Grammar Induction for Machine Translation

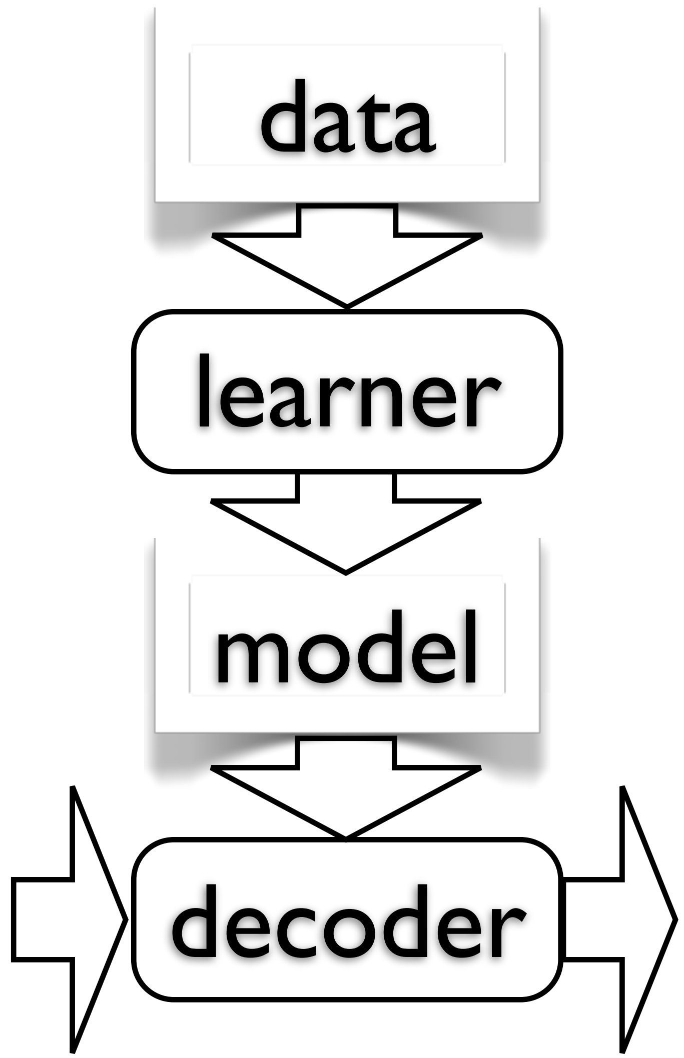
Taro Watanabe
taro.watanabe @ nict.go.jp

NICT

Machine Translation

"I fire linguists"

黑山头口岸联检部门将原来要二至三天办完的出入境手续改为一天办完。



"I'm not crude"

The United Inspection Department of Heishantou Port has shortened the procedures for leaving and entering the territory from originally 2 - 3 days to 1 day.

- A data-driven approach to MT (or, "crude force of computer")
- We learn parameters from data assuming a "model"

Bilingual Data

1.上海浦东开发与法制建设同步
2.新华社上海二月十日电（记者谢金虎、张持坚）
3.上海浦东近年来颁布实行了涉及经济、贸易、建设、规划、科技、文教等领域的七十一件法规性文件，确保了浦东开发的有序进行。
4.浦东开发开放是一项振兴上海，建设现代化经济、贸易、金融中心的跨世纪工程，因此大量出现的是以前不曾遇到过的新情况、新问题。
5.对此，浦东不是简单的采取“干一段时间，等积累了经验以后再制定法规条例”的做法，而是借鉴发达国家和深圳等特区的经验教训，聘请国内外有关专家学者，积极、及时地制定和推出法规性文件，使这些经济活动一出现就被纳入法制轨道。
6.去年初浦东新区诞生的中国第一家医疗机构药品采购服务中心，正因为一开始就比较规范，运转至今，成交药品一亿多元，没有发现一例回扣。

1.The development of Shanghai's Pudong is in step with the establishment of its legal system
2.Xinhua News Agency, Shanghai, February 10, by wire (reporters Jinhu Xie and Chijian Zhang)
3.In recent years Shanghai's Pudong has promulgated and implemented 71 regulatory documents relating to areas such as economics, trade, construction, planning, science and technology, culture and education, etc., ensuring the orderly advancement of Pudong's development.
4.Pudong's development and opening up is a century-spanning undertaking for vigorously promoting Shanghai and constructing a modern economic, trade, and financial center. Because of this, new situations and new questions that have not been encountered before are emerging in great numbers.
5.In response to this, Pudong is not simply adopting an approach of "work for a short time and then draw up laws and regulations only after waiting until experience has been accumulated." Instead, Pudong is taking advantage of the lessons from experience of developed countries and special regions such as Shenzhen by hiring appropriate domestic and foreign specialists and scholars, by actively and promptly formulating and issuing regulatory documents, and by ensuring that these economic activities are incorporated into the sphere of influence of the legal system as soon as they appear.
6.Precisely because as soon as it opened it was relatively standardized, China's first drug purchase service center for medical treatment institutions, which came into being at the beginning of last year in the Pudong new region, in operating up to now, has concluded transactions for drugs of over 100 million yuan and hasn't had one case of kickback.

(part of LDC2007T02, English translation of Chinese treebank)

A Process of Translation

I do not want to work

?

Je ne veux pas travailler

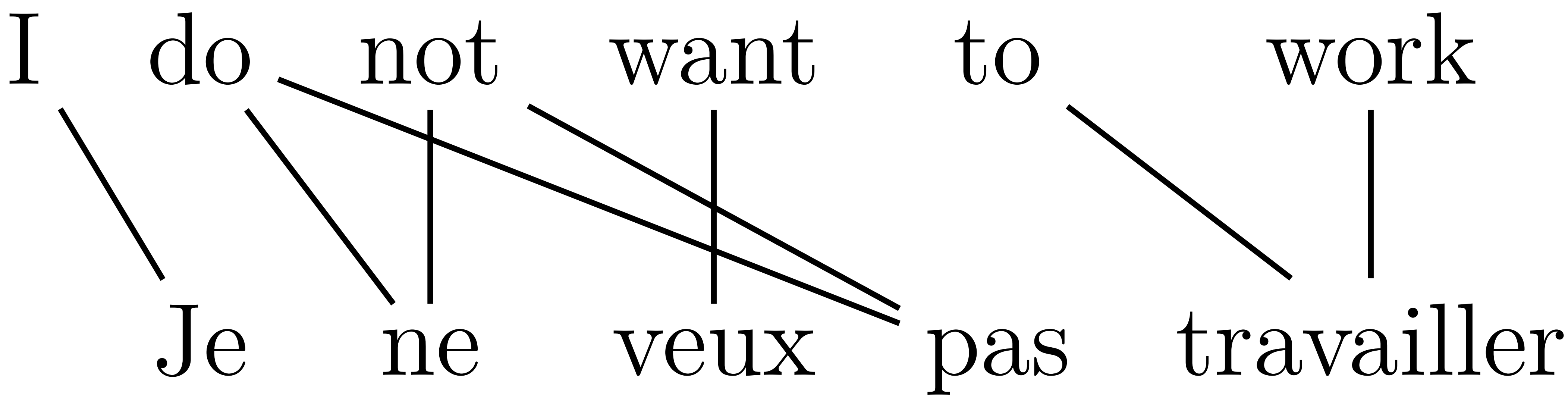
Generative Story

$$
\begin{aligned}
\hat{e} &= \arg\max_{e} Pr(e|f) \\
&= \arg\max_{e} Pr(f|e) Pr(e) \\
&= \arg\max_{e} \sum_{d} Pr(f, d|e) Pr(e)
\end{aligned}
$$

- d: a derivation which "encodes" a process of translation from e to f

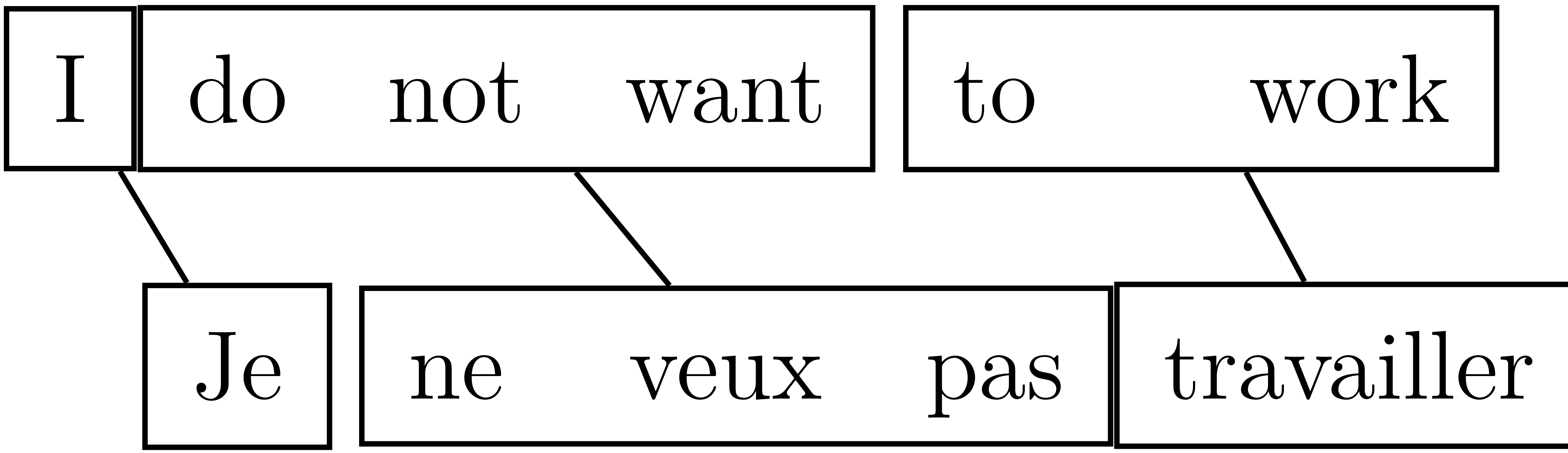
I will skip the Vauquois $\triangle$

$\mathbf{d}$ = Word Alignment



(Brown et al., 1993)

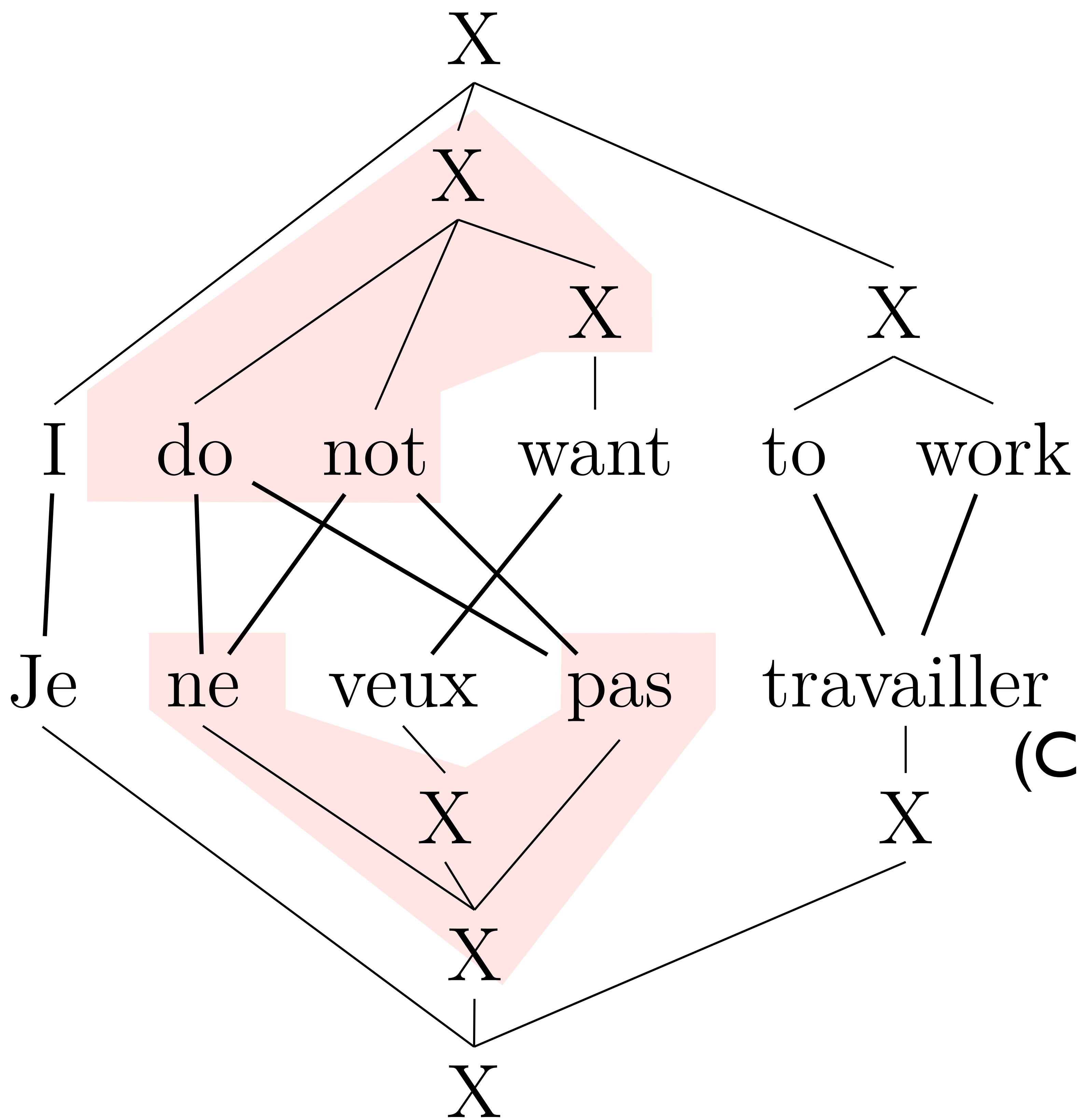
d = Phrase Pairs



(Koehn et al., 2003)

