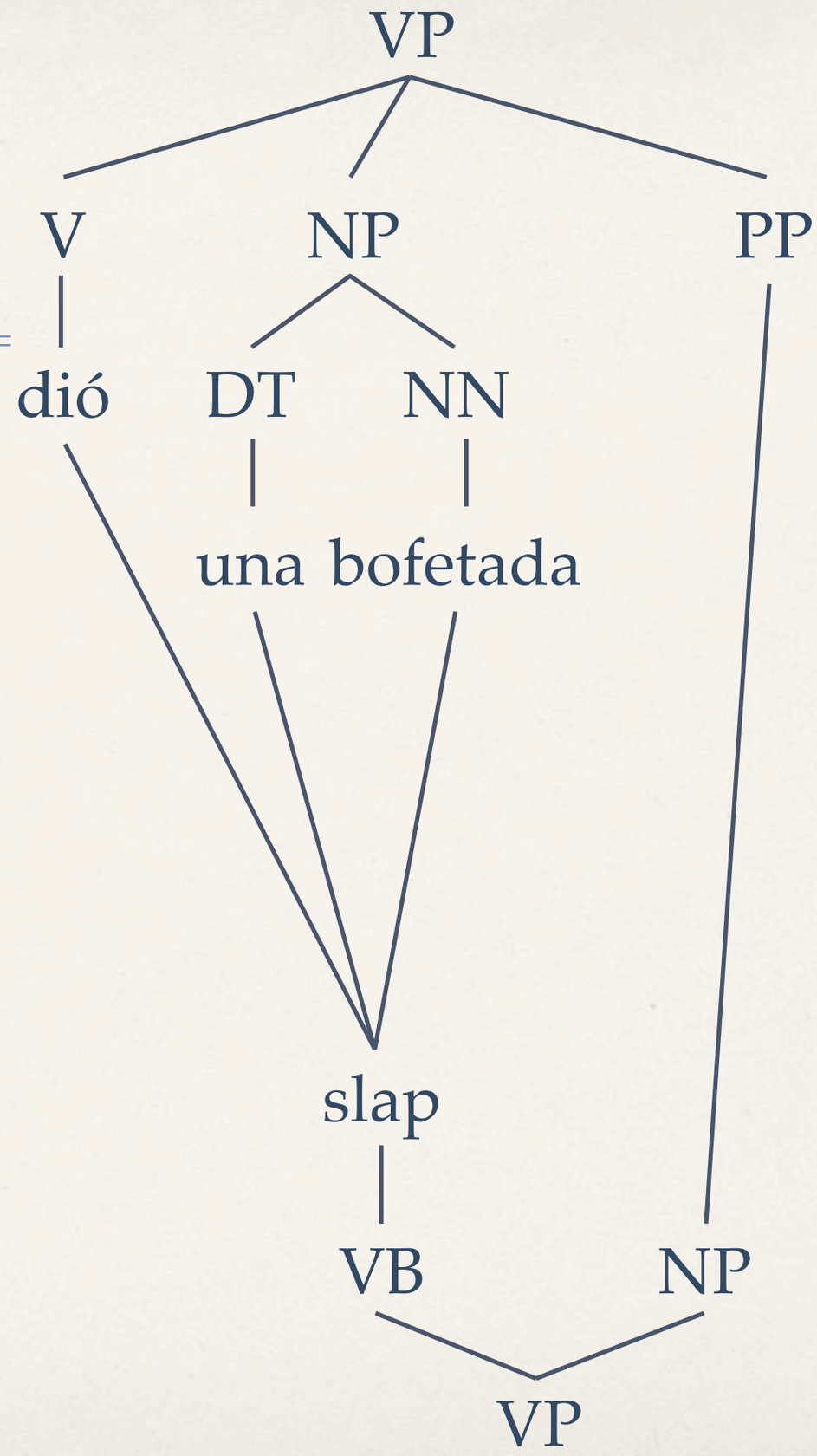
   - respect word alignments
   - are syntactic constituents on *both* sides

2. Phrase pairs form rules

3. Subtract phrases to form rules

STSG extraction

1. Phrases
   - respect word alignments
   - are syntactic constituents on *both* sides
2. Phrase pairs form rules
3. Subtract phrases to form rules



STSG translation

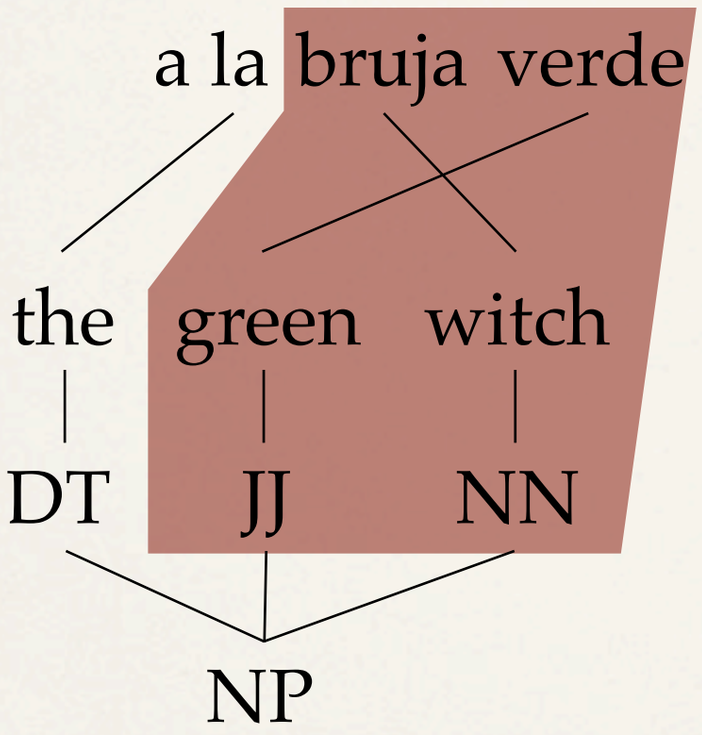
Liu et al, 2009	
phrase	23.66
STSG	20.21

Zhang et al, 2008	
phrase	23.86
STSG	24.71

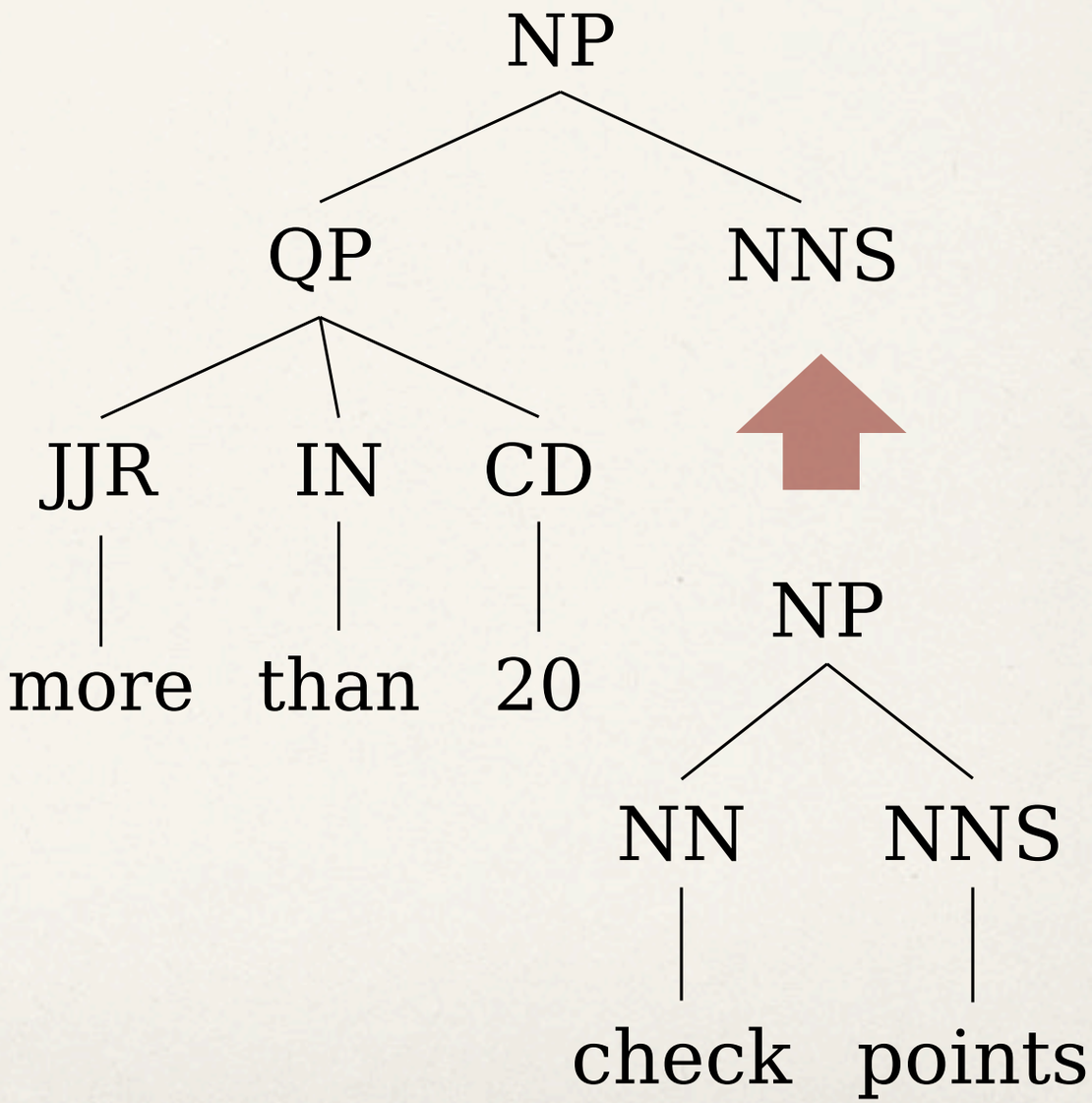
Ambati and Lavie, 2008	
phrase	30.18
STSG	22.23

