Joint Parsing and Alignment with Weakly Synchronized Grammars

David Burkett, John Blitzer, & Dan Klein

Statistical MT Training Pipeline

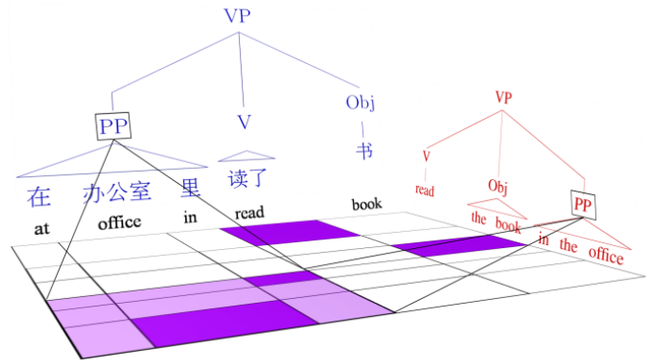
1) Align sentence pairs (GIZA++)

2) Parse English sentences (Berkeley parser)
   Parse Foreign sentences

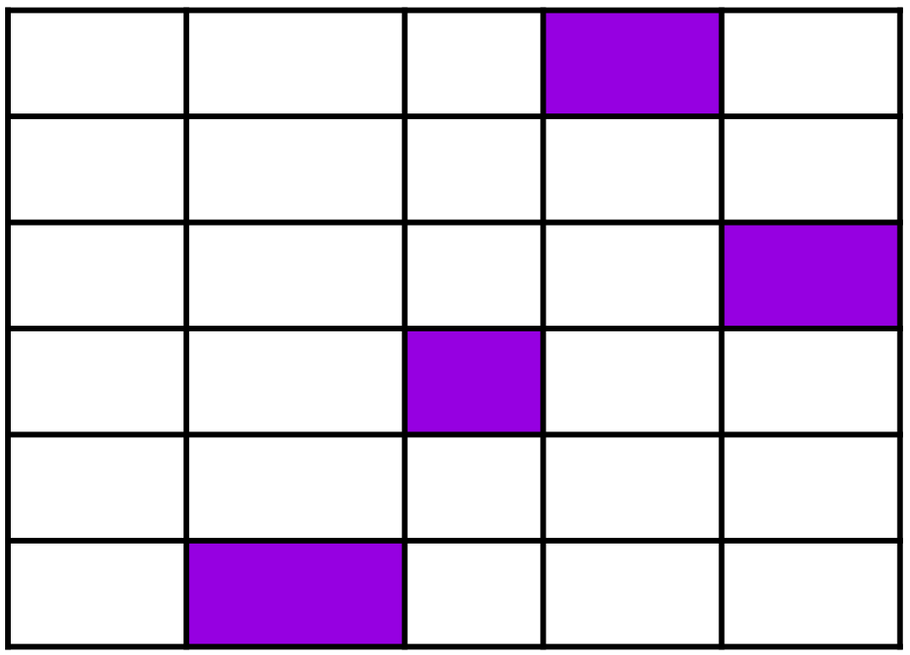
3) Extract rules (Galley et al. 2006)

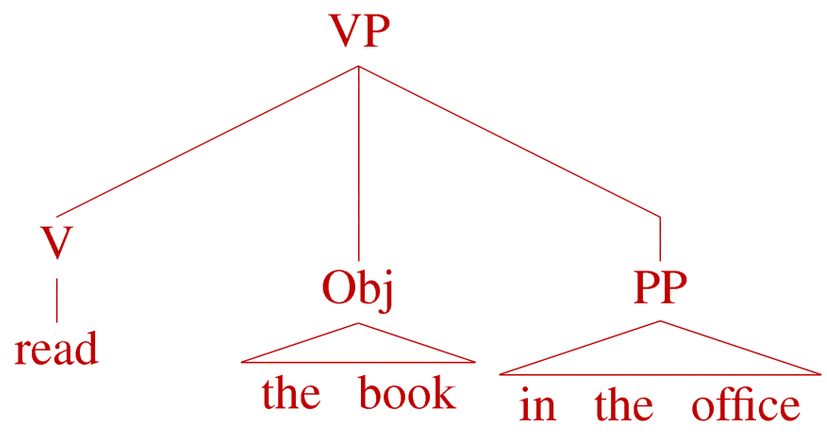
4) Tune discriminative parameters

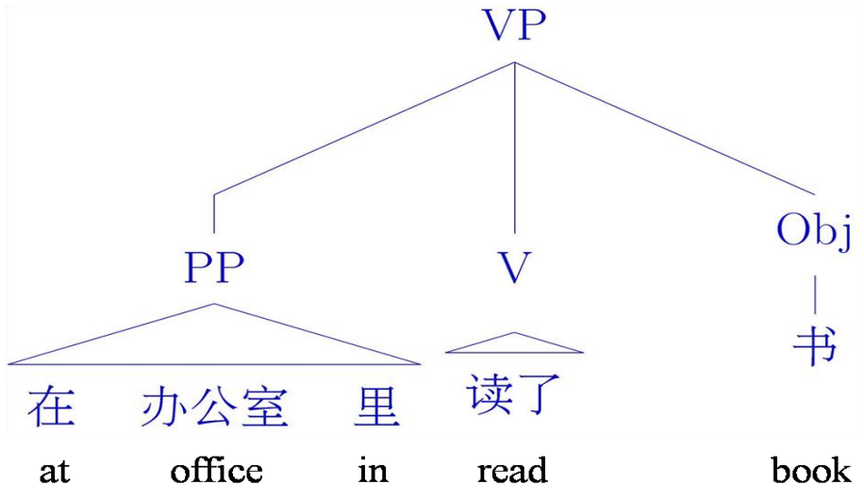
Joint model for (1) & (2)



