Large-Scale English-Japanese Simultaneous Interpretation Corpus: Construction and Analyses with Sentence-Aligned Data

Kosuke Doi[1], Katsuhito Sudoh[1, 2], Satoshi Nakamura[1, 2]

[1]Nara Institute of Science and Technology

[2]AIP, RIKEN

Our paper

- Constructing a new large-scale English↔Japanese Simultaneous interpretation (SI) corpus
    - Over 300 hours
    - Some lectures have SI data from 3 interpreters with different amounts of experience [Shimizu+ 2014]
- Analyzed the corpus: (1) latency, (2) quality, (3) word order
    - Experienced interpreters controlled latency and quality better
    - Large latency hurt SI quality
- Release a part of the corpus at:

https://dsc-nlp.naist.jp/data/NAIST-SIC/

Outline

- Introduction
- Corpus construction
- Corpus analysis
- Results
  - Latency
  - Quality
  - Human evaluation
  - Word order
- Conclusion

Background

- Simultaneous interpretation (SI)
  - Translating speech in real-time

- Various studies on automatic speech translation including SI
- Speech Translation corpora vs. SI corpora [Zhang+ 2021]

Speech Translation	Features	SI
Based on complete audio data or transcripts	Translations	Based on actual SIs
Available many	# of corpora	Remains very limited

→ Construct a new SI corpus

Outline

- Introduction
- Corpus construction
- Corpus analysis
- Results
  - Latency
  - Quality
  - Human evaluation
  - Word order
- Conclusion

Overview of our corpus

- Over 300 hours
- * = interpreted by interpreters from all 3 ranks (4h x 3 interpreters)
- Others = interpreted by either an S- or A-rank interpreter
- About half of the SIs have been transcribed

Direction	Source	2018	2019	2020
En → Ja	TED	67+12*	50	50
Jp → En	TEDx	12*	40	0
	CSJ	33	0	0
	JNPC	4	36.5	0
Total		128	126.5	50
Cum.		128	254.5	304.5

Experience	Rank
15 years	S-rank
4 years	A-rank
1 years	B-rank

Source

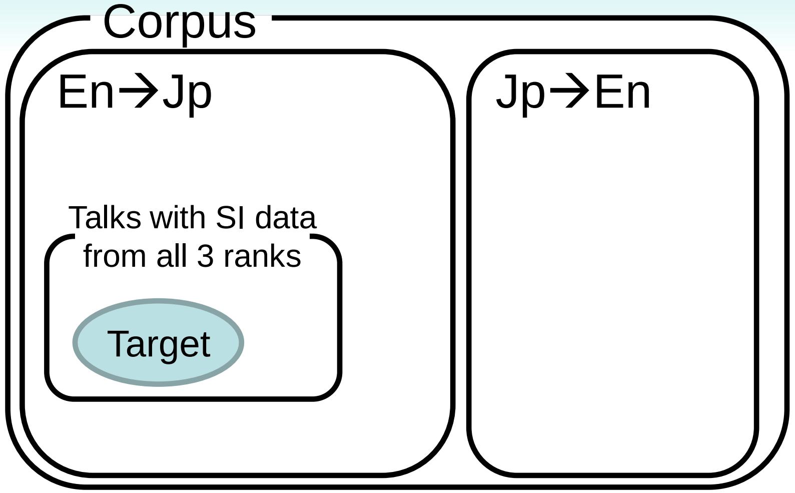
- English → Japanese
  - TED: various topics from science to culture
- Japanese → English
  - TEDx: the same format as TED
  - CSJ: academic lectures and speeches on everyday topics
  - JNPC: press conferences