Why is tree-to-tree hard?

too few rules

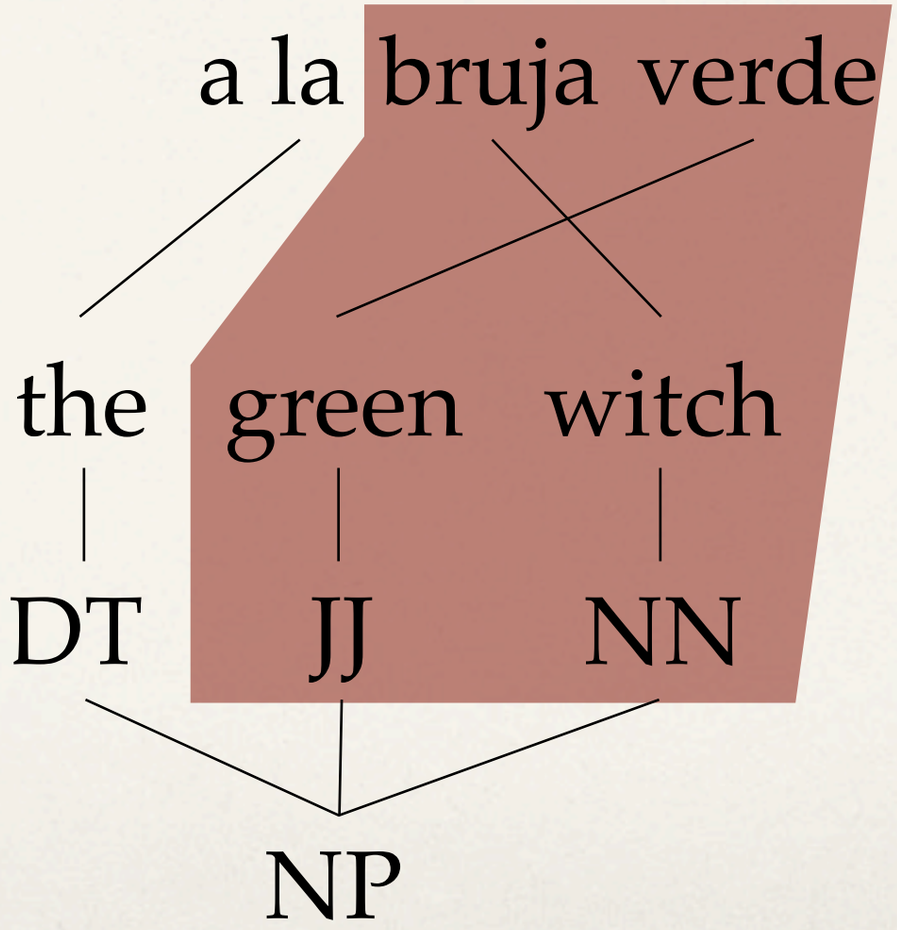


too few derivations

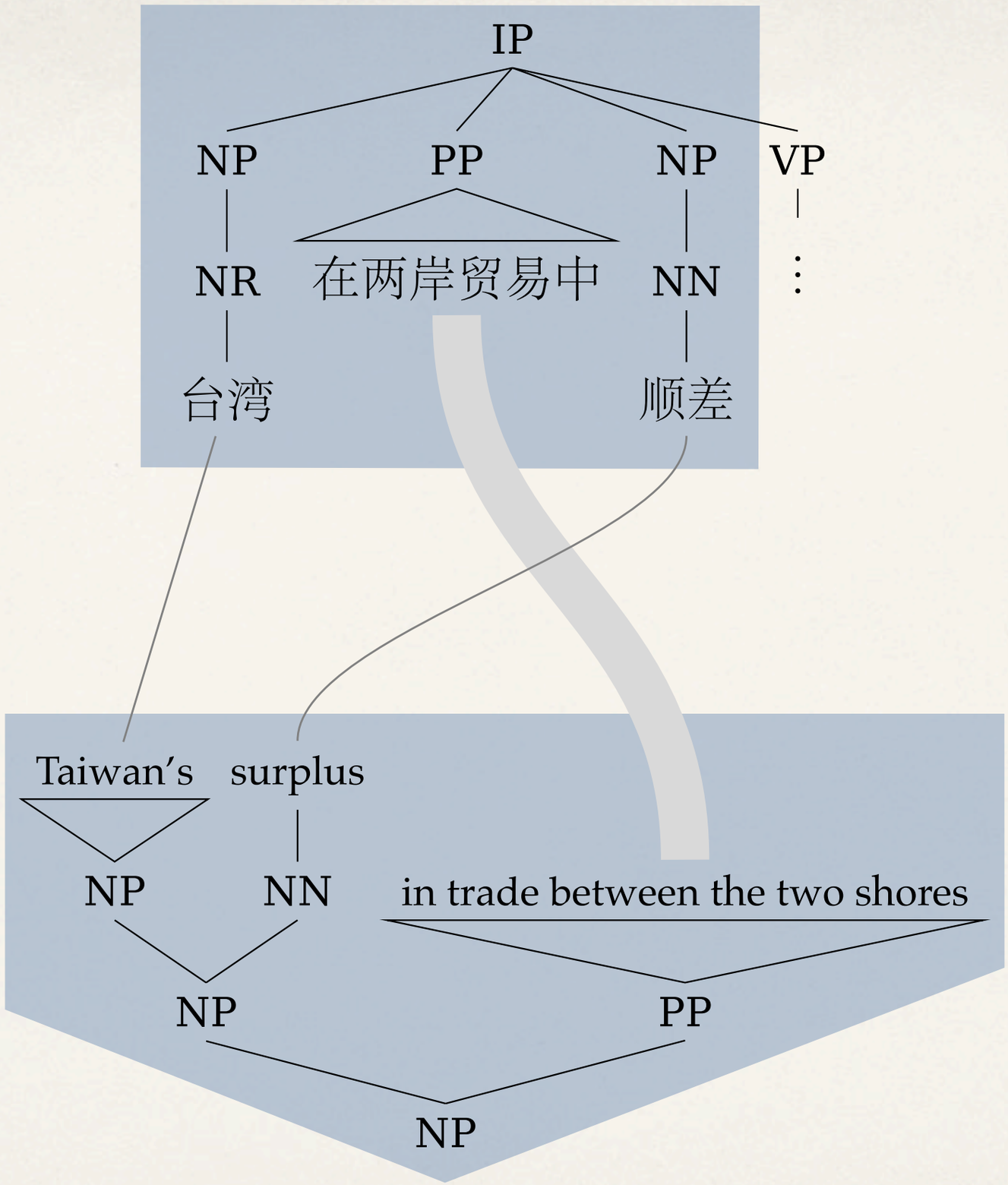


Why is tree-to-tree hard?

