Word Order in English-Japanese Simultaneous Interpretation: Analyses and Evaluation using Chunk-wise Monotonic Translation

Kosuke Doi[1], Yuka Ko[1], Mana Makinae[1], Katsuhito Sudoh[1, 2], Satoshi Nakamura[1, 3]

[1]Nara Institute of Science and Technology,
[2]Nara Women's University, [3]The Chinese University of Hong Kong, Shenzhen
{doi.kosuke.de8, sudoh, s-nakamura}@is.naist.jp

Abstract

This paper analyzes the features of monotonic translations, which follow the word order of the source language, in simultaneous interpreting (SI). Word order differences are one of the biggest challenges in SI, especially for language pairs with significant structural differences like English and Japanese. We analyzed the characteristics of chunk-wise monotonic translation (CMT) sentences using the NAIST English-to-Japanese Chunk-wise Monotonic Translation Evaluation Dataset and identified some grammatical structures that make monotonic translation difficult in English-Japanese SI. We further investigated the features of CMT sentences by evaluating the output from the existing speech translation (ST) and simultaneous speech translation (simulST) models on the NAIST English-to-Japanese Chunk-wise Monotonic Translation Evaluation Dataset as well as on existing test sets. The results indicate the possibility that the existing SI-based test set underestimates the model performance. The results also suggest that using CMT sentences as references gives higher scores to simulST models than ST models, and that using an offline-based test set to evaluate the simulST models underestimates the model performance.

1 Introduction

Simultaneous interpreting (SI) is the task of translating speech from a source language into a target language in real time. SI is cognitively demanding, and human simultaneous interpreters employ such strategies as segmentation, summarization, and generalization (He et al., 2016). Maintaining word order in a source language is another important strategy, especially for language pairs whose word order differs (*e.g.*, English and Japanese), to shorten delays and reduce cognitive load. Because of these features, SI sentences are different from offline translation sentences, although most automatic SI studies (Oda et al., 2014; Ma et al., 2019; Liu et al., 2020; Papi et al., 2023) have used offline translation corpora (*e.g.*, MuST-C; Di Gangi et al., 2019) for both training and evaluatng models due to the limited amount of simultaneous interpretation corpora (SICs).

For English-Japanese language pairs, several SICs have been constructed (Toyama et al., 2004; Shimizu et al., 2014; Matsushita et al., 2020; Doi et al., 2021). Based on the NAIST Simultaneous Interpretation Corpus (NAIST-SIC; Doi et al., 2021), Zhao et al. (2024)[1] created an automatically-aligned parallel SI dataset: NAIST-SIC-Aligned. Since its sentences are aligned at the sentence level, they can be used for model training. Actually, Ko et al. (2023) and Zhao et al. (2024) trained SI models using SI data from NAIST-SIC-Aligned. Their model performances were evaluated through automatic evaluation metrics such as BLEU (Papineni et al., 2002) using a small test set curated based on SI sentences generated by professional human simultaneous interpreters.

Although the scores reported in Ko et al. (2023) and Zhao et al. (2024) were relatively low, the test set used in both studies might have underestimated the model performance. Since human simultaneous interpreters use such strategies as summarization and generalization, phrases that do not affect the main idea are not necessarily translated into the target language. If an SI model generates translations for phrases that a human interpreter did not, the output sentence might not be evaluated properly, even when it is a *correct* translation.

Fukuda et al. (2024) pointed out the difficulty for SI models to learn which phrases in source speech are less important and advocated constructing SI models that only employ a strategy that maintains the word order in a source language. As a first step, they created the NAIST English-to-

[1] The dataset was released in 2023 (see version 3 of the paper).

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 302–312
August 15-16, 2024 ©2024 Association for Computational Linguistics

