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
{doi.kosuke.de8, sudoh, s-nakamura}@is.naist.jp

Abstract

This paper describes the construction of a new large-scale English-Japanese Simultaneous Interpretation (SI) corpus and presents the results of its analysis. A portion of the corpus contains SI data from three interpreters with different amounts of experience. Some of the SI data were manually aligned with the source speeches at the sentence level. Their latency, quality, and word order aspects were compared among the SI data themselves as well as against offline translations. The results showed that (1) interpreters with more experience controlled the latency and quality better, and (2) large latency hurt the SI quality.

1 Introduction

Simultaneous interpretation (SI) is a task of translating speech from a source language into a target language in real-time. Unlike consecutive translation, where the translation is done after the speaker pauses, in SI the translation process starts while the speaker is still talking. With recent developments in machine translation and speech processing, various studies have been conducted aiming at automatic speech translation (Pino et al., 2020; Wu et al., 2020; Inaguma et al., 2021; Bahar et al., 2021), including SI (Oda et al., 2014; Zheng et al., 2019; Arivazhagan et al., 2019; Zhang et al., 2020; Nguyen et al., 2021), based on speech corpora.

Existing speech corpora can be classified into *Speech Translation* corpora or *Simultaneous Interpretation* corpora, as defined by Zhang et al. (2021). Table 1 lists publicly-available SI corpora. Although a large number of *Speech Translation* corpora have been published, the number of SI corpora remains very limited. Both types of corpora are comprised of audio data and their corresponding translations, although how the translations are generated is different. For *Speech Translation* corpora, a translation is based on complete audio data or transcripts; for SI corpora, human interpreters actually do SI. SI corpora are useful not only for the construction of automatic SI systems but also for translation studies.

Corpora	Language	Hours
Toyama et al. (2004)	En↔Jp	182
Paulik and Waibel (2009)	En↔Es	217
Shimizu et al. (2014)	En↔Jp	22
Zhang et al. (2021)	Zh→En	68
Ours	En↔Jp	304.5

Table 1: Existing SI corpora and ours

To facilitate research in the field of SI, we are constructing a new large-scale English↔Japanese SI corpus[1]. We recorded the SIs of lectures and press conferences and amassed over 300 hours of such data. Some lectures have SI data generated by three interpreters with different amounts of experience, as in Shimizu et al. (2014), which enables comparisons of SI differences based on experience.

In this paper, we describe the construction of a new corpus and present the results of its analysis. Its design follows the framework of Shimizu et al. (2014). The analysis was conducted on a subset of lectures that have SI data from three interpreters. In some parts of the data, the source speech and the SI data were manually aligned at the sentence level to compare the following properties: latency, quality, and word order, all of which are typically investigated in translation studies. We compared those SI data among them as well as against translations that are generated offline. Importantly, we adopt an automatic metric and a manual analysis to evaluate the SI quality.

[1]A part of the corpus is available at https://dsc-nlp.naist.jp/data/NAIST-SIC/

2 Related Work

2.1 Existing SI Corpora

Despite their usefulness, the number of SI corpora is very limited (Table 1). The Simultaneous Interpretation Database (SIDB) is an English↔Japanese SI corpus, which consists of over 180 hours of recordings, including both monologues (lectures) and dialogues (travel conversations).

Shimizu et al. (2014) also constructed an English↔Japanese SI corpus. It is a relatively small corpus (22 hours), and has the following two notable features: (1) all the speeches have SI data from three interpreters with different amounts of experience; and (2) offline translations are available for some of the speeches. The features allow comparisons among the SI data themselves as well as with the translation data.

In language pairs other than English↔Japanese, Paulik and Waibel (2009) developed an SI system using SI data collected from European Parliament Plenary Sessions (EPPS), which are broadcast live by satellite in the various official languages of the European Union. Zhang et al. (2021) proposed the first large-scale Chinese→English *Speech Translation* and SI corpus.

2.2 Translation Studies

In translation studies, SI characteristics have typically been investigated from the aspects of latency, quality, and word order. For evaluating latency by human interpreters, Ear-Voice Span (EVS) is commonly used as a metric. EVS denotes the lag between the original utterances and the corresponding SIs.