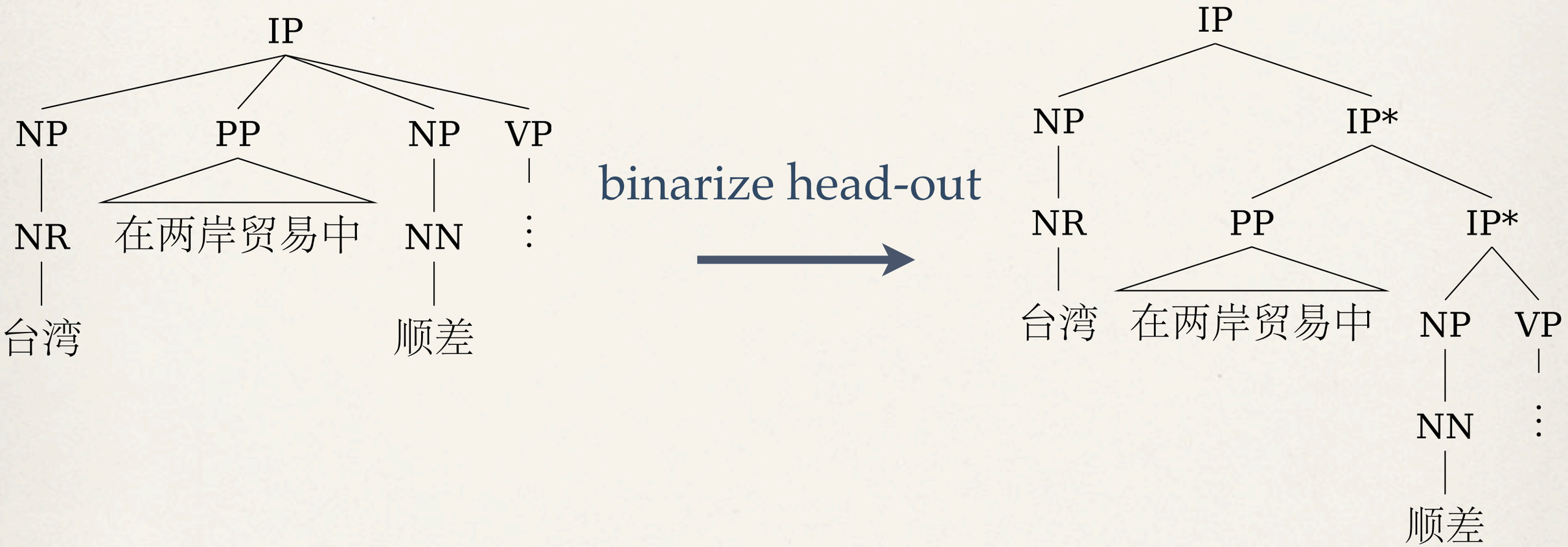
too few rules



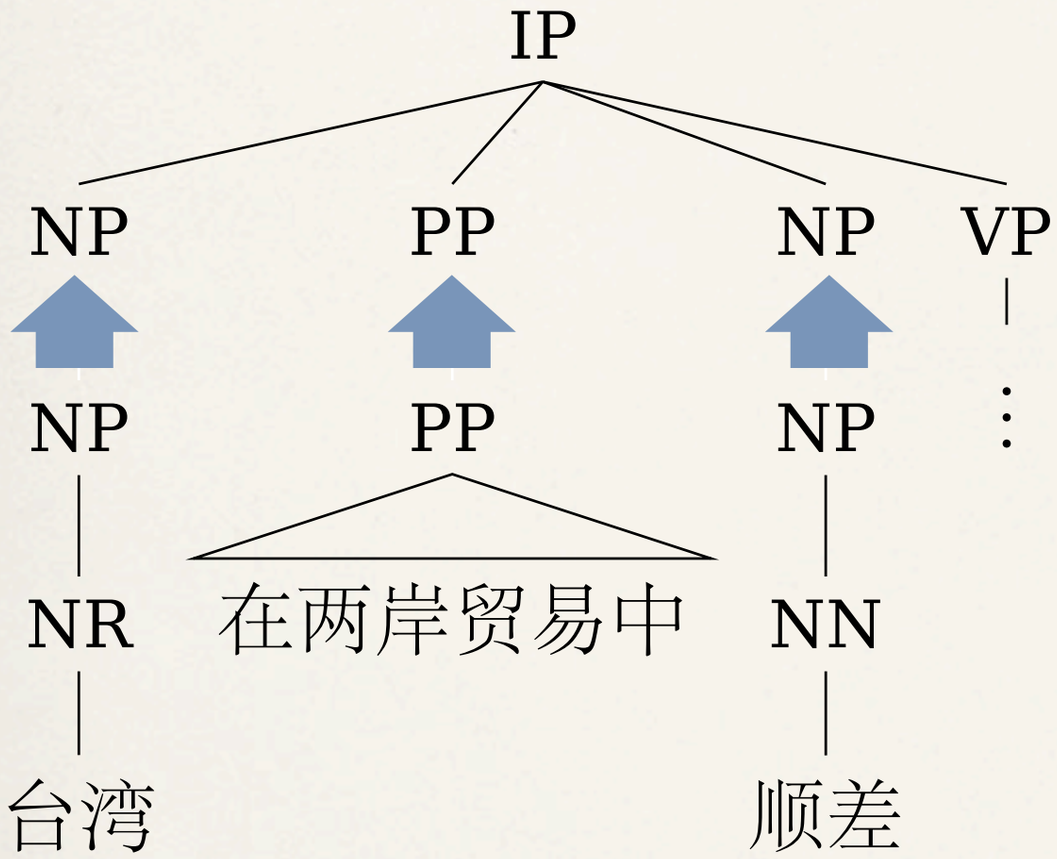

IP
NP
PP
NP
VP
NR
在两岸贸易中
NN
台湾
顺差
Taiwan's
surplus
NP
NN
in trade between the two shores
NP
PP
NP


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

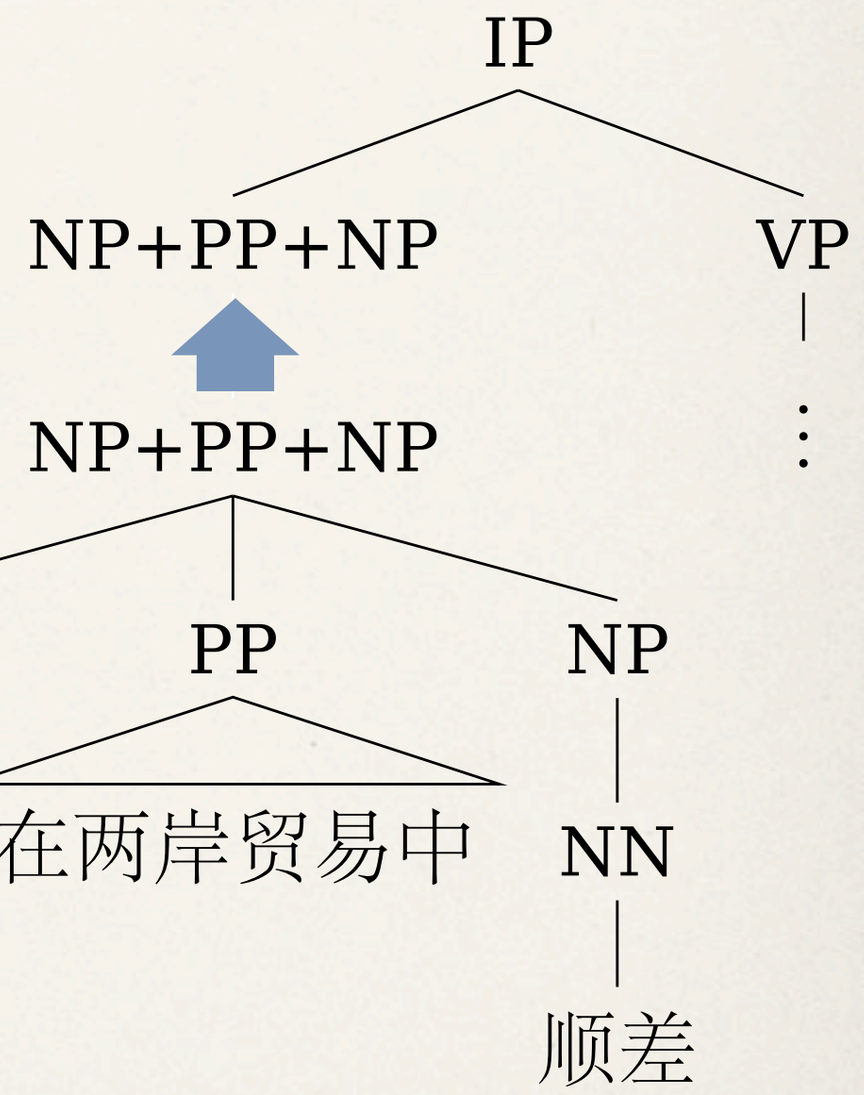
Extracting more rules



Extracting more rules



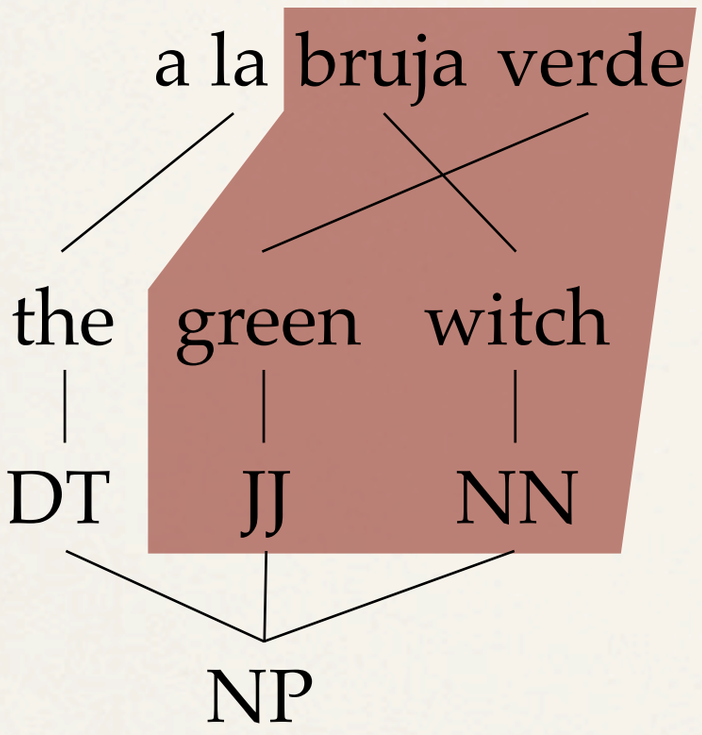
Tree-Sequence Substitution Grammar (M. Zhang et al., 2008)