在	办公室	里	读了	书
at	office	in	read	book



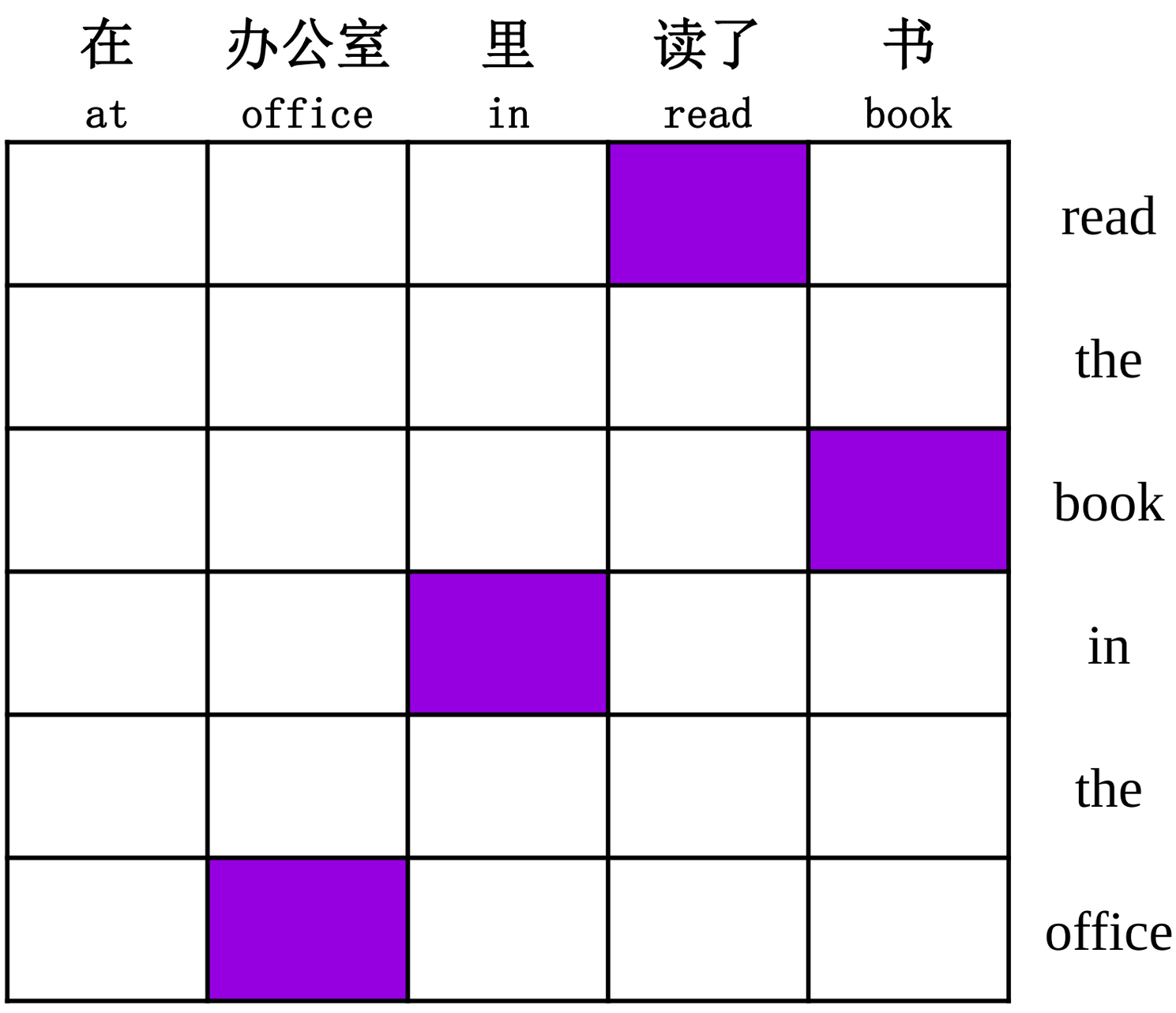




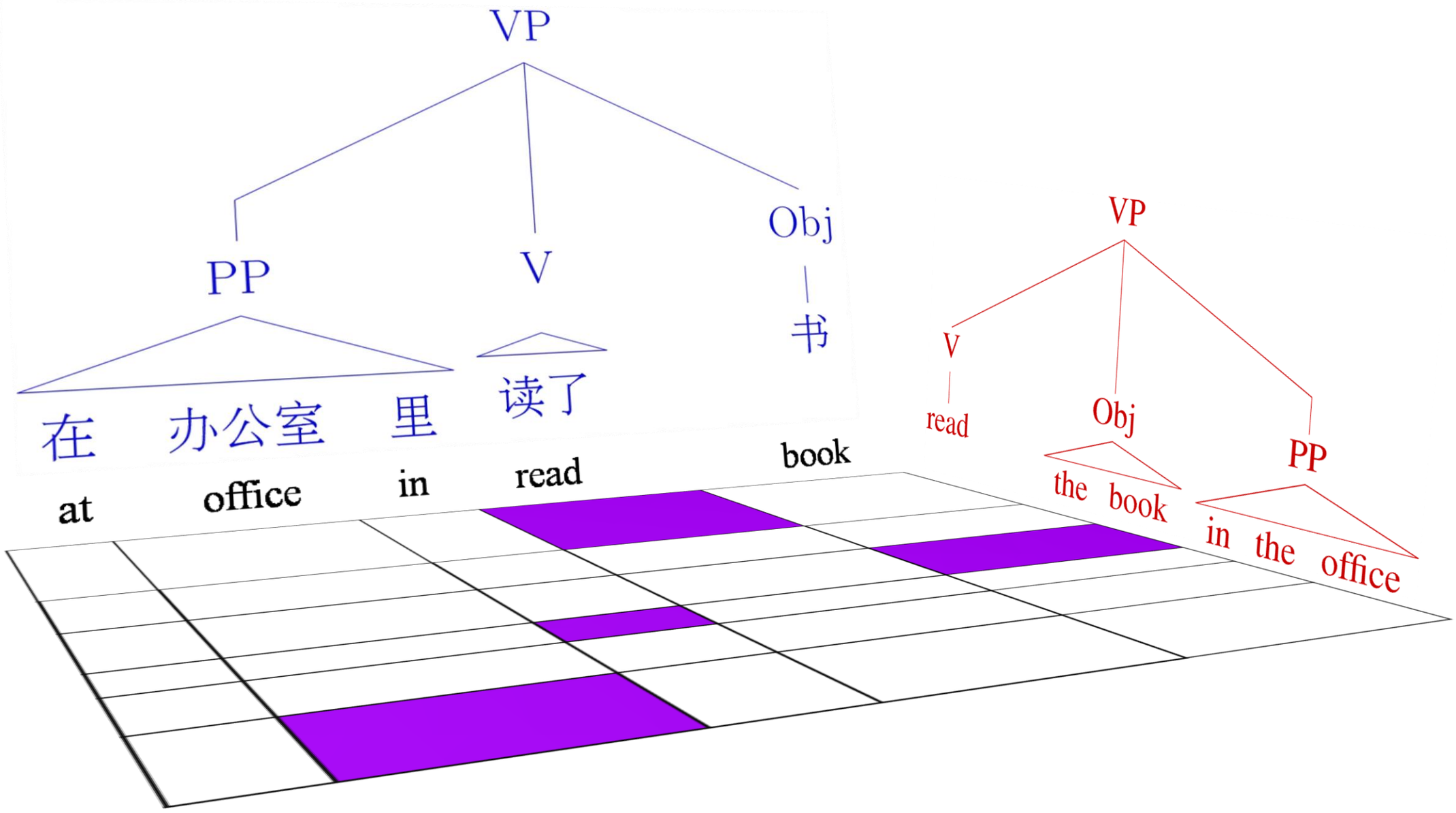
Data Setting for Joint Models

English WSJ	Chinese CTB	Unlabeled parallel text	Parallel, Aligned CTB
(EN;▲)	(中文;▲)	(EN; 中文)	(EN,中文;▲▲▦)
(EN;▲)	(中文;▲)	(EN; 中文)	(EN,中文;▲▲▦)
.	.	.	.
.	.	.	.
.	.	.	.
(EN;▲)	(中文;▲)	(EN; 中文)	(EN,中文;▲▲▦)

Word alignment grids

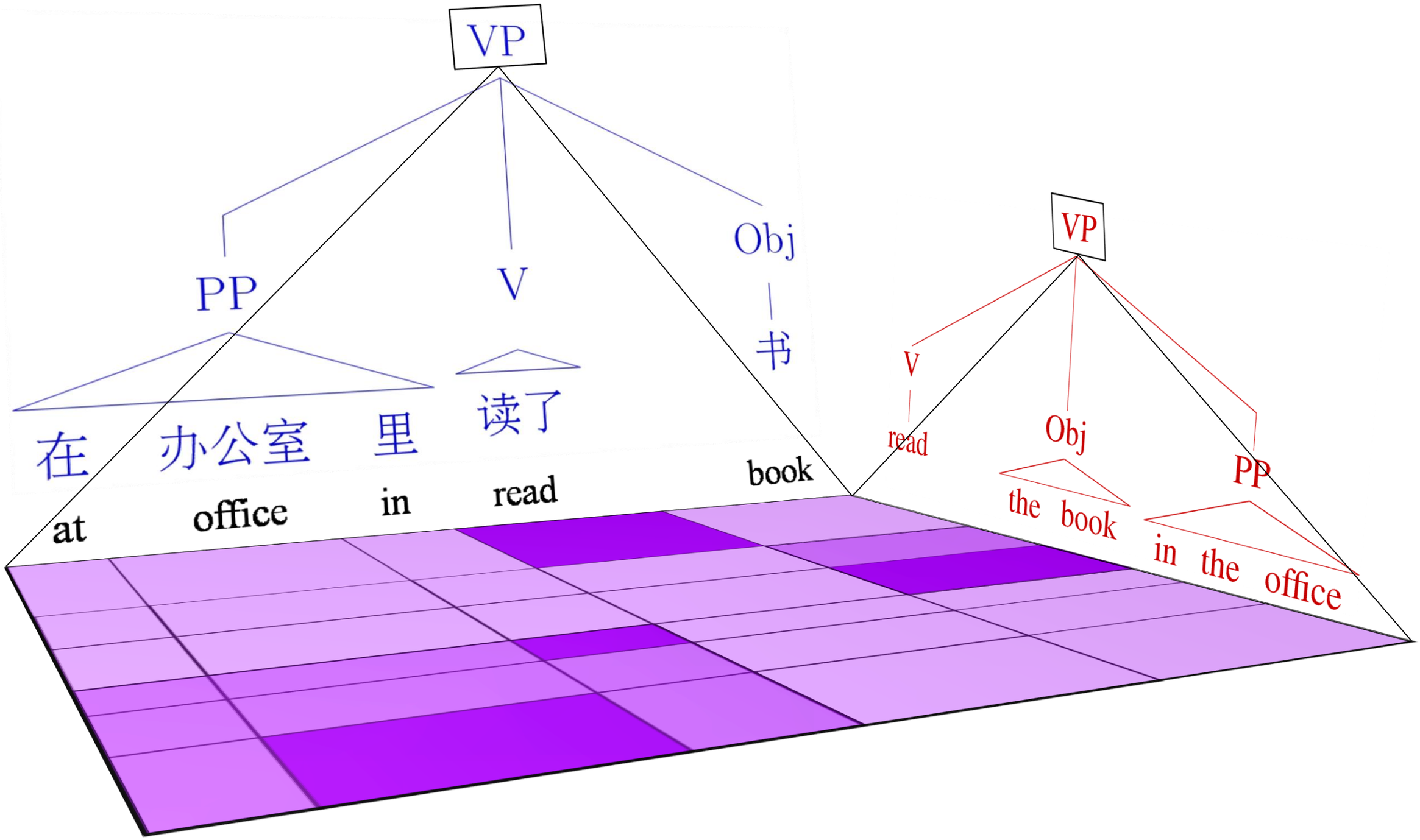


Syntactic Correspondences

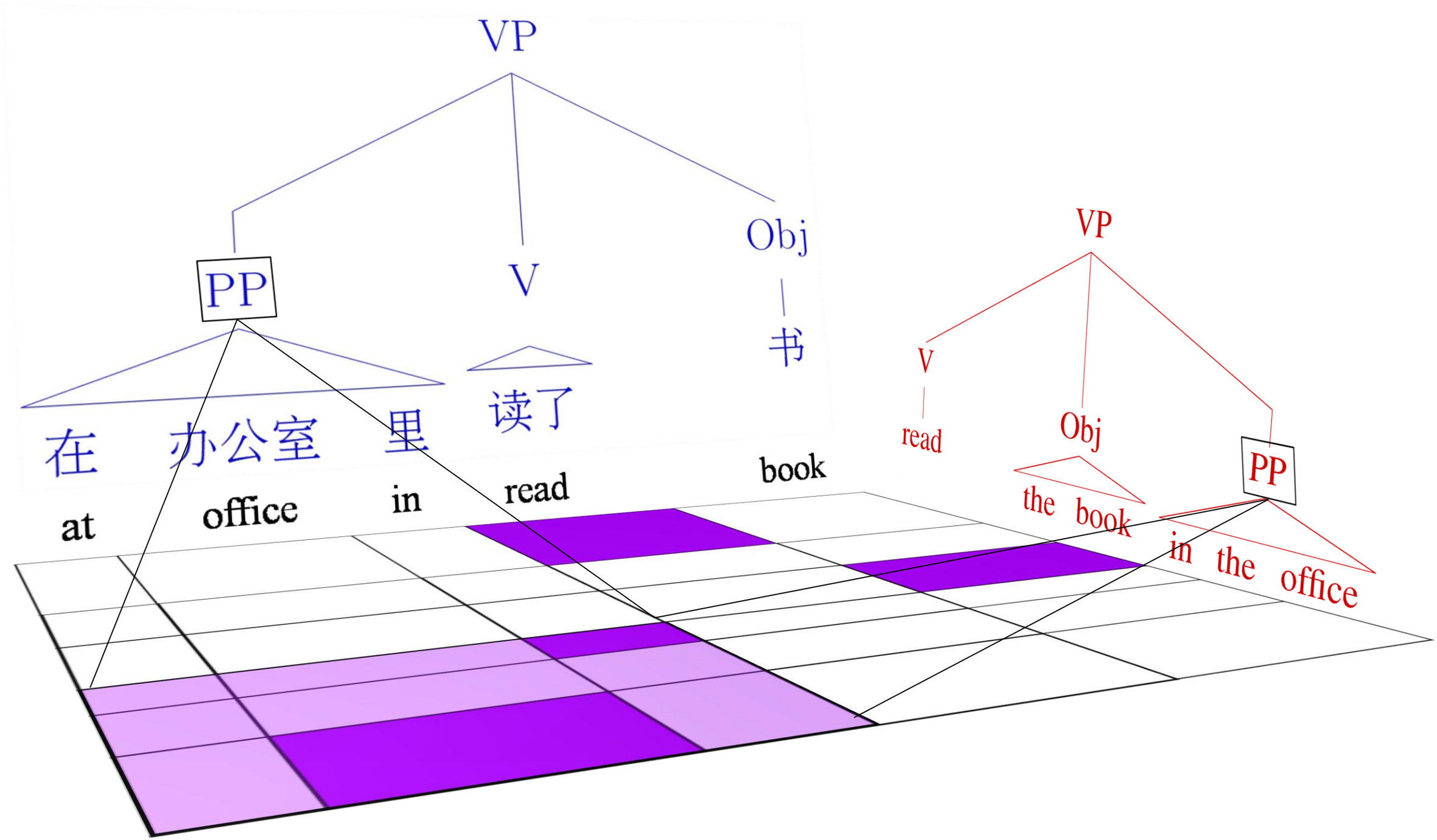


Build a model p_θ (▲, ▲, ▦ | 中文, EN)