The analysis of quality often relies on a manual evaluation of the corpus data (Fantinuoli and Prandi, 2021). Ino and Kawahara (2008), for example, investigated SI faithfulness based on manual annotation of the data. SI aims to translate a source speech with low latency and high quality, where the two factors are in a trade-off relationship. However, previous studies (*e.g.,* Lee, 2002) argued that a longer latency negatively affects SI quality.

Word order has also been intensively studied in the field. Recent research by Cai et al. (2020) demonstrated a statistical study based on SIDB and compared word order between translation and SI.

3 Corpus Construction

3.1 Material

Our corpus consists of the SIs of four kinds of materials. For the English→Japanese direction, the interpreters interpreted TED talks[2].

TED: TED offers short talks on various topics from science to culture. The videos of the talks are available on its website. More importantly, TED talks have been manually transcribed and translated by volunteers, and Japanese translations (*i.e.,* subtitles) are available for many talks.

For the Japanese→English direction, the interpreters interpret speech from the following materials.

TEDx: TEDx is an event where local speakers present topics to local audiences. The events are held under a license granted by TED, and the talks follow the format of TED talks. The videos are available on YouTube as well as on the TED website.

CSJ: The Corpus of Spontaneous Japanese (Maekawa, 2003) consists of academic lectures and speeches on everyday topics. It contains audio data and their transcripts with linguistic annotations.

JNPC: The Japan National Press Club (JNPC) annually organizes about 200 press conferences involving Japanese and foreign guest speakers from politicians to business representatives. The press conferences are video-recorded and available online[3]. For some of them, transcripts are provided on its website.

3.2 Recording

Professional simultaneous interpreters with different amounts of experience participated in the recordings. Each interpreter was assigned a rank based on length of experience, as in Shimizu et al. (2014) (Table 2). The recordings were made from 2018 to 2020.

Interpreters wore a headset and interpreted speech while watching video on a computer. They only listened to the audio when interpreting the CSJ speech because no videos were available. The interpreters were provided in advance documents related to the speech to improve the SI quality. In

[2] https://www.ted.com/
[3] https://www.jnpc.or.jp/

Amount of experience	Rank
15 years	S-rank
4 years	A-rank
1 years	B-rank

Table 2: Ranks of simultaneous interpreters

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

Table 3: Recorded hours of our SI corpus. Figures with asterisk (*) indicate parts with SI data generated by three interpreters with different amounts of experience (*i.e.,* 4 hours × 3 interpreters).

fact, related information or materials (*e.g.,* presentation slides) are usually provided to them in their actual work. The following are the details of the documents given in our recording procedures:

- TED, TEDx (2018): Summary of talk; referenceable during SI.
- TED (2019-): English transcripts from TED website; *not* referenceable during SI.
- TEDx (2019-): Japanese subtitles generated by YouTube; *not* referenceable during SI.
- CSJ: 10% summary of Japanese transcripts; referenceable during SI.
- JNPC: No documents provided.

Table 3 shows the details of the recorded hours of our corpus. In spontaneous speech, sentence boundaries are ambiguous, and it is difficult to provide the number of sentences included in our corpus. A total of four hours of TED and TEDx recorded in 2018 were interpreted by interpreters from all three ranks (4 hours × 3 interpreters = 12 hours; marked with asterisk). The other talks were interpreted by either an S-rank or an A-rank interpreter. About half of the recorded SIs have been manually transcribed. The whole corpus consists of SIs of more than 1200 talks. The average talk length by materials is the following: TED 11.20 minutes, TEDx 15.85 minutes, CSJ 13.55 minutes, and JNPC 84.33 minutes.

```
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。
```

Figure 1: Example of an SI transcript: Preceding each utterance, IDs and start/end times are annotated. Some discourse tags are used: F: fillers, (?): unintelligible, ⟨H⟩: prolongations.

4 Corpus Analyses

4.1 Data

The English→Japanese SI data from 14 TED talks were analyzed based on three properties: latency, quality, and word order. The talks were a subset of 12 hours of recordings of SI data from interpreters of each rank (see Table 3).