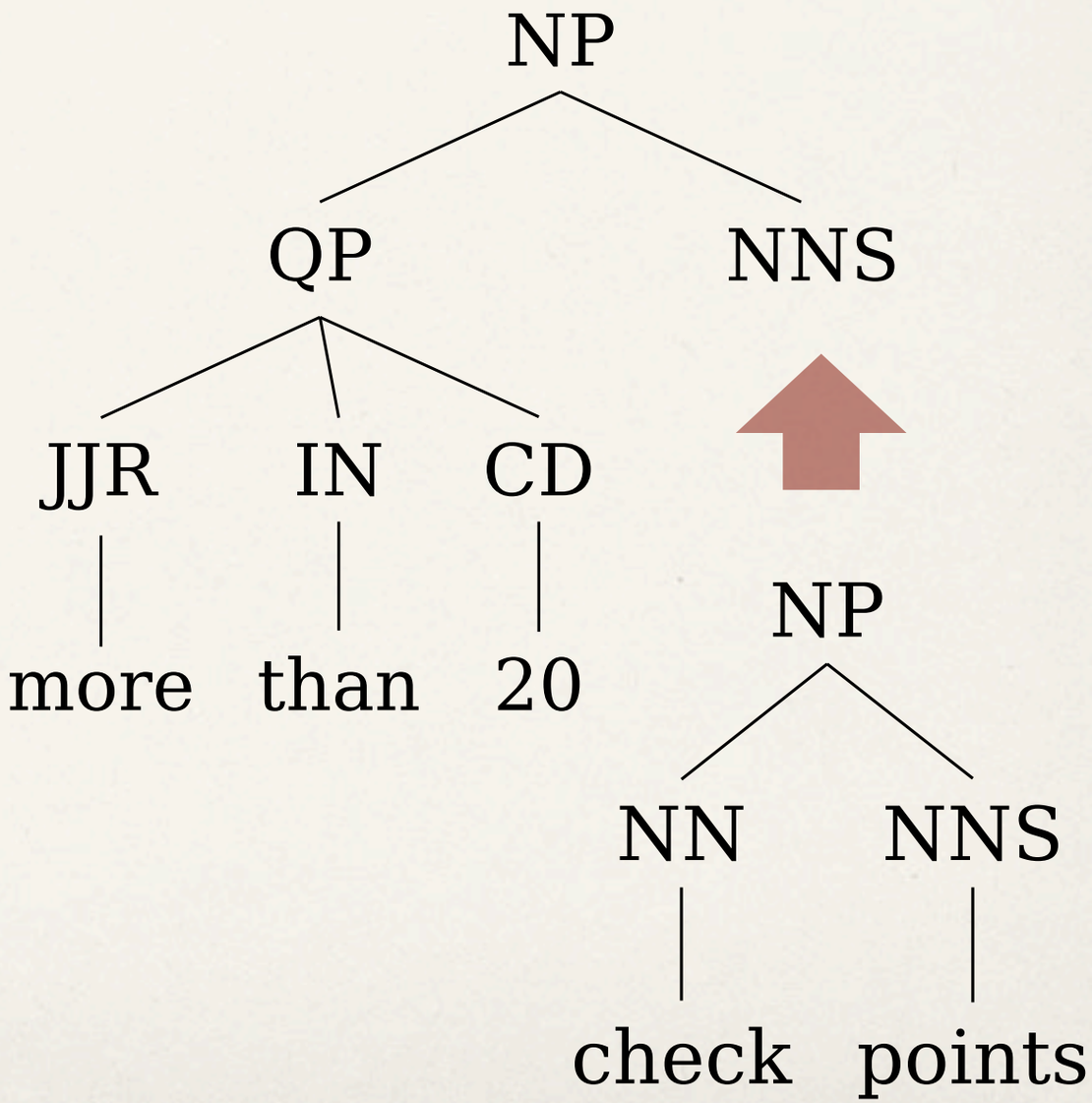
Syntax-Augmented Machine Translation (Venugopal & Zollmann)

Why is tree-to-tree hard?