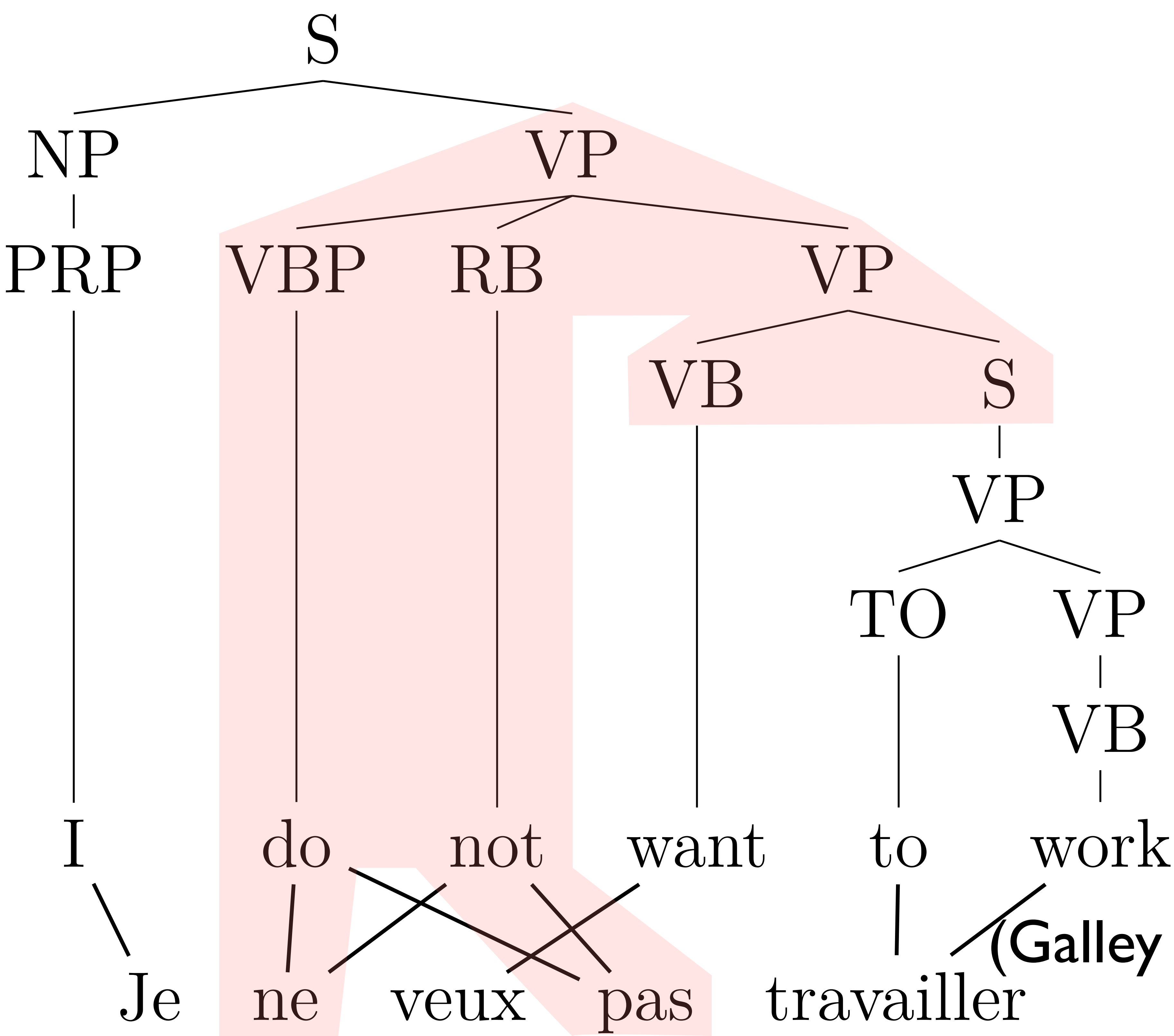
“No linguistic intuition”

d = Hierarchical Phrases



(Chiang, 2007)

d = Tree Substitution



(Galley et al., 2004)

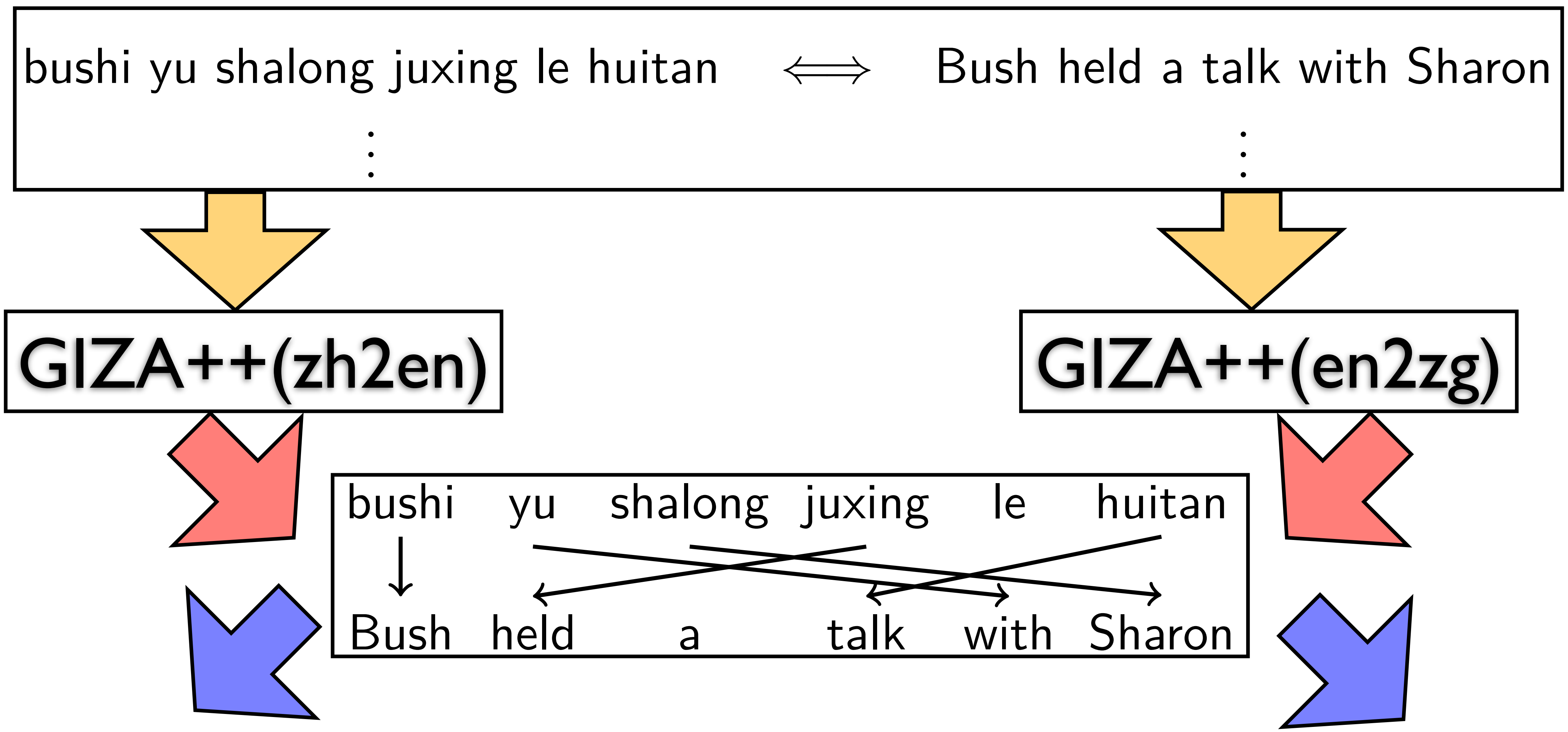
Research on MT @ NICT

- Model, Search, Optimization
- Today's Focus: Inducing structural relations (d), i.e. grammars, from the pairs of (f, e)
  - Phrase pair induction
  - Label induction

Phrase Pair Induction

An Unsupervised Model for Joint Phrase Alignment and Extraction
Graham Neubig, Taro Watanabe, Eiichiro Sumita, Shunsuke Mori, Tatsuya Kawahara. In *ACL 2011*.

“Traditional” (S)MT



bushi	|||	Bush
yu	|||	with
shalong	|||	Sharon
yu shalong	|||	with Sharon
juxing le huitan	|||	held a talk

⋮

$$\mathrm{X} \rightarrow \langle \text{bushi } \mathrm{X}_{\boxed{1}}, \text{Bush } \mathrm{X}_{\boxed{1}} \rangle$$

$$\mathrm{X} \rightarrow \langle \mathrm{X}_{\boxed{1}} \text{ yu shalong } \mathrm{X}_{\boxed{2}}, \mathrm{X}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}} \text{ with S}$$

$$\mathrm{X} \rightarrow \langle \mathrm{X}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}} \text{ le huitan}, \mathrm{X}_{\boxed{2}} \text{ a talk } \mathrm{X}_{\boxed{1}} \rangle$$

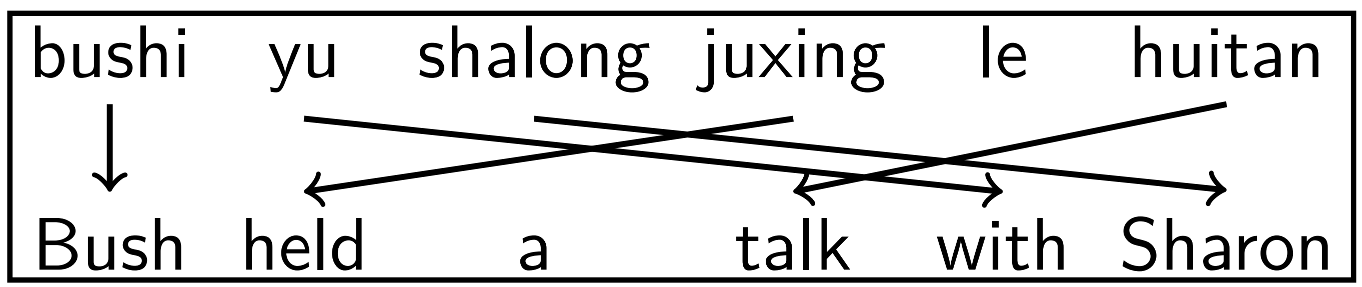
⋮

Simpler

bushi yu shalong juxing le huitan	$\Longleftrightarrow$	Bush held a talk with Sharon
$\vdots$		$\vdots$

A model to capture many-to-many alignment

(Zhang et al., 2008; DeNero et al., 2008; Blunsom et al., 2010)



bushi	|||	Bush
yu	|||	with
shalong	|||	Sharon
yu shalong	|||	with Sharon
juxing le huitan	|||	held a talk
	$\vdots$	

$$X \rightarrow \langle \text{bushi } X_{\boxed{1}}, \text{Bush } X_{\boxed{1}} \rangle$$

$$X \rightarrow \langle X_{\boxed{1}} \text{ yu shalong } X_{\boxed{2}}, X_{\boxed{1}}\ X_{\boxed{2}} \text{ with S}$$

$$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}} \text{ le huitan}, X_{\boxed{2}} \text{ a talk } X_{\boxed{1}} \rangle$$

$$\vdots$$

Single, Direct Model

bushi yu shalong juxing le huitan $\iff$ Bush held a talk with Sharon

$\vdots$ $\vdots$

bushi	|||	Bush
yu	|||	with
shalong	|||	Sharon
yu shalong	|||	with Sharon
juxing le huitan	|||	held a talk

$\vdots$

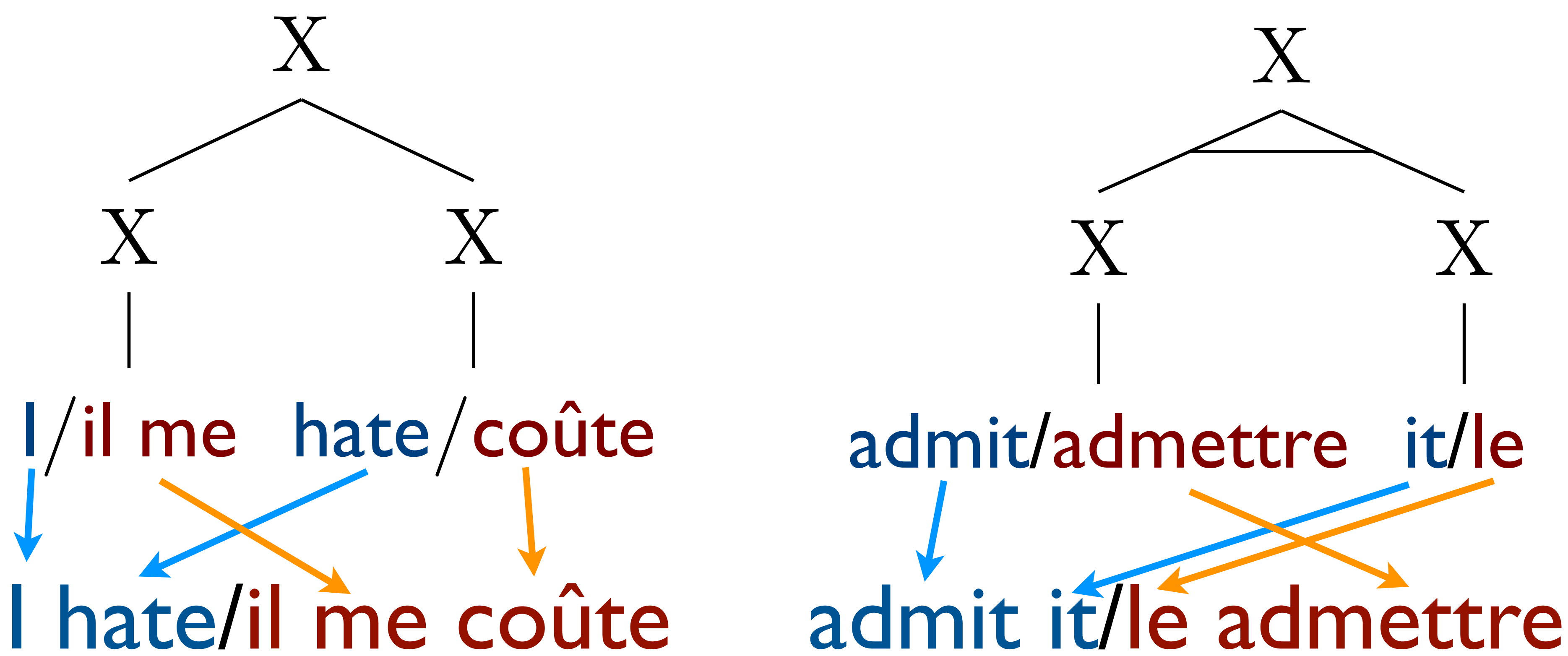
$$\mathrm{X} \rightarrow \langle \text{bushi } \mathrm{X}_{\boxed{1}}, \text{Bush } \mathrm{X}_{\boxed{1}} \rangle$$

$$\mathrm{X} \rightarrow \langle \mathrm{X}_{\boxed{1}} \text{ yu shalong } \mathrm{X}_{\boxed{2}}, \mathrm{X}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}} \text{ with S}$$

$$\mathrm{X} \rightarrow \langle \mathrm{X}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}} \text{ le huitan}, \mathrm{X}_{\boxed{2}} \text{ a talk } \mathrm{X}_{\boxed{1}} \rangle$$