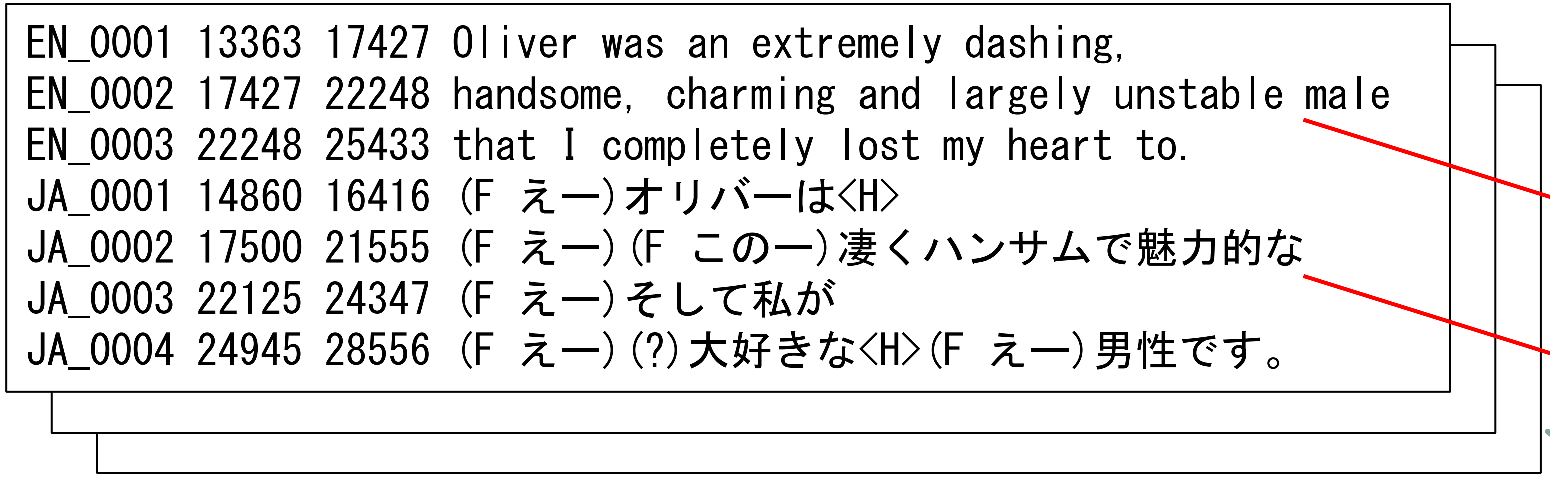
Outline

- Introduction
- Corpus construction
- Corpus analysis
- Results
  - Latency
  - Quality
  - Human evaluation
  - Word order
- Conclusion

Data

- Target of the analysis: 14 TED talks
  - Subset of talks that have En→Jp SI data from interpreters of all 3 ranks





```
EN_0001 13363 17427 Oliver was an extremely dashing,
EN_0002 17427 22248 handsome, charming and largely unstable male
EN_0003 22248 25433 that I completely lost my heart to.
JA_0001 14860 16416 (F えー)オリバーは<H>
JA_0002 17500 21555 (F えー)(F このー)凄くハンサムで魅力的な
JA_0003 22125 24347 (F えー)そして私が
JA_0004 24945 28556 (F えー)(?)大好きな<H>(F えー)男性です。
```

En subtitle provided by TED

SI

(En + SI) x 3

```
1089 5153  オリバーはとても威勢が良く、
5153 9974  ハンサムで魅力的で、じっとしていることがない。
9974 13159 私が完全に心奪われた男性でした。
```

Offline translation
(Jp subtitle provided by TED)

Data

- SI and offline translation
  - Divided into *bunsetsus** using Juman++ Japanese morphological analyzer [Morita+ 2015] and the KNP parser [Kawahara+ 2006]

- Segment-aligned data
  - Not necessarily match among the interpreters

  → Converted to sentence-level alignment

*A bunsetsu is a basic unit of dependency in Japanese.

Sentence alignment: Rules (1)

- Manually aligned with the source speeches
  - Based on English sentences
    - Ending with a period (.) or a period + a double quotation mark (.”)
    - Ending with a question mark (?) or a question mark + a double quotation mark (?”)
    - Ending with a closed parenthesis

```
EN_0177 469789 471829 I've got two questions for you.
JA_0116 XXXXXX 473315 二つの質問がありますよ。

EN_0178 471829 473469 (Laughter)
JA_0000 XXXXXX XXXXXX __null__

EN_0179 473469 476069 You know what's coming now, right?
JA_0117 474778 476197 質問分かってるんですね。
```

Sentence alignment: Rules (2)

- Words/phrases not interpreted: ignored
- Segments corresponding to multiple sentences
  - Divided at the boundary
  - Marked xxxxx for end/start times

```
EN_0177 469789 471829 I've got two questions for you.
JA_0116 XXXXXX 473315 二つの質問がありますよ。

EN_0178 471829 473469 (Laughter)
JA_0000 XXXXXX XXXXXX __null__

EN_0179 473469 476069 You know what's coming now, right?
JA_0117 474778 476197 質問分かってるんですね。
```

JA_0116 466888 473315
(F え)こちらの方が匿名でやってるので誰も見られていない、そしてお金が入ってくる訳{です||で}。二つの質問がありますよ。

Sentence alignment: Rules (3)

- Sentences not interpreted: __drop__
- Sentences not interpreted intentionally: __skip__

  e.g., Thank you.

```
EN_0135 354509 357749 It's just wrong to lie, for example.
JA_0079 359795 362350 例えば、嘘をつくのは悪いと言う事です。

EN_0136 357749 360909 So, meet my friend Immanuel here.
JA_0000 XXXXXX XXXXXX __drop__

EN_0137 360909 363749 He knows that the sausage is very tasty,
EN_0138 363749 366229 but he's going to turn away because he's a good dog.
JA_0080 362690 367930 こちらの犬ですが、ソーセージがとてもおいしいと分かっているけれども、いい犬なので、
JA_0081 368775 370830 あそこに飛び上がろうとはしません。
```

Sentence alignment: Rules (4)

- Sentences that do not need to be interpreted: __null__
- No corresponding English segments: __null__