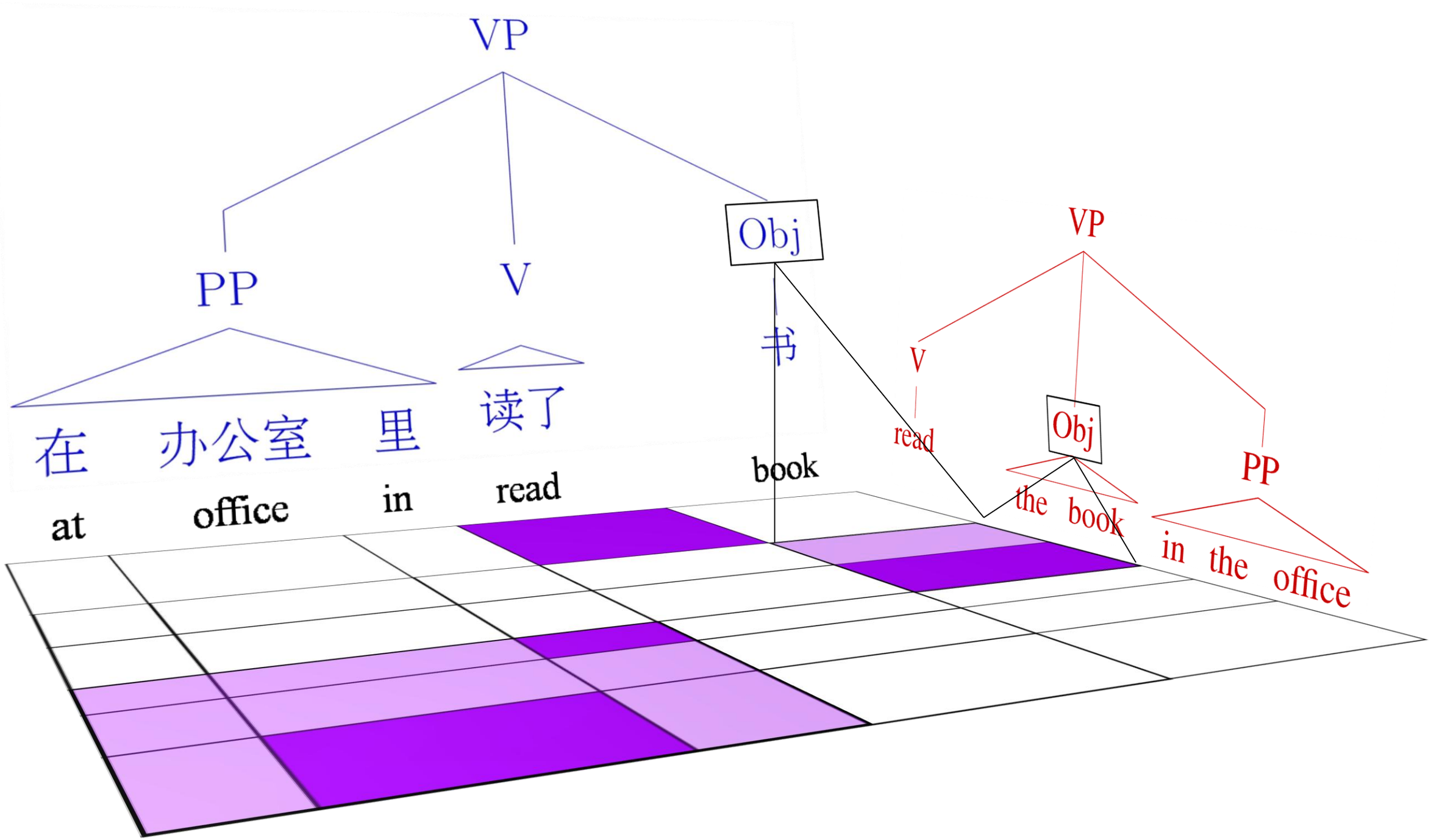
Correspondence via Synchronous Grammars



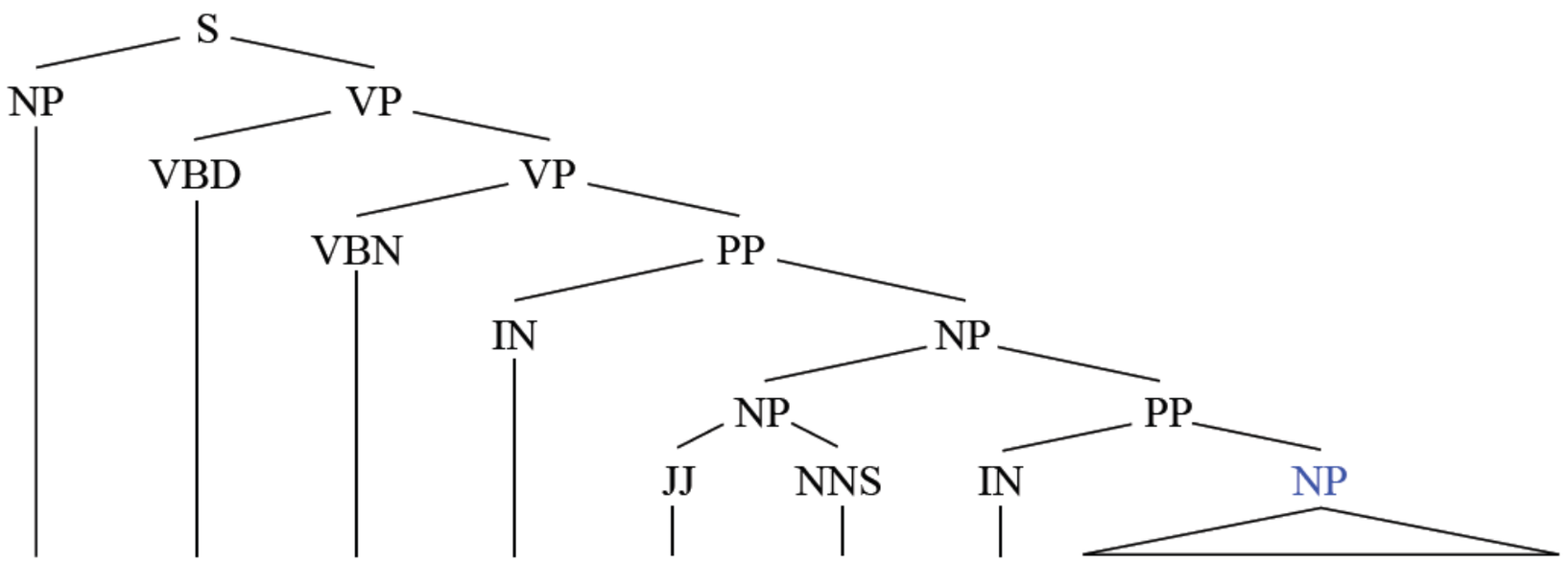
Synchronous derivation



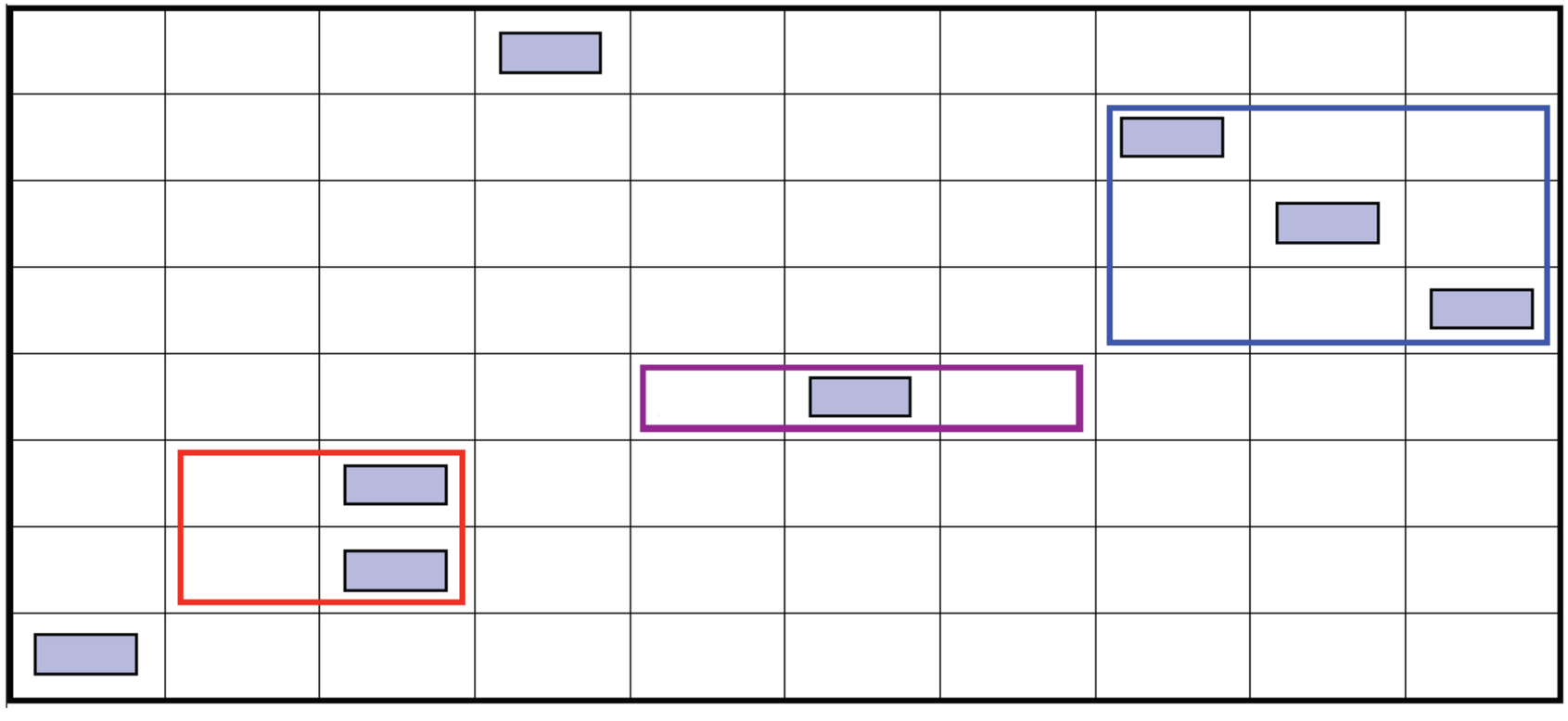
Synchronous Derivation



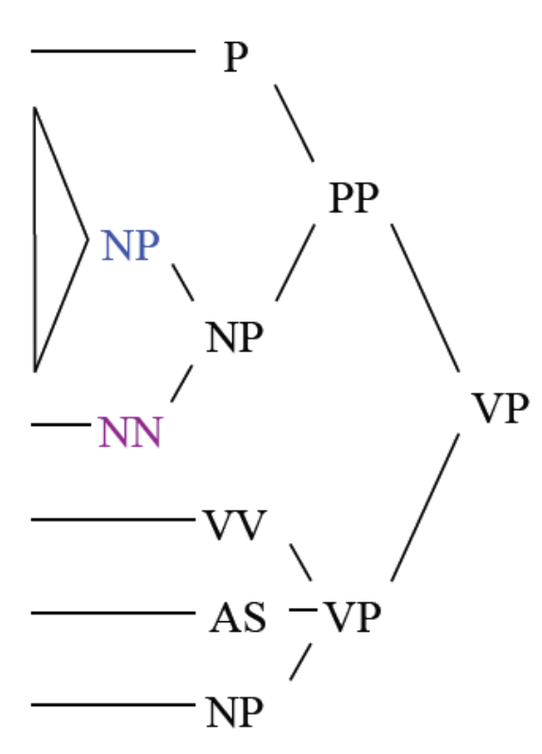
Weakly Synchronized Example



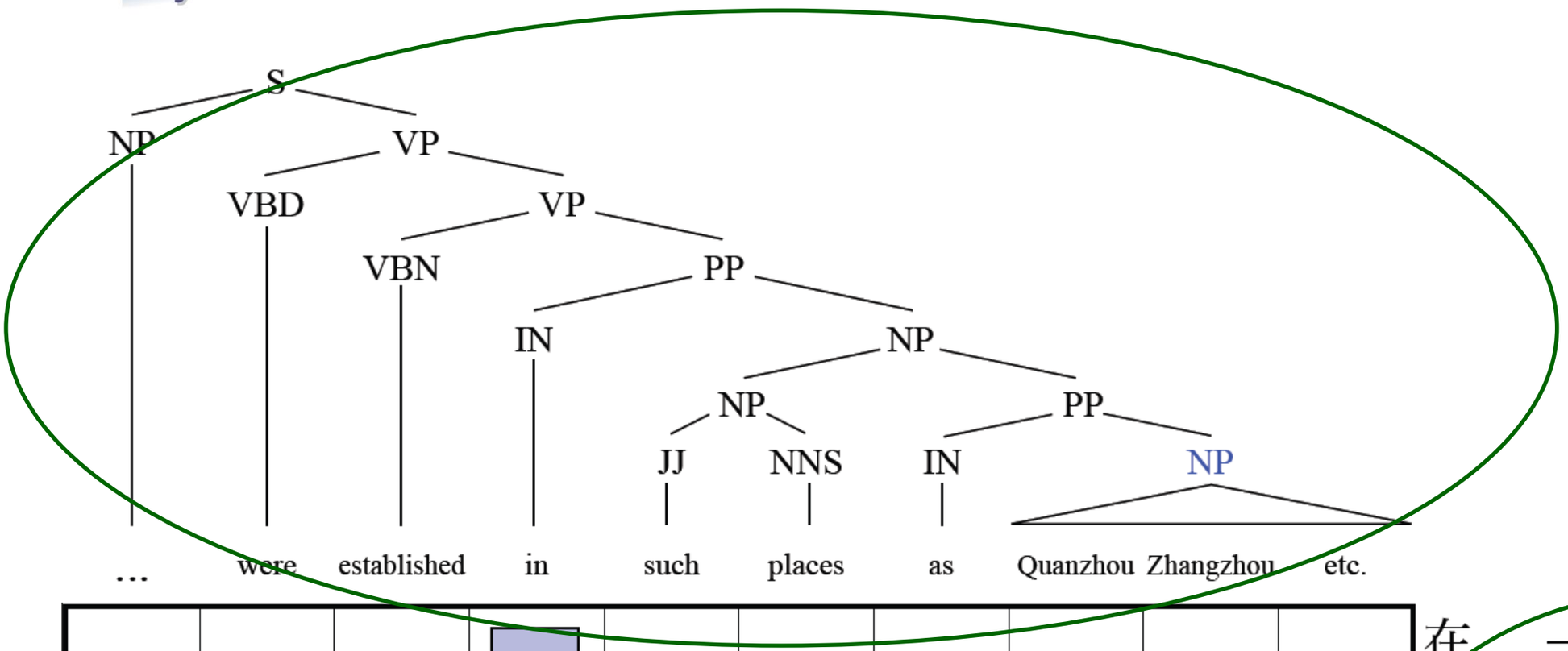
… were established in such places as Quanzhou Zhangzhou etc.