The SI data were aligned to the source speech based on segments. A transcript example is shown in Fig. 1. Each segment is annotated with an ID, start/end times, and discourse tags (*e.g.,* fillers, slips of the tongue, pauses). A segment does not necessarily correspond to a sentence.

In addition to the SI data, offline translation data (*i.e.,* Japanese subtitles) were used to examine the SI quality and word order. Disfluencies in the SI data were removed with the help of discourse tags. Then the SI and translation data were automatically divided into *bunsetsus*[4] using the Juman++ Japanese morphological analyzer[5] (Morita et al., 2015) and the KNP parser (Kawahara and Kurohashi, 2006).

4.2 Sentence Alignment

For subsequent corpus analyses, the SI data of 14 talks were manually aligned at the sentence level with the source speeches by the first author to fairly compare the data of the interpreters of each rank. Since the segments in the SI transcripts were based on the interpreters' utterances, they did not necessarily match among the interpreters. Thus, we gave sentence alignments based on the sentences of the English transcripts segmented using the following rules:

[4] A *bunsetsu* is a basic unit of dependency in Japanese that consists of one or more content words and the following zero or more function words (Kawahara and Kurohashi, 2006).

[5] We used Juman++ ver.1.02 rather than the development version of Juman++ V2 (Tolmachev et al., 2018).

```
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。
```

Figure 2: Example of sentence-level alignment

- segments ending with a period (.) or a period + a double quotation mark (.")
- segments ending with a question mark (?) or a question mark + a double quotation mark (?")
- segments ending with a closed parenthesis

Japanese segments were aligned to English sentences by the following rules[6]:

- Words/phrases that are not interpreted: ignored.
- Sentences that are not interpreted: marked as __drop__ in Japanese segments.
- Sentences that are not interpreted *intentionally*: marked as __skip__ in Japanese segments. (*e.g.,* Thank you.)
- Sentences that do not need to be interpreted: marked as __null__ in Japanese segments. (*e.g.,* (Laughter))
- No corresponding English sentence: add __null__ to English segments.
- Japanese segments that correspond to multiple English sentences: divide where it corresponds to the boundary of English sentences. Mark XXXXX for end/start times of Japanese segments.
- English segments that consist of multiple sentences: divide at sentence boundary. Mark XXXXX for end/start times of segments.

An example of the data aligned at the sentence level is shown in Fig. 2. Each sentence is delimited by one blank line.

4.3 Metrics

Latency: As a latency metric, EVS was calculated for each sentence. Since the start/end times of the transcribed speech segments are available in our data, we separately calculated EVS at the beginning and the end of a sentence[7]:

$$EVS_{start} = start\ time_{JP} - start\ time_{EN}$$
$$EVS_{end} = end\ time_{JP} - end\ time_{EN}.$$

However, we failed to calculate EVS in some sentences because some segments were divided into multiple segments during the sentence-level alignment, and the start/end times were unavailable. Furthermore, EVS at the end of sentences can become negative if the interpreter quit interpreting in the middle of a sentence. These cases were excluded from our analyses.

Quality: To evaluate the SI quality, we calculated two metrics[8].

The first one was BERTScore (Zhang et al., 2019), which is also used to evaluate machine translations (*e.g.,* Edunov et al., 2020). It is based on contextualized subword embeddings and is expected to capture meanings rather than surface forms like BLEU (Papineni et al., 2002). It would be appropriate for evaluating the aspects of SIs used by interpreters, including anticipation, summarization, and generalization. BERTScores were calculated between SIs (candidates) and offline translations (references) for each sentence.

The other quality metric was the *bunsetsu-level semantic preservation score* (BSPS), which evaluated the faithfulness of the SIs against the translations. An example is shown in Fig. 3. Similar to Ino and Kawahara (2008), each *bunsetsu* that appeared in the translation was considered a unit of ideas. Then we counted the number of *bunsetsus* in the SI that conveyed the ideas. If a *bunsetsu* in the SI successfully conveyed its idea in the translation, it got one point. If the *bunsetsu* in the SI partially conveyed an idea, it got half a point. The BSPS for a given sentence was calculated by adding the points and dividing by the number of ideas in the translation.