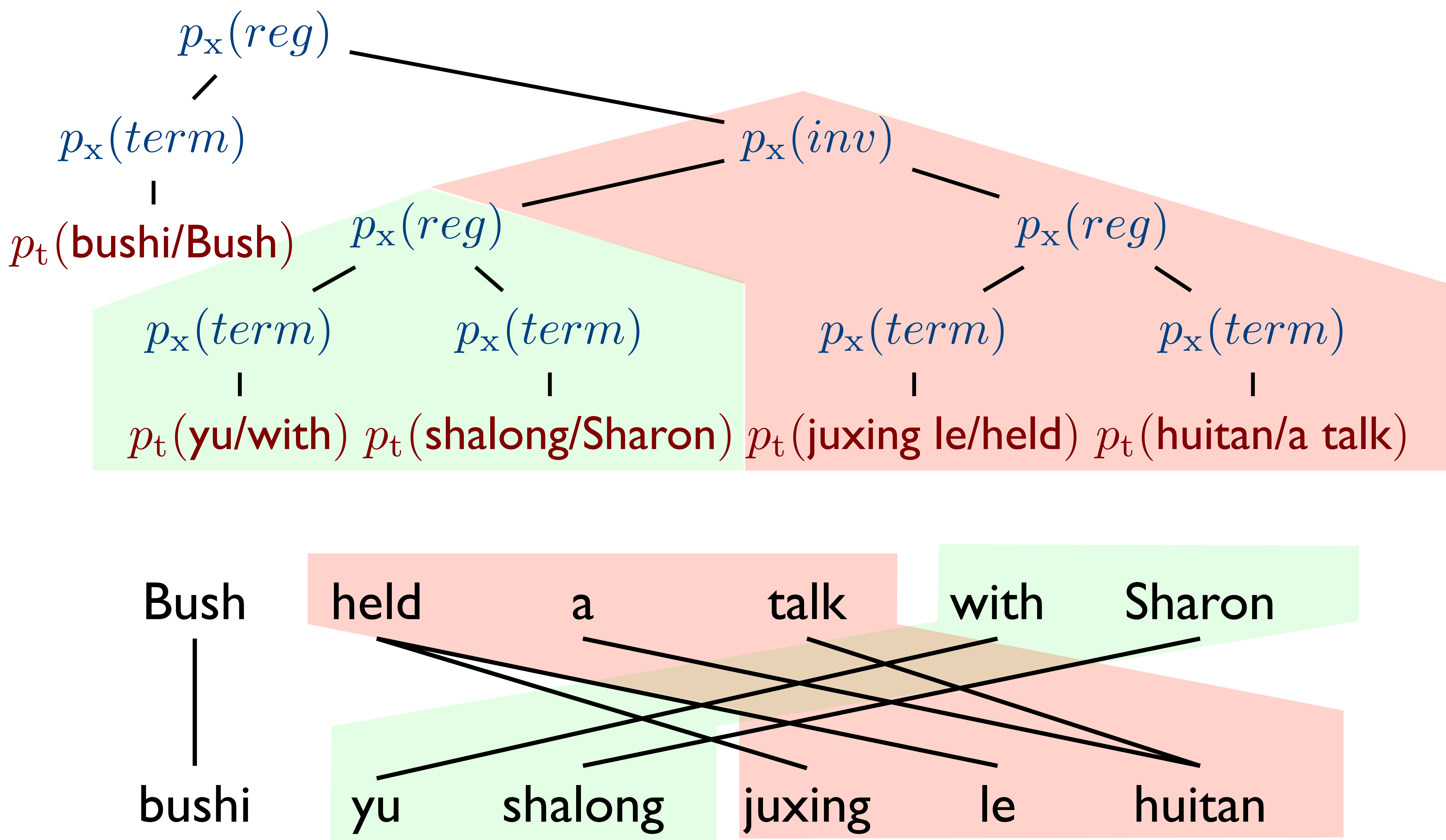
$\vdots$

ITG

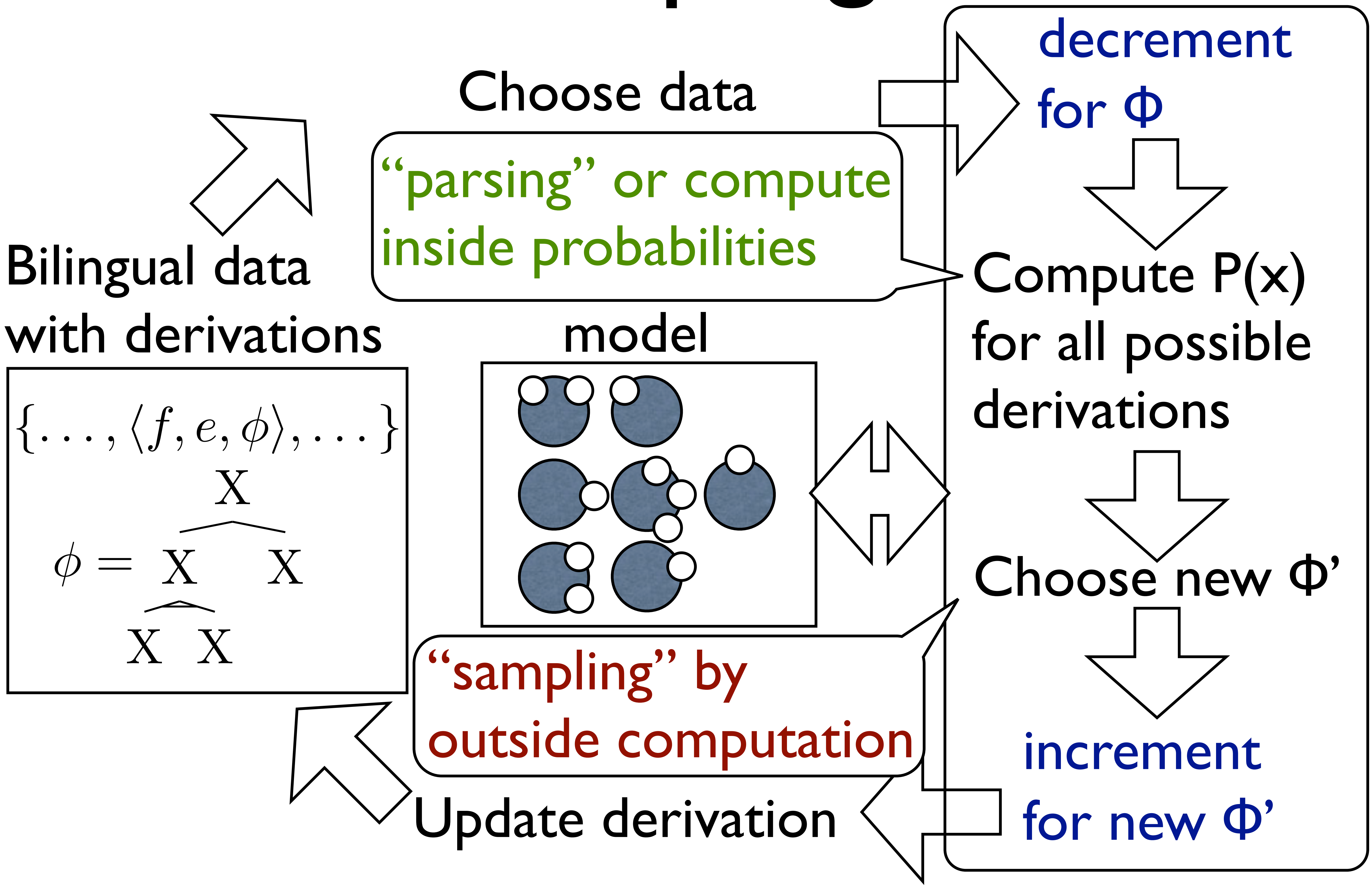


- Inversion Transduction Grammar (ITG) (Wu, 1997) is a CFG over two languages:
  - single non-terminal + regular/inverted production
  - single pre-terminal + terminals: phrase pairs

Alignment by Biparsing



Sampling



Choose data
"parsing" or compute
inside probabilities
Bilingual data
with derivations
$\{\ldots, \langle f, e, \phi \rangle, \ldots\}$
$\phi =$ X X X X X
model
decrement
for Φ
Compute P(x)
for all possible
derivations
Choose new Φ'
increment
for new Φ'
"sampling" by
outside computation
Update derivation

Parsing

● = parsed source/target span

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

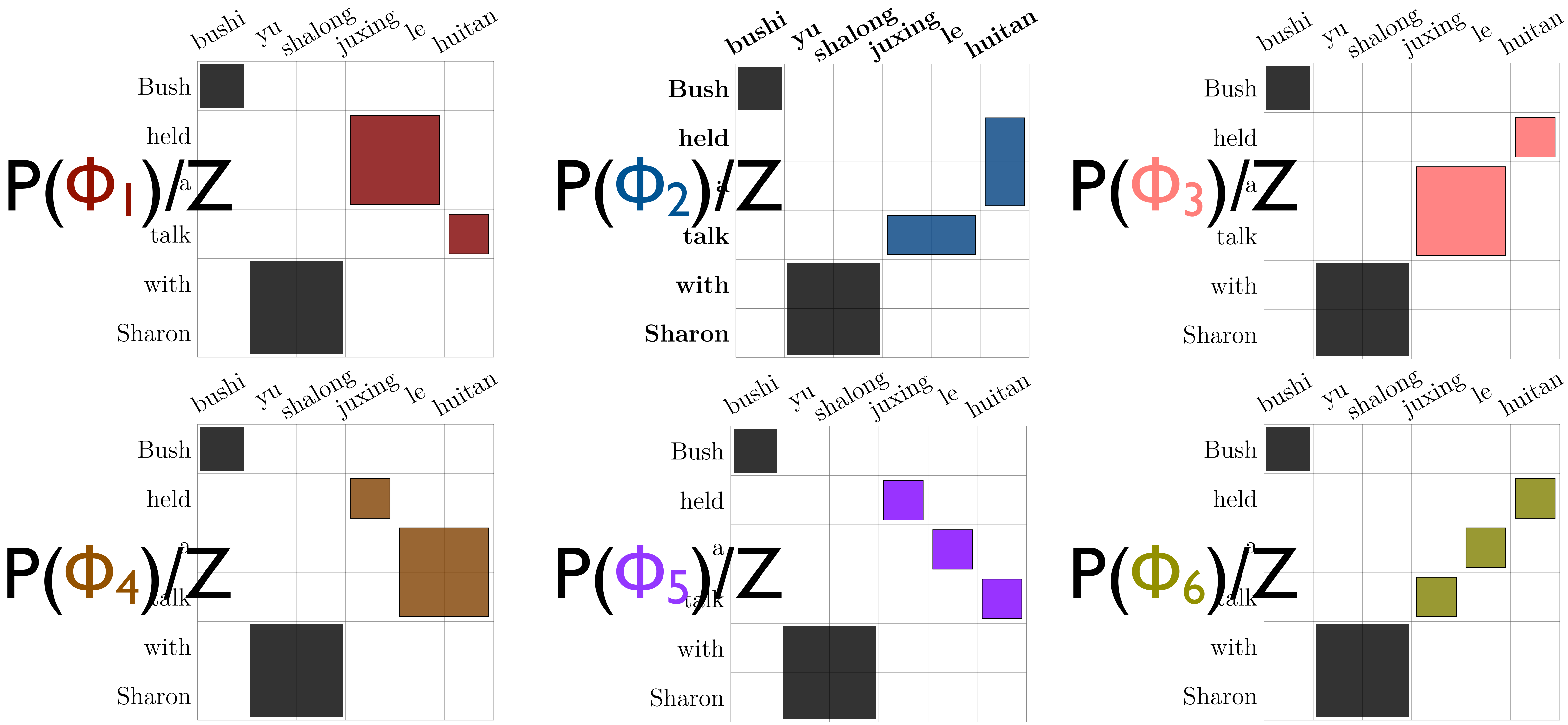
synchronized by the # of words parsed

continue until: ●●●●●●,●●●●●●

no chart, no stack

(Saers et al., 2009)

Block Sampling



$$Z = P(\Phi_1) + P(\Phi_2) + P(\Phi_3)$$
$$+ P(\Phi_4) + P(\Phi_5) + P(\Phi_6)$$

- Sample new “sentence-wise” derivations Φ’