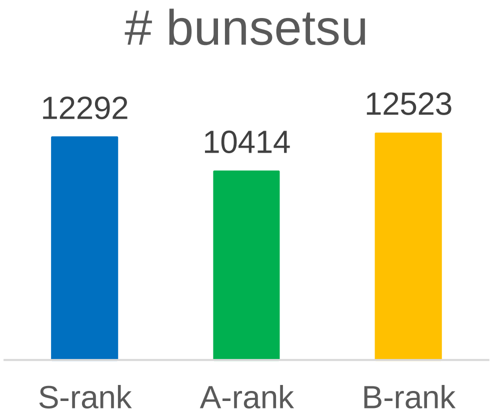
```
EN_0019 71325 72688 (Laughter)
JA_0000 XXXXX XXXXX __null__

EN_0000 XXXXX XXXXX __null__
JA_0024 72267 72882 突然

EN_0020 72688 74839 It’s a curious shape for a normal condition.
JA_0025 73687 74417 これは面白い
JA_0026 74820 77850 形ですね、{||こどげ}普通の状況と考えるには。
```

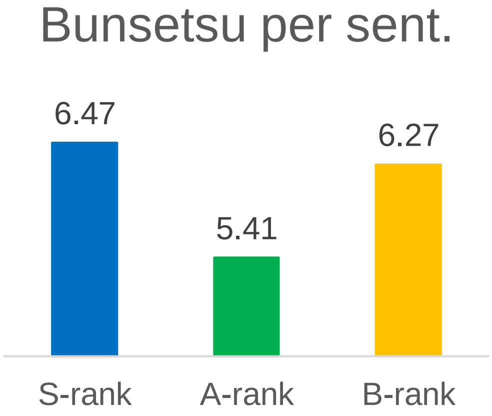
Outline

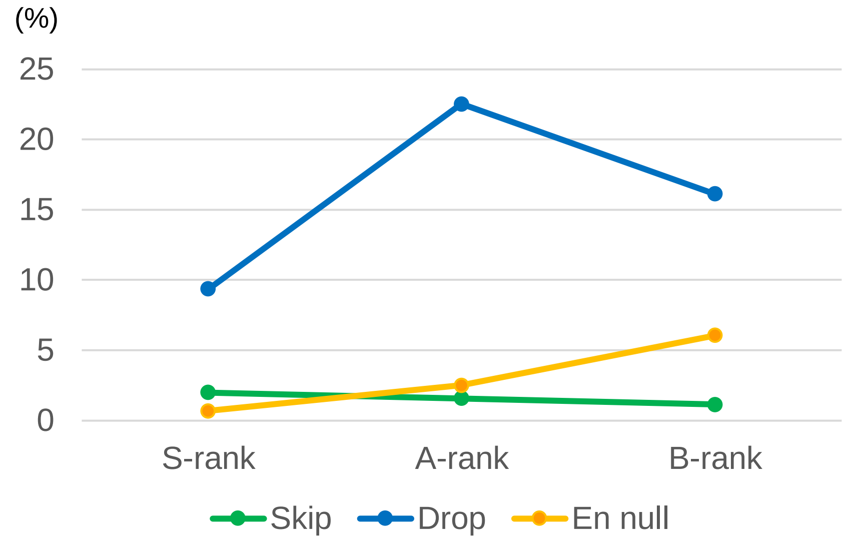
- Introduction
- Corpus construction
- Corpus analysis
- Results
  - Latency
  - Quality
  - Human evaluation
  - Word order
- Conclusion

Overall trend

- Higher ranked interpreters had higher quality
  - SI length: B > S > A
  - En null: B > A > S
  - Drop: A > B > S
  - SI length per sentence
    - S = B > A ($p < 0.001$)
  - Skip: not significant







Latency metrics

- Ear-Voice Span (EVS)
  - The lag between the original utterances and the corresponding SIs

```
EN_0006 31877 33347 But I like to say,
EN_0007 33347 36951 okay, let's look at the modern human condition.
JA_0005 32643 33450 でも私は、
JA_0006 35761 36901 (F え)(F この一)
JA_0007 37951 39960 現代の人間の状態と言うのを見てみましょう。
```

EVS_{start}: the lag at the beginning

$$32643 - 31877 = 766$$

EVS_{end}: the lag at the end

$$39960 - 36951 = 3009$$

  - The following cases were excluded from the analyses:

```
EN_0009 38643 41045 This is the normal way for things to be.
JA_0012 XXXXX 42236 これが
JA_0013 42465 44085 極普通の状況です。
```

Start/end times is unavailable

```
EN_0033 115144 118838 And Ed Witten unleashed the second
                      superstring revolution.
JA_0044 117516 118421 エドウイットの
```

EVS_{end} becomes negative
(= quit interpreting in the middle)