To calculate BSPS, we manually created *bunsetsu* level alignments for three talks, which were selected based on the following procedures:

- Assign a score of 1-3 to the SI data (14 talks $\times$ 3 interpreters) based on the overall quality.

[6]Subjectively judged by the authors, except for the boundaries of the English sentences.

[7]Due to the limitations of our data, we calculated a simplified EVS, which was different from that in previous studies.

[8]We focused on faithfulness in this paper, although other factors may affect SI quality (*e.g.,* grammaticality, delivery).

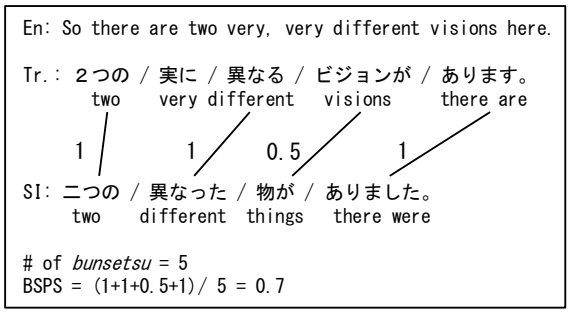

Figure 3: Example of calculating BSPS

- Calculate the average for each talk and assign a label of `high`, `mid`, or `low`.
- Choose one talk from each label.

The talks labeled `high` are those that are easy to interpret, and the talks labeled `low` are difficult. We chose three talks: AlexanderWagner_2016X (`Ale`), NickBostrom_2015 (`Nic`), and LaurelBraitman_2014S (`Lau`), for easy, medium, and difficult levels.

Word Order: To examine the differences in word order between SI and offline translation, we computed Kendall's K distance (Kendall, 1938), ranging [0, 1], and equaling 0 if the two lists are identical and 1 if one list is the reverse of the other. The metric, which captures pairwise disagreements between two lists, can measure the degree of reordering. K was calculated based on the *bunsetsu* level alignment shown in Fig. 3.

4.4 Results

4.4.1 Overall Trend

Table 4 provides basic statistics for the SI data of the 14 TED talks. B-rank interpreters produced the longest SIs (`# Bunsetsu`), but they frequently added something that the original speaker did not say (`en null`). The ratio of `en null` decreased as the amount of experience became longer. In addition, the ratio of `drop` for S-rank interpreters (9.22) was lower than that for the others (A-rank: 21.67, and B-rank: 15.69). These results suggest that the SI generated by higher ranked interpreters tends to have higher overall quality.

At the sentence level, S-rank interpreters produced the most *bunsetsus* (`Bunsetsu. per sent.`). A one-way ANOVA detected significant differences among groups ($F(2, 5818) = 21.881, p < 0.001$), and the following Tukey's test showed that S- and B-rank interpreters produced significantly more *bunsetsus* than A-rank interpreters ($p < 0.001$). Although the difference between S- and B-ranks is not significant, the results suggest that interpreters with more experience also did better at the sentence level. This point is discussed below in Section 4.4.3.

In Table 4, we can also see that higher ranked interpreters tended to have higher `skip` ratios. However, the differences among the groups were not statistically significant based on a one-way ANOVA ($F(2, 39) = 0.5172, p = 0.6002$).

4.4.2 Latency

Table 5 compares the latency measured by EVS. A-rank interpreters had the largest latency both at the beginning and at the end of sentences, followed by B- and S-rank interpreters. The amount of latency ranged from 2 to 4 seconds, which was consistent with the majority of previous studies (see Robbe, 2019).

However, a relatively great number of EVS took large values (> 5 seconds). The relationship between EVS and sentence length in the source language is shown in Fig. 4. As Pearson's correlation coefficient indicates ($r = 0.2584, 0.1206$, respectively), sentence length in the target language did not seem to affect EVS, which did not match the results reported in Lee (2002).

EVS_{start} became large because interpreters sometimes did not interpret the earlier part of the sentence, as in this example:

> (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.]