Model

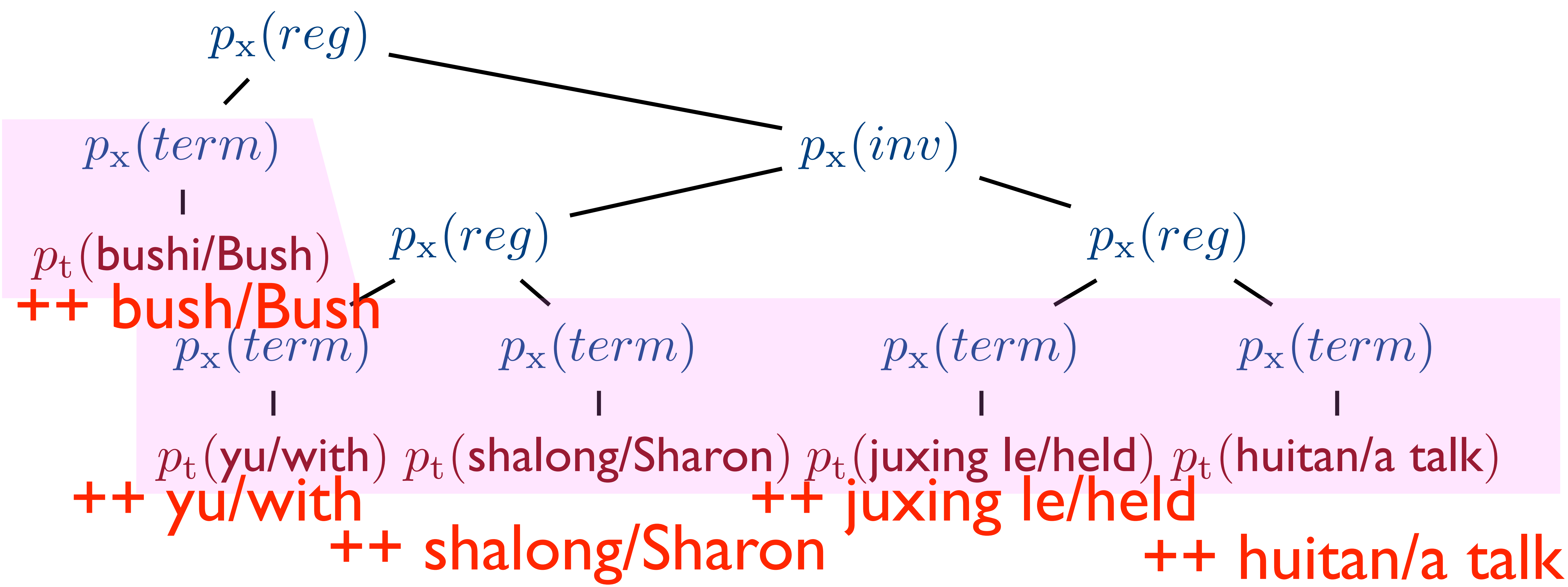
$$
\begin{aligned}
P_{\mathrm{t}} &\sim \mathrm{PY}(d, s, P_{base}) \\
P_{\mathrm{x}} &\sim \mathrm{Dirichlet}(\alpha, 1/3)
\end{aligned}
$$

- Bayesian approach:
  - Terminal probabilities for phrase pair (f, e) come from Pitman-Yor Process
  - Branch probabilities (x = reg,inv,term) come from Dirichlet distribution

$$
\begin{aligned}
P_{\mathrm{t}}(f, e) &= \frac{c(f, e)}{c(\_) + s} + \frac{s}{c(\_) + s} P_{base}(f, e) \\
P_{\mathrm{x}}(x) &= \frac{c(x)}{c(\_) + \alpha_x} + \frac{\alpha_x / 3}{c(\_) + \alpha_x}
\end{aligned}
$$

(omit d for brevity)

Minimum Phrases

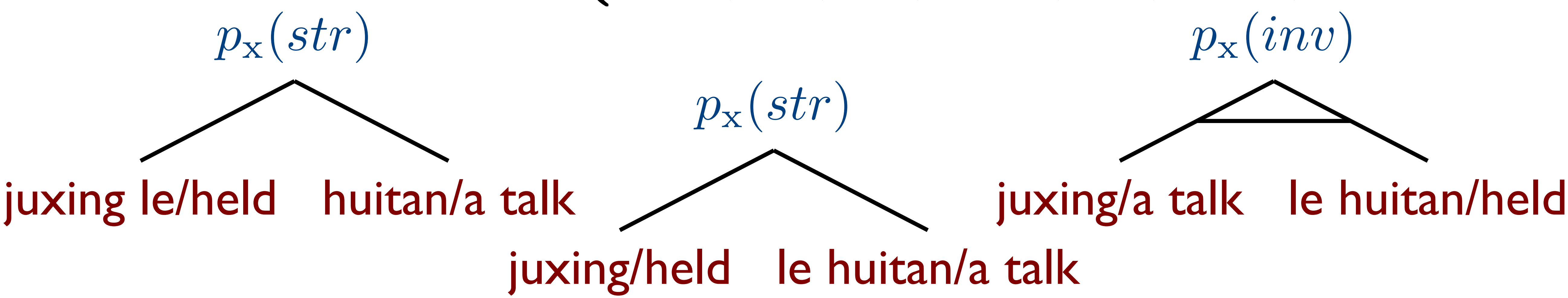


- Generate from a root until we reach terminals
- Sampled derivations contain only minimum phrases

Fallback Modeling

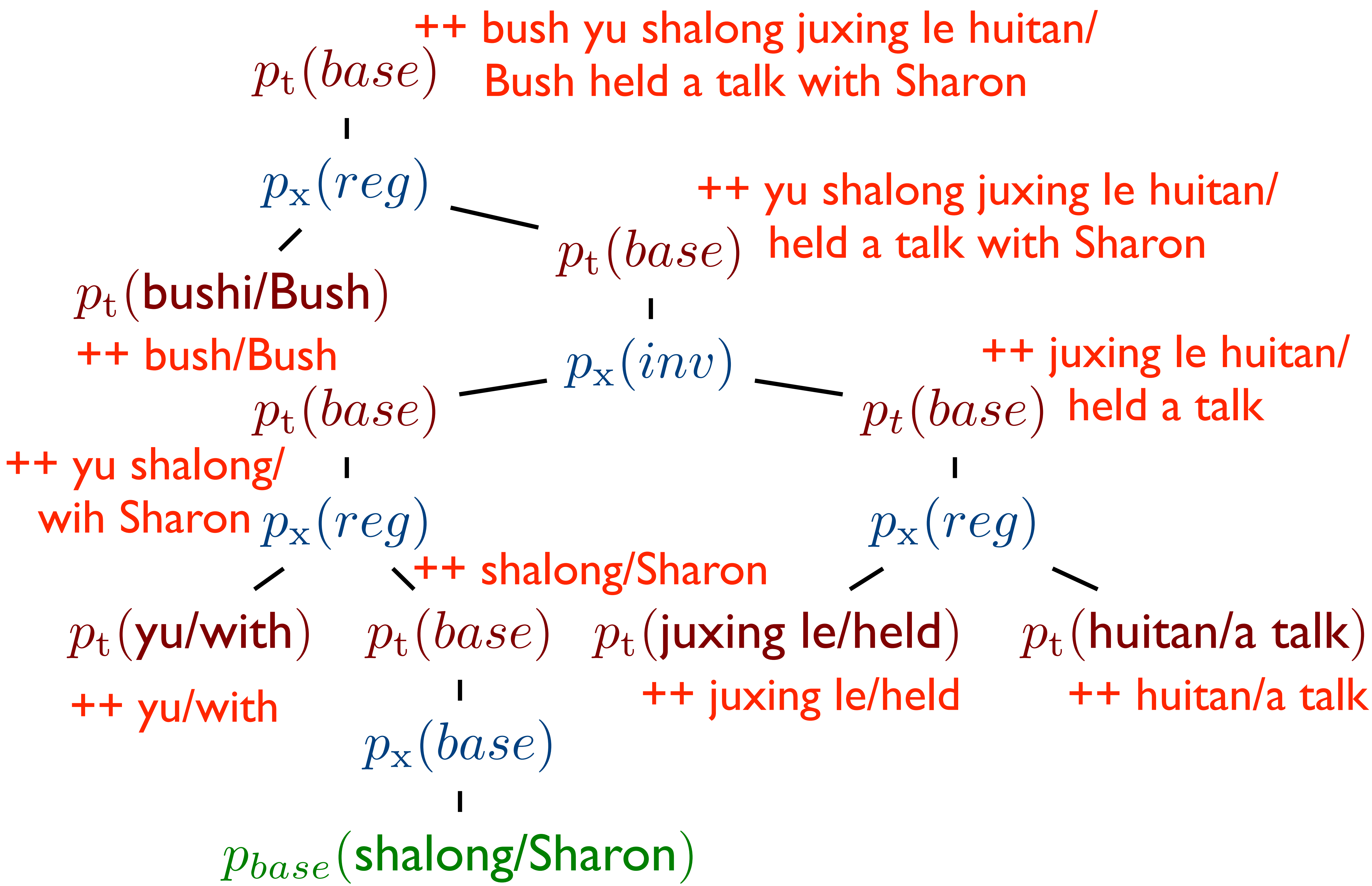
$$P_{\mathrm{t}}(f,e) = \frac{c(f,e)}{c(\_) + s} + \frac{s}{c(\_) + s} P_{dac}(f,e)$$

$$P_{dac}(f,e) = \begin{cases} P_{\mathrm{x}}(base) P_{base}(f,e) \\ P_{\mathrm{x}}(str) P_{\mathrm{t}}(f',e') P_{\mathrm{t}}(f'',e'') \\ P_{\mathrm{x}}(inv) P_{\mathrm{t}}(f',e'') P_{\mathrm{t}}(f'',e') \end{cases}$$



- Compute terminal (phrase) probabilities, first
- If not in the model, split and divide-and-conquer

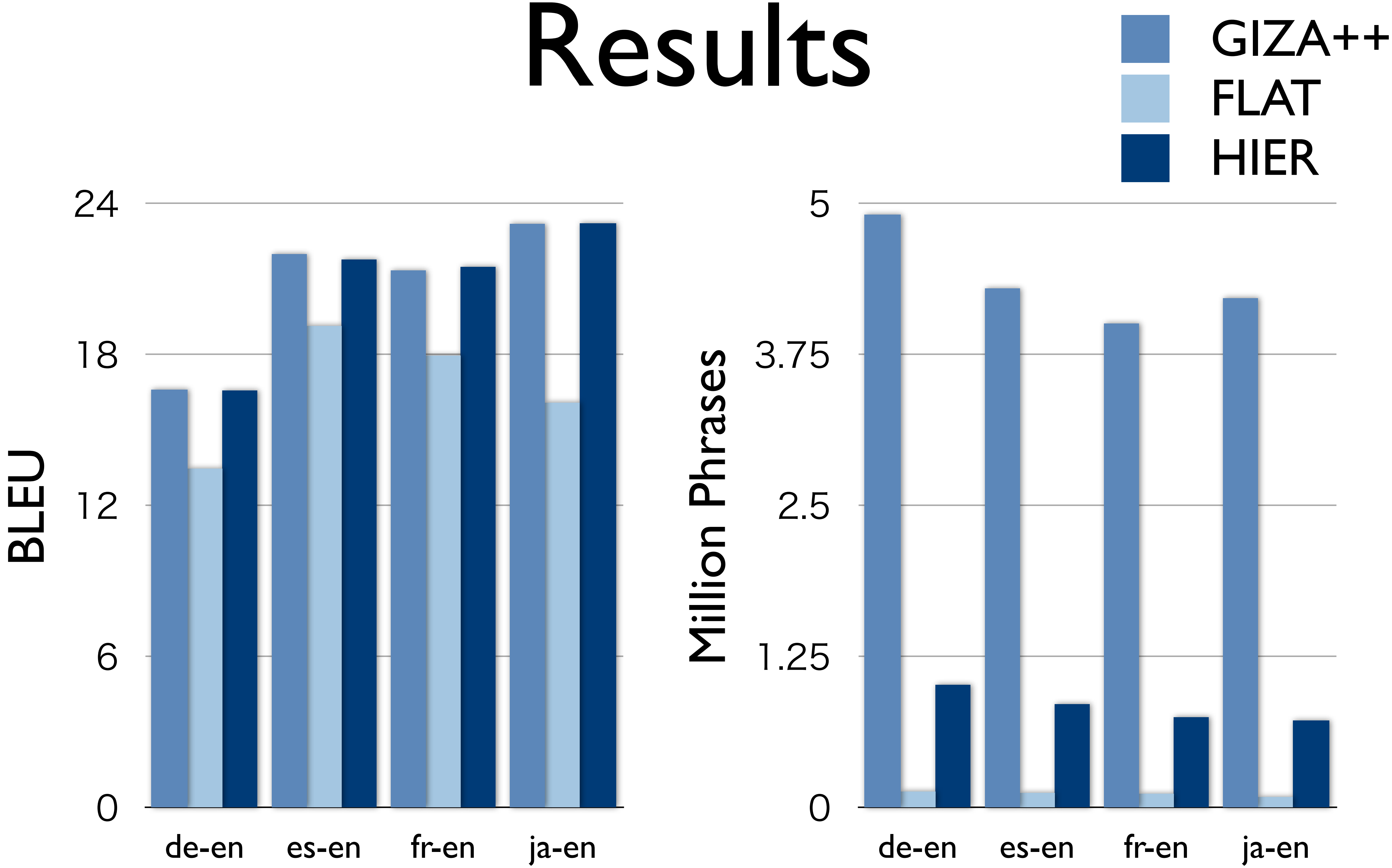
Exhaustive ITG Phrases



Experiments

- WMT10 news-commentary: de-en, es-en, fr-en
- NTCIR-8 patent translation: ja-en
- Moses for decoding:
  - Heuristics (GIZA++)
  - ITG (FLAT)
  - Fallback ITG (HIER)
  - FLAT and HIER employ phrases in the model

Results



GIZA++
FLAT
HIER
BLEU
24
18
12
6
0
Million Phrases
5
3.75
2.5
1.25
0
de-en
es-en
fr-en
ja-en
de-en
es-en
fr-en
ja-en