Source	(1) The US Secret Service, / (2) two months ago, / (3) froze the Swiss bank account / (4) of Mr. Sam Jain right here, / (5) and that bank account / (6) had 14.9 million US dollars in it / (7) when it was frozen.
Offline	(1) 米国のシークレットサービスは / (2) 2ヶ月前に / (4) サム・ジェイン氏の / (3) スイス銀行口座を凍結しました / (5) その口座には / (6) 米ドルで1490万ドルありました [The US Secret Service / two months ago / Mr. Sam Jain's / froze the Swiss bank account / that bank account / had 14.9 million US dollars]
SI	(1) アメリカのシークレッドサービスが、/ (3) スイスの銀行の口座を凍結しました。 / (4) サムジェインのものです。 / (5) この銀行口座の中には、/ (6) 一千四百九十万ドルが入っていました。 [The US Secret Service / froze the Swiss bank account / it is Sam Jain's one / in this bank account / had 14.9 million dollars]
CMT	(1) アメリカ合衆国シークレットサービスは、 / (2) 2ヶ月前に、 / (3) スイスの銀行口座を凍結しました、 / (4) ここにいるサム・ジェイン氏の口座です、 / (5) そしてその銀行口座には / (6) 490万米ドルが入っていました、 / (7) 凍結された時。 [The US Secret Service / two months ago / froze the Swiss bank account / the account of Mr. Sam Jain right here / and that bank account / had 14.9 million US dollars in it / when it was frozen]

Table 1: Comparison of target sentences in each translation mode. Examples of offline, SI, and CMT are respectively from subtitles of TED talks, NAIST-SIC, and NAIST English-to-Japanese Chunk-wise Monotonic Translation Evaluation Dataset. "/" shows boundaries of chunks. Numbers preceding chunks in source sentence represent appearance order. Numbers preceding chunks in target sentences correspond to numbers in source sentence.

Japanese Chunk-wise Monotonic Translation Evaluation Dataset[2]. The source sentences in the test set used in Ko et al. (2023) were automatically segmented into chunks, each of which was translated in a way that did not include the content of subsequent chunks. Unlike in SI sentences by human interpreters, where not all the information in the source sentences is translated, chunk-wise monotonic translation (CMT) sentences[3] were translated so that all the information is translated (Table 1)[4]. Fukuda et al. (2024) have investigated the quality of the CMT sentences in their dataset through human evaluation, although they have not analyzed its characteristics. Nor have they conducted any evaluation experiments in which model outputs are evaluated on their dataset.

In this paper, we qualitatively and quantitatively analyze CMT sentences in the NAIST English-to-Japanese Chunk-wise Monotonic Translation Evaluation Dataset. In the process of generating CMT sentences for the dataset, it was allowed to `repeat`, `defer`, and `omit` phrases in the source sentences to maintain the translation's fluency. We assume the presence of factors (*e.g.*, syntactic structures) that prevent monotonic translation if phrases were `repeated`, `deferred`, or `omitted` in the CMT sentences since they were translated without time constraints. In addition, we evaluate the output from an existing speech translation (ST) model and two simultaneous speech translation (simulST) models (See 5.2). Both the ST[5] and simulST models are used to investigate the differences in scores when evaluating translations with different characteristics. The contributions of this paper are as follows:

- We analyze CMT sentences and show that they tend to be longer than offline translations primarily because of `repetition`.

- We investigate what causes the phrases in source sentences to be `repeated`, `deferred`, and `omitted` and show that most cases occur because of particular grammatical structures. When a phrase in a chunk is a dependent of a phrase in the preceding chunk, the head phrase is typically `repeated` or `deferred`.

- We evaluate the output from three different models on the NAIST English-to-Japanese Chunk-wise Monotonic Translation Evaluation Dataset: (1) an ST model trained on offline data, (2) a simulST model trained on offline data, and (3) a simulST model trained on both offline and SI data. The results suggest that the existing SI-based test set (Ko et al., 2023; Zhao et al., 2024) underestimates the model performance. The results also suggest that using CMT sentences as references gives higher scores to simulST models than ST models, while using an offline-based test set for evaluating simulST models underestimates the model performance.

[2] https://dsc-nlp.naist.jp/data/NAIST-SIC/Aligned-Chunk_Mono-EJ/

[3] CMT refers to the task of segmenting a source sentence into chunks and translating it in the order of the chunks. A CMT sentence is a target sentence generated through CMT.

[4] Precisely, omissions that maintained the fluency of the sentence were allowed. See Section 3.1 for the details about the dataset.

[5] A ST model generates translations after the utterances are completed.

2 Related Work

2.1 Simultaneous Interpretation Corpora

SICs are valuable resources both for developing automatic SI models and analyzing SI's characteristics. For English-Japanese language pairs, several SICs are publicly available (Toyama et al., 2004; Shimizu et al., 2014; Matsushita et al., 2020; Doi et al., 2021), although the amount of such corpora is very limited compared to offline translation corpora.