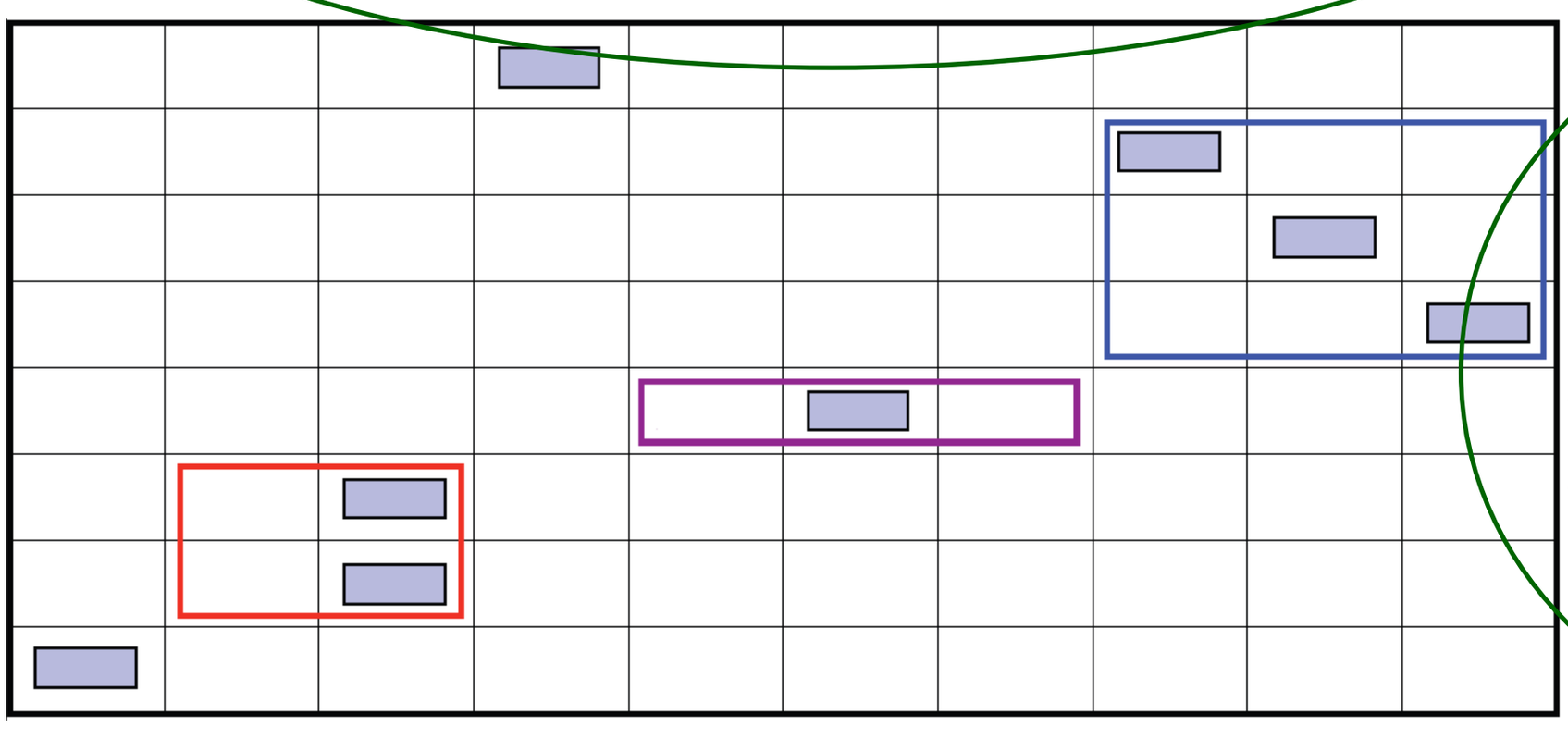
在 泉州 漳州 等 地 设立 了 …

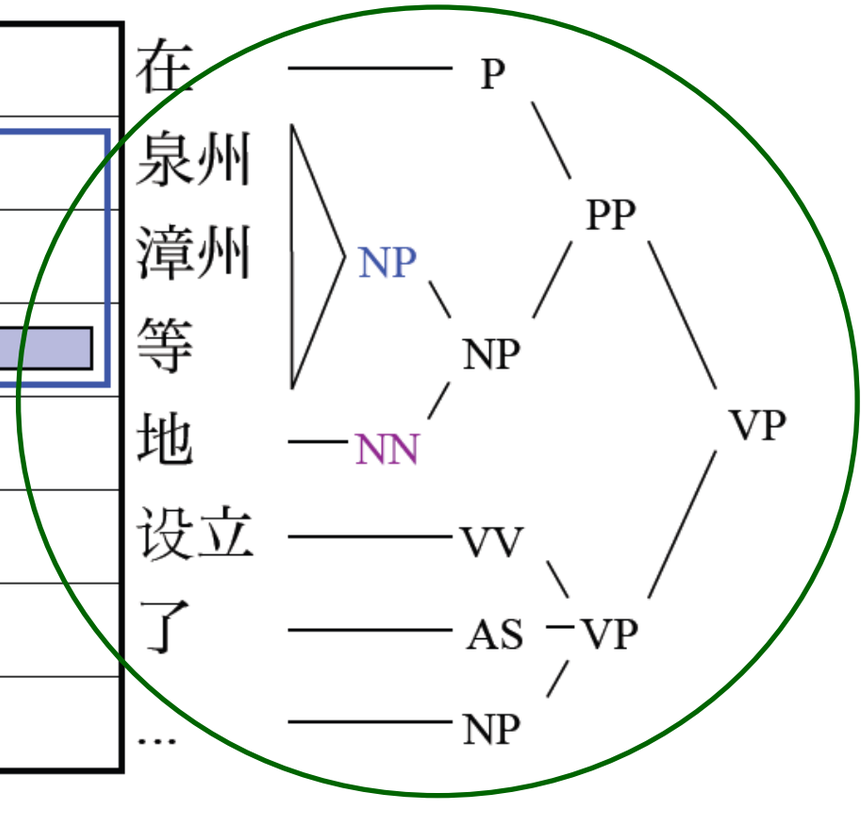


Weakly Synchronized Example

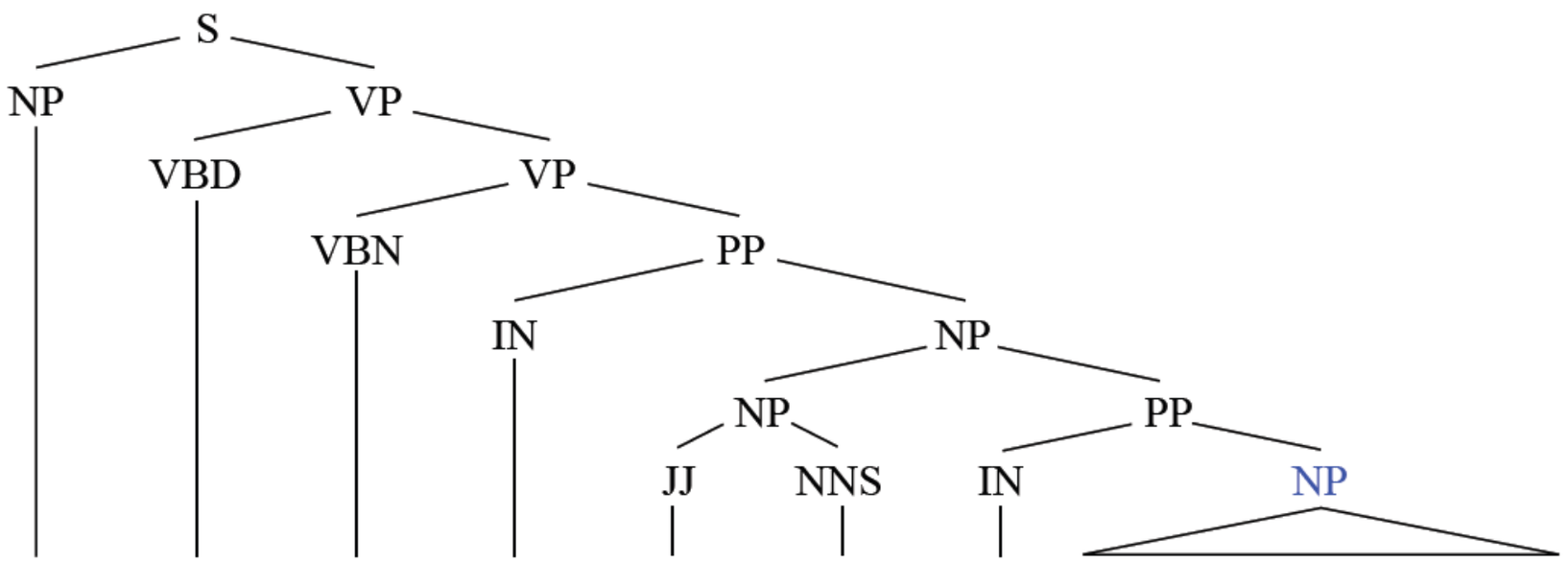


Separate PCFGs

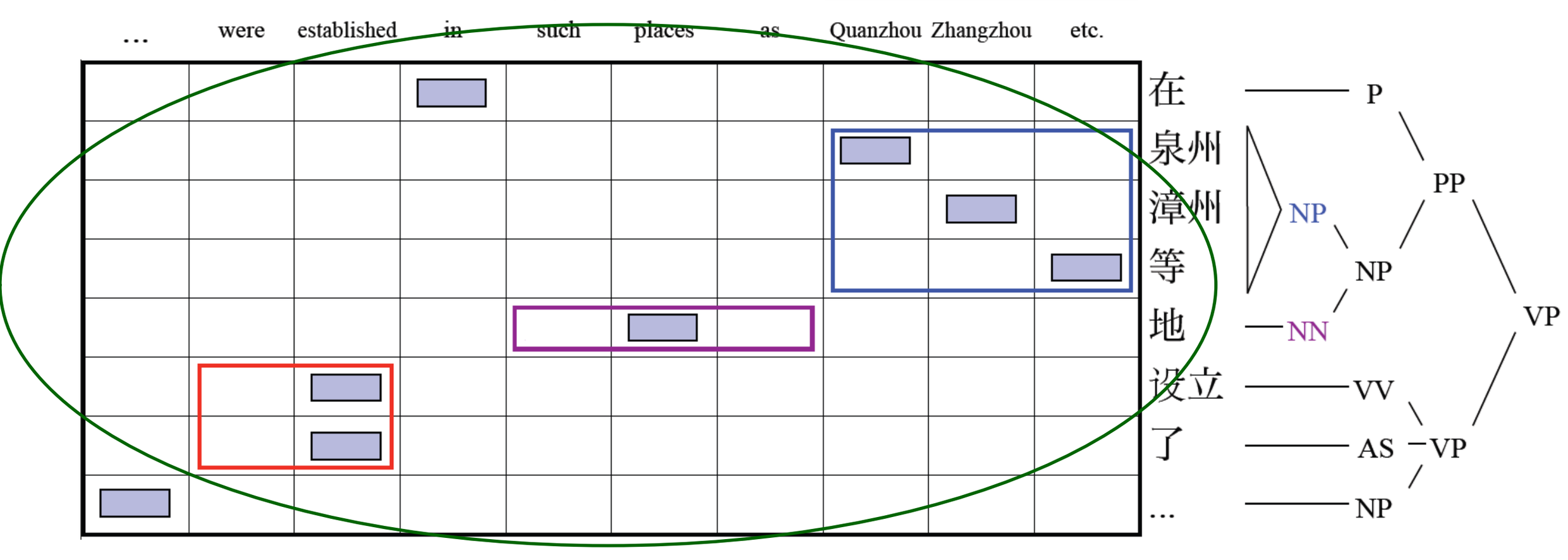




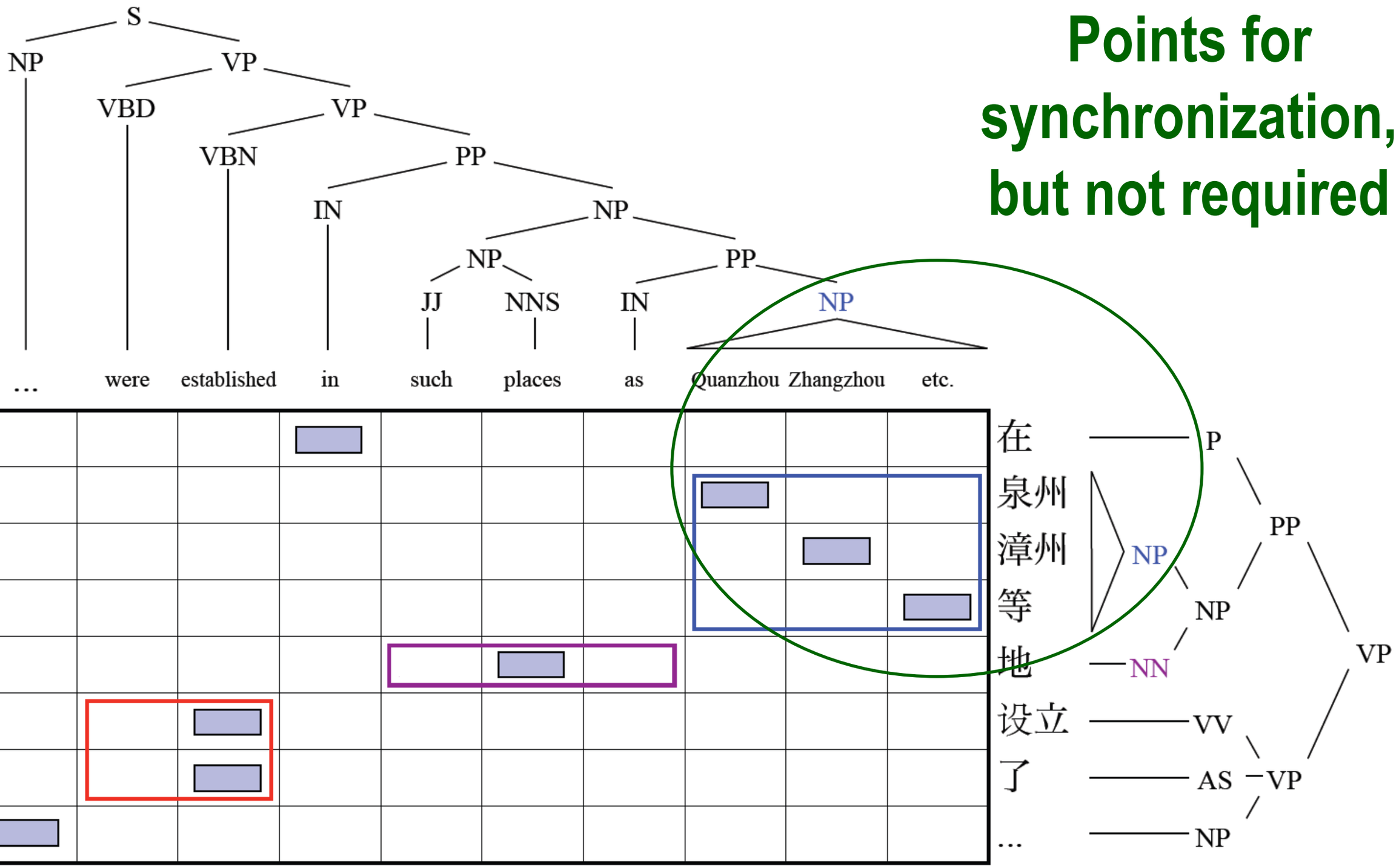
Weakly Synchronized Example



ITG alignment



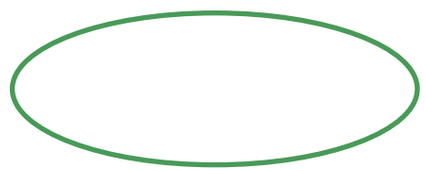
Weakly Synchronized Example

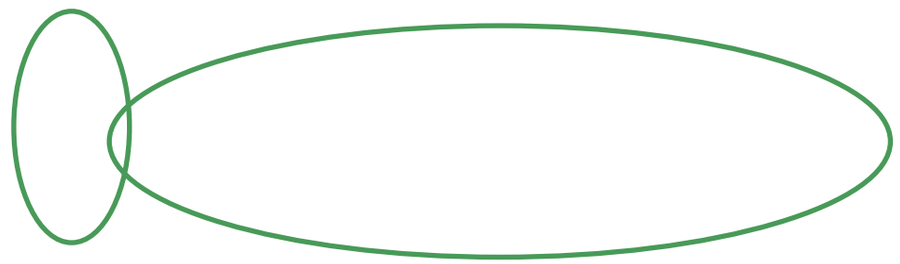


Points for synchronization, but not required

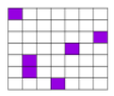
Correspondence Model & Feature Types

p_θ (, , | EN , 中文)





Feature type 1: Word Alignment

ϕ ( , EN , 中文)

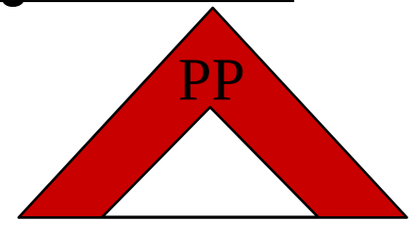
办公室

office

[HBDK09]

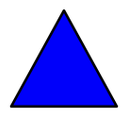
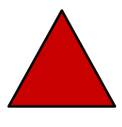
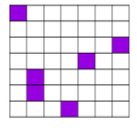
Feature type 2: Monolingual Parser

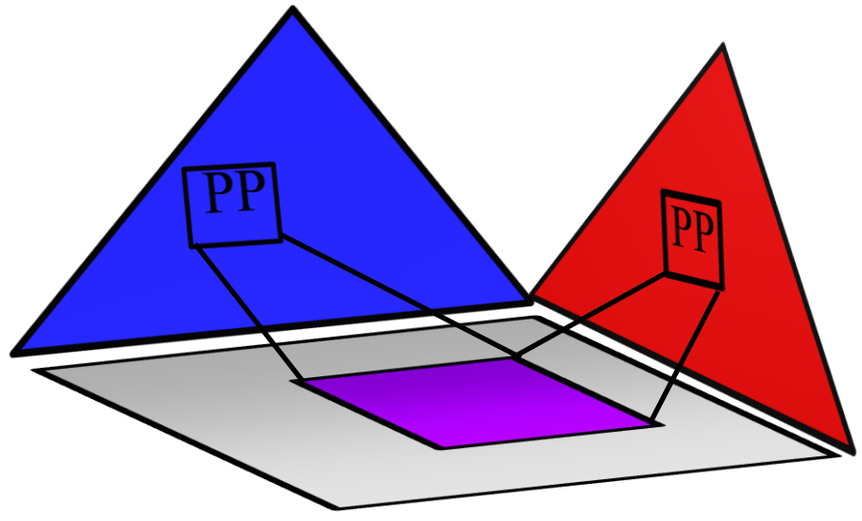
ϕ (, EN)



in the office

Feature type 3: Correspondence

ϕ ( ,  , )