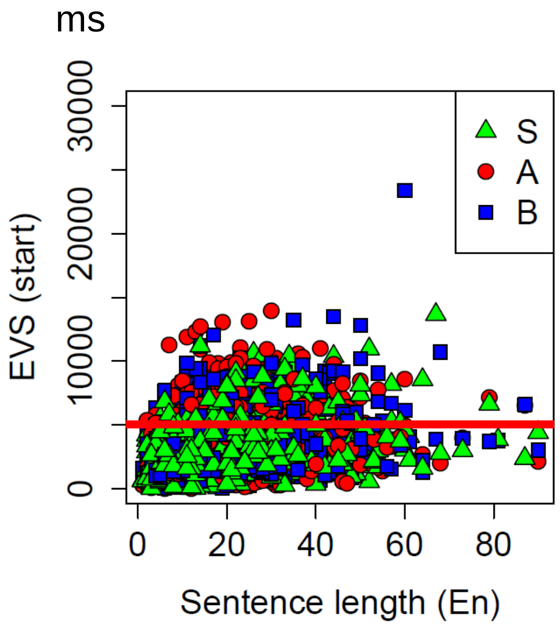
$$118421 - 118838 = -417$$

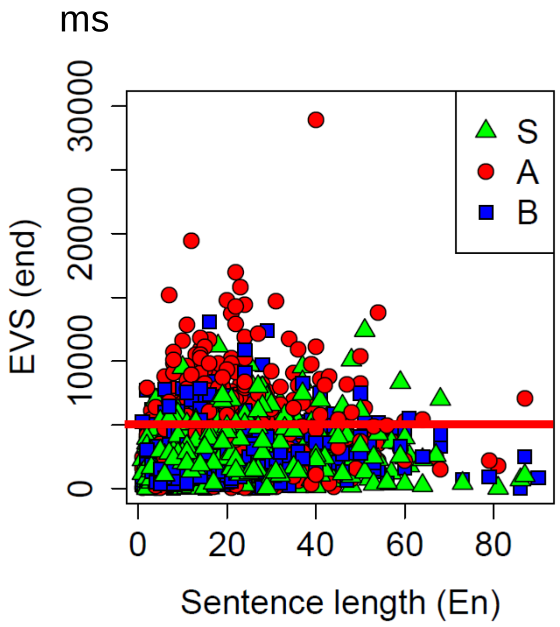
Latency results

- EVS results
  - A-rank > B-rank > S-rank
  - Ranged 2-4 seconds
  - Consistent with previous studies

Interpreter	Start	End
S-rank	2.95	2.48
A-rank	**3.57**	**3.89**
B-rank	3.46	2.79

- Found large EVS (> 5 seconds)
  - Sentence length of the original speech did not affect EVS
  - $r = 0.2584, 0.1206$
  - Did not match [Lee 2002]





Why some EVS took large values?

- EVS_{start}
  - Sometimes did not interpret the earlier part of the sentence

    (En) A week later, Ping was discovered in the apartment alongside the body of her owner, and the vacuum had been running the entire time.

    (A-rank) そしてずっと掃除機がオンになったまま残されていたんですけれども

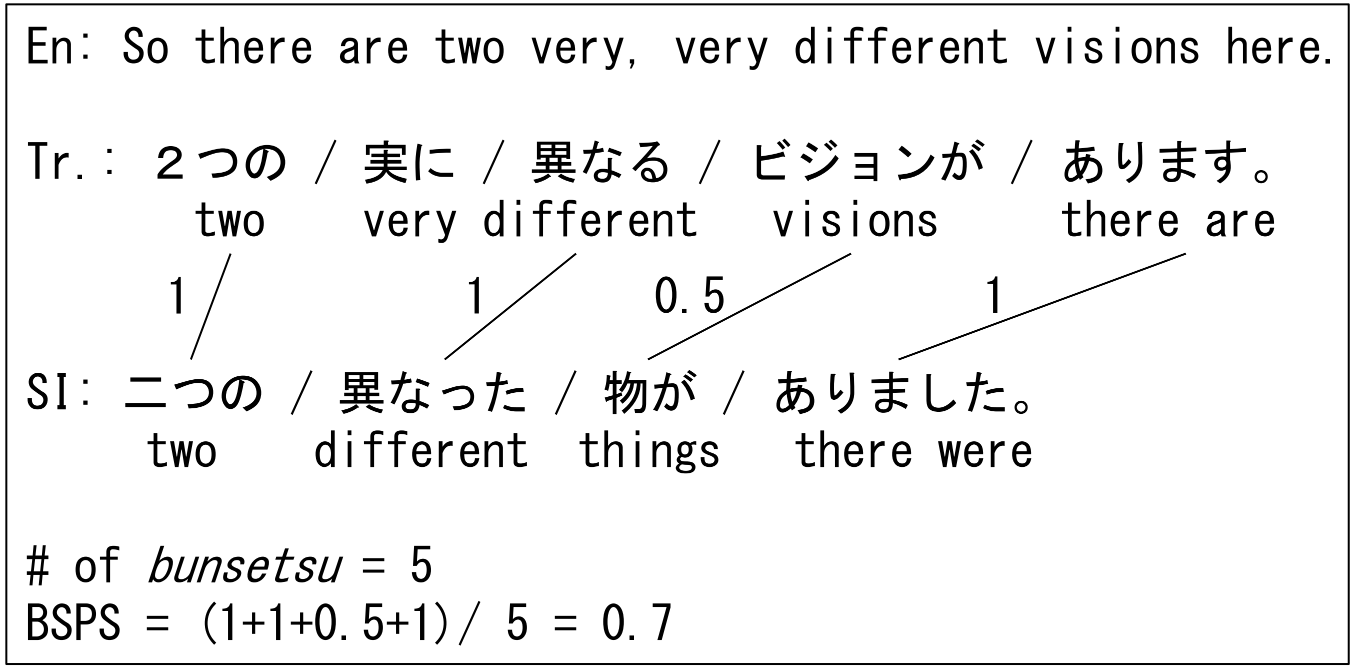
    [And the vacuum had been running the entire time.]