Using these corpora, SI sentences have been analyzed from various perspectives, such as strategies and interpreting patterns used by interpreters, latency, translation quality, and word order (Tohyama and Matsubara, 2006; Ono et al., 2008; Cai et al., 2018, 2020; Doi et al., 2021). SI models have also been developed using SICs (Ryu et al., 2004; Shimizu et al., 2013; Ko et al., 2023).

2.2 Word Order in Simultaneous Interpreting

When dealing with language pairs whose sentence structures are different, including English and Japanese (SVO/head-initial vs. SOV/head-final), reducing the word order differences between the source and the target languages is crucial for minimizing delays.

Murata et al. (2010) segmented source sentences into semantically meaningful units with a maximum length of 4.3 seconds and translated those units from an SI viewpoint. He et al. (2015) designed syntactic transformation rules for Japanese-English simultaneous machine translation. By applying the rules to target language sentences (*i.e.*, English), they generated more monotonic translations, while preserving the meaning of source sentences and maintaining the grammaticality of the target language. In English-Japanese SI, Futamata et al. (2020) reordered Japanese sentences to make the word order closer to the original English sentences. They further applied style transfer to increase the fluency and obtained sentences close to SI sentences by human interpreters. Han et al. (2021) proposed an algorithm to reorder and refine the target sentences so that the target sentences were aligned largely monotonically. They trained SI models for four language pairs, including English-Japanese. Nakabayashi and Kato (2021) segmented sentences into chunks and created bilingual pairs of such chunks with explicit annotations of context information. The SI model trained on the data translated the source sentences while referencing the preceding chunks although naturally connecting chunks remained a challenge. Higashiyama et al. (2023) constructed a large-scale English ↔ Japanese SIC with the information of chunk boundaries in source and target sentences and phrases that can be omitted in target sentences. The NAIST English-to-Japanese Chunk-wise Monotonic Translation Evaluation Dataset (Fukuda et al., 2024), which is similar to Higashiyama et al.'s (2023), is relatively small and intended for the evaluation purposes.

The word order differences among different translation modes have also been investigated. Okamura and Yamada (2023) quantitatively compared the degree to which the word order of the source sentences was maintained and found that SI sentences retained the order better than consecutive interpreting and offline translation sentences. Cai et al. (2020) found syntactic and non-syntactic factors that affect interpreters' word order decisions through the statistical analyses of an SIC. In this paper, we analyze what makes monotonic translation difficult. While Cai et al. (2020) analyzed the actual SI data generated by human simultaneous interpreters, we use CMT sentences, which were generated without time constraints, and in which all the information in the source sentences is translated into target sentences.

3 Chunk-wise Monotonic Translation

In SI between language pairs with different sentence structures, interpreters segment source sentences into chunks and translate them from chunk to chunk[6] (Okamura and Yamada, 2023). This section describes the details of the NAIST English-to-Japanese Chunk-wise Monotonic Translation Evaluation Dataset, which is used in our analyses.

3.1 Data

The NAIST English-to-Japanese Chunk-wise Monotonic Translation Evaluation Dataset consists of 511 pairs of source sentences and their corresponding chunk-wise monotonic translation (CMT) sentences with information of chunk boundaries in the source and target sentences. Source (*i.e.*, English) sentences, which were used as the test set in Ko et al. (2023), were segmented following the five rules that reflected the interpreter's strategies[7]

[6] Reordering may occur within a chunk.

[7] See the original paper or the code for chunking: `https://github.com/ahclab/si_chunker`

Data	Sum	Per Sent.±SD
# sentence pairs	511	–
# chunks	1,677	3.28±2.12
# source words	8,104	15.86±10.16
# target words	13,981	27.36±18.55

Table 2: Statistics for chunk-wise monotonic translation data. Standard deviations and number of words in target sentences were calculated by us. Other values are cited from Fukuda et al. (2024).

based on the syntactic analysis results from spaCy. The source sentences come from eight TED talks.

Translators were provided with source sentences with chunk boundaries and asked to (1) translate them in the order of the chunks while (2) naturally connecting the chunks and (3) not including the content of subsequent chunks. They were allowed to (1) repeat, (2) defer, and (3) omit phrases in the source sentences to keep translation fluency, although they were instructed to minimize their use of the operations with larger number as much as possible (*e.g.*, defer should not be used when repeat can handle the situation) . Data statistics are shown in Table 2.