too few rules

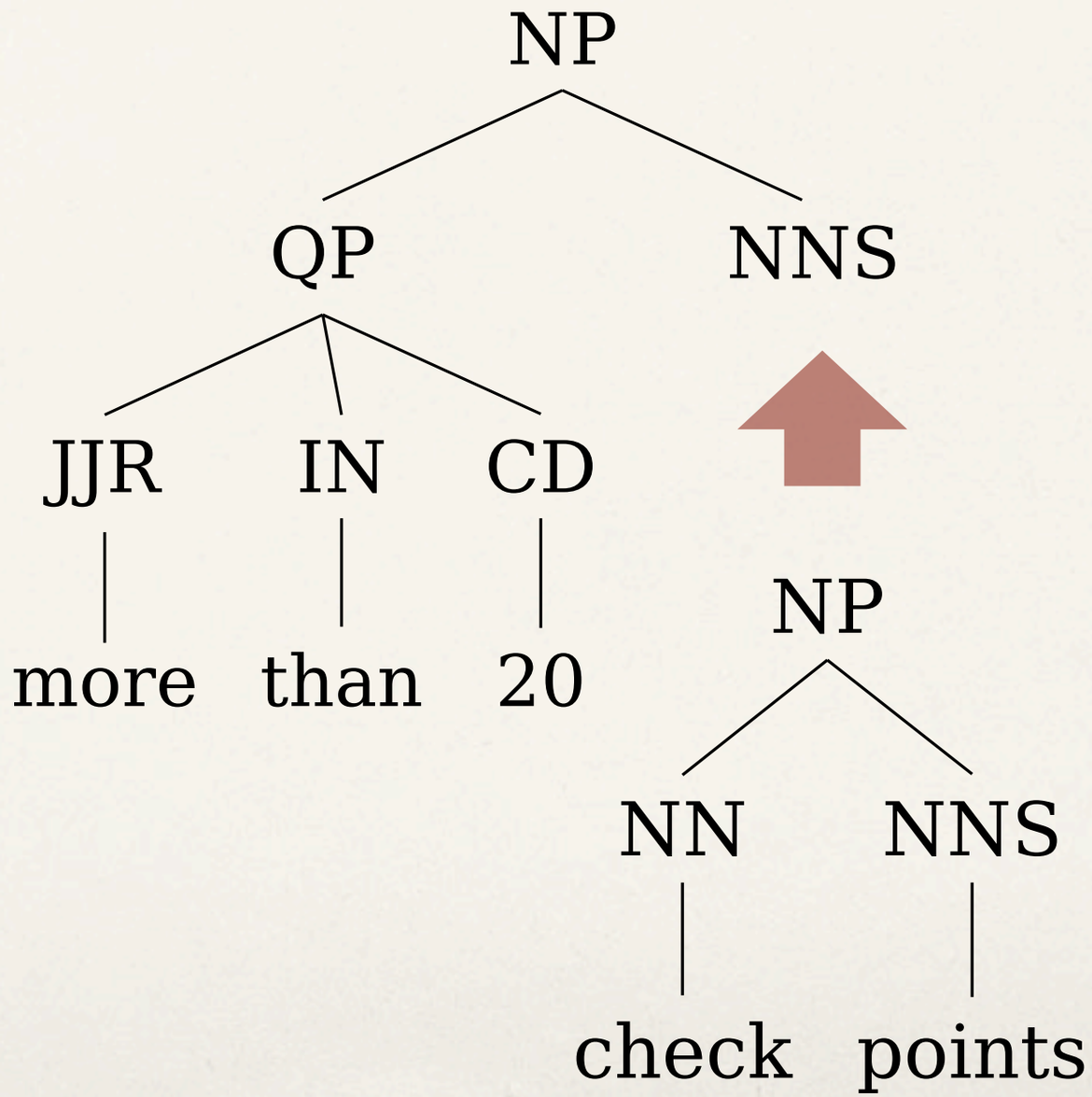


too few derivations

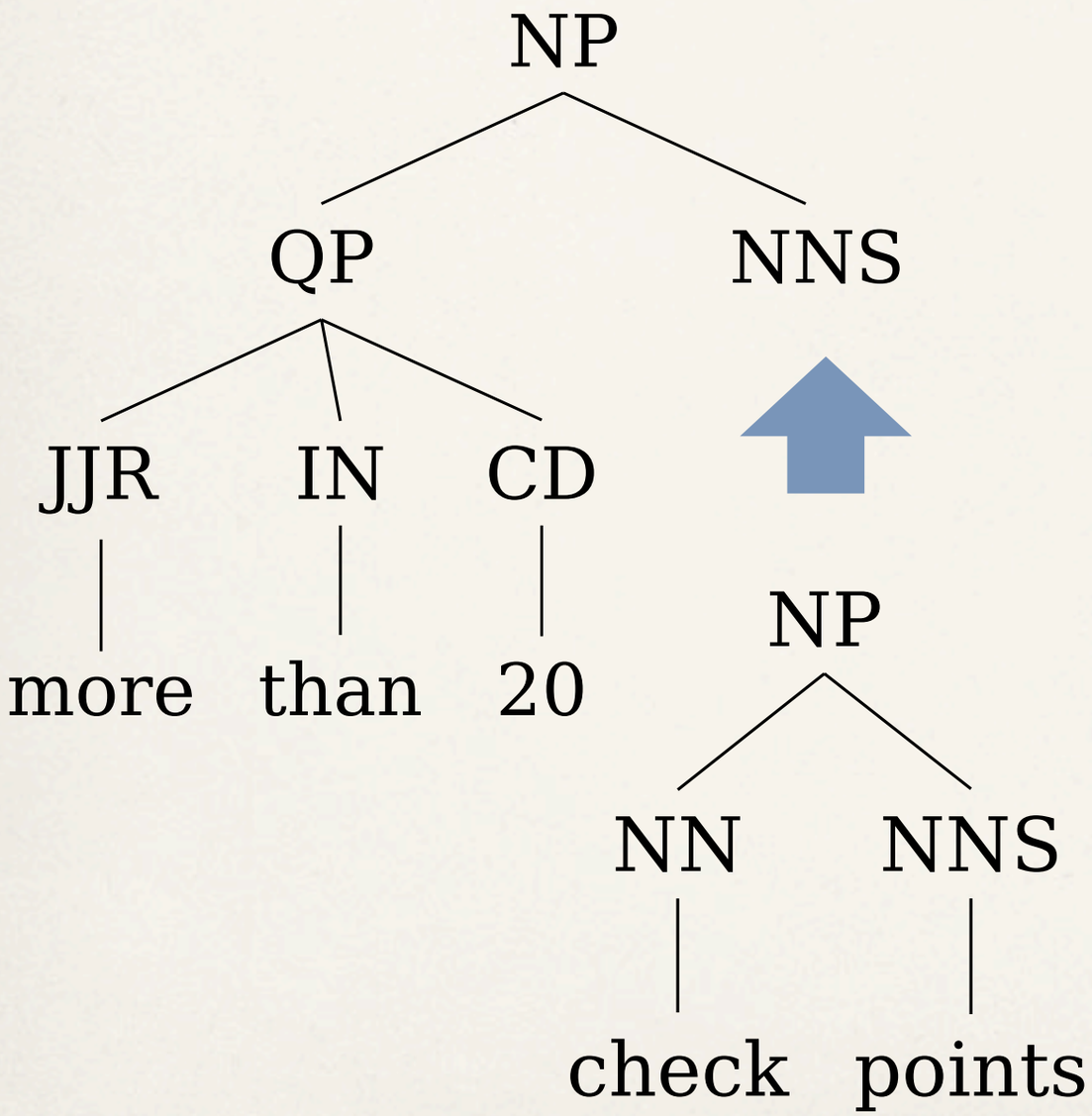


Why is tree-to-tree hard?

too few derivations

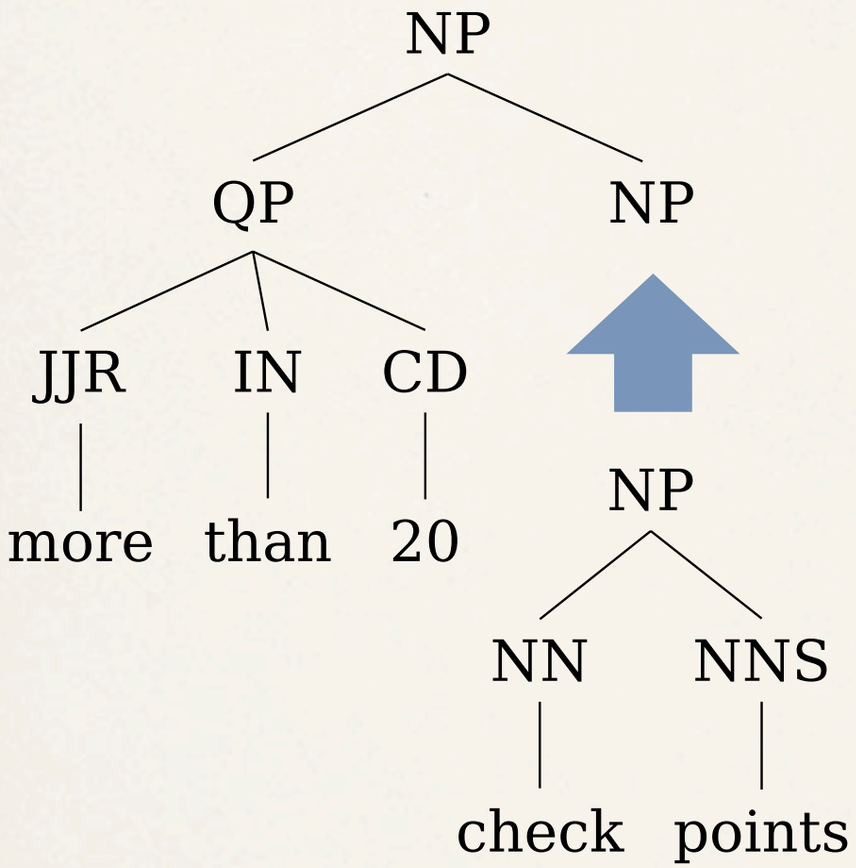


Allow more derivations



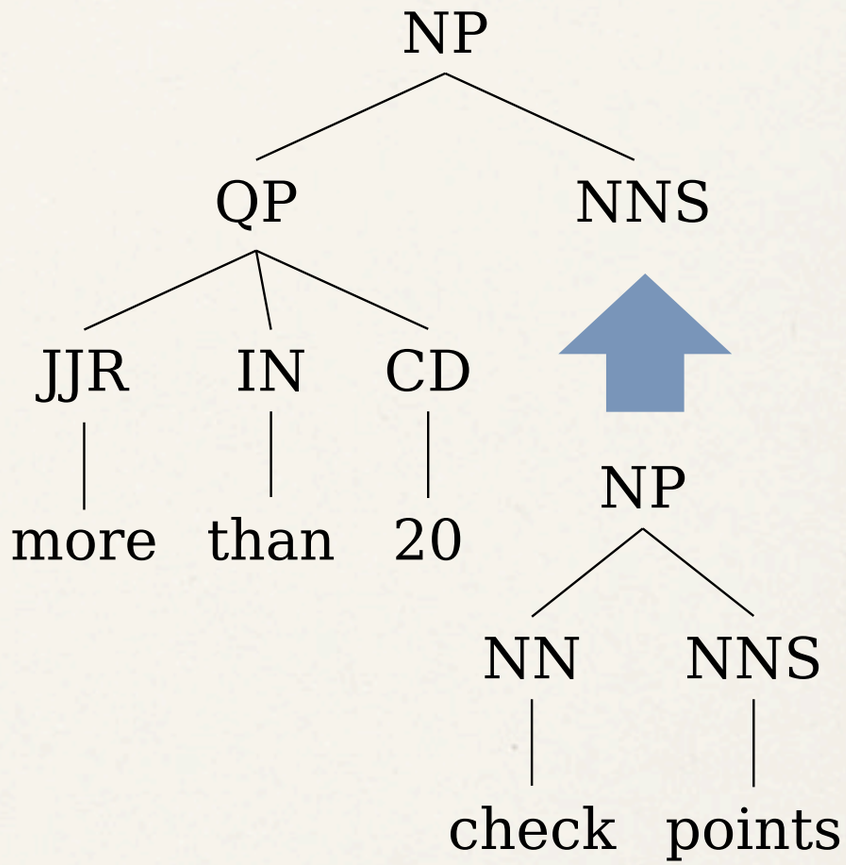
- STSG: allow only matching substitutions
- Hiero-like: allow any substitutions
- Let the model learn to choose:
  - matching substitutions
  - mismatching substitutions
  - monotone phrase-based

