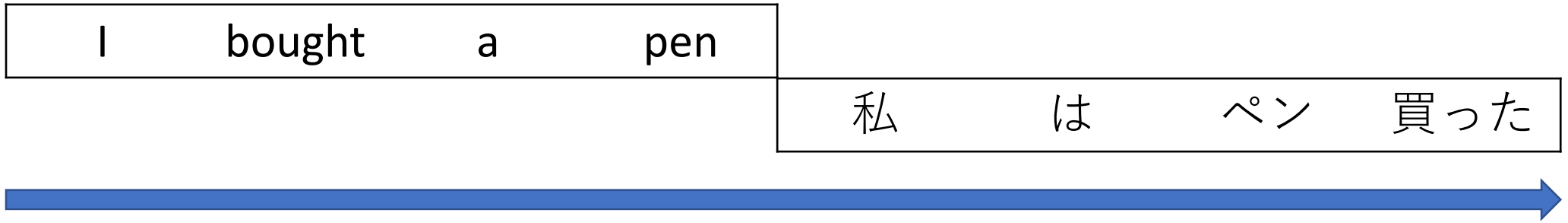
Average Token Delay:

A Latency Metric for Simultaneous Translation

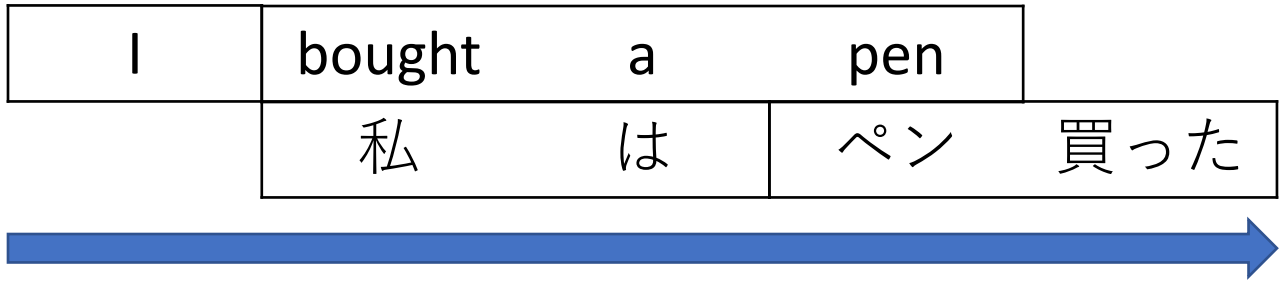
Yasumasa Kano, Katsuhito Sudoh, Satoshi Nakamura
Nara Institute of Science and Technology, Japan
kano.yasumasa.kw4@is.naist.jp

Simultaneous Translation

Standard translation

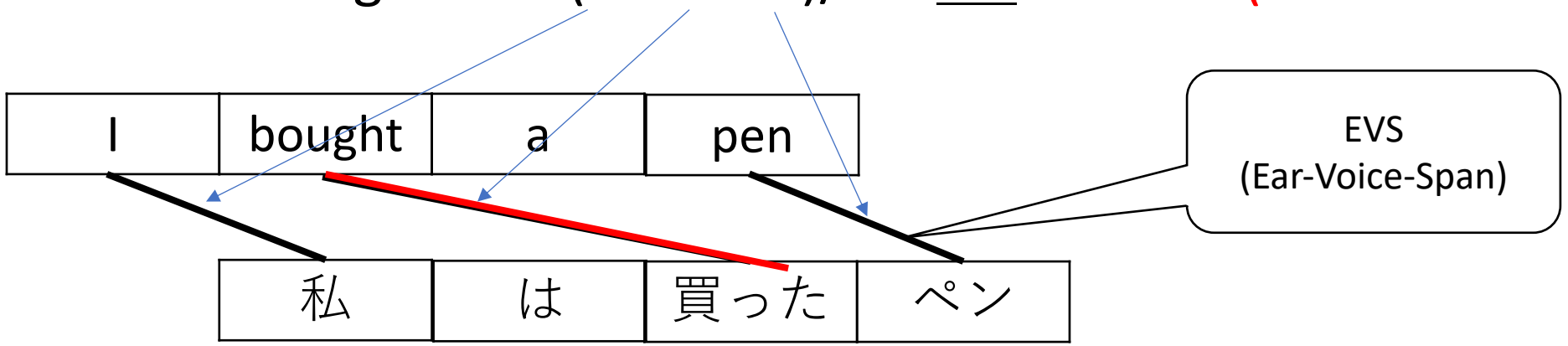


Simultaneous translation

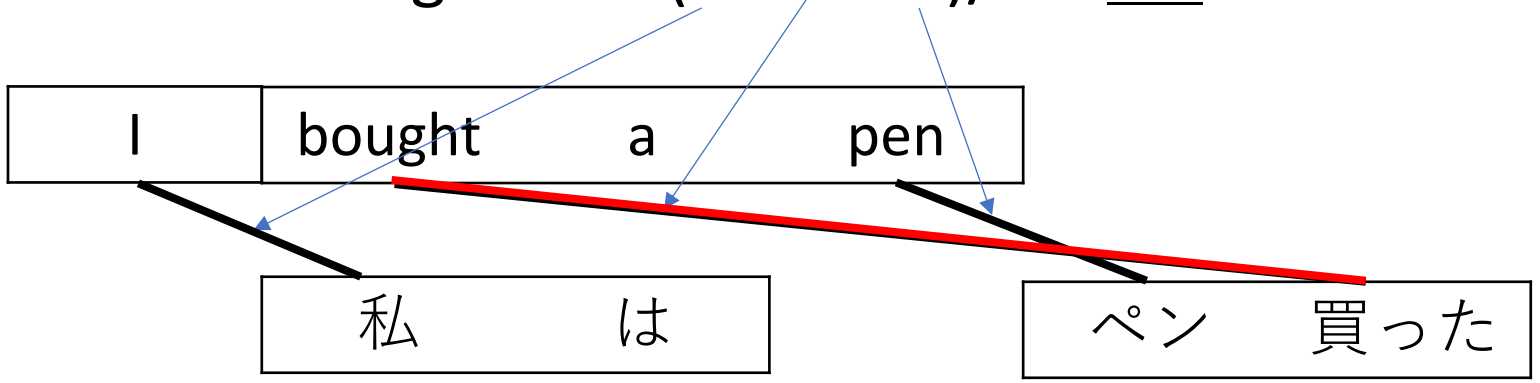


Latency of Simultaneous Translation

Case 1: Avg. EVS = (1 + 2+ 1)/3 = **1.3** (Smaller Latency)