The EVS_{end} results suggest that S- and B-rank interpreters might wrap up the sentence to a certain extent when the next sentence started, but A-rank interpreters might cling to the sentence, resulting in larger EVS_{end}. A large EVS_{end} seemed to negatively impact the SI of the subsequent sentence, as reported in Lee (2002). Focusing on the top 10% of sentences whose EVS_{end} was large ($N = 187$), 56.68% of their subsequent sentences were not interpreted at all (*i.e.,* `drop`) by A-rank interpreters.

Interpreter	# Seg.	# Sent.	# Bunsetsu	Bunsetsu. per sent.	Skip (%)	Drop (%)	En null (%)
S-rank	2750	1902	12292	**6.47**	**2.00**	9.36	0.68
A-rank	2609	1948	10414	5.41	1.58	**22.54**	2.50
B-rank	**3077**	**1998**	**12523**	6.27	1.13	16.13	**6.05**
avg.	2812	1949.33	11743.00	6.05	1.57	16.01	3.08

Table 4: Comparison of SI data among interpreters with different amounts of experience

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

Table 5: Comparison of EVS (seconds) among interpreters with different amounts of experience. Figures are averages of each sentence.

Interpreter	Pre.	Rec.	F1
S-rank	**0.6544**	**0.6396**	**0.6465**
A-rank	0.5374	0.5221	0.5292
B-rank	0.6238	0.6115	0.6171

Table 6: Comparison of BERTScores among interpreters with different amounts of experience. Scores are averages of each sentence, where 0 is assigned to `drop` and `skip`.

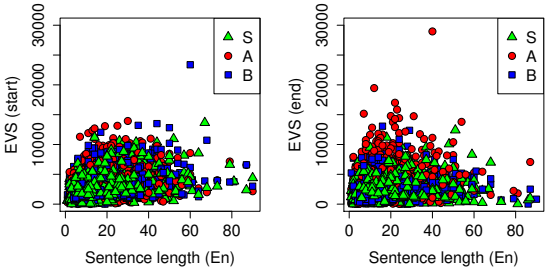

Figure 4: Relationship between EVS and sentence length of original speech

4.4.3 Quality

BERTScore: The quality of the SI data measured by BERTScore is shown in Table 6. Precision was higher than Recall in all three interpreter ranks. The results match our intuition because simultaneous interpreters sometimes summarize or generalize the content of the original speech to handle latency, and not all the content is interpreted. BERTScore captured the quality of SI well in the following example:

> (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.]

The F1 score of the example was 0.8325. Although the wording that corresponds with "did" is different between the translation (Ref) and the interpretation, BERTScore captured the similarity of the meaning. On the other hand, as shown in the next example, BERTScore did not always do well, especially when interpreters used a strategy:

> (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.]

The F1 score of the example was 0.5519. The interpreter adopted a strategy (summarization) and conveyed the core ideas of the original utterance, although BERTScore struggled to capture them.

Comparing the three interpreter ranks, S-rank interpreters achieved the highest scores in Precision, Recall, and F1. A one-way ANOVA detected significant differences among groups ($F(2, 5045) = 65.802, 70.095, 68.386$ for Precision, Recall, and F1, $p < 0.001$), and the following Tukey's test showed that the differences among all the groups were significant ($p < 0.05$). The scores of the A-rank interpreters were probably lower than those of B-rank interpreters because of the high `drop` ratio.

***Bunsetsu-level Semantic Preservation Score*:** BSPS was calculated for the three talks, `Ale` (easy), `Nic` (medium), and `Lau` (difficult). The results in Table 7 indicate that the higher ranked interpreters achieved higher BSPS, except for `Ale`. In fact, the low ratio of `drop` and `en null` (8.33 and 0.00) suggest that the B-rank interpreter did well on `Ale`, which matched the human evaluation results. One of the human evaluators remarked that key words such as proper nouns were well translated or ap-

Talk	Interpreter	BSPS
Ale	S-rank	0.5671
(easy)	A-rank	0.4316
	B-rank	**0.5871**
Nic	S-rank	**0.4471**
(medium)	A-rank	0.3715
	B-rank	0.3411
Lau	S-rank	**0.4130**
(difficult)	A-rank	0.3618
	B-rank	0.3207

Table 7: Comparison of BSPS among three talks and interpreter's rank

propriately rephrased to corresponding Japanese words.