- EVS_{end}
  - Clang to the sentence though the next sentence started
  - (A-rank) top 10% of large EVS_{end}
    - 56.68% of their subsequent sentences were __drop__

→ Large EVS_{end} seemed to negatively impact the SI quality

Quality metrics

- BERTScore [Zhang+ 2019]
  - Based on contextualized subword embeddings
  - Expected to capture meaning
- Bunsetsu-level semantic preservation score (BSPS) [Ino+ 2008]
  - – Evaluate the faithfulness of the SIs against the translations
    - Calculated on 3 talks (Ale, Nic, Lau)



En: So there are two very, very different visions here.

Tr.: ２つの / 実に / 異なる / ビジョンが / あります。

SI: 二つの / 異なった / 物が / ありました。

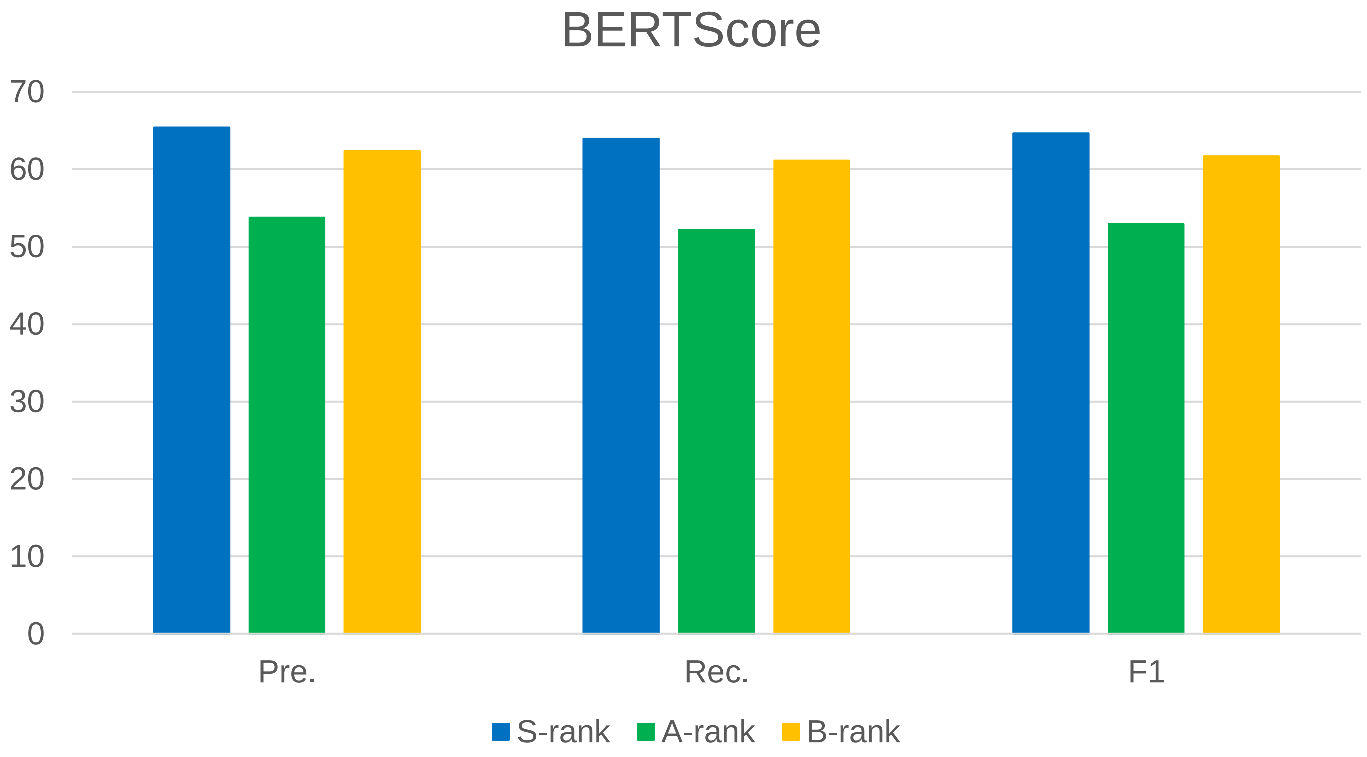
# of *bunsetsu* = 5

BSPS = (1+1+0.5+1)/ 5 = 0.7

Quality: BERTScore

- Precision > Recall
  - Strategies such as summarization and generalization

- S > B > A ($p < 0.05$)
  - High drop ratio of the A-rank interpreter



Quality: BERTScore

- Good example: F1 = 0.8325

(En) We did this experiment for real.
(Ref) 実際にこの実験を行ってみました。
(A-rank) これを実際にしました。［Did this for real.］