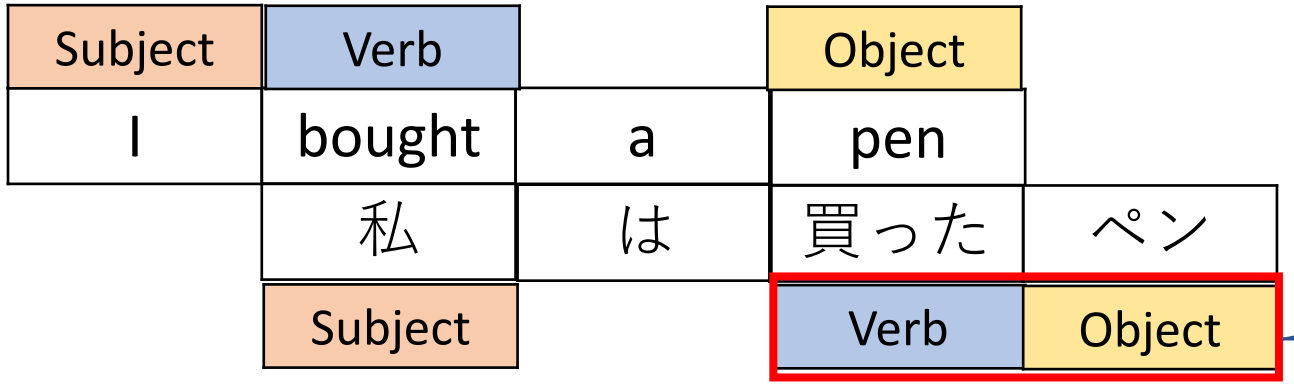


Case 2: Avg. EVS = (1 + 4+ 1)/3 = **2.0**



Quality of Simultaneous Translation

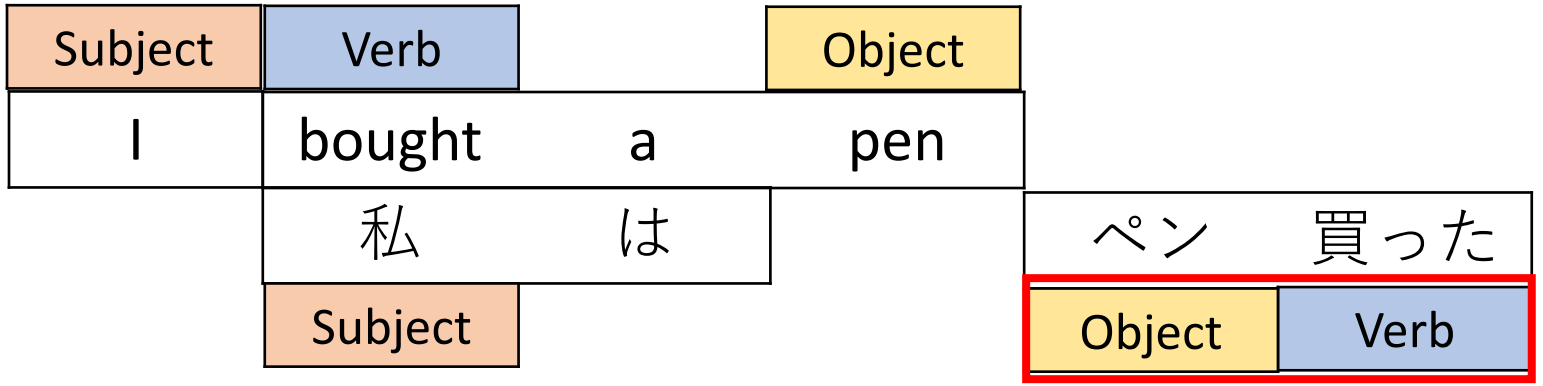
Case 1



Japanese word order:
SOV (Subject-Object-Verb)

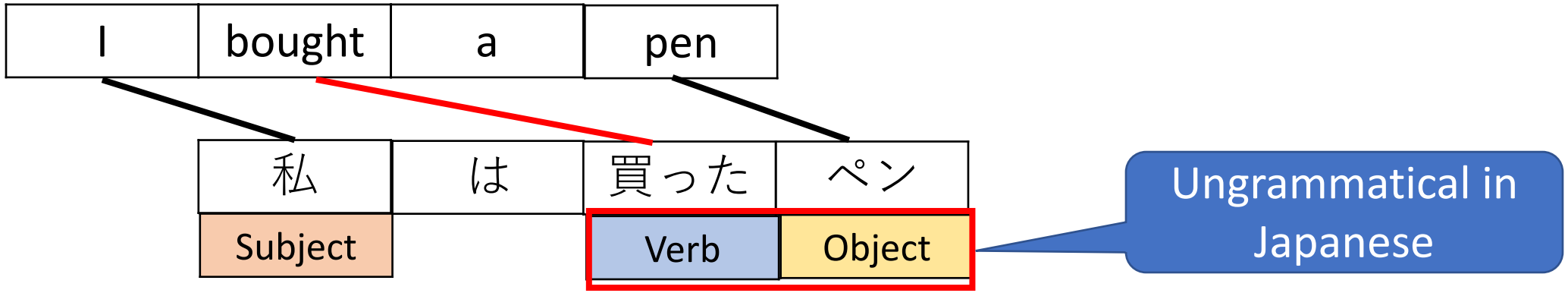
Ungrammatical in Japanese

Case 2 (Higher quality)