Allow more derivations



fire *subst*:NP→NP

fire *subst:match*

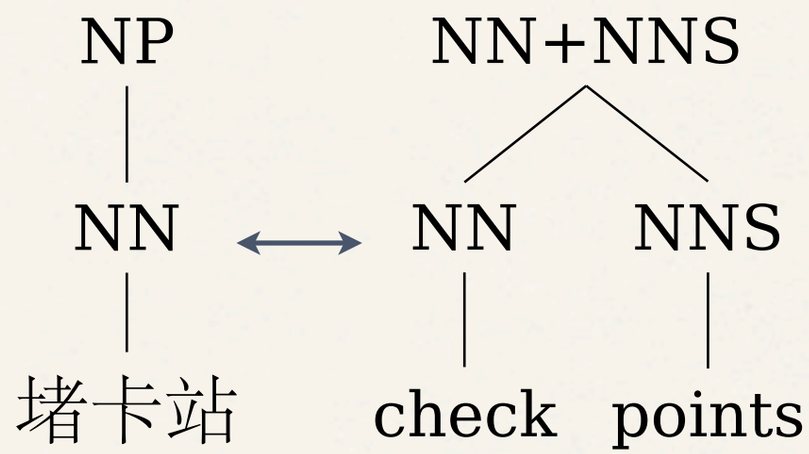


fire *subst*:NNS→NP

fire *subst:unmatch*

Allow more derivations

Cross-lingual features



fire *root*:NP,NN+NNS

suggested by Adam Pauls

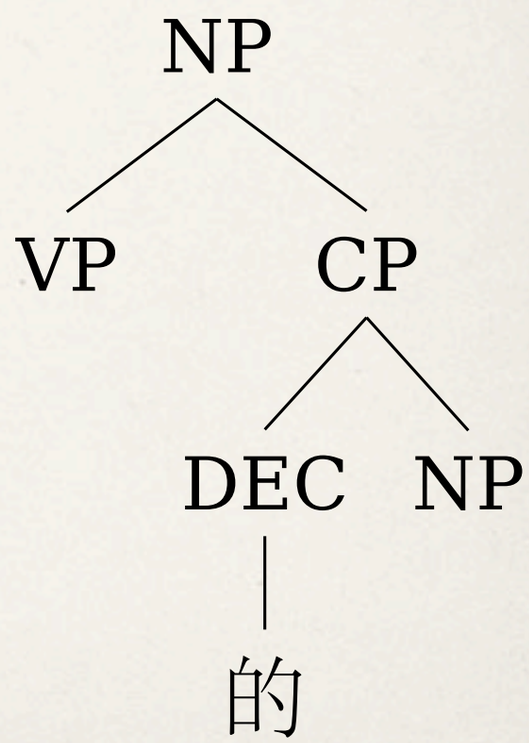
Allow more derivations

Hiero-like decoding

$$\frac{[X,i,j] \qquad [X,j+1,k]}{[X,i,k]} \qquad X \rightarrow X \text{ 的 } X$$

STSG decoding

$$\frac{[VP,i,j] \qquad [NP,j+1,k]}{[NP,i,k]}$$



fuzzy STSG decoding

$$\frac{[A,i,j] \qquad [B,j+1,k]}{[NP,i,k]}$$

Experiments

	Chinese-English	Arabic-English
parallel text	240M+260M	190+220M
language model	2G	
parser (source)	800k	600k
parser (target)	2.1M	

Results

	Chinese-English			Arabic-English		
extraction	rules	feats	BLEU	rules	feats	BLEU
Hiero	440M	1k	23.7	790M	1k	48.9
fuzzy STSG	50M	5k	23.9	38M	5k	47.5
fuzzy STSG +binarize	64M	5k	24.3	40M	6k	48.1
fuzzy STSG +SAMT	440M	160k	24.3	790M	130k	49.7

Example tree-to-tree translation

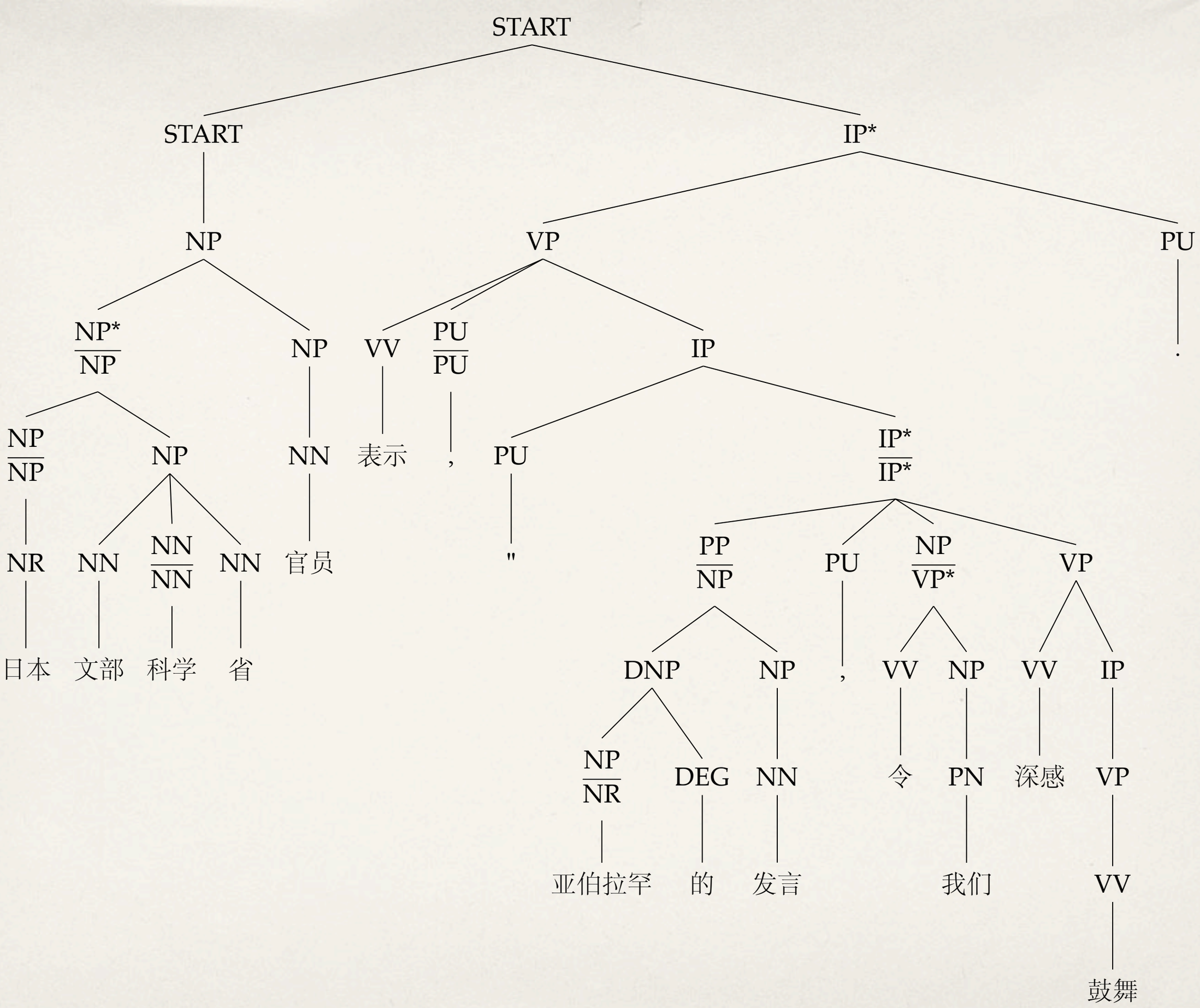
日本 文部 科学 省 官员 表示 , " 亚 伯拉罕 的 发言 , 令 我们 深感 鼓舞
Japan MEXT official said , “ Abraham ’s comment make us deeply-feel courage

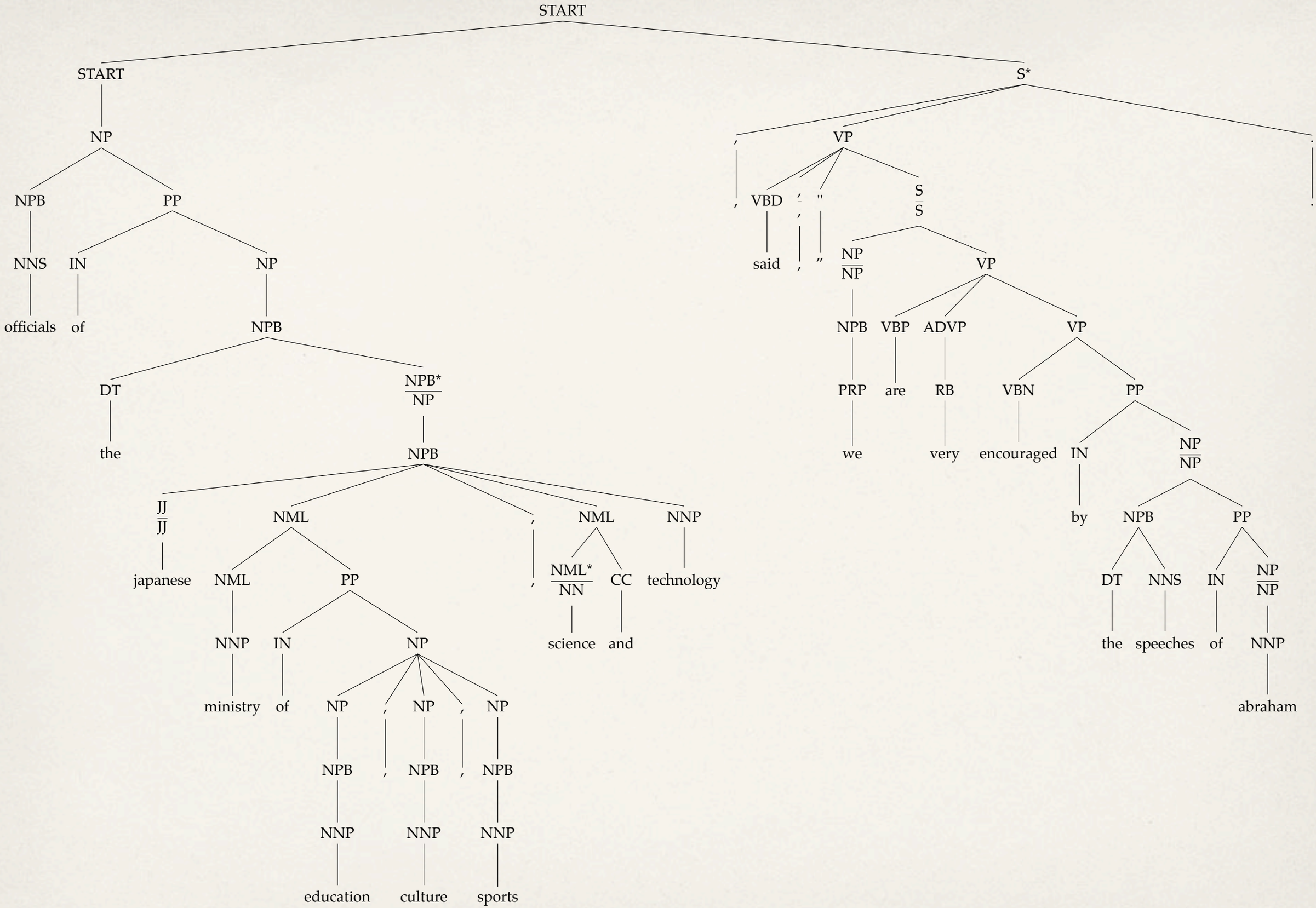
reference: An official from Japan ’s science and technology ministry said , " We are highly encouraged by Abraham ’s comment .