The BSPS results imply that higher ranked interpreters generated better SIs at the sentence level. The metric captured how many ideas, which were presented in the original speech, were actually covered in each sentence of the SIs. S-rank interpreters produced the most *bunsetsus* per sentence (Table 4), probably because they reproduced more of the ideas presented in the original speech.

Relationship between latency and quality: Since previous studies have shown that higher latency damages quality (*e.g.,* Lee, 2002), we investigated the relationship between them based on EVS_{start}. In Section 4.4.2, the negative effect of a large EVS_{end} on the following sentence was discussed; in this section, we examine whether a large EVS_{start} hurts the quality of the sentence being processed.

Fig. 5 shows the relationship between EVS_{start} and the number of *bunsetsus* in SIs. When the latency increased ($>$ 5 seconds), few SIs had large numbers of ($>$ 15) *bunsetsu*. The large EVS_{start} indicated that the original sentence was long, which expected a longer SI. A similar tendency was found for BERTScore and BSPS. From Figs. 6 and 7, SIs with a large EVS_{start} tended to get low scores.

The relationship between EVS_{start} and the quality metrics of `Ale`, `Nic`, and `Lau` is shown in Figs. 6 and 7. When the talk was easy to interpret (`Ale`), the standard deviation was smaller than the other talks (`Ale`= 1.33, `Nic`= 2.25, `Lau`= 2.16). Furthermore, the S-rank interpreters' standard deviation was smaller than that of the others (*e.g.,* S= 1.06, A= 1.68, B= 1.27 for `Ale`).

The above results suggest that a large EVS_{start} negatively affected the quality of the sentence being processed.

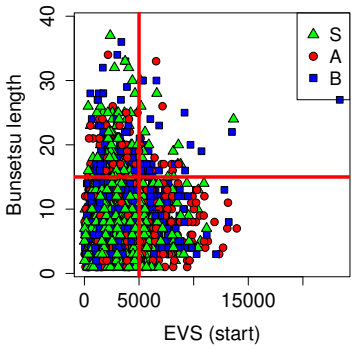

Figure 5: Relationship between EVS_{start} and the number of *bunsetsus* in SIs

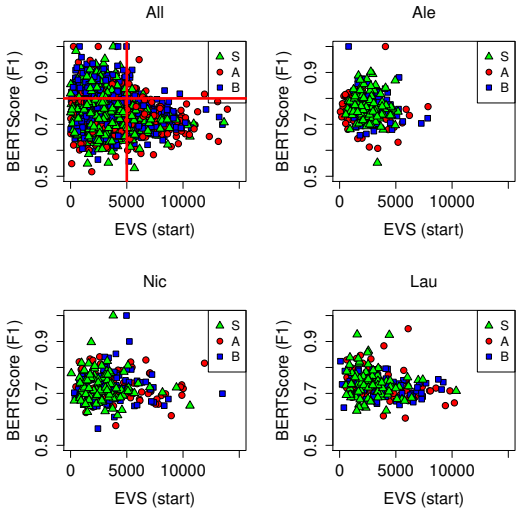

Figure 6: Relationship between EVS_{start} and BERTScore (F1)

4.5 Human Evaluation

The quality of the SI data was further examined through human evaluations. Three professional translators (*i.e., not* interpreters) subjectively evaluated the faithfulness of each sentence on a scale of 1 (incomprehensible), 2 (poor), 3 (minor errors), and 4 (acceptable). Table 8 shows that higher ranked interpreters received higher scores, which matched the BERTScore and BSPS results. The B-rank interpreter interpreted `Ale` well, which was mentioned in the overall comments by the translators. Individual differences of interpreters (*e.g.,* background knowledge) could affect the SI quality because not necessarily the same interpreters interpreted the three talks.