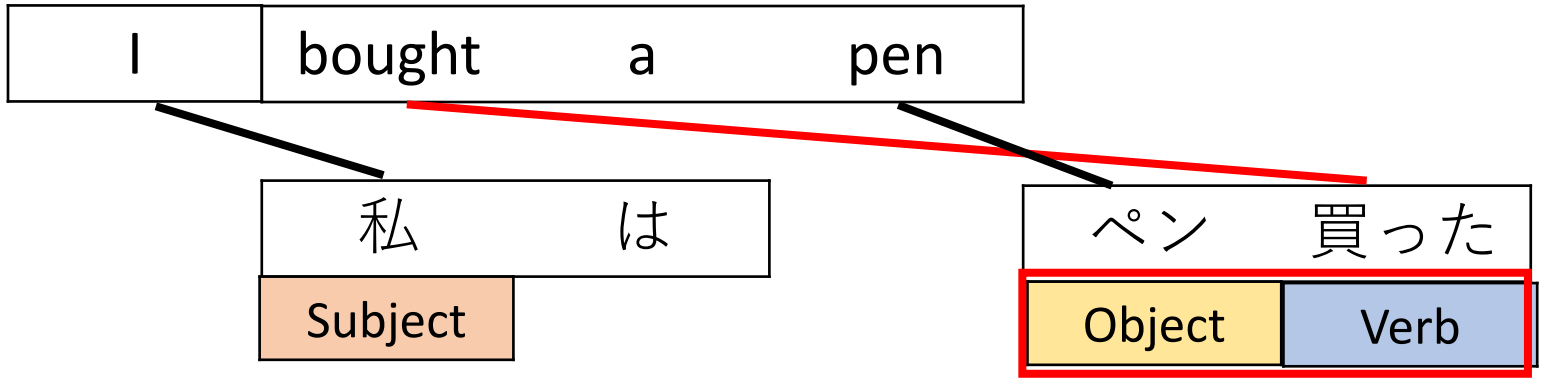


Quality-Latency Trade-off

Case 1 Latency: Small Quality: Low



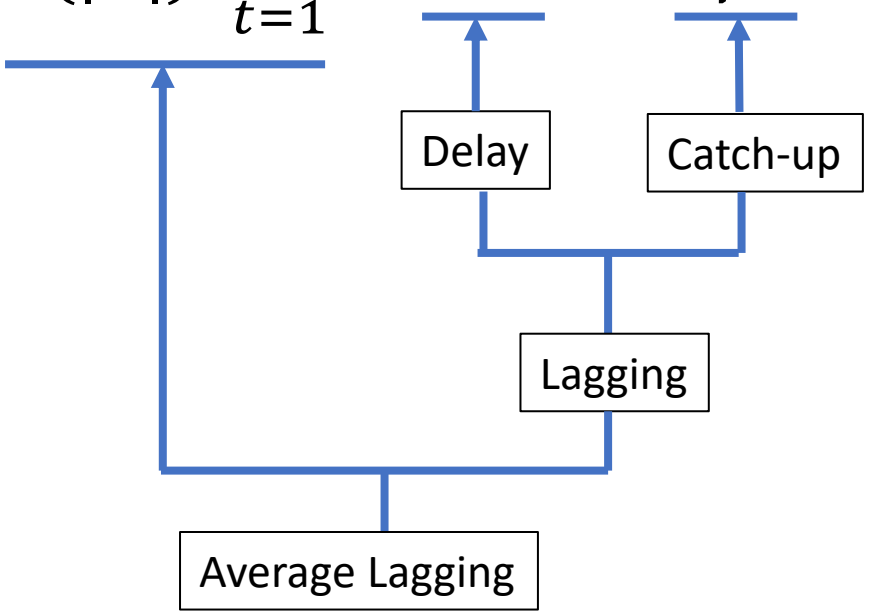
Case 2 Latency: Large Quality: High



Previous Latency Metric: AL (Average Lagging)

[Ma+, 2019]

$$AL = \frac{1}{\tau(|\boldsymbol{x}|)} \sum_{t=1}^{\tau(|\boldsymbol{x}|)} \left(g(t) - \frac{t-1}{\gamma} \right)$$