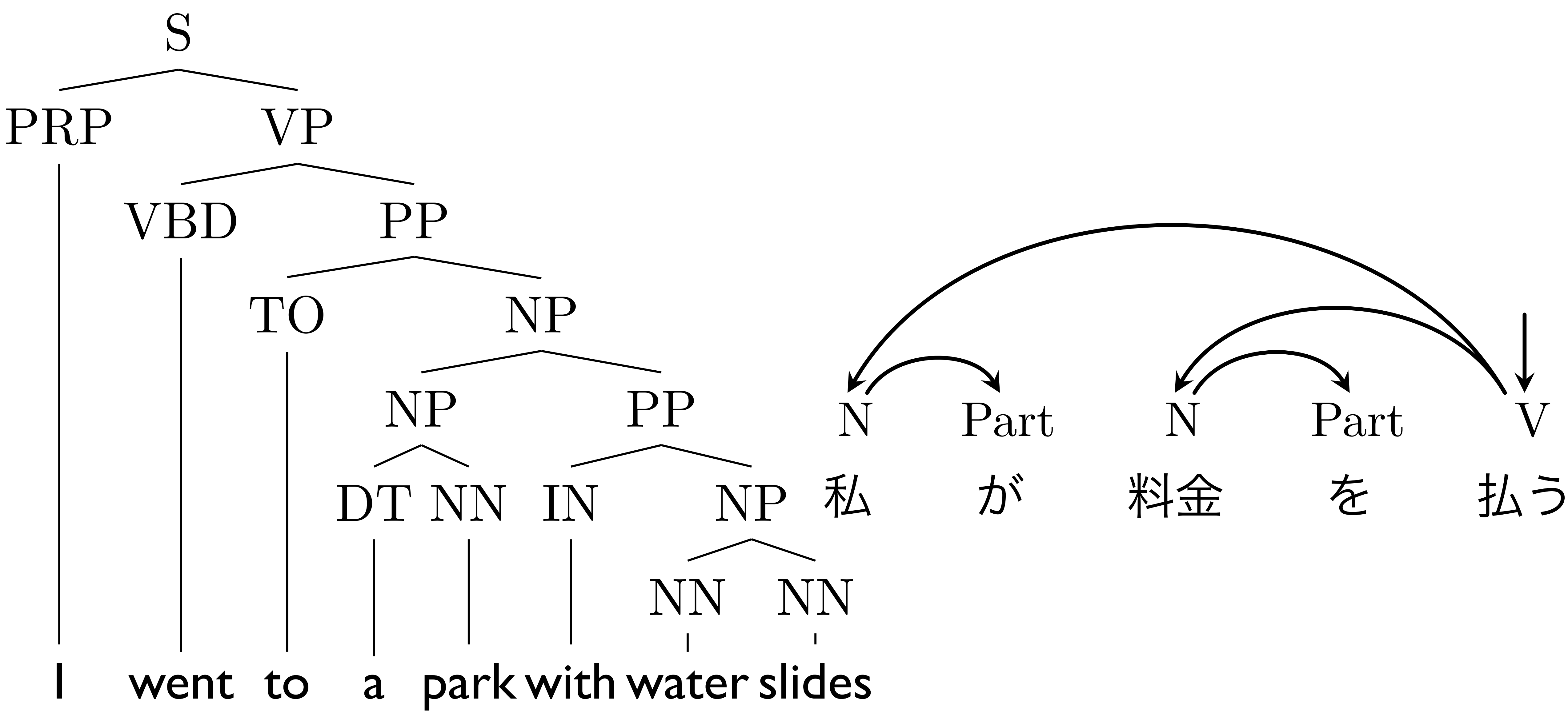
Summary

- Bilingual version of Adaptor Grammars (Johnson et al., 2007)
- Competitive accuracy with GIZA++ baseline with significantly smaller phrase table size
- No more heuristics: a single model captures bilingual relation
- Open sourced: pialign

POS Induction in Dependency Trees

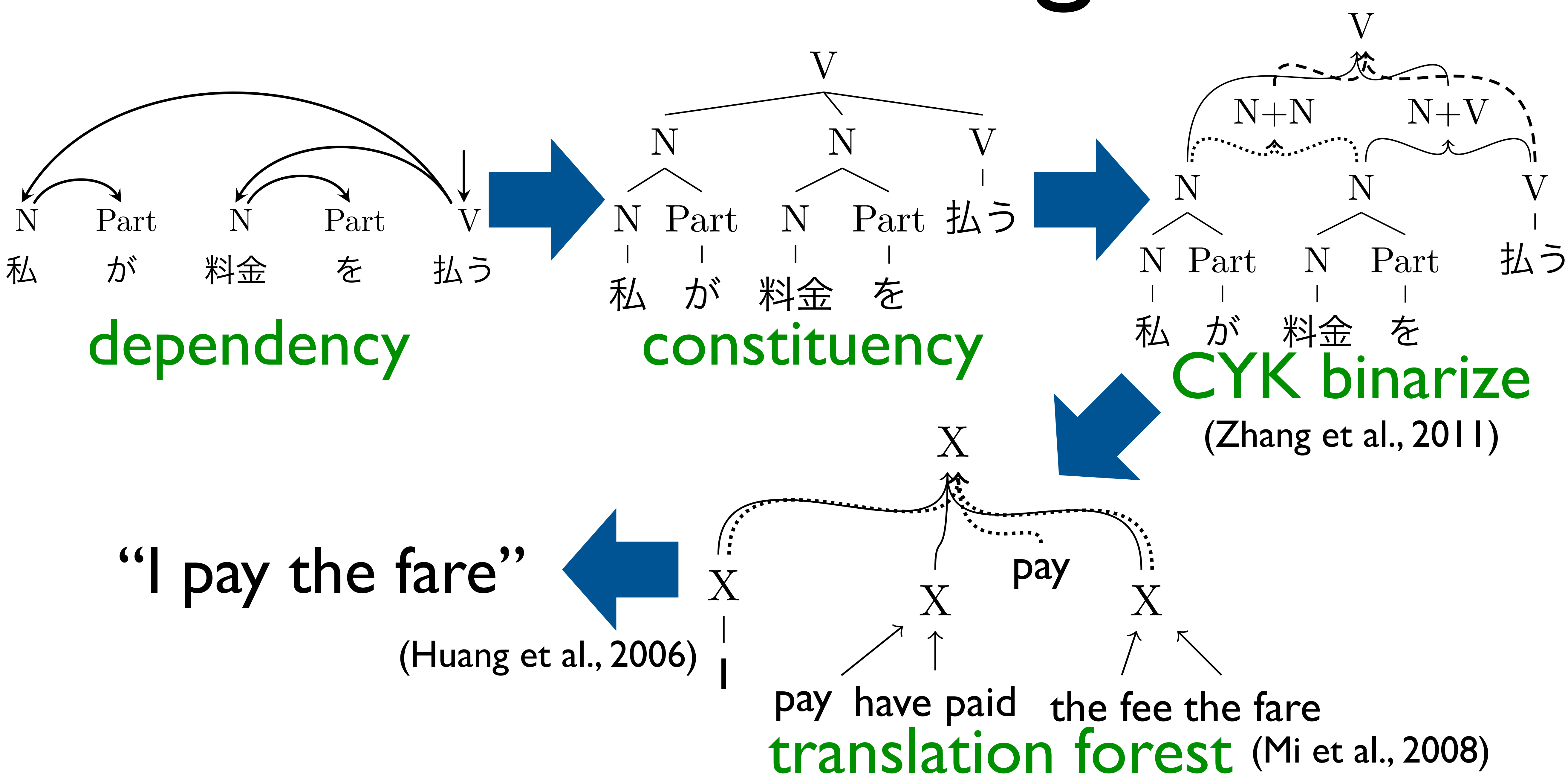
POS Induction in Dependency Trees for SMT
Akihiro Tamura, Taro Watanabe, Eiichiro Sumita, Hiroya Takamura, Manabu Okumura. In *ACL 2013*.

Syntax for MT



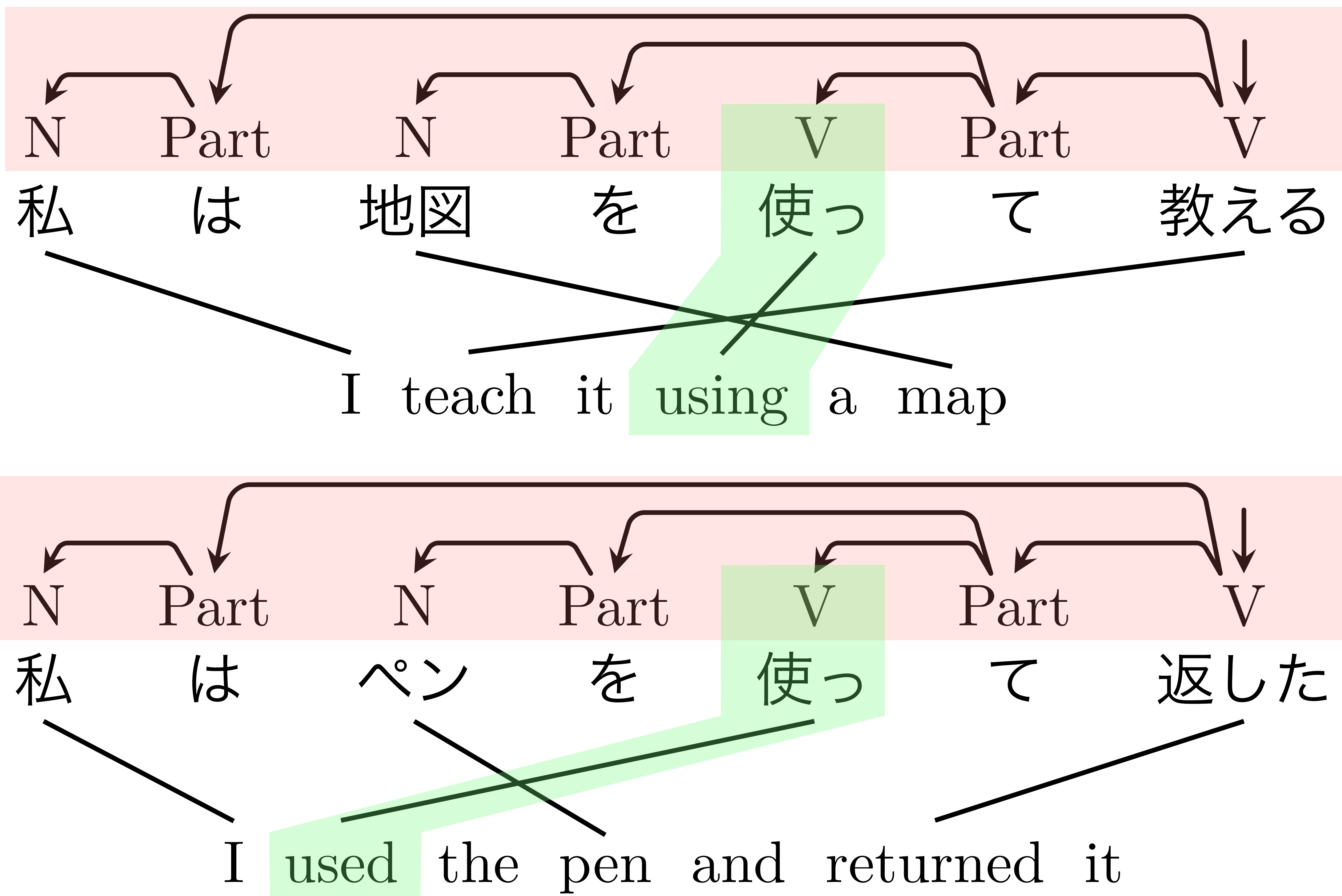
- How syntax can help MT?

Forest-to-String MT



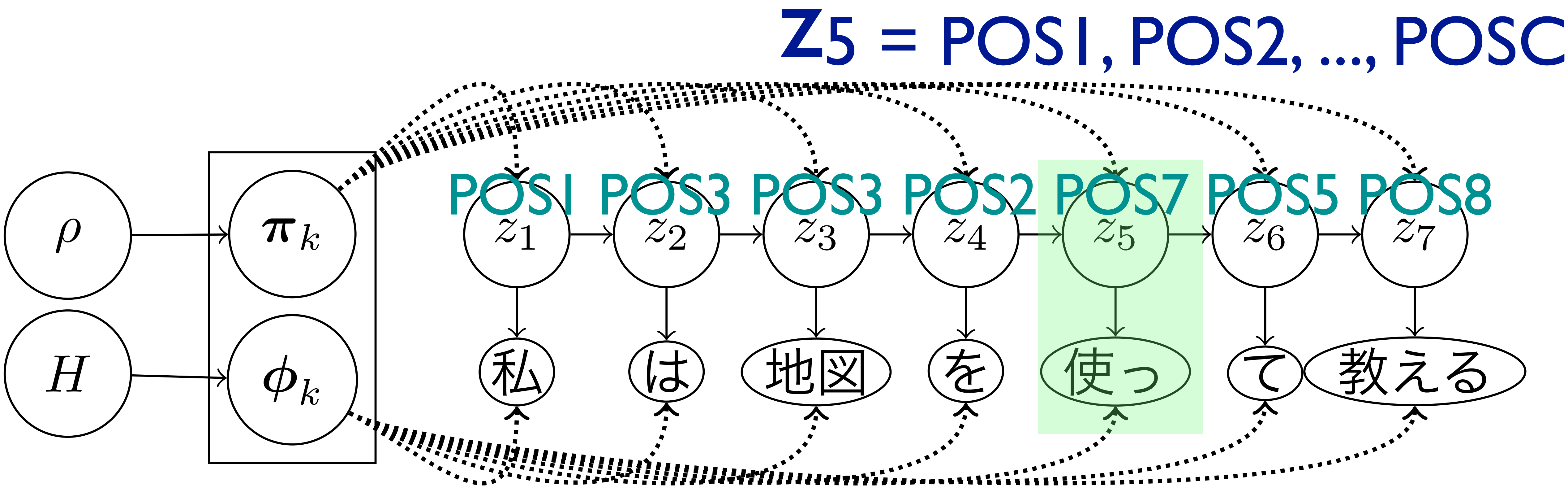
- A very strong syntax based MT: CYK binarized forest-to-string (Zhang et al., 2011)