4 Data Analysis

Fukuda et al. (2024) have examined the quality of CMT sentences through human evaluation but have not analyzed the characteristics of them. We suppose that factors exist that prevent monotonic translation if a phrase in the source sentences is repeated, deferred, or omitted since the CMT sentences were generated without time constraints. Therefore we qualitatively and quantitatively analyze the CMT sentences with these operations and reveal such factors.

To better understand the characteristics of the CMT sentences, we also compare them with SI sentences from NAIST-SIC and NAIST-SIC-Aligned as well as offline translation sentences from the subtitles of TED talks. Since the SI sentences in NAIST-SIC were not aligned at the sentence level, we manually align them. Some source sentences did not match across the datasets, and we excluded those ten sentences from the analyses. In addition, 25 sentences were not translated in NAIST-SIC[8], which were also excluded from the analyses. As a result, 476 sentences were used for our analyses.

4.1 Annotations

To analyze the characteristics of the CMT sentences, we annotated tags to the source and CMT sentences. The list of tags is shown in Table 3. Prior to the annotations, we tokenized the English sentences using spaCy[9] and the Japanese sentences using MeCab (Kudo et al., 2004) with unidic. Then, we concatenated the source and CMT sentences with a special token [SEP] and annotated them using an open-source data labeling tool, doccano.[10]

We identified spans (*i.e.*, words or phrases) that are repeated, deferred, or omitted and annotated the span tags. In addition, we annotated ahead, add, and error tags for analyses of problematic translations. The corresponding span tags in the source and CMT sentences were associated using relation tags. Annotation examples are shown in Figure 1.

The first and second authors collaboratively annotated the first 50 examples while discussing their decisions. Since sufficient agreement was assumed, the remainder of the data were just annotated by the first author.

4.2 Analysis Results

4.2.1 Comparison among Different Translation Modes

We compared the sentence lengths of the four datasets. Fukuda et al. (2024) also conducted similar comparisons based on the number of characters. However, since variations in spelling and differences in transcription systems (*e.g.*, numbers) were found, we made comparisons based on the number of words segmented by MeCab.

Table 4 shows that the CMT sentences were the longest, followed by offline, NAIST-SIC, and NAIST-SIC-Aligned. These results matched those reported in Fukuda et al. (2024). Long translated sentences can pose some problems. As discussed in Fukuda et al. (2024), the lenght may increase the cognitive load on the listeners/readers of the translations. In addition, longer output may cause a delay even though CMT aims to reduce it.

4.2.2 Factors that Lengthen CMT Sentences

To reveal what factor lengthened the CMT sentences, we first analyzed them qualitatively. Our analyses suggest that (1) CMT sentences contain

[8] For example, due to time constraints, interpreters might have been unable to translate a whole sentence. See Doi et al. (2021).

[9] https://spacy.io/

[10] https://github.com/doccano/doccano

Type	Tag	Meaning
Span	repeat	Phrases that are repeated
	zero-repeat	Target phrases that are repeated; No corresponding source phrases (*e.g.*, zero that-clause)
	defer	Phrases that are not translated within the current chunk but in a subsequent chunk
	omit	Source phrases that are not translated in the target sentences
	ahead	Target phrases translated using the subsequent chunks
	add	Target phrases that have no corresponding source phrases
	error	Phrases with translation errors
	sep	Boundaries of source and target sentences
Relation	rel-repeat	Connect source and target phrases with repeat tags
	repeat_d#	Connect target phrases with repeat tags
	defer_d#	Connect source and target phrases with defer tags
	ahead_d#	Connect source and target phrases with ahead tags
	rel-err	Connect source and target phrases with error tags

Table 3: List of tags used for annotations. “d#” (#=1, 2, ...) represents distance between chunks.