$\boldsymbol{x} = x_1, x_2, \cdots, x_{|\boldsymbol{x}|}$: Input tokens

$\boldsymbol{y} = y, y_2, \cdots, y_{|\boldsymbol{x}|}$: Output tokens

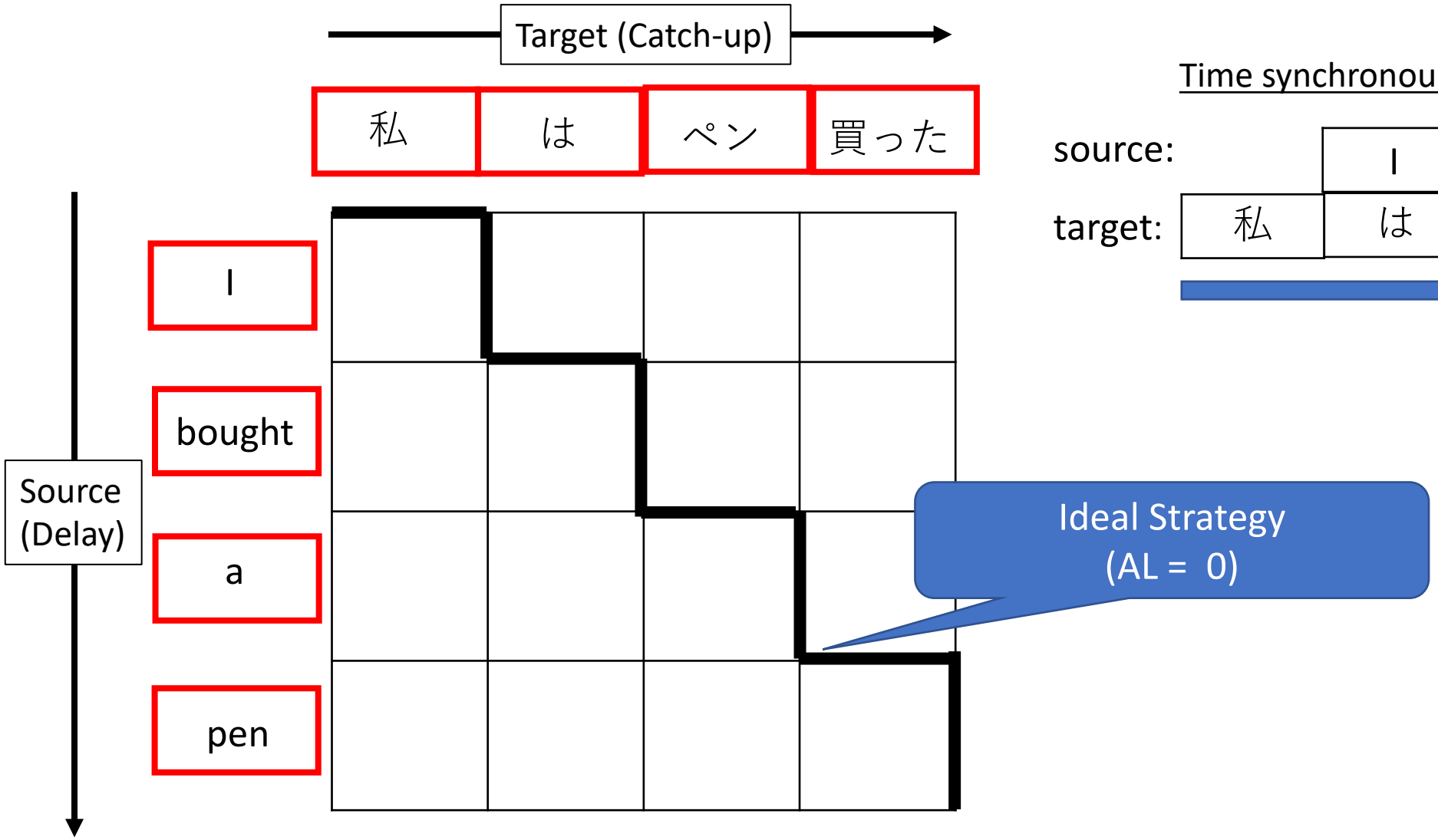
$\gamma = \frac{|\boldsymbol{y}|}{|\boldsymbol{x}|}$

t : Target token index

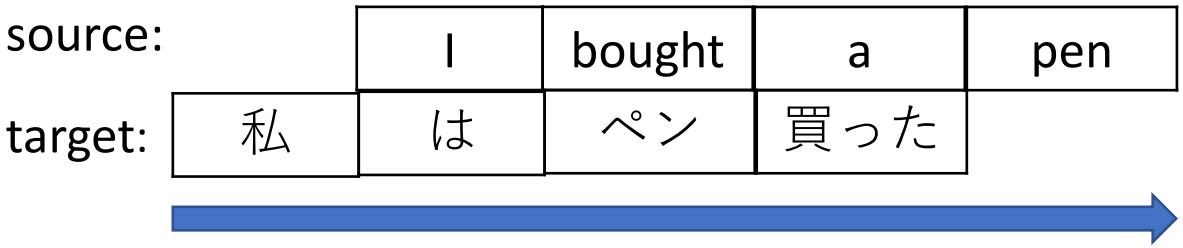
$g(t)$: Number of input tokens read to output t th target token

$\tau(|\boldsymbol{X}|) = \min(t|\ g(t) = |\boldsymbol{x}|)$

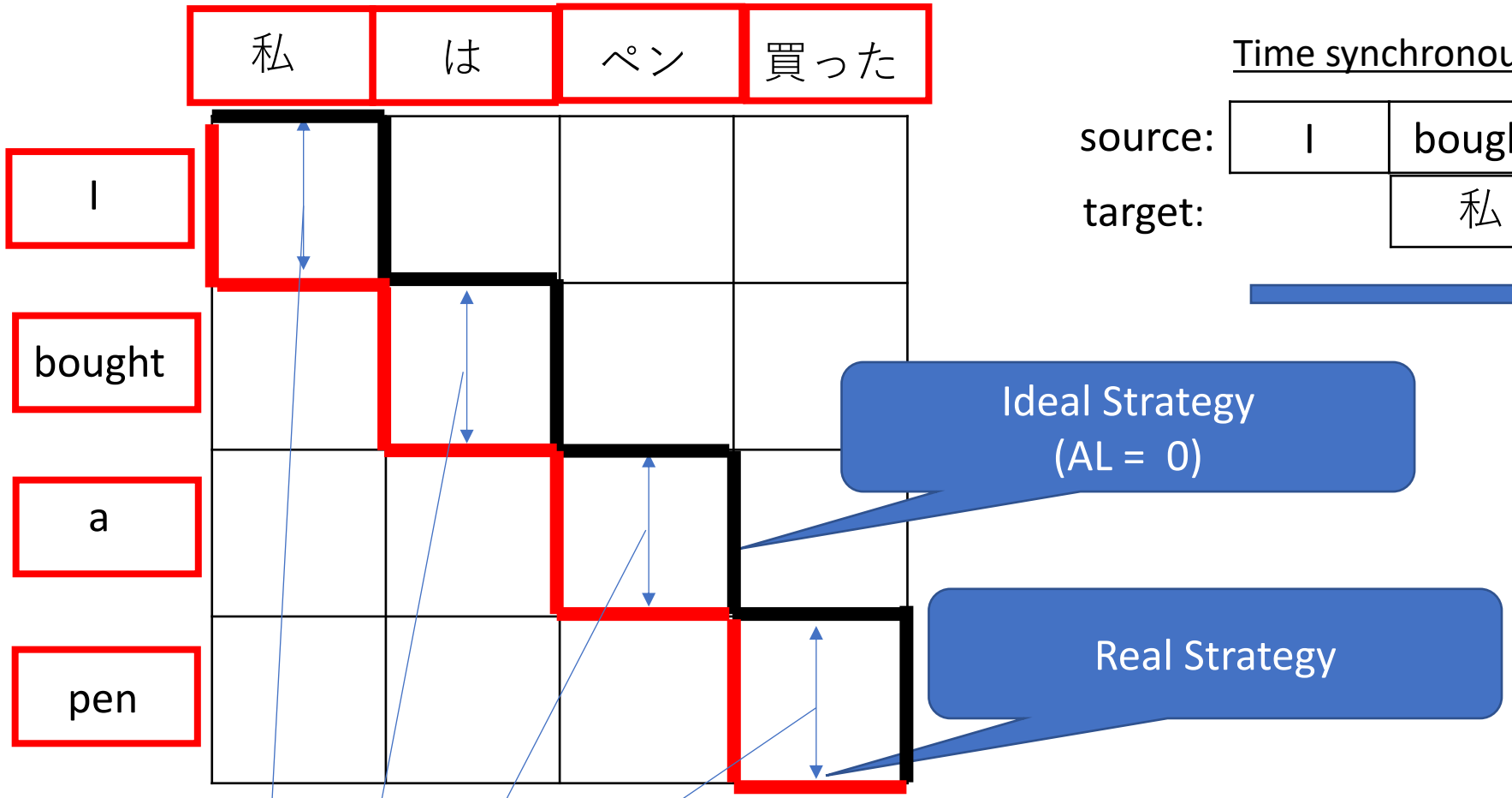
Calculation of AL (Ideal Strategy)