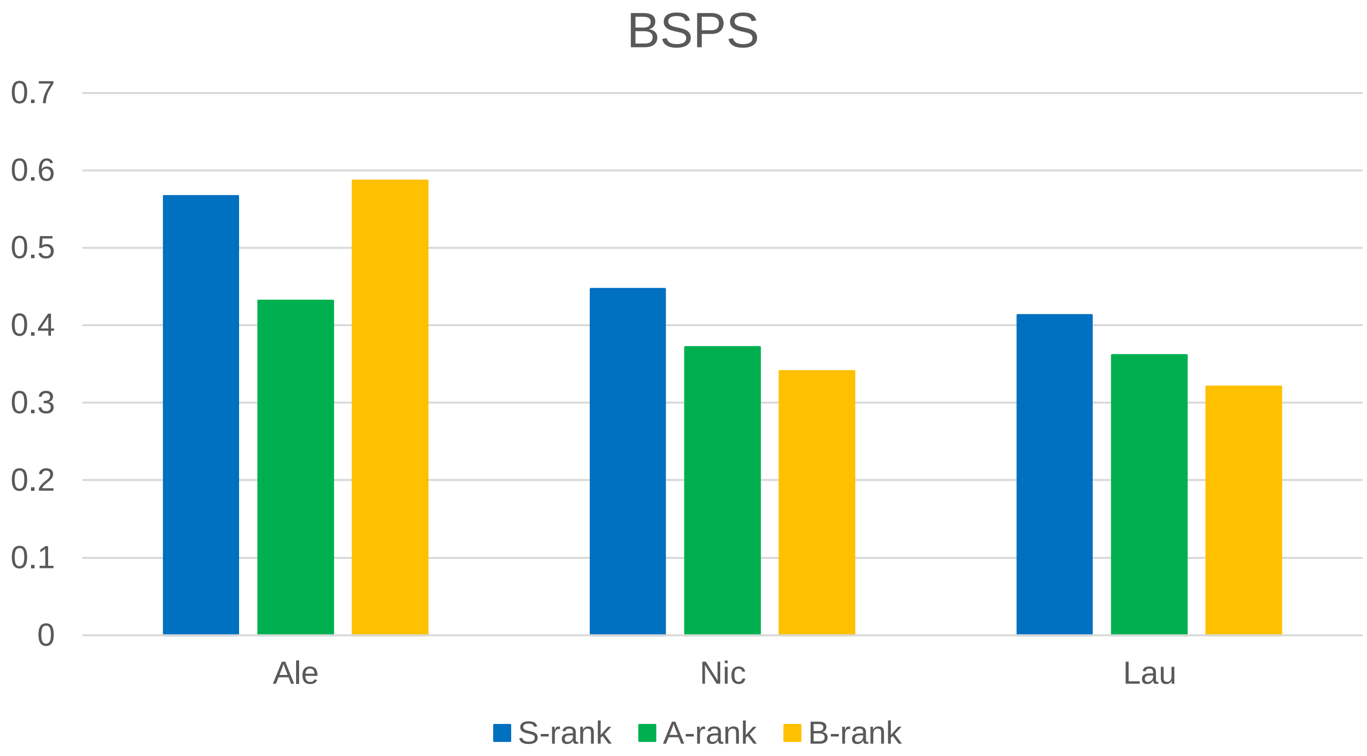
- Bad example: F1 = 0.5519

(En) We can all think of some examples, right?
(Ref) 例を挙げる事ができると思います。
(S-rank) 例えば、［For example,］

Quality: BSPS

- Calculated for the three talks
  - `Ale` (easy), `Nic` (medium), and `Lau` (difficult)

  - S > A > B (except `Ale`)
    - `Ale`
      - Low drop/en null ratio of the B-rank interpreter
      - Matched the human evaluation results