Estimating θ

- Set θ to maximize the log-likelihood of the correct parses & alignments

$$\log p_\theta\left(\blacktriangle, \blacktriangle, \text{⚄} \,\middle|\, \text{EN}, \text{中文}\right) = \langle \theta, \phi(\blacktriangle, \blacktriangle, \text{⚄}, \text{EN}, \text{中文}) \rangle - \log Z(\text{EN}, \text{中文})$$

- $Z(\text{EN}, \text{中文})$ normalizes p_θ to sum to 1

$$Z(\text{EN}, \text{中文}) = \sum_{\blacktriangle, \blacktriangle, \text{⚄}} \exp\left\{\langle \theta, \phi(\blacktriangle, \blacktriangle, \text{⚄}, \text{EN}, \text{中文}) \rangle\right\}$$

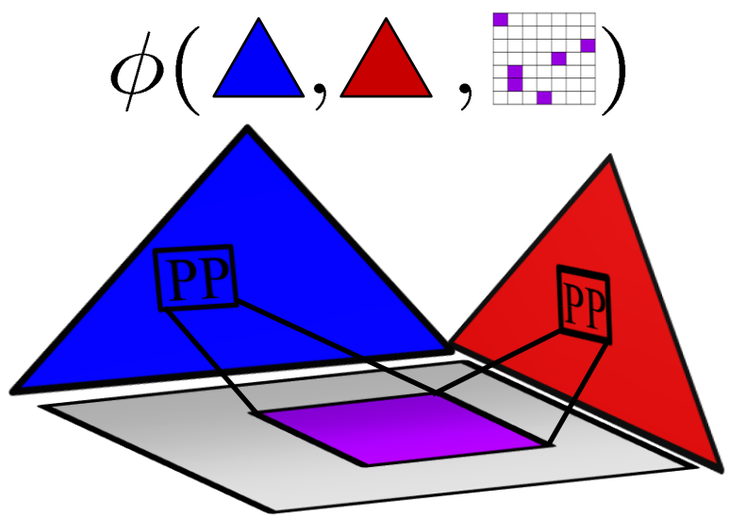
Computing $Z(\text{EN}, \text{中文})$

$$Z(\text{EN}, \text{中文}) = \sum_{\blacktriangle, \blacktriangle, \blacksquare} \exp\{\langle \theta, \phi(\blacktriangle, \blacktriangle, \blacksquare, \text{EN}, \text{中文}) \rangle\}$$

Individual $\sum_{\blacktriangle}$, $\sum_{\blacktriangle}$, $\sum_{\blacksquare}$ have polynomial-time dynamic programming algorithms