Time synchronous view



Calculation of AL (Case 1)



Time synchronous view (Case 1)

source:	I	bought	a	pen	
target:		私	は	ペン	買った

Ideal Strategy
(AL = 0)

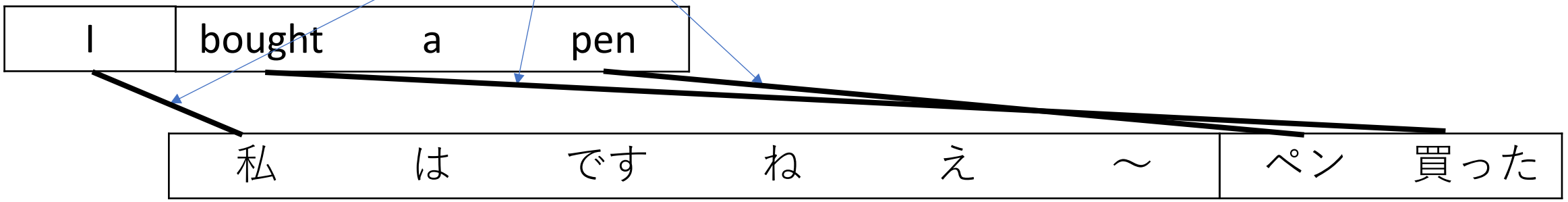
Real Strategy

AL = (1 + 1 + 1 + 1) / 4 = **1**

Case 3: Long Translation Output

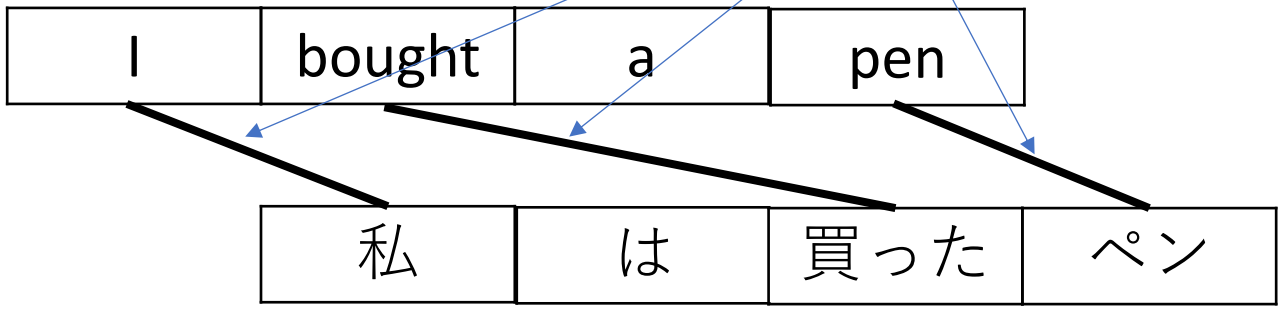
Case 3: Avg. EVS = (1 + 7+ 4)/3 = **4.0**

(Larger Latency)



Review

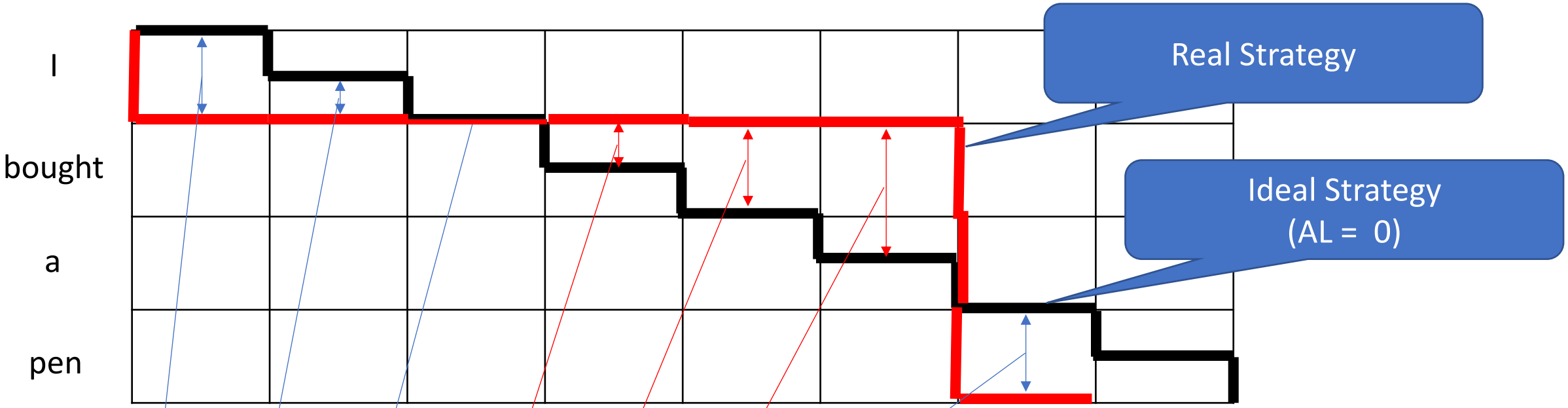
Case 1: Avg. EVS = (1 + 2+ 1)/3 = **1.3**