Relationship between latency and quality

- Large EVS_{start} hurt the quality of the sentence being processed

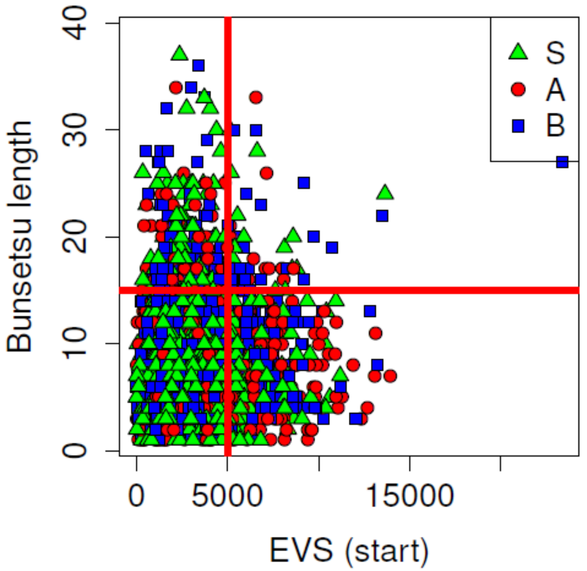
  → Few SIs with more than 15 bunsetsus

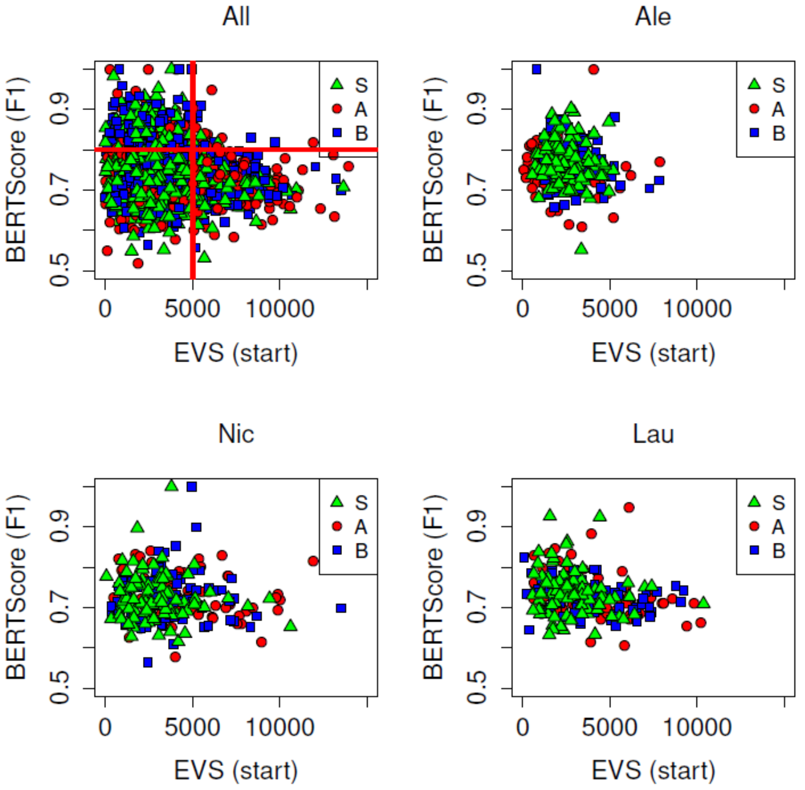
  → Low BERTScore/BSPS

- SD of EVS_{start}

Ale	Nic	Lau
1.33	2.25	2.16

S-rank	A-rank	B-rank
1.06	1.68	2.16

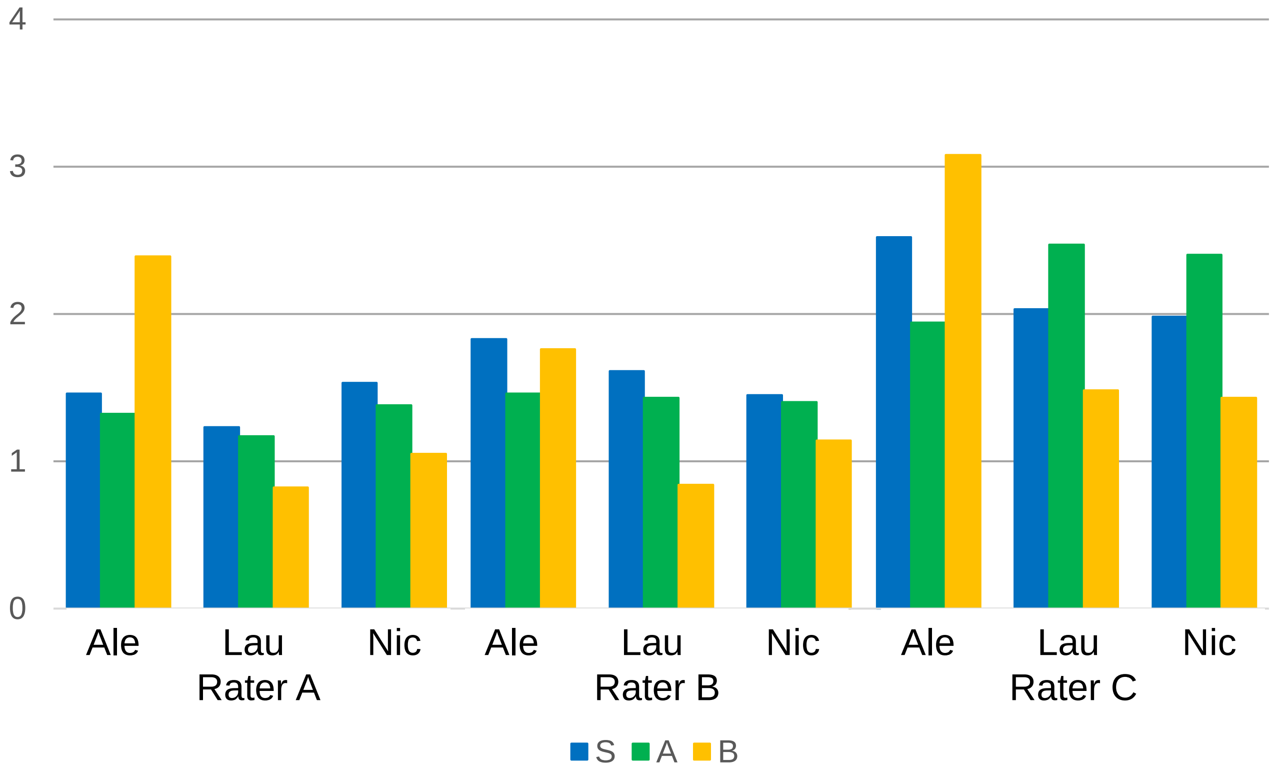




Human evaluation

- Subjectively evaluated by 3 professional translators
  - Focused on only faithfulness (i.e., *not* delay, grammaticality, etc.)
  - 1 (incomprehensible), 2 (poor), 3 (minor errors), 4 (acceptable)
- Higher ranked interpreters received high scores
  - Except B-rank on Ale
  - Individual differences

    e.g., background knowledge
- Scores were low



Why human evaluation scores were low?