Hiero: Officials of the Japanese ministry of education and science , " said Abraham speeches , we are deeply encouraged by .

string-to-tree: Japan ’s ministry of education , culture , sports , science and technology , " Abraham ’s statement , which is most encouraging , " the official said .

Fuzzy STSG, binarize: Officials of the Japanese ministry of education , culture , sports , science and technology , said , " we are very encouraged by the speeches of Abraham .


START
START
IP*
NP
VP
PU
NP*
NP
NP
VV
PU
PU
IP
.
NP
NP
NP
NN
表示
,
PU
IP*
IP*
NR
NN
NN
NN
官员
"
PP
NP
PU
NP
VP*
VP
日本
文部
科学
省
DNP
NP
,
VV
NP
VV
IP
NP
NR
DEG
NN
令
PN
深感
VP
亚伯拉罕
的
发言
我们
VV
鼓舞


START
START
S*
NP
,
VP
.
NPB
PP
,
VBD
"
,
"
S
S
NNS
IN
NP
said
,
"
NP
NP
VP
officials
of
NPB
NPB
VBP
ADVP
VP
DT
NPB*
NP
PRP
are
RB
VBN
PP
the
NPB
we
very
encouraged
IN
NP
NP
JJ
JJ
NML
,
NML
NNP
by
NPB
PP
japanese
NML
PP
,
NML*
NN
CC
technology
DT
NNS
IN
NP
NP
NNP
IN
NP
science
and
the
speeches
of
NNP
ministry
of
NP
,
NP
,
NP
abraham
NPB
,
NPB
,
NPB
NNP
NNP
NNP
education
culture
sports

officials of NPB

DT NPB*/NP

the NPB

JJ/JJ NML , NML NNP

japanese NML PP , NML*/NN CC technology

NNP IN NP science and

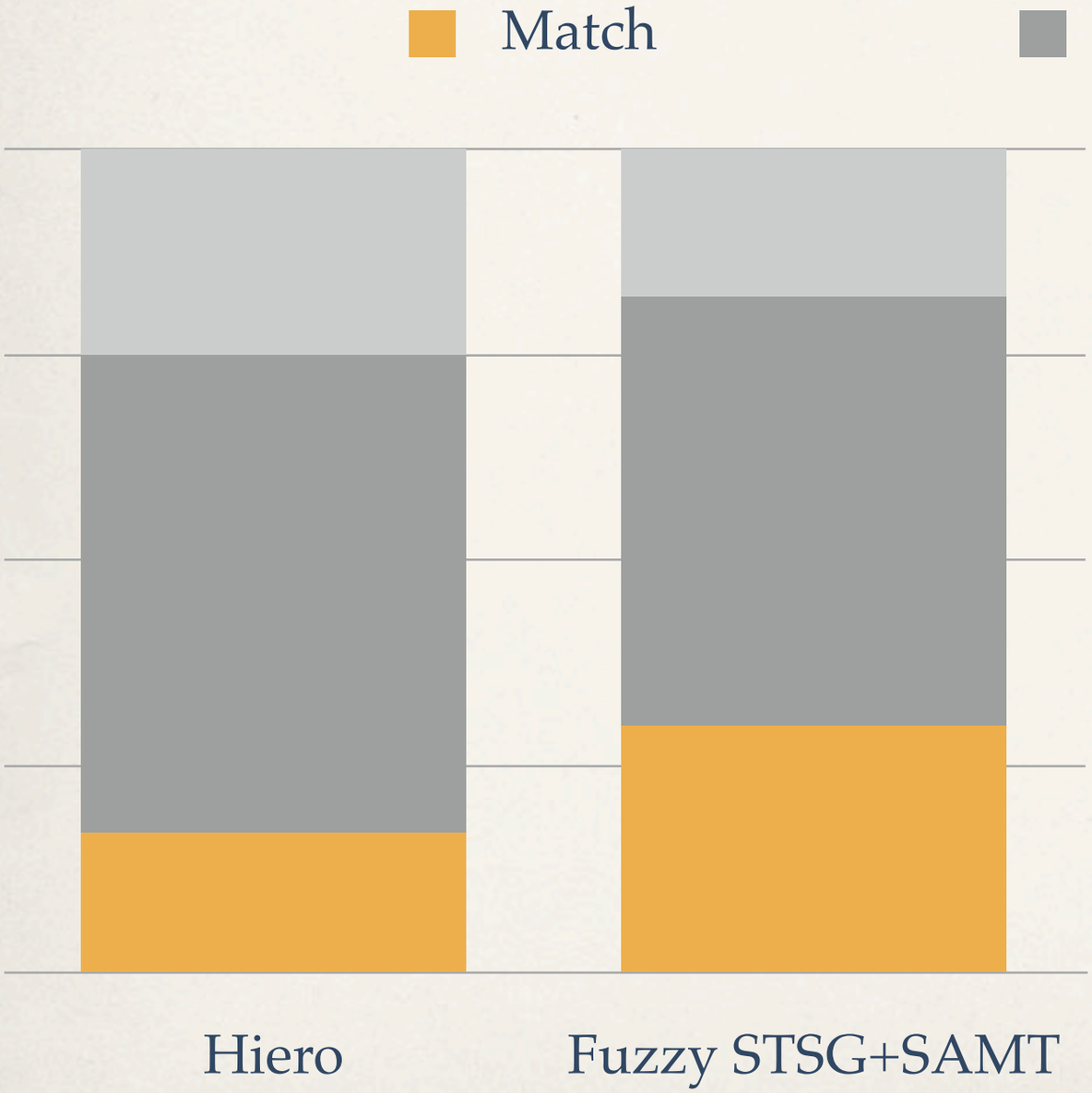
ministry of NP , NP , NP

NPB , NPB , NPB

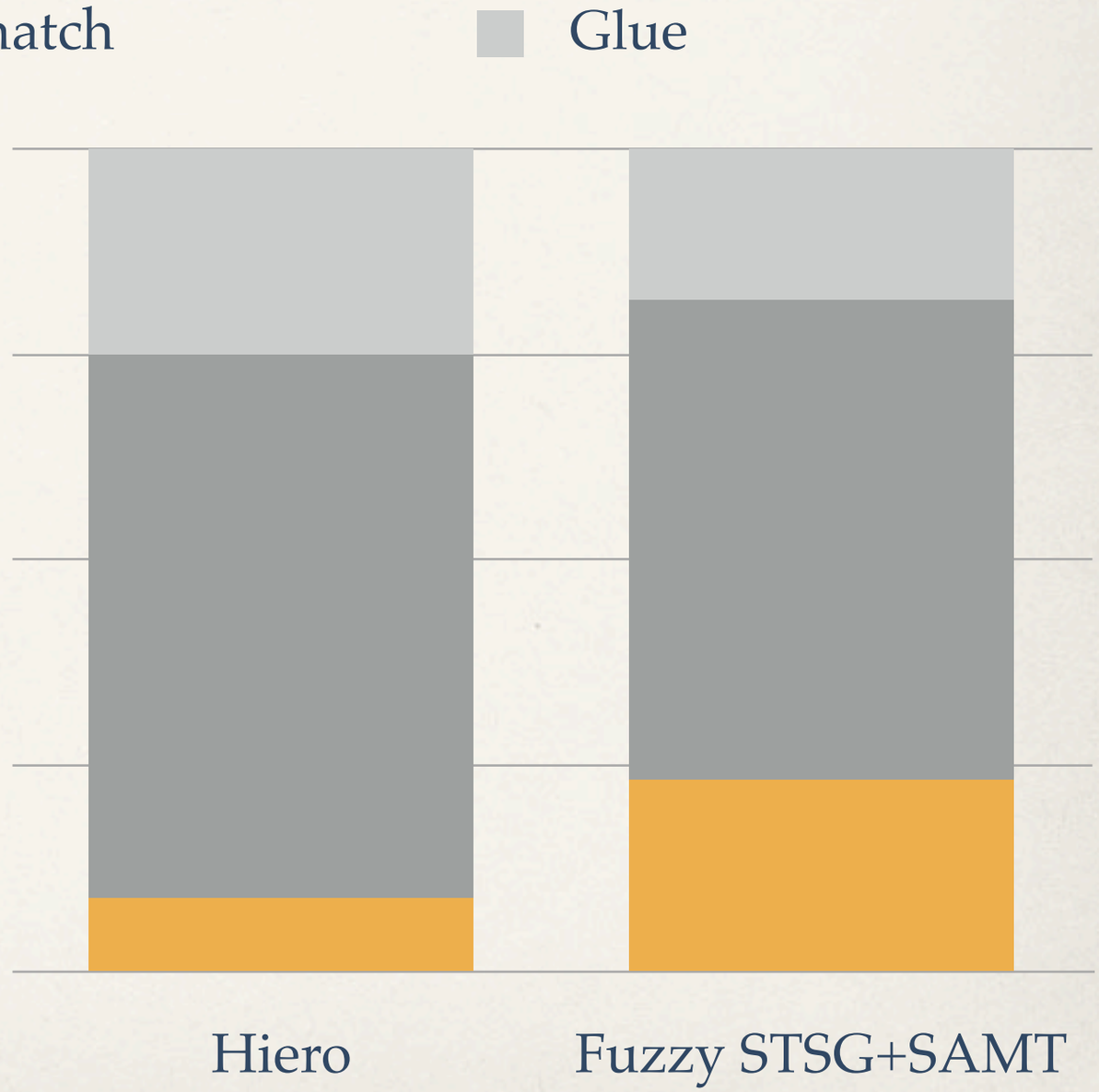
NNP NNP NNP

education culture sports

Rule usage (Chinese-English)