Calculation of AL (Case 3)

Time synchronous view (Case3)

I	bought	a	pen						
	私	は	です	ね	え	～	ペン	買った	



AL = (1.0 + 0.5 + 0.0 – 0.5 – 1.0 – 1.5 + 1.0) / 7 = **-0.07**

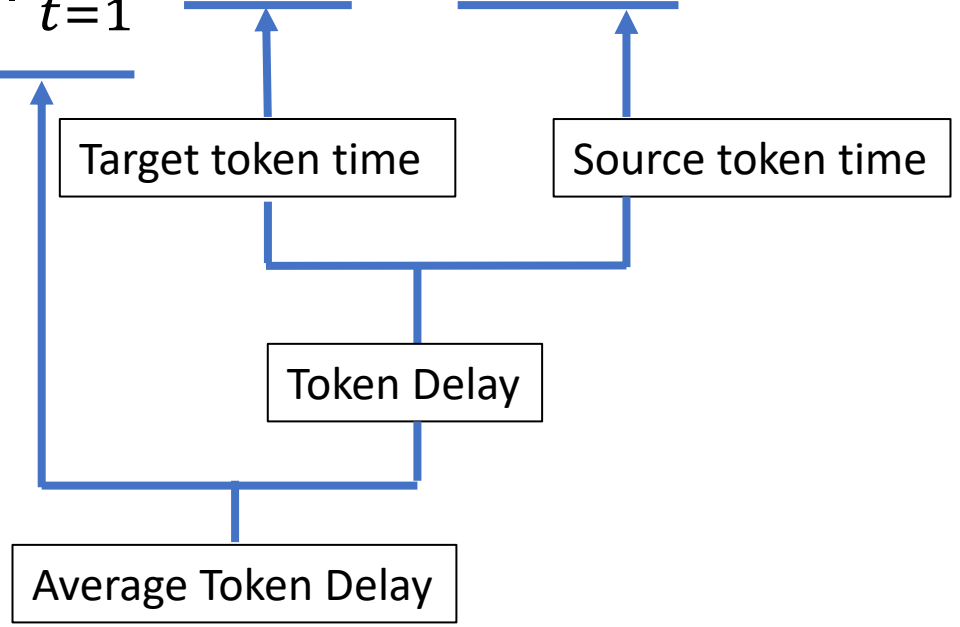
Problems of AL

- Longer translation output → Smaller latency
- Negative latency value

➡ Propose latency metrics to solve the problems

Proposed Metric: ATD (Average Token Delay)

$$ATD = \frac{1}{|\boldsymbol{y}|}\sum_{t=1}^{|\boldsymbol{y}|}\left(T(y_t) - T(x_{a(t)})\right)$$



$\boldsymbol{x} = x_1, x_2, \cdots, x_{|x|}$: Input tokens

$\boldsymbol{y} = y, y_2, \cdots, y_{|x|}$: Output tokens

t : Target token index

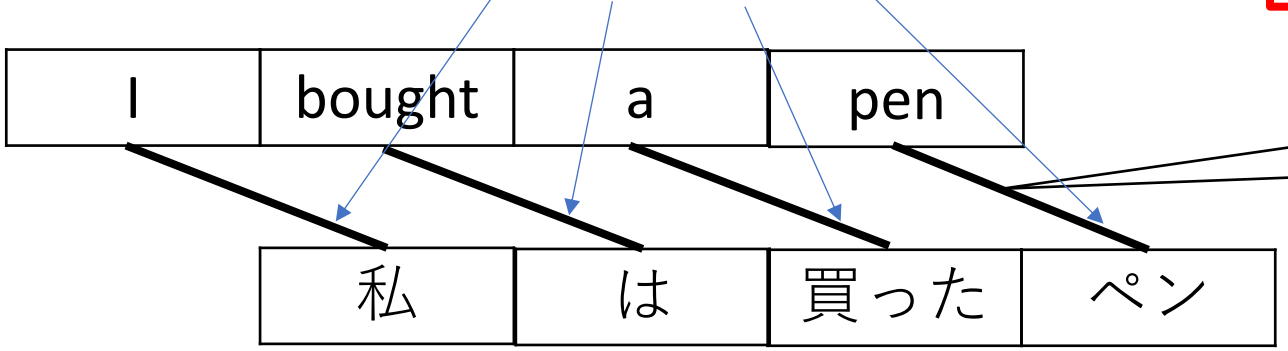
$a(t)$: Source token index corresponding to target token index t

$T(*)$: Ending time of each token

Calculation of ATD (Inspired by EVS)