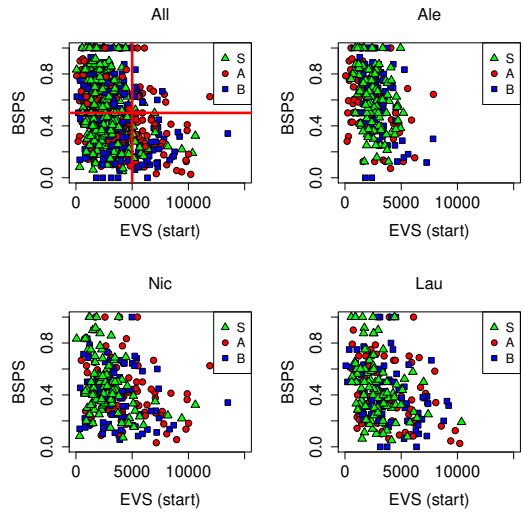

Figure 7: Relationship between EVS_{start} and BSPS

From Table 8, human evaluation scores were low, most often less than 2. One possible reason is that the translators were strict about the sentence structure in the source language, as in this example:

> (En) People are motivated by 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.]
> (Human evaluation scores) 1, 3, 2

The verb phrase (are motivated) was interpreted with a noun (motivation) to maintain the word order of the English sentence, while the rater A indicated the disagreement in his overall comment and assigned one point. Future work will involve human evaluation with simultaneous interpreters.

Pearson's correlation coefficient was calculated between the human evaluation scores and the two metrics. BSPS achieved relatively higher correlations with human judgments than BERTScore (Table 9). However, if the correlations were examined talk by talk, BSPS correlated poorly with the human evaluations in `Nic_S` (ranging around $r = 0.3$), and the correlation between BERTScore (F1) and human evaluation was relatively high (ranging around $r = 0.45$). Further research is needed on the behavior of the metrics.

4.5.1 Word Order

The differences in word order between the SI data and the offline translations measured by Kendall's K distance are shown in Table 10. Because of the difference between English (SVO and head-initial) and Japanese (SOV and head-final), the difference between SI and translation (*i.e.,* large K) suggests that the interpreters adopted a strategy of maintaining the word order of the source language. However, differences due to interpreter ranks were not clear, and we observed sentences with relatively large K (> 0.7).

An example is shown in Table 11, whose K was 0.75. In the translation (Ref), the word order was almost reversed from the English sentence, although the simultaneous interpreter successfully interpreted in the first-in-first-out manner. The example matched the word order patterns reported in Cai et al. (2020), who found that simultaneous interpreters often preferred maintaining the word order in the original speech when interpreting nominal modifiers and dependent clauses.

Talk_Rank	Rater A	Rater B	Rater C
Ale_S	1.46	**1.83**	2.52
Ale_A	1.32	1.46	1.94
Ale_B	**2.39**	1.76	**3.08**
Lau_S	**1.23**	**1.61**	2.03
Lau_A	1.17	1.43	**2.47**
Lau_B	0.82	0.84	1.48
Nic_S	**1.53**	**1.45**	1.98
Nic_A	1.38	1.40	**2.40**
Nic_B	1.05	1.14	1.43

Table 8: Comparison of subjective evaluations by three professional translators

Metric	Rater A	Rater B	Rater C
BSPS	0.4724	0.4640	0.4372
BERTScore (P)	0.2696	0.2281	0.2658
BERTScore (R)	0.3326	0.2966	0.3380
BERTScore (F1)	0.3125	0.2728	0.3131

Table 9: Correlation between human evaluations and quality metrics

Interpreter	Ale	Nic	Lau
S-rank	0.1118	0.0987	0.0832
A-rank	0.1467	0.1023	0.0767
B-rank	0.1347	0.0796	0.0985

Table 10: Comparison of Kendall's K distance among three talks and interpreter ranks

Source	Example
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]

Table 11: Example of interpretations with large K

5 Conclusion

We described the construction of a new large-scale English↔Japanese SI corpus that contains SI data generated by simultaneous interpreters with different amounts of experience (S-, A-, and B-ranks) from identical lectures. Focusing on latency, quality, and word order, we compared the SI data among interpreter ranks and against offline translations. The S-rank interpreters controlled latency and quality better than the other two ranks. We strongly believe that our new corpus will be a useful resource for further research in translation studies and for the construction of automatic SI systems.