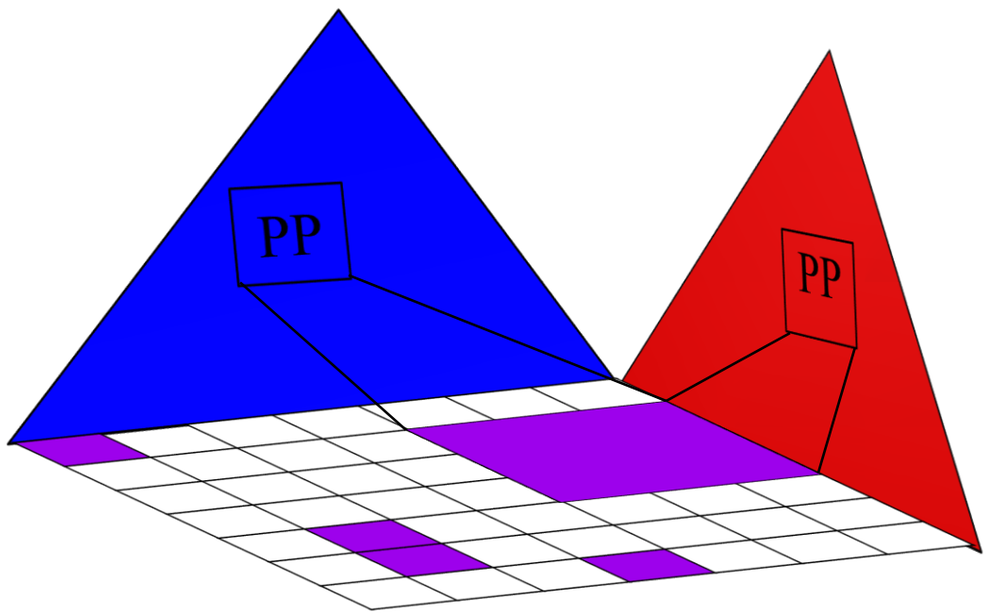
Correspondence features tie pieces together

Computing $Z(\text{EN}, \text{中文})$ exactly is intractable



Approximating $Z(\text{EN},\text{中文})$: Mean Field

- Exploit tractability in individual models: $\sum_{\blacktriangle} \sum_{\blacktriangle} \sum_{\blacksquare}$
- Factored approximation: $p_\theta(\blacktriangle,\blacktriangle,\blacksquare \mid \text{EN},\text{中文}) \approx q(\blacktriangle)q(\blacktriangle)q(\blacksquare)$
- Set q to minimize $KL(\ q(\blacktriangle)q(\blacktriangle)q(\blacksquare),\ \ p_\theta(\blacktriangle,\blacktriangle,\blacksquare \mid \text{EN},\text{中文})\ \ )$



Algorithm

1) Initialize $q(\blacktriangle)\ q(\blacktriangle)\ q(\blacksquare)$ separately

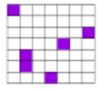
2) Iterate:

$$q(\blacktriangle) \propto \exp\{\langle\theta, \phi(\blacktriangle, E_q(\blacktriangle), E_q(\blacksquare))\rangle\}$$
$$q(\blacktriangle) \propto \exp\{\langle\theta, \phi(E_q(\blacktriangle),\blacktriangle, E_q(\blacksquare))\rangle\}$$
$$q(\blacksquare) \propto \exp\{\langle\theta, \phi(E_q(\blacktriangle), E_q(\blacktriangle), \blacksquare)\rangle\}$$

Large scale inference

We can approximate $Z(\text{EN}, \text{中文})$ in polynomial time, but . . .

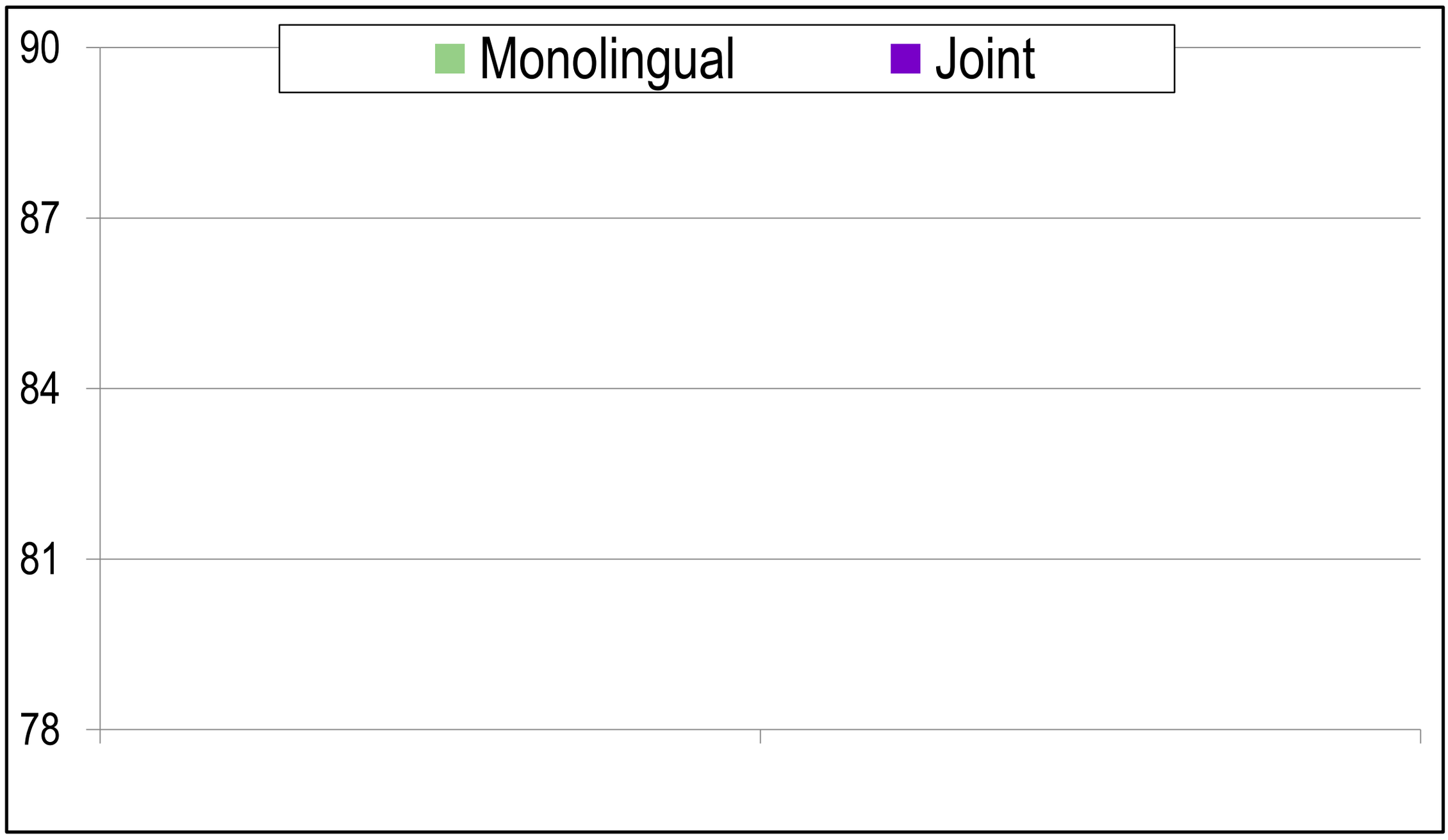
$\sum$ Sum over possible alignments is an $O(n^6)$ algorithm.



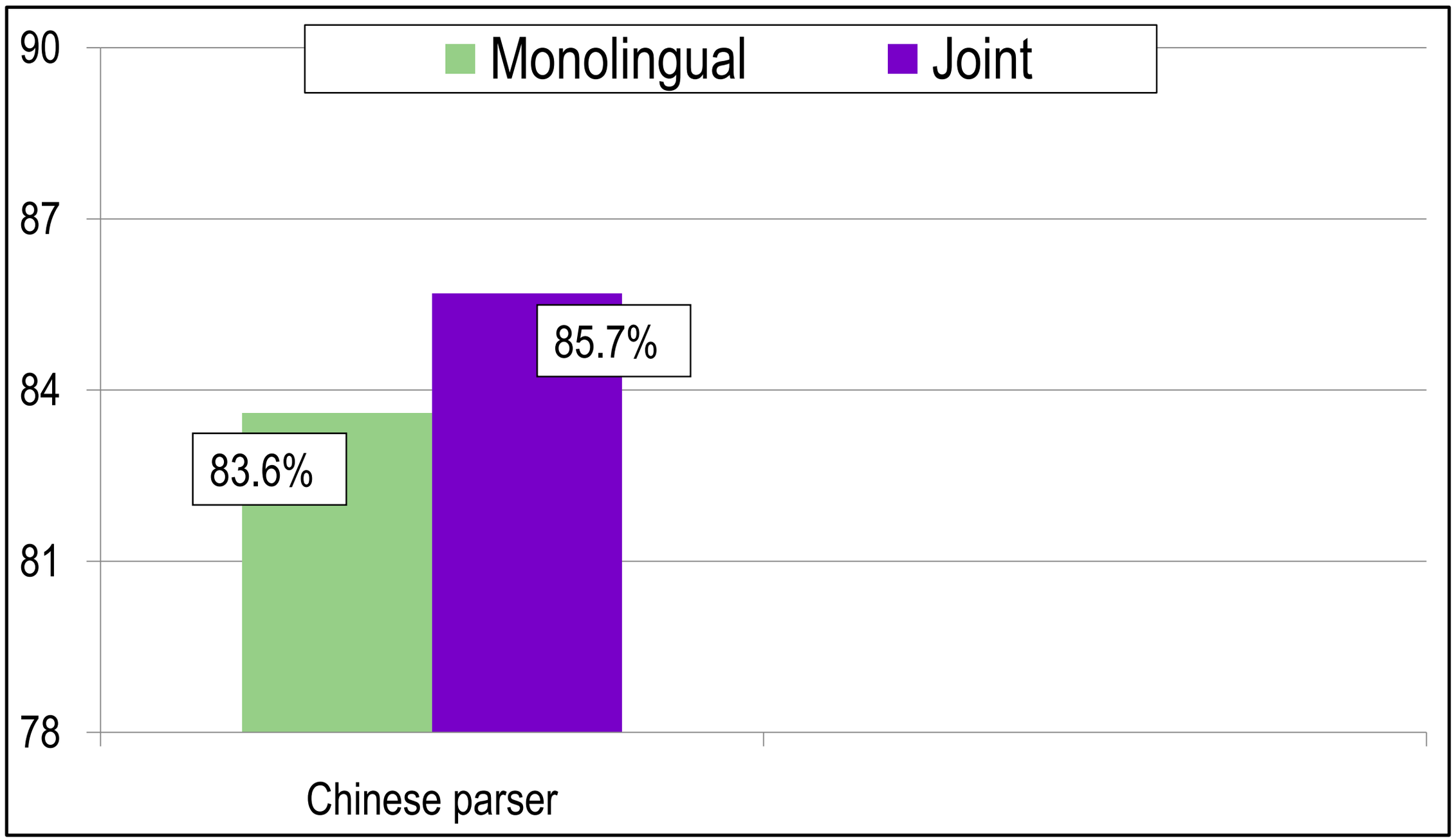
But computers are fast, right?