“Monolingual” Labels



POS Induction

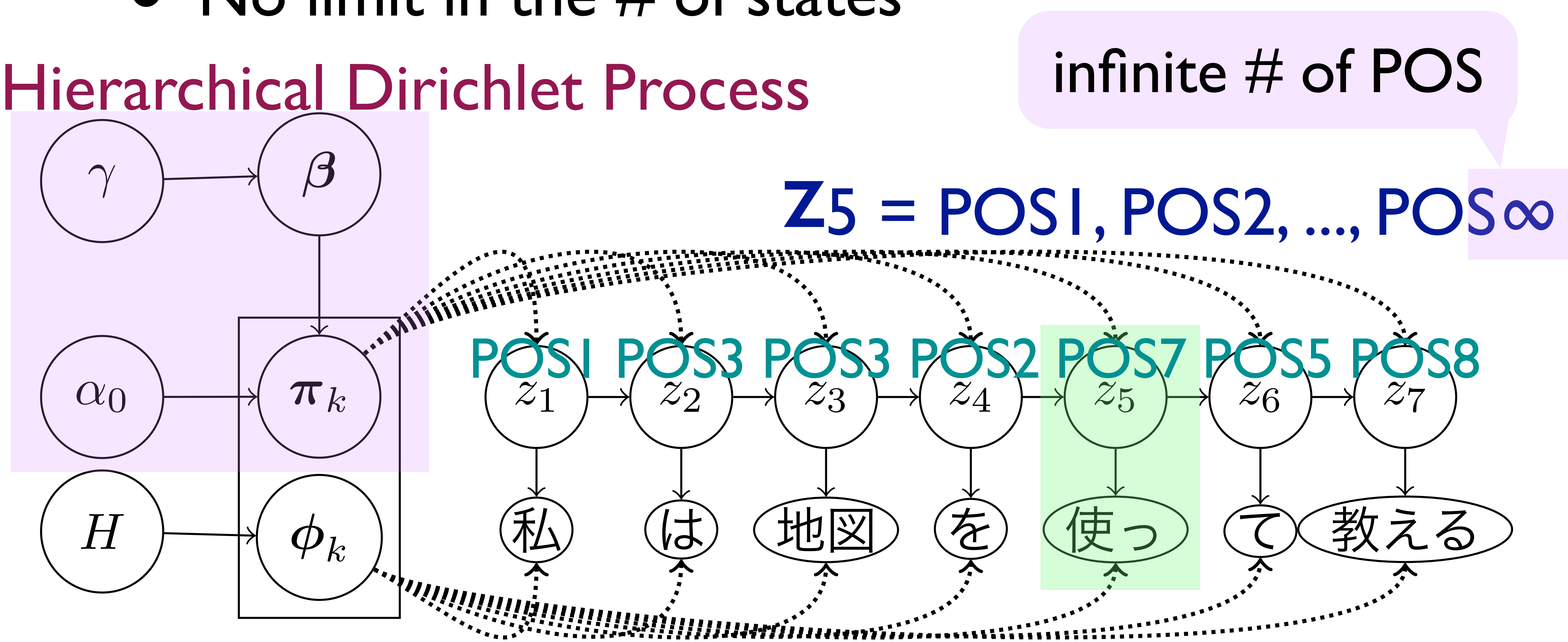
- Hidden states as POS labels



(Gao et al,. 2008)

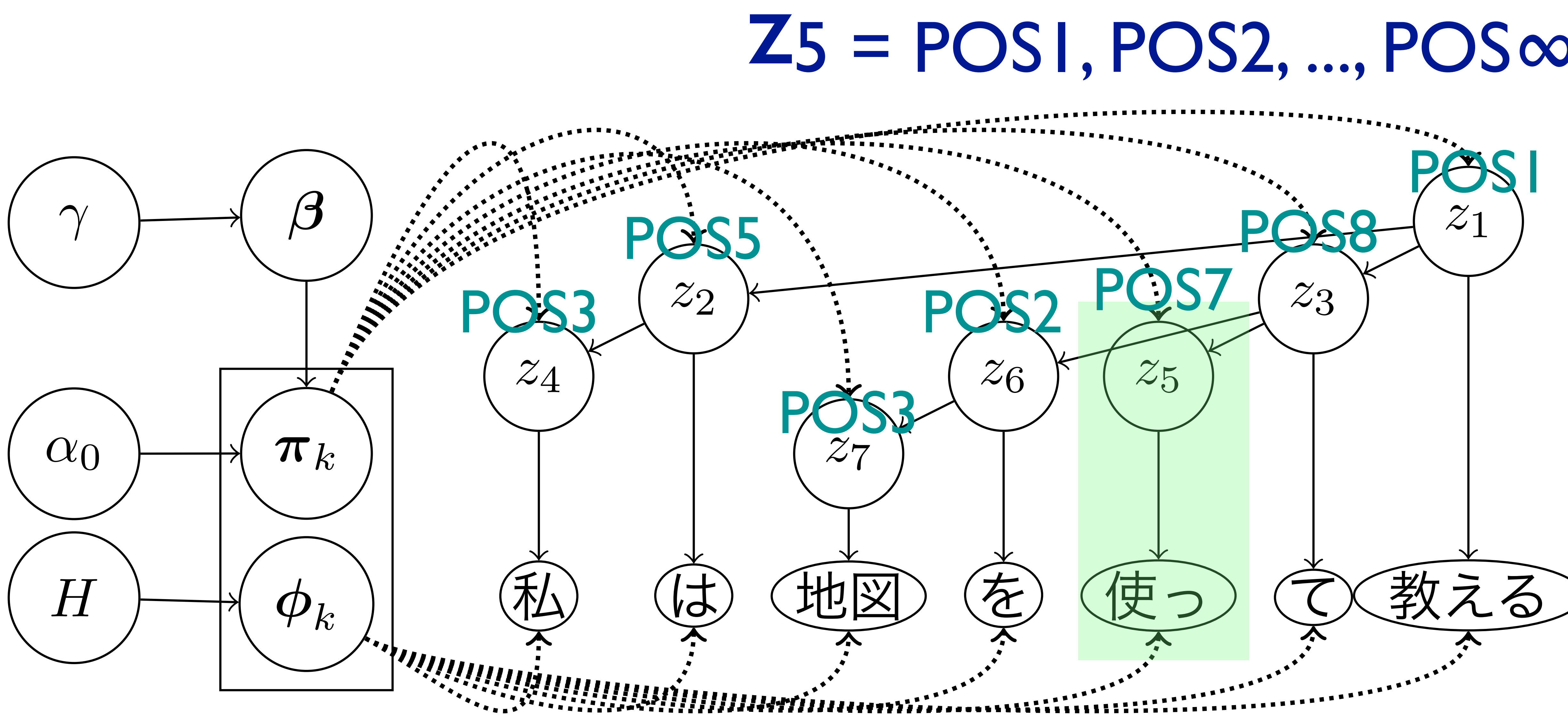
Infinity

- No limit in the # of states



(Gael et al,. 2009)

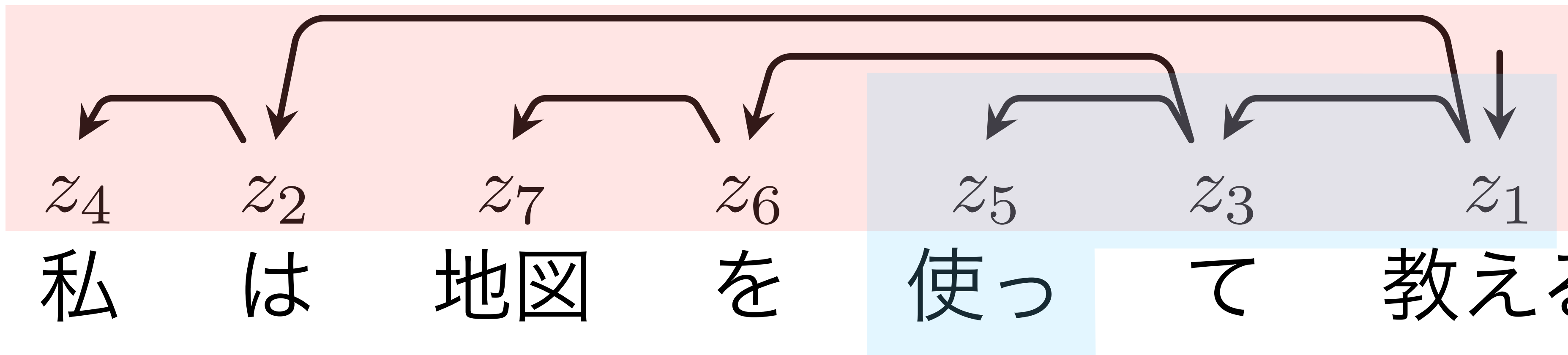
Infinity Over Tree

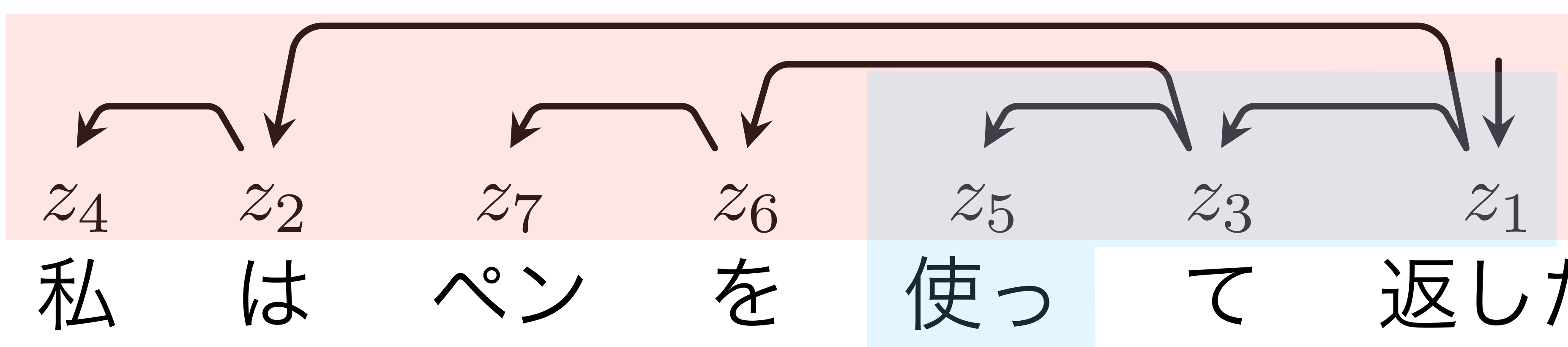


(Finkel et al,. 2007)

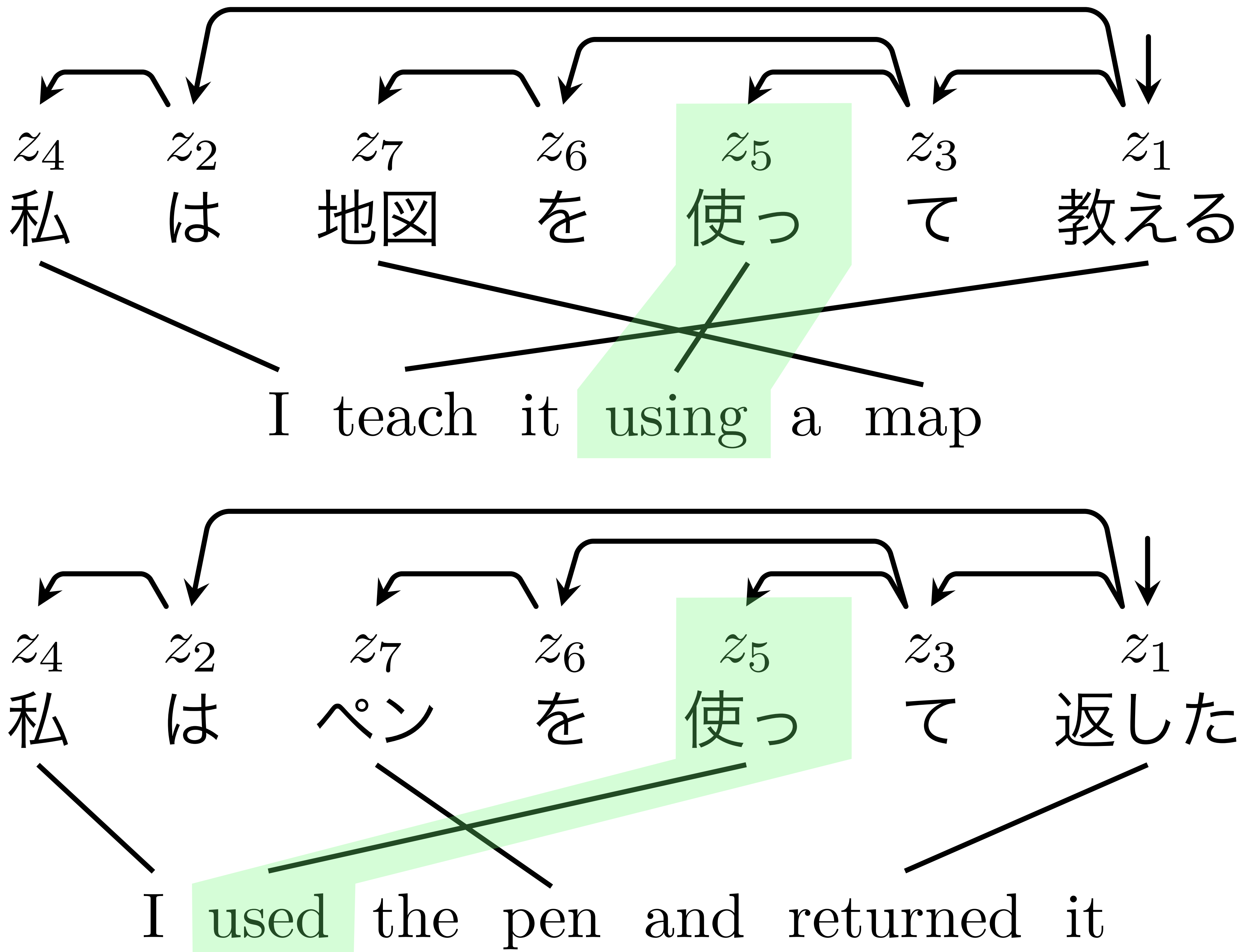
- Instead of HMM, we assume dependency tree is given