- Translators were strict about the sentence structure in the source language

(En) People are motivated by the different values perhaps.

(A-rank) 人のモチベーションは／違う物によって／起こってきます

[People's motivation / by different things / is raised.]

(Scores) 1, 3, 2

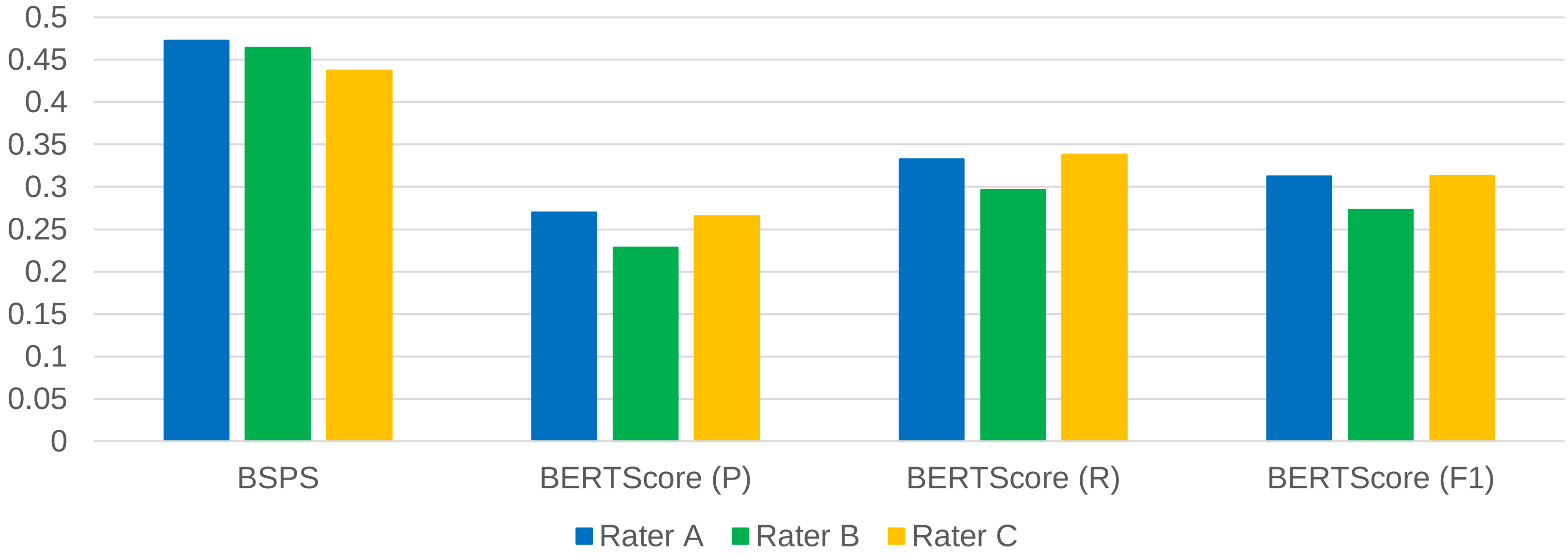
→ Future work: Human evaluation with simultaneous interpreters

Human evaluation

- Correlation between human evaluations and quality metrics
  - (Overall) BSPS > BERTScore

⇔ individual talks

- Nic_S: BSPS (≈ 0.3) < BERTScore (≈ 0.45)



Word order metric

- Kendall’s K distance [Kendall 1938]
  - Ranges [0, 1]
  - 0 if the two lists are identical
  - 1 if one list is the reverse of the other

$$d_K(\pi, \sigma) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} z_{ij}}{n(n-1/2)}$$

$$\text{where } z_{ij} = \begin{cases} 1 & \text{if } \pi(i) < \pi(j) \text{ and } \sigma(i) > \ \sigma(j) \\ 0 & \text{otherwise} \end{cases}$$

Word order

- No clear differences due to interpreter ranks
- Example: K=0.75

En	That’s a huge problem if you think about, especially, an economy like Switzerland, which relies so much on the trust put into its financial industry.
Ref	金融業界の／信用に／大きく依存する／スイスのような／経済を／考えると／これは巨大な問題です。 [put into financial industry / the trust / which relies so much on / like Switzerland / an economy / if you think about / that’s a huge problem]
B-rank	これは、大きな問題です。／特に、／スイスの様な／経済を／考えてみると／そうでしょう。／金融業界に対する／信頼／によって成り立っている／国だからです。 [that’s a huge problem / especially / like Switzerland / an economy / if you think about / it’s true / on its financial industry / the trust / based on / it’s a country]

Conclusion

- Constructing a new large-scale English↔Japanese SI corpus
  - Contains SI data from 3 interpreters with different amounts of experience (S-, A-, and B-ranks)

- Analyzed the SI data among interpreter ranks and against offline translations
  - Interpreters with more experience controlled the latency and quality better
  - Large latency hurt the SI quality