Case 1: ATD = (1 + 1 + 1 + 1)/4 = **1.0**

Avg. EVS: **1.3**

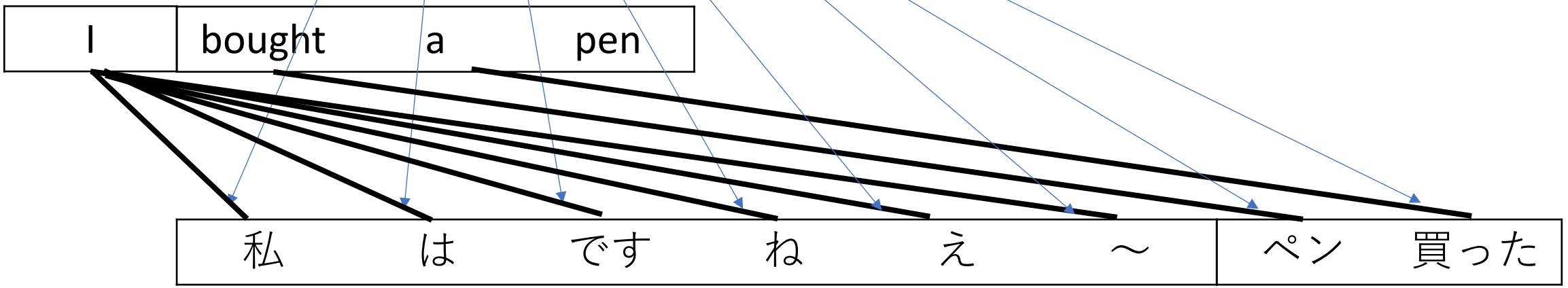


Token Delay: Without Semantic Correspondence

EVS: With Semantic Correspondence

Case 3: ATD = (1 + 2 + 3 + 4 + 5 + 6 +6 + 6)/8 = **4.1**

Avg. EVS: **4.0**



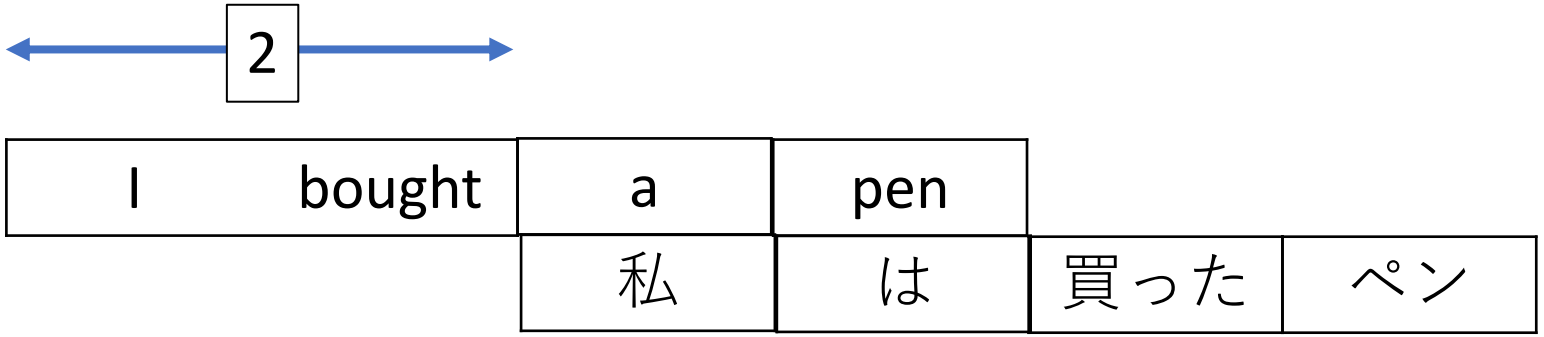
Comparison of AL and ATD

- AL
  - Longer translation output → Smaller Latency
  - Negative latency value
- ATD
  - Close to EVS
  - Longer translation output → Larger Latency
  - Non-negative latency value

ATD solves the problems of AL

Experiment of Simultaneous Translation

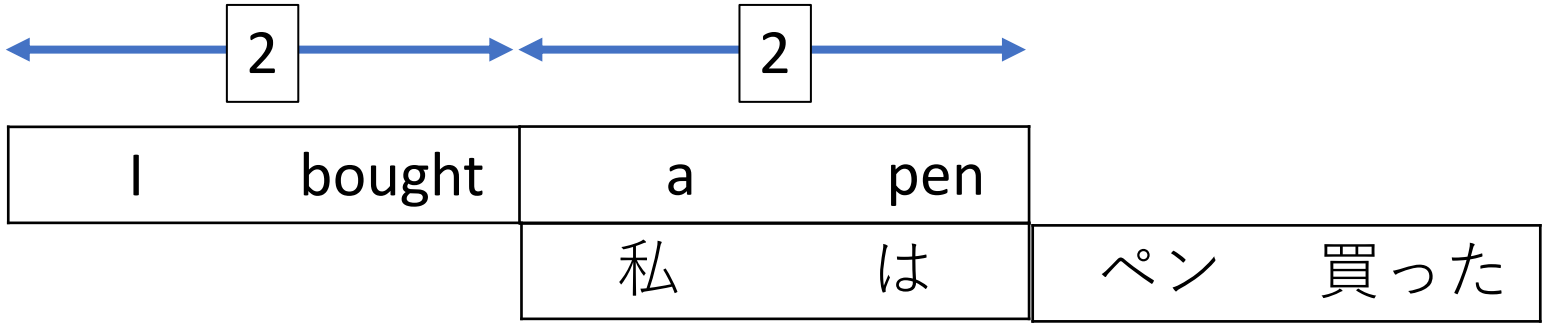
1.Wait-k [Ma+, 2019]　k=2



To adjust latency:
k = [2, 4, 6, …, 30]