Monolingual Induction

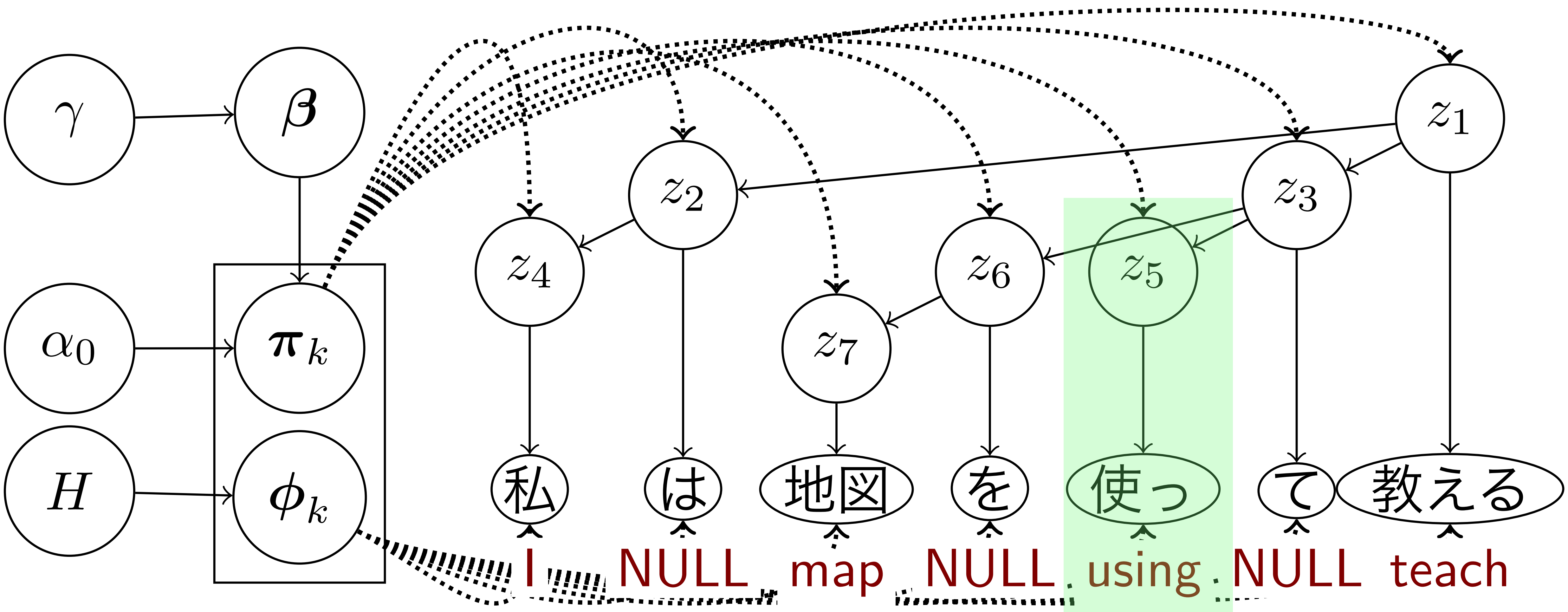




Think Bilingually

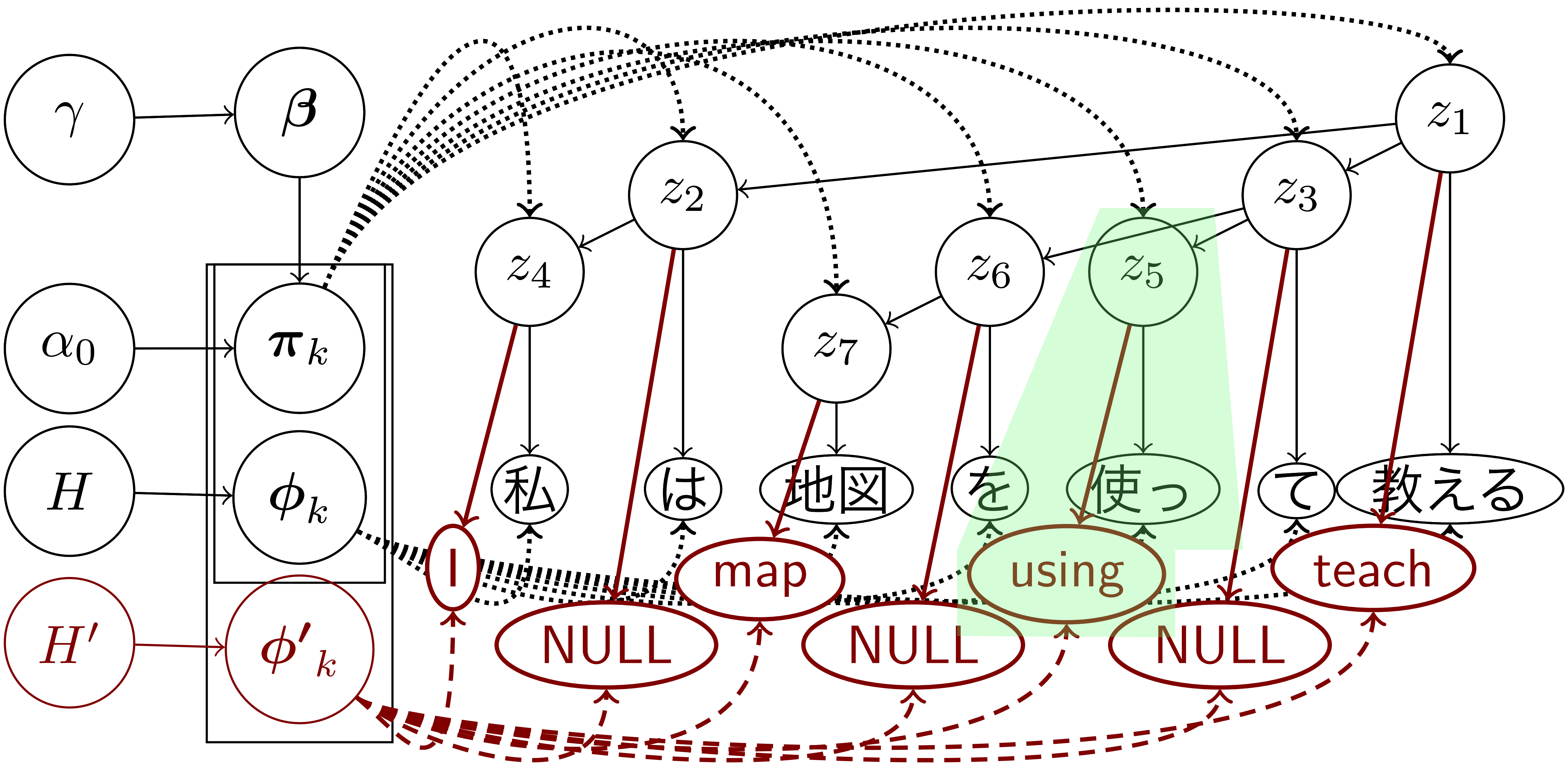


Bilingual Induction



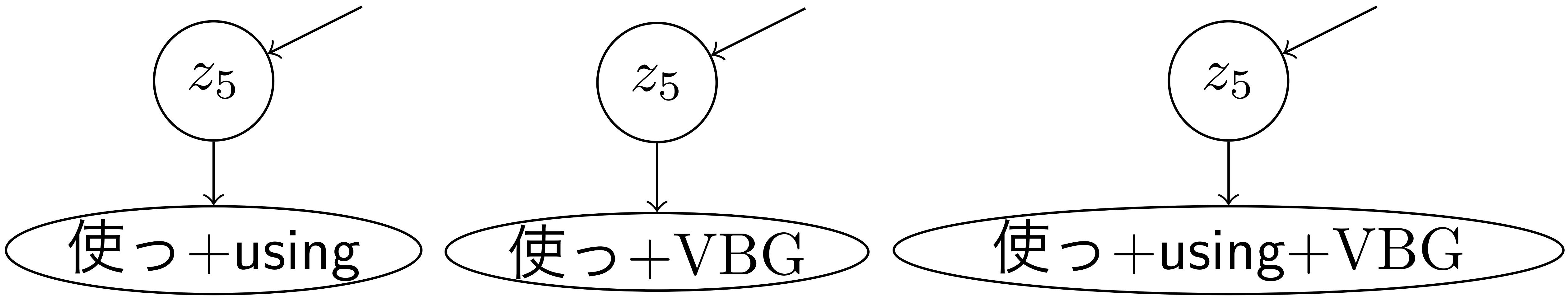
- Jointly emit both of the source and target terminals

Independent Model

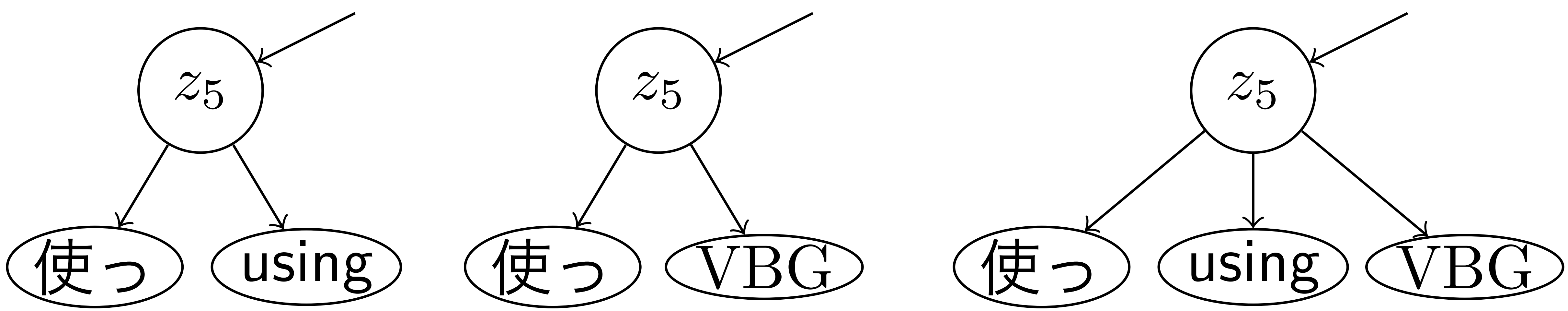


Variation in Emission

Joint Model

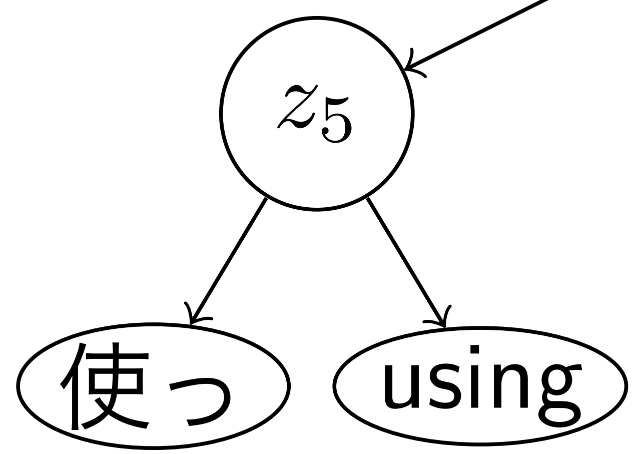


Independent Model



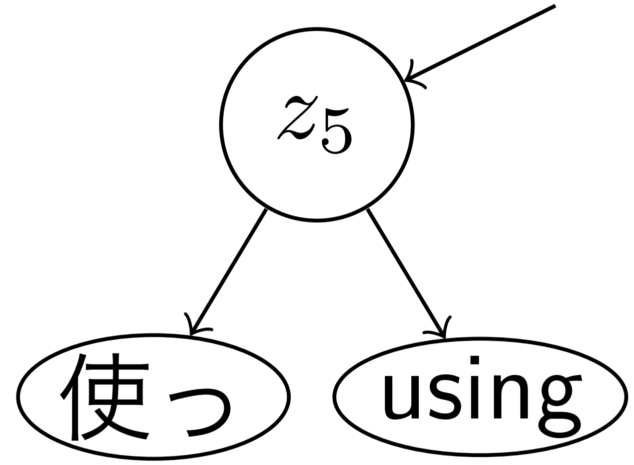
Refinement

Z5 = POS1, POS2, ..., POS∞



The tags, i.e. V, come from the original parse trees

Z5 = V1, V2, ..., V∞



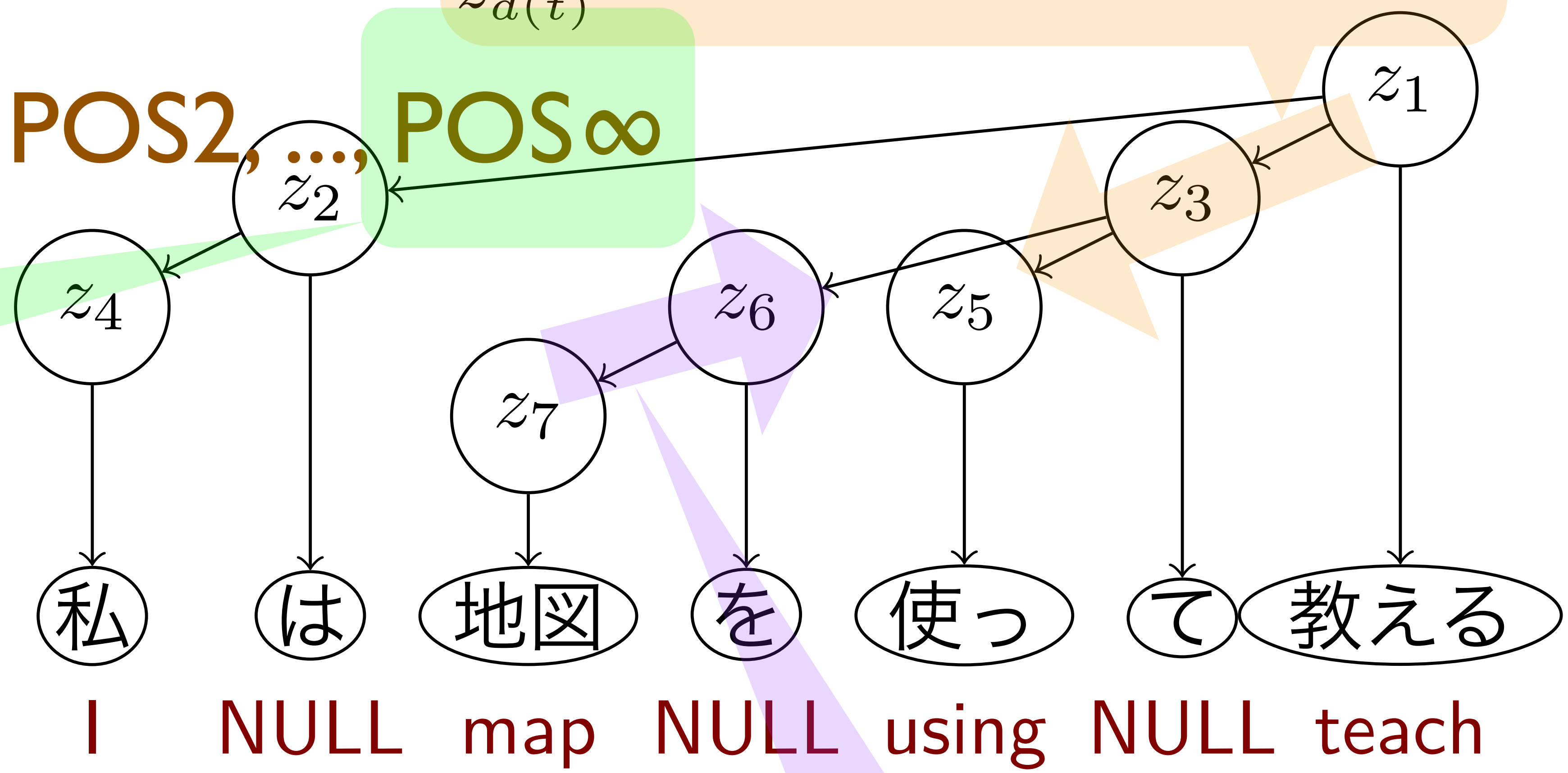
Inference

forward filtering

$$p(z_t|x_{\sigma(t)}, u_{\sigma(t)}) \propto p(x_t|z_t) \sum_{z_{d(t)}} p(z_{d(t)}|x_{\sigma(d(t))}, u_{\sigma(d(t))})$$

z_5 = POS1, POS2, ..., POS∞

how to handle this?