- Medium-length sentences are 50 words long
- Small translation data sets are 250,000 sentences
- ~4 quadrillion operations (See **[BBK10, HBDK09]** for speedup details)

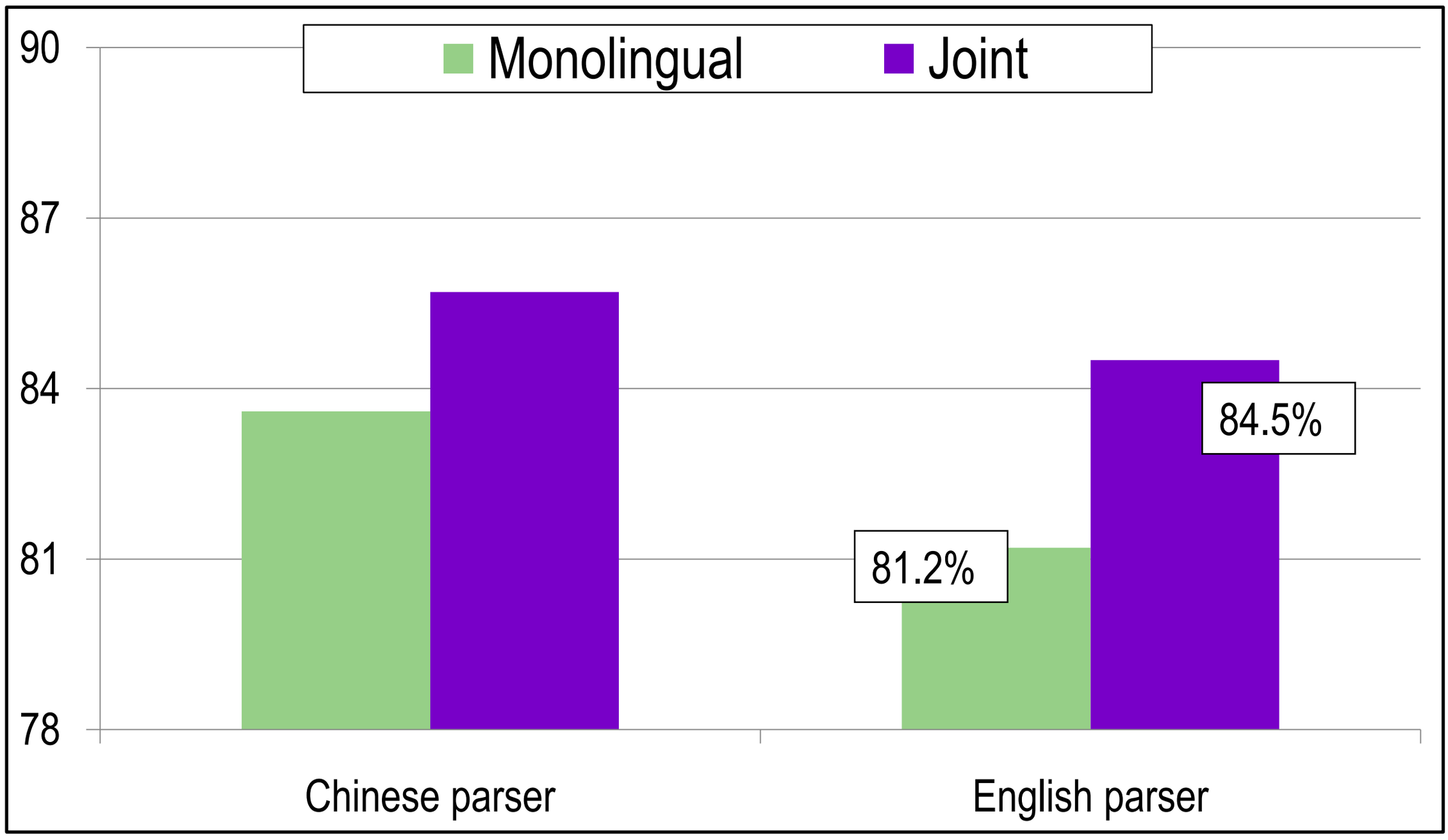
Quantitative Results: Parsing



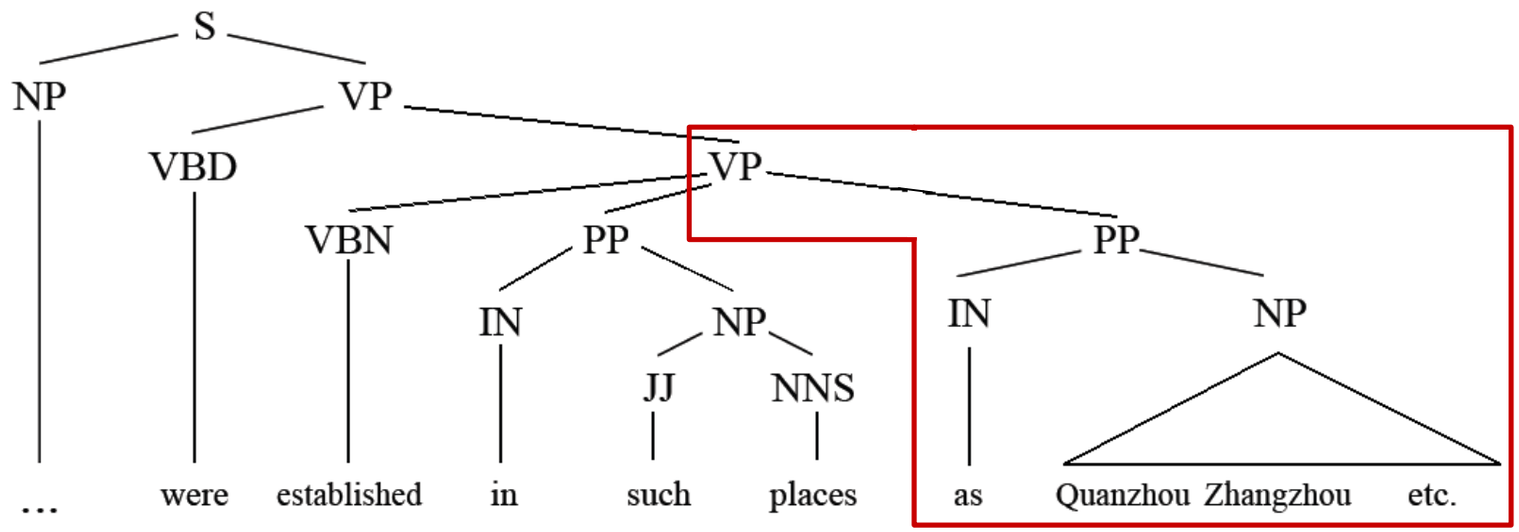
Quantitative Results: Parsing

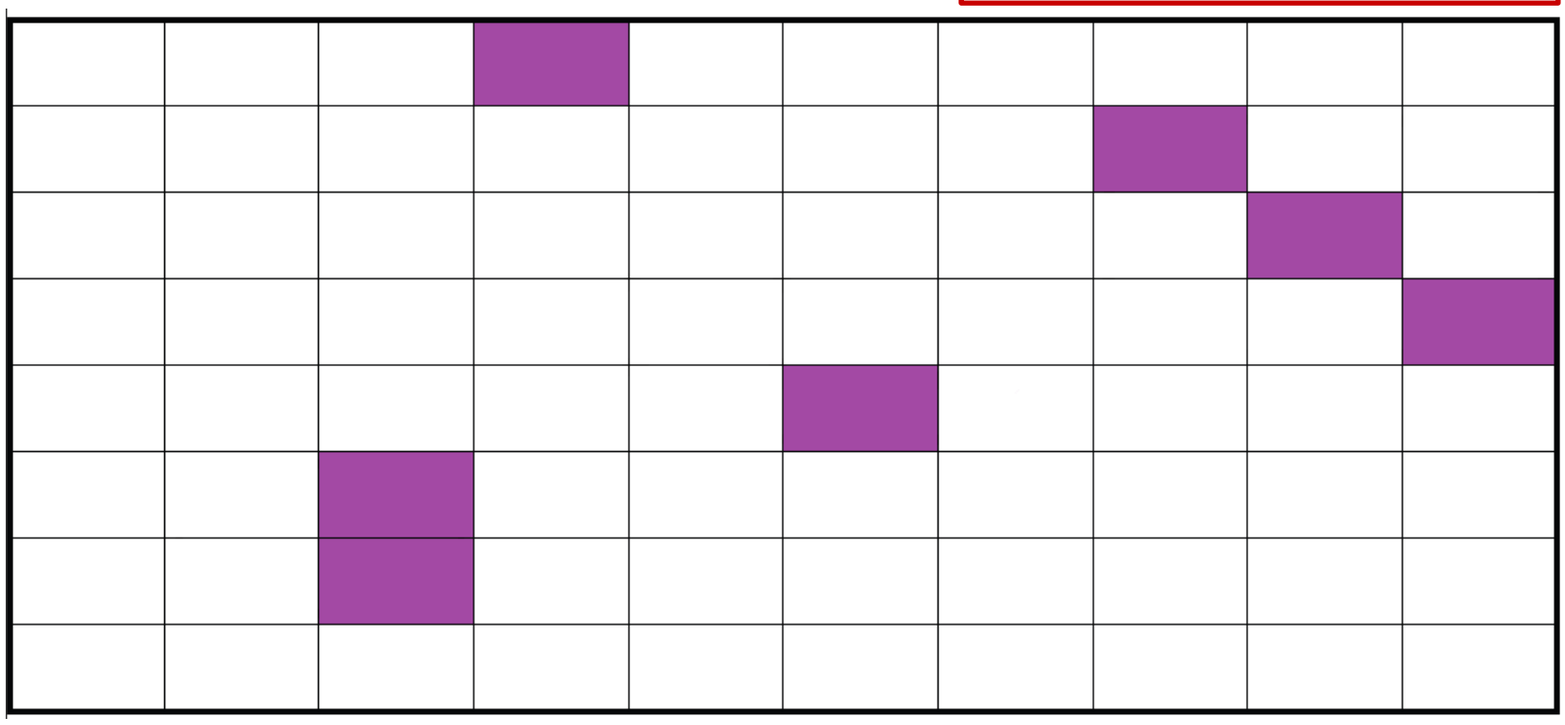


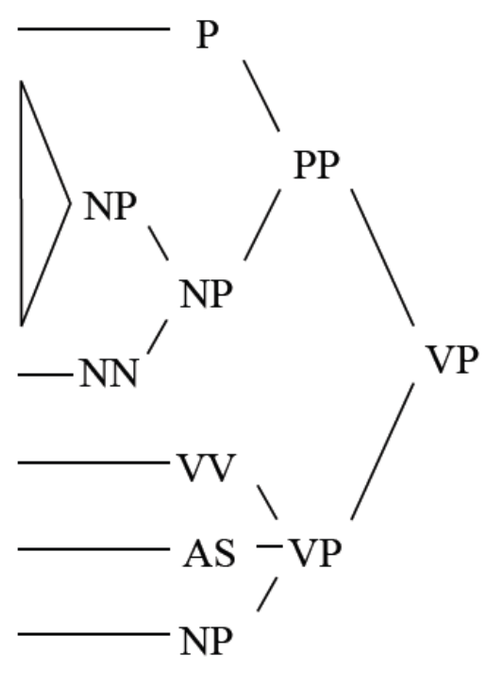
Quantitative Results: Parsing



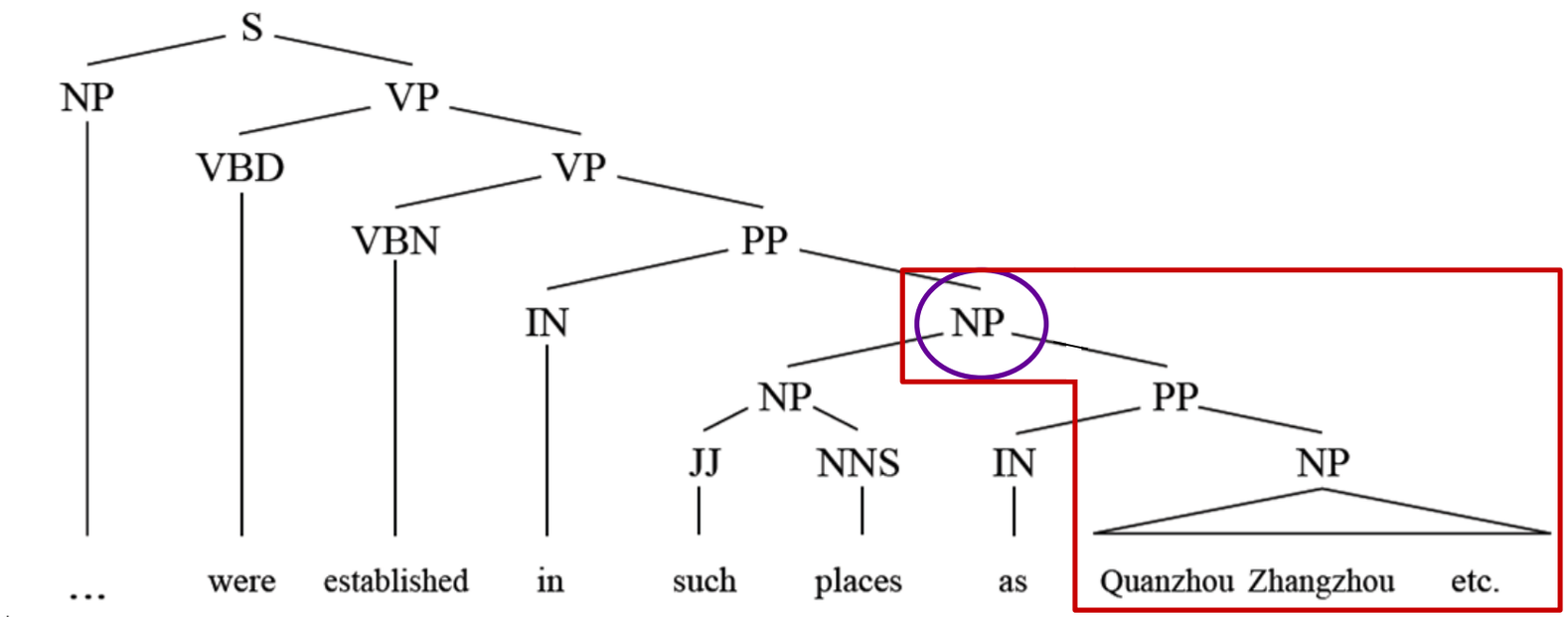
Incorrect English PP Attachment

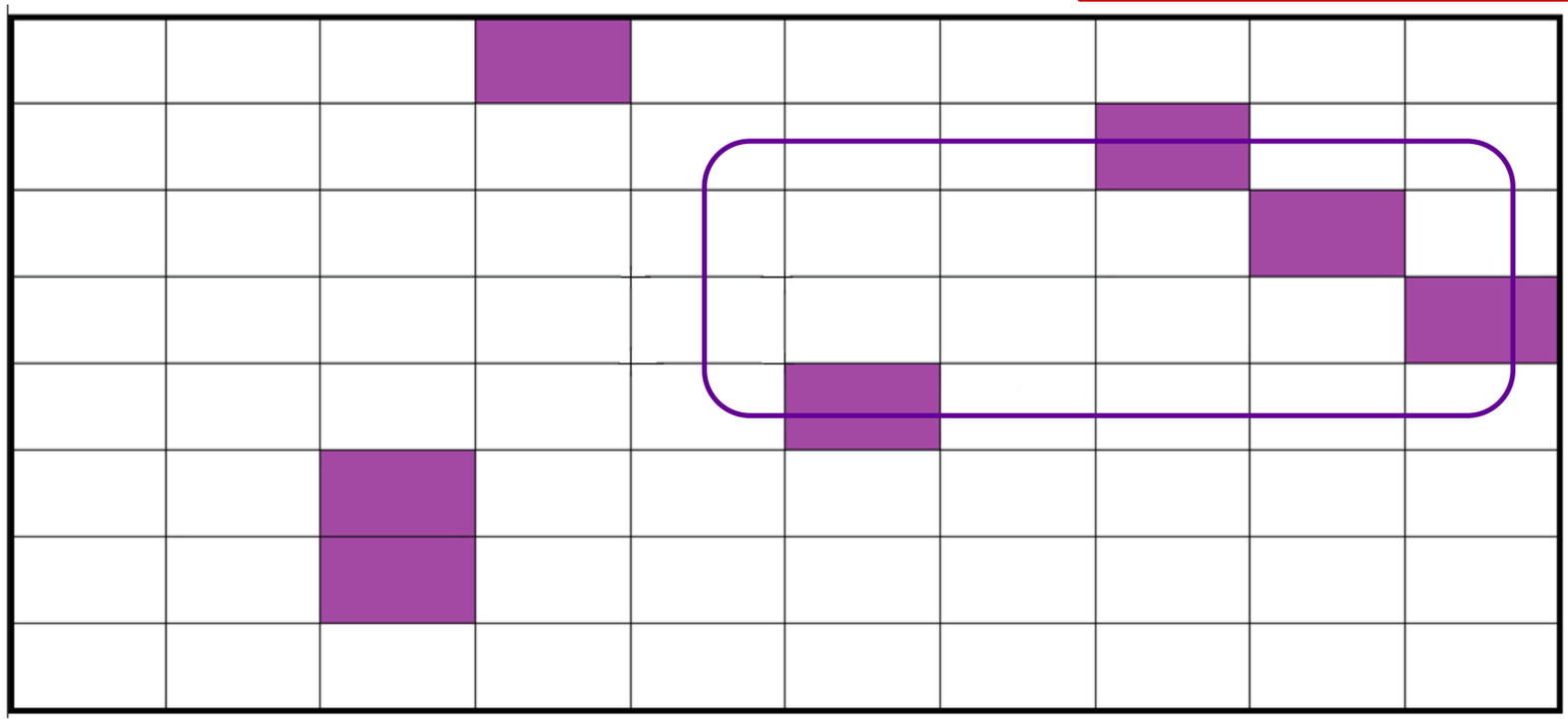


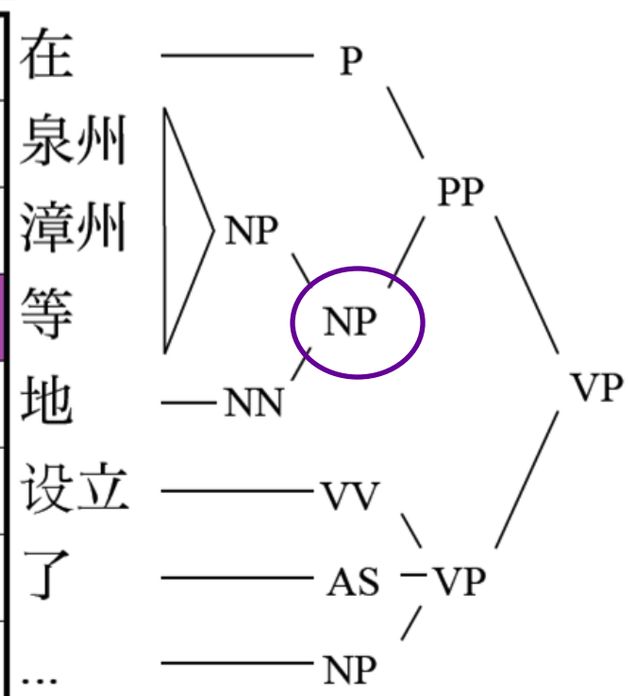




Corrected English PP Attachment

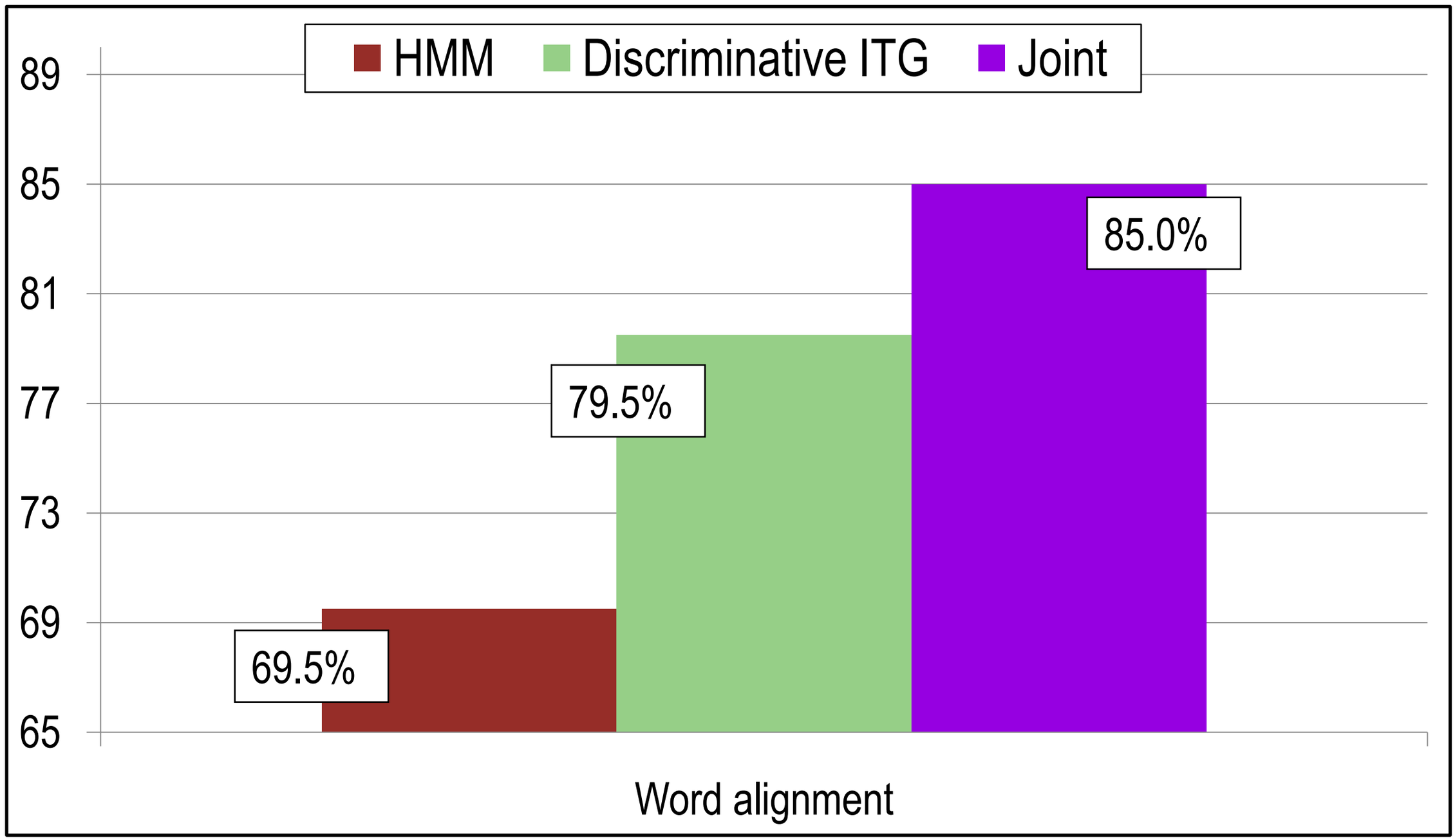






Quantitative Results: Translation

BLEU improvement from 29.4 to 30.6



Better Translations with Bilingual Adaptation

目前	导致	飞机	相撞	的	原因	尚	不	清楚,	当地	民航	部门	将	对此	展开	调查
Cur-rently	cause	plane	crash	DE	reason	yet	not	clear,	local	civil aero-nautics	bureau	will	toward	open	investi-gations

Reference

At this point the cause of the plane collision is still unclear. The local caa will launch an investigation into this .

Baseline (GIZA++)

The cause of planes is still not clear yet, local civil aviation department will investigate this .

Bilingual Adaptation Model

The cause of plane collision remained unclear, local civil aviation departments will launch an investigation .

University of California
C A L
N L P
Berkeley

Thanks