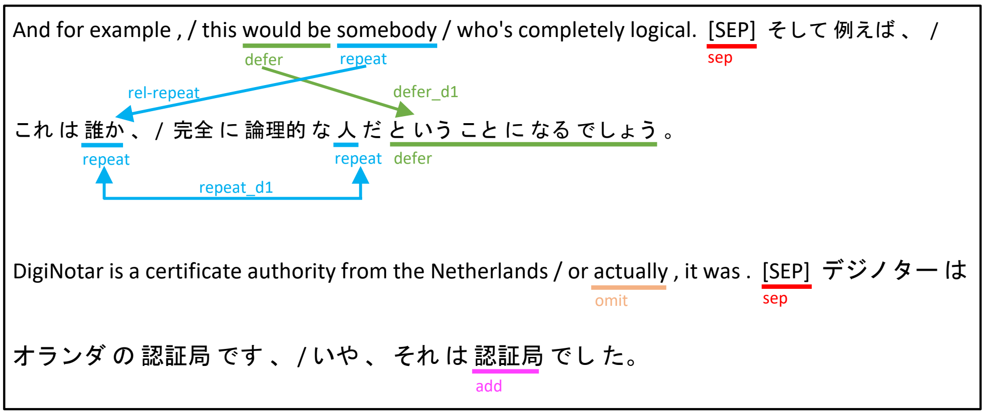
And for example , / this would be somebody / who's completely logical. [SEP] そして 例えば 、 /
defer
repeat
sep
rel-repeat
defer_d1
これ は 誰か 、 / 完全 に 論理的 な 人 だ と いう こと に なる でしょう 。
repeat
repeat
defer
repeat_d1
DigiNotar is a certificate authority from the Netherlands / or actually , it was . [SEP] デジノター は
omit
sep
オランダ の 認証局 です 、 / いや 、 それ は 認証局 でし た。
add

Figure 1: Annotation examples. Repeat tags were assigned even if strings did not exactly match but referred to same entity or had same meaning.

Dataset	Sum	Per Sent.±SD
CMT	13,508	28.38±18.66
NAIST-SIC	8,914	18.73±12.08
NAIST-SIC-Aligned	8,072	16.96±11.52
Offline	9,907	20.81±12.62

Table 4: Comparison of number of words in different translation modes

many formulaic expressions for the end of sentences as they are segmented into small chunks, and (2) the words that are often omitted in Japanese (*e.g.*, pronouns) are explicitly translated since translators were instructed to avoid omitting phrases in the source language, as in the following examples:

> (En) It’s when we warmed it up, / and we turned on the lights / and looked inside the box, / we saw that the piece metal / was still there in one piece.
> (CMT) 私たちがそれを暖めるときです、/電気をつけて、/箱の中を見たときです、/私たちはその金属片を見たんです、/それはまだ一つの塊としてそこにありました。
> (offline) 物体を暖め明かりをつけ 箱の中を見たところ金属片はまだそこに存在していました
>
> (En) He only has use of his eyes.
> (CMT) 彼は目だけを使えます。
> (offline) 目だけしか動かせません

In addition to the above characteristics, we observed many repetitions in particularly *long* sentences. To verify this, we further analyzed particularly *long* and *short* sentences, chosen based on the length ratio of the CMT sentences to the offline ones. The *long* and *short* sentences were defined as those with a length ratio greater/smaller than the average ± 0.5 standard deviations (avg.=1.39, SD=0.43). We subjectively judged whether these sentences contained many repetitions. We also identified sentences whose offline translations were short.

Table 5 shows that *long* sentences contained more repetitions than *short* ones. The offline translation sentences were short, probably because they were originally subtitles, for which limited space was allowed. We also quantitatively checked them

Type	*N*	Repeat (subjective)	Repeat (tag)	Short offline
Long	131	54 (41.22%)	3.35	53
Short	171	22 (12.87%)	0.81	–

Table 5: Comparison between *long* and *short* CMT sentences. Number of repeat tags is denoted per sentence.

Operation	*N*
repeat	301
defer	173
omit	36

Table 6: Comparison of number of operations used in CMT sentences

using the number of assigned `repeat` tags and found that the frequency of `repetition` tags was higher in *long* sentences (Table 5).

4.2.3 Omission in SI Sentences

To find techniques for shortening translations, we analyzed the SI sentences in NAIST-SIC. Based on the length ratio of SI sentences to the offline ones, we defined SI sentences that might have reasonable omissions (`omission`; $0.6 \leq$ ratio < 0.9) and SI sentences that probably failed to fully convey the meaning of the source sentences (`undertranslation`; ratio < 0.6), following the criteria in Higashiyama et al. (2023).