i.e. z_5 = POS7

backward sampling

$$p(z_t|z_{c(t)}, x_{1:T}, u_{1:T}) \propto p(z_t|x_{\sigma(t)}, u_{\sigma(t)}) \prod_{t' \in c(t)} p(z_{t'}|z_t, u_{t'})$$

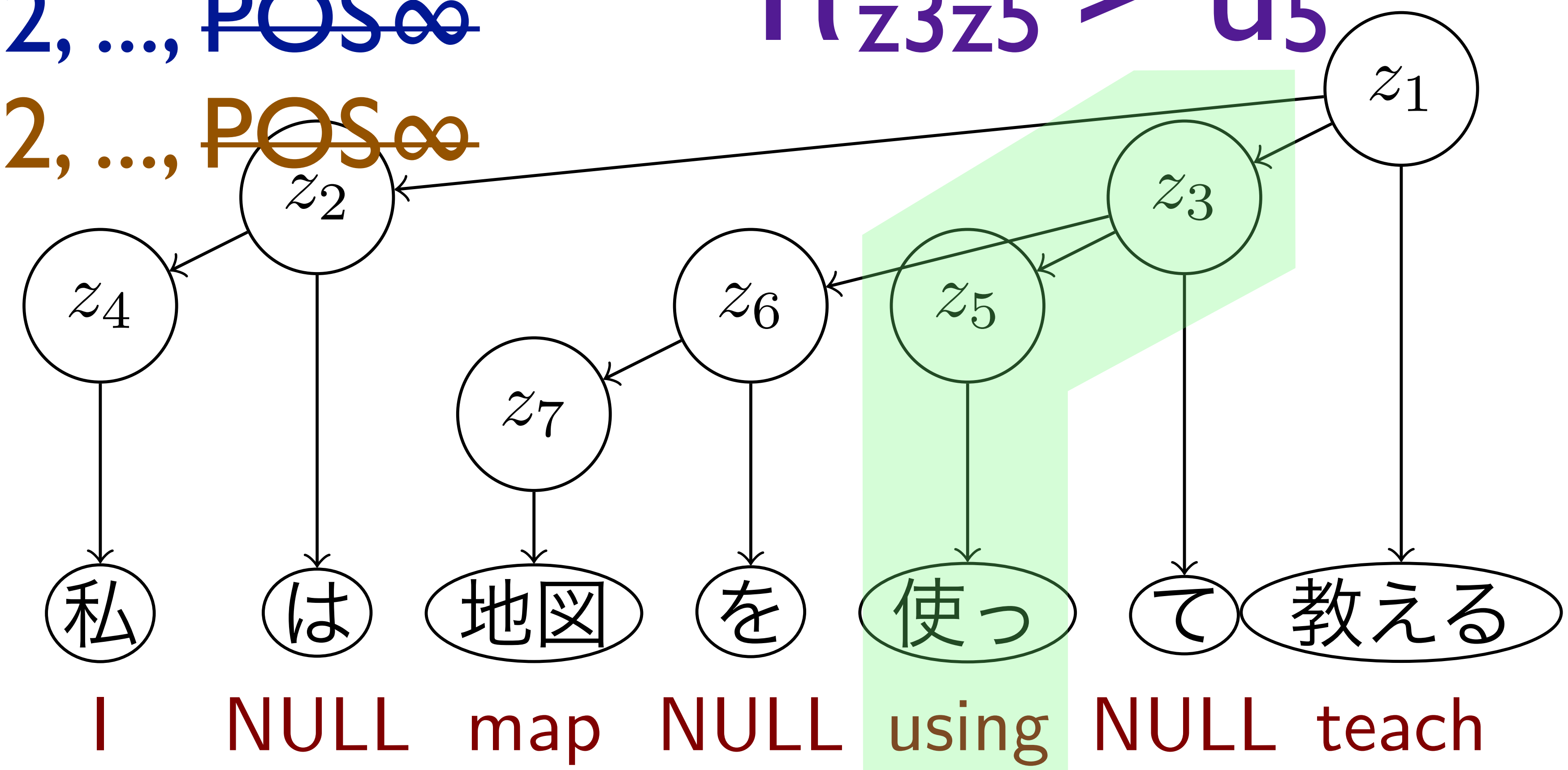
Limit Infinity: Beam Sampling

$$u_t \sim \mathrm{Uniform}(0, \pi_{z_{d(t)} z_t})$$

z_3 = POS1, POS2, ..., ~~POS∞~~

z_5 = POS1, POS2, ..., ~~POS∞~~

$\pi_{z_3 z_5} > u_5$

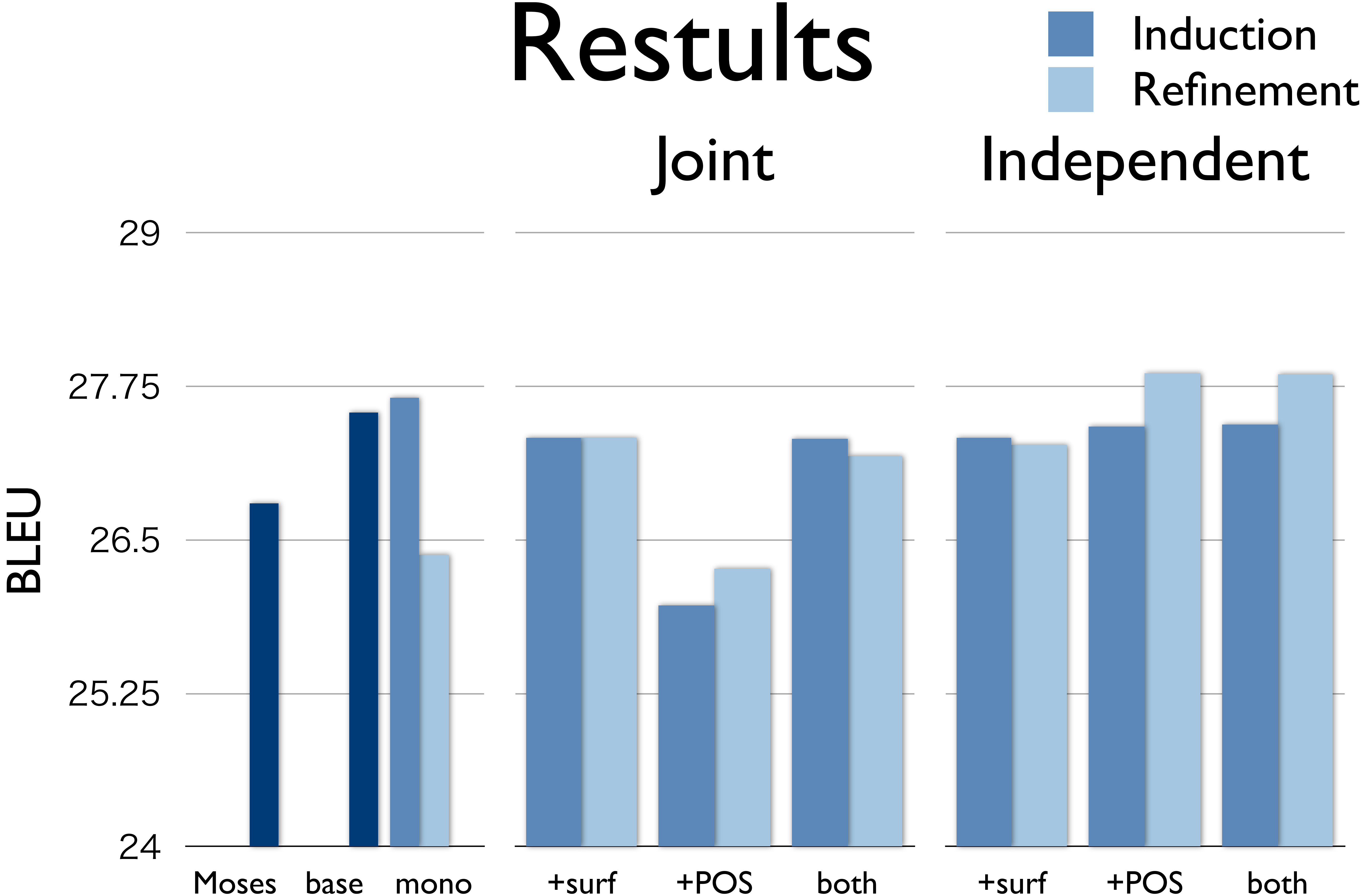


- u_t: an auxiliary variable to limit the infinity
- Choose pairs which satisfies: $\pi_{z_{t'} z_t} > u_t$

Experiments

- NTCIR-9 patent translation: ja-en
- Learn labels on subset (10K)
  1. Tagging/parsing/alignment (MeCab/CaboCha/GIZA++)
  2. Learn new labels
  3. Learn a joint tagger/parser (corbit)
  4. Parse all the data (3M)
- Forest-to-string translation (cicada)

Restults


Induction
Refinement
Joint
Independent
BLEU
29
27.75
26.5
25.25
24
Moses
base
mono
+surf
+POS
both
+surf
+POS
both

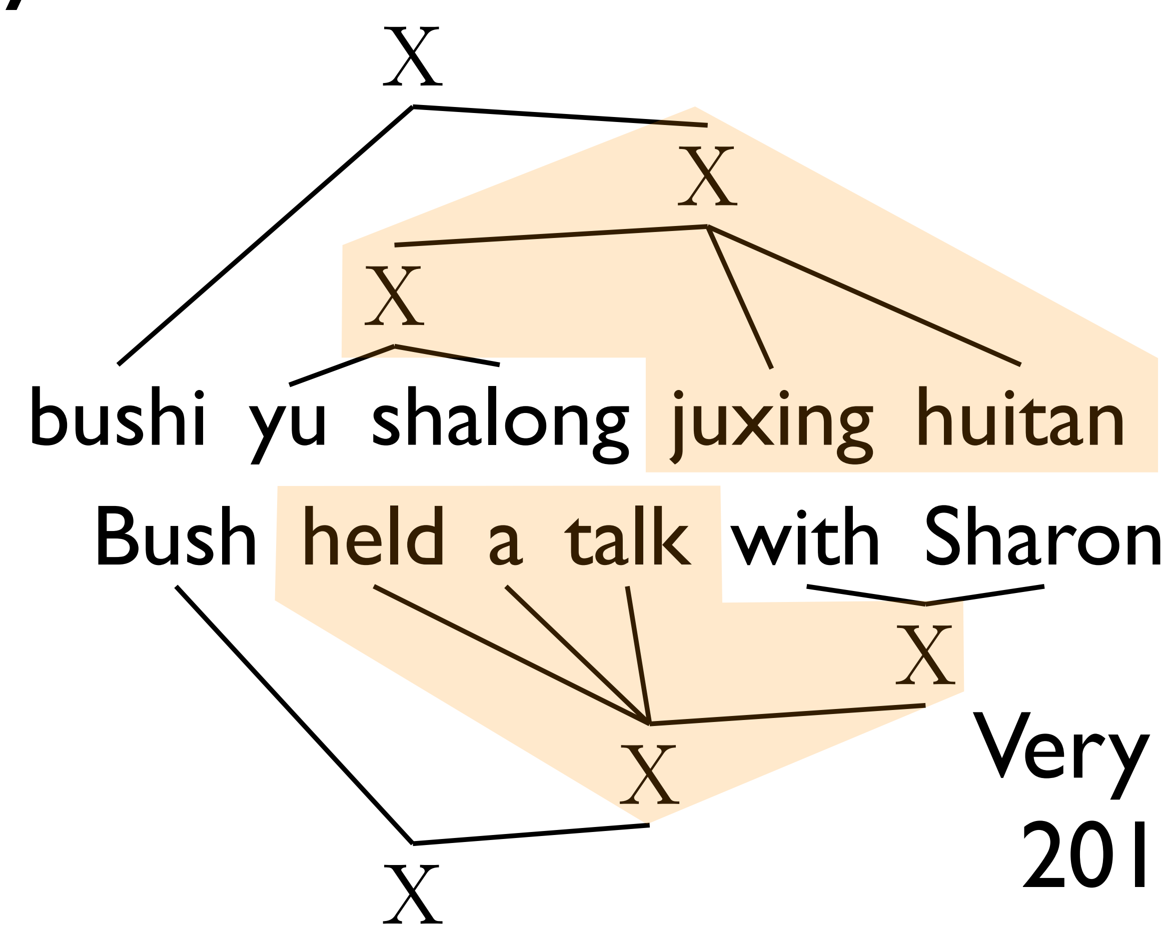
Summary

- Induce labels which consider bilingual relation in the other language
- Improved performance:
    - Better than monolingual induction
    - Independent model alleviates the sparsity problems

Conclusion

Grammar Induction

- Currently, limited to phrase-pairs under ITG
- Future work: more complex arbitrary synchronous-CFG with arbitrary labels



Very good work by (Xiao et al, 2012; Xiao and Xiong, 2013)

Dependency Induction

- Currently, limited to POS given dependency trees
- Future work: jointly induce tree structures + POS