Acknowledgments

This research was supported by JSPS KAKENHI Grant Number JP17H06101.

References

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019. Monotonic infinite lookback attention for simultaneous machine translation. In *Proceedings of ACL*, pages 1313–1323.

Parnia Bahar, Tobias Bieschke, Ralf Schlüter, and Hermann Ney. 2021. Tight integrated end-to-end training for cascaded speech translation. *arXiv*, arXiv:2011.12167.

Zhongxi Cai, Koichiro Ryu, and Shigeki Matsubara. 2020. What affects the word order of target language in simultaneous interpretation. In *Proceedings of IALP*, pages 135–140.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of ACL*, pages 2836–2846.

Claudio Fantinuoli and Bianca Prandi. 2021. Towards the evaluation of simultaneous speech translation from a communicative perspective. *arXiv*, arXiv:2103.08364.

Hirofumi Inaguma, Tatsuya Kawahara, and Shinji Watanabe. 2021. Source and target bidirectional knowledge distillation for end-to-end speech translation. *arXiv*, arXiv:2104.06457.

Kinuyo Ino and Kiyoshi Kawahara. 2008. Comparative analysis of simultaneous mode and prepared mode in broadcast interpreting. *Interpreting and Translation Studies*, 8:37–55. (in Japanese).

Daisuke Kawahara and Sadao Kurohashi. 2006. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *Proceedings of HLT-NAACL*, pages 176–183.

Maurice Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Tae-Hyung Lee. 2002. Ear voice span in english into korean simultaneous interpretation. *Meta*, 47(4):596–606.

Kikuo Maekawa. 2003. Corpus of spontaneous japanese: Its design and evaluation. In *Proceedings of ISCA/IEEE Workshop on Spontaneous Speech*, pages 7–12.

Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. 2015. Morphological analysis for unsegmented languages using recurrent neural network language model. In *Proceedings of EMNLP*, pages 2292–2297.

Ha Nguyen, Yannick Estève, and Laurent Besacier. 2021. An empirical study of end-to-end simultaneous speech translation decoding strategies. *arXiv*, arXiv:2103.03233.

Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Optimizing segmentation strategies for simultaneous speech translation. In *Proceedings of ACL*, pages 551–556.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Matthias Paulik and Alex Waibel. 2009. Automatic translation from parallel speech: Simultaneous interpretation as mt training data. In *Proceedings of ASRU*, pages 496–501.

Juan Pino, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. 2020. Self-training for end-to-end speech translation. In *Proceedings of Interspeech*, pages 1476–1480.

Elisa Robbe. 2019. *Ear-voice span in simultaneous conference interpreting en-es and en-nl: A case study*. Doctoral dissertation, Ghent University.

Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Collection of a simultaneous translation corpus for comparative analysis. In *Proceedings of LREC*, pages 670–673.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of EMNLP*, pages 54–59.

Hitomi Toyama, Shigeki Matsubara, Koichiro Ryu, Nobuo Kawaguchi, and Yasuyoshi Inagaki. 2004. Ciair simultaneous interpretation corpus. In *Proceedings of Oriental COCOSDA*.

Anne Wu, Changhan Wang, Juan Pino, and Jiatao Gu. 2020. Self-Supervised Representations Improve End-to-End Speech Translation. In *Proc. Interspeech 2020*, pages 1491–1495.

Ruiqing Zhang, Xiyang Wang, Chuanqiang Zhang, Zhongjun He, Hua Wu, Zhi Li, Haifeng Wang, Ying Chen, and Qinfei Li. 2021. Bstc: A large-scale chinese-english speech translation dataset. *arXiv*, arXiv:2104.03575. Version 3.

Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2020. Learning adaptive segmentation policy for simultaneous translation. In *Proceedings of EMNLP*, pages 2280–2289.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv*, arXiv:1904.09675. Version 3.

Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019. Simultaneous translation with flexible policy via restricted imitation learning. In *Proceedings of ACL*, pages 5816–5822.