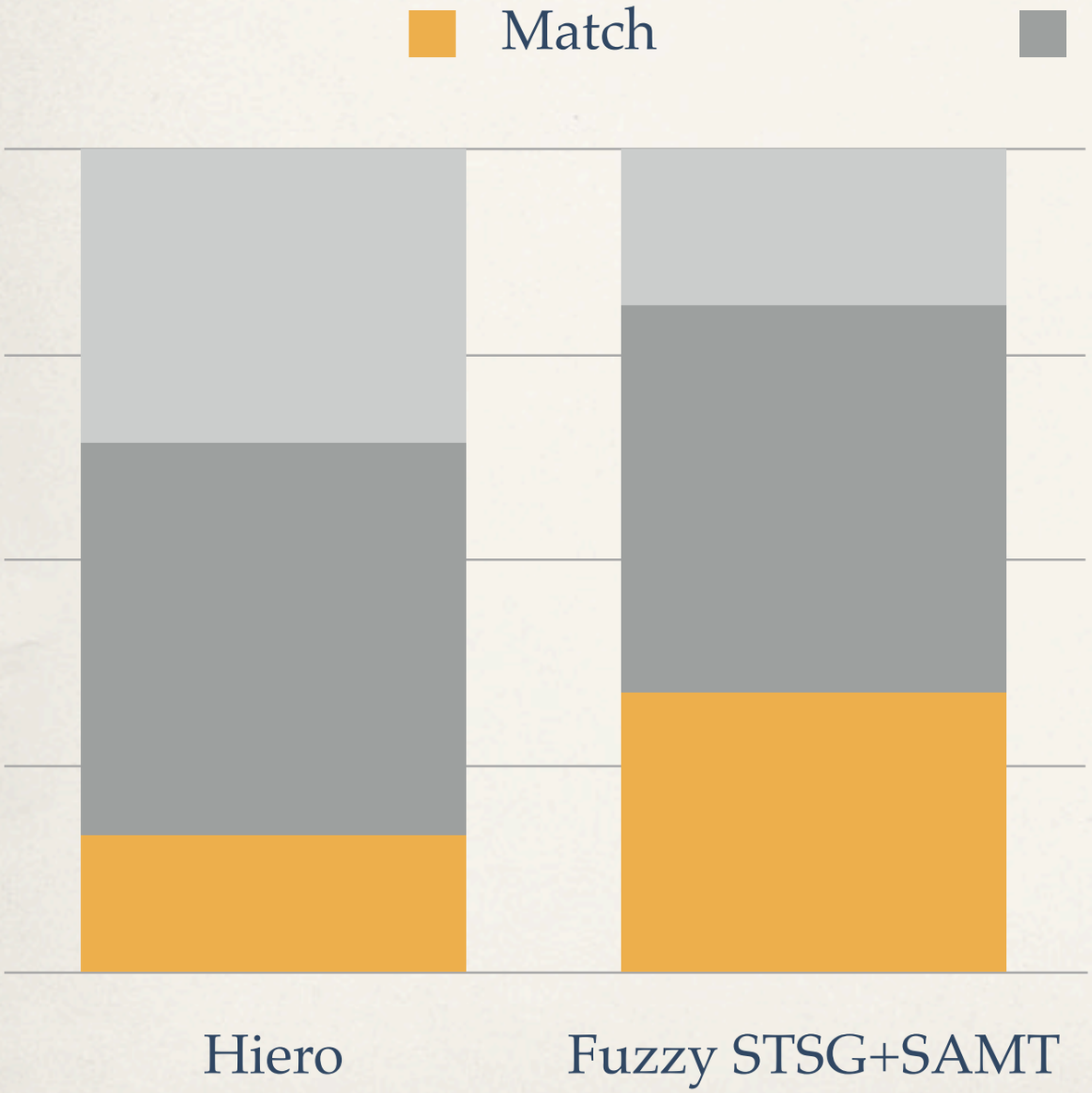


Chinese side

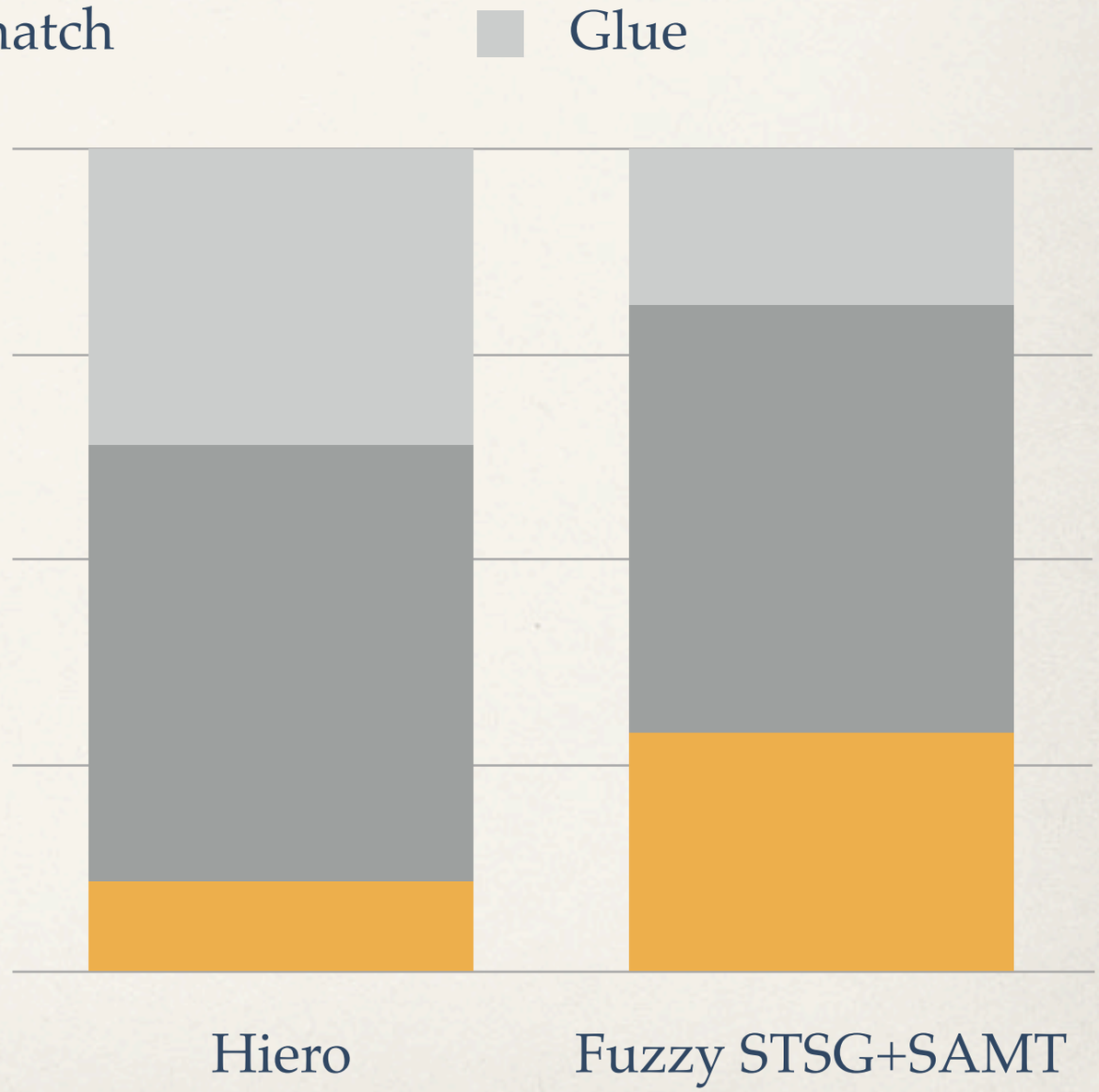


English side

Rule usage (Arabic-English)



Arabic side



English side

Conclusions

- Why is tree-to-tree translation hard?
  - Too few rules
  - Too few derivations
- How can we make it better?
  - Extract more rules: even simple binarization works
  - Allow more derivations: let model learn how much syntax to use