Larger k,
Larger latency

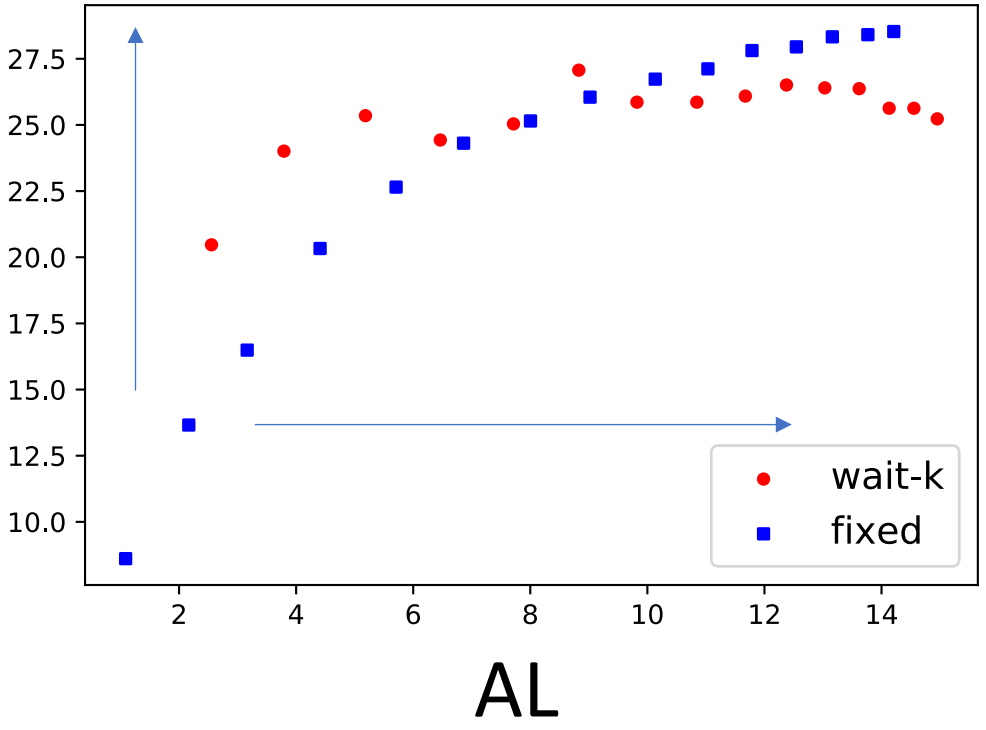
2.Fixed-size segmentation　k=2



Result

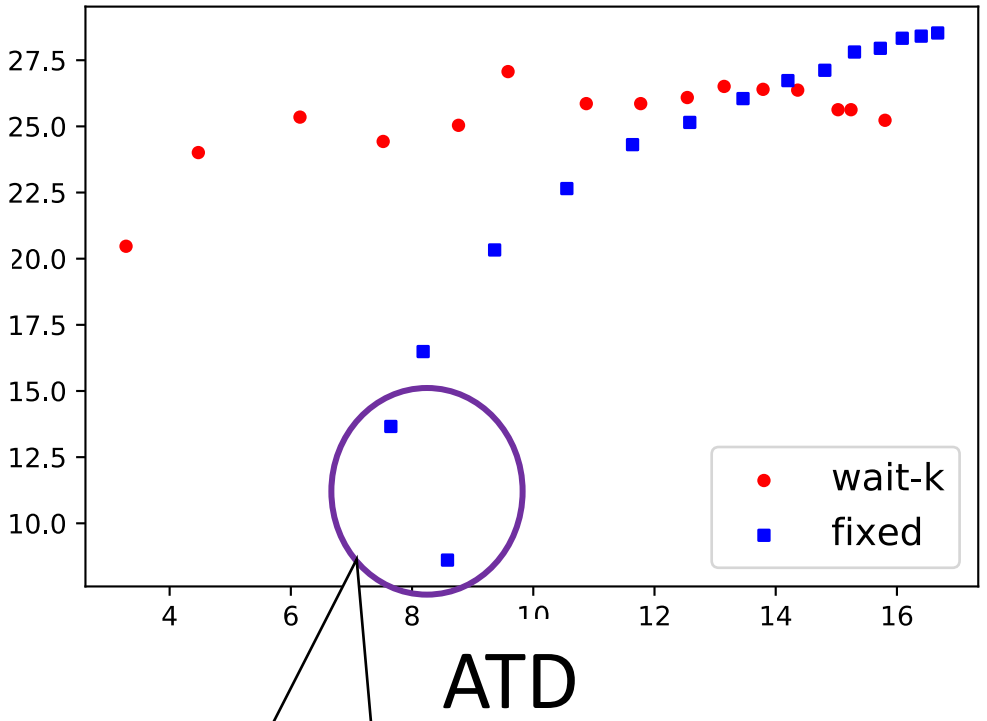
BLEU

[Papineni+, 2022]



Larger Delay, Higher Quality

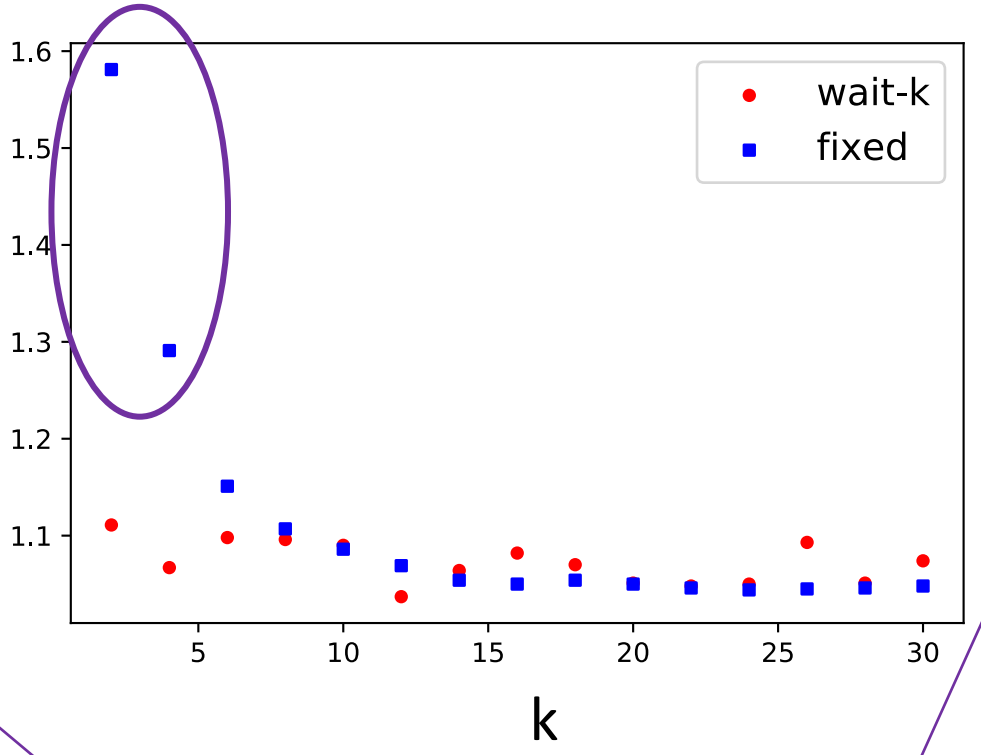
BLEU



Larger Delay, Lower Quality

Analysis: Length ratio

$$length\ ratio = \frac{predicted\ translation\ length}{reference\ length}$$



Length Ratio

Unnecessary words

Quality

Omission

Larger

Latency

Smaller

ATD reflects this characteristic

Conclusion

- Problems of AL
  - Longer translation output → Smaller latency
  - Negative latency value
- Proposed latency metric: ATD
  - Solved the problems above
- Future work
  - Investigate correlation between latency metrics and latency scores evaluated by human