Although we expected to identify some trends (*e.g.*, part-of-speech) in the phrases that were omitted, we did not do so. In addition, we found a certain number of *unacceptable* translations in both categories (43.12% and 60.00% for `omission` and `undertranslation`, respectively). The results suggest that human simultaneous interpreters judge the importance of phrases based on context and decide whether to translate them; some judgements are correct, and some are not.

4.2.4 Factors that Make Monotonic Translation Difficult

With the help of tags annotated to the source and CMT sentences, we analyzed the factors that make monotonic translation difficult. Table 6 shows the number of source phrases that were `repeated`, `deferred`, or `omitted`. The values are based on the number of `rel-repeat`, `defer_d#`, and `omit` tags. We counted the relation tags for `repeat` and `defer` because the span tags for those two operations were assigned to both the source and CMT sentences. The results show that the translators used `repeat` most frequently, followed by `defer` and `omit`, as they were instructed (see Section 3.1).

For these phrases, we explored what makes monotonic translations difficult. Our analyses revealed that most cases of `repeat` and `defer` were caused by particular grammatical structures. Table 7 lists the major structures along with their frequencies in the data and examples. In these structures, a phrase in a chunk is typically a dependent of a phrase in the preceding chunk. In the example of a post-modifier (Table 7), the relative pronoun clause is a dependent of the noun phrase *a device*, which is in a preceding chunk. When phrases with a dependency relation exist across multiple chunks, CMT is difficult because Japanese is a strongly head-final language. The examples in Table 7 show how human translators address these structures by repeating or deferring some phrases in subsequent chunks.

Prepositions, post-modifiers, and dependent clauses have also been identified as syntactic factors that affect interpreters' word order decisions in Cai et al. (2020). Human interpreters find these structures challenging for SI and adopt a strategy to maintain the word order of the source language.

In addition, we observed that inappropriate segmentation was addressed by `repeating` and `deferring` the phrases. Most inappropriate segmentation was found in phrasal verbs, verbal gerunds, and to-infinitives.

In the SI data, we also found that human interpreters `repeat` phrases to maintain the word order of the source language. For example, in the example in Table 1, a noun modified by a preposition phrase is `repeated`:

(En) ... / froze the Swiss bank account / of Mr. Sam Jain right here, / ...

(SI) ... / スイスの銀行の口座を凍結しました。 / サムジェインのものです。 / ...
[... / froze the Swiss bank account / it is Sam Jain's one / ...]

In addition, Okamura and Yamada (2023) reported that the order of the chunks is shuffled about once on average in an SI sentence. These things suggest that human interpreters address the word order differences that make monotonic translation difficult by `repeating` and `deferring` some phrases.

Structures	# repeat	# defer	Examples
Noun with a post-modifier	88	12	And now we've created a device / that has absolutely no limitations. さて、私たちはデバイスを作り出しました、/ 全く制限のないものです。 [And we've created a device / one that has absolutely no limitations]
Head followed by multiple dependents	35	6	... / allows for deep squats, / crawls and high agility movements. ... / 深いスクワットを可能にし、/ クロールや高い敏捷性の動きを可能にします。 [... / allows for deep squats / allows for crawls and high agility movements]
Dependent conjunction	26	6	... / when he's covered / in four feet of snow. ... / 彼が / 四フィートの雪に覆われてしまっていた時には。 [he / when was covered in for feet of snow]
Chunk boundary before a clause	15	13	... / you know that this isn't / how it normally goes. / あなたは分かるはずです、これは / 通常の進行ではないということが。 [you know that this / isn't how it normally goes]
Chunk boundary before a preposition	10	7	... / providing totalitarian governments with tools / to do this / against their own citizens. ... / 全体主義政府にツールを提供しているということです、/ これを行うための、/ 自国の市民に対してこれを行うためのツールを。 [providing totalitarian governments with tools / to do this / tools to do this against their own citizens]

Table 7: Syntactic factors that prevent monotonic translations. Cases involving multiple structures were classified separately as *compound factors*.

5 Evaluation Using CMT sentences

To investigate the impact of using CMT sentences for evaluating translation quality, we evaluated the output from existing ST and simulST models using the NAIST English-to-Japanese Chunk-wise Monotonic Translation Evaluation Dataset as well as existing test sets.

5.1 Data

We used the following four datasets as references for the automatic evaluation metrics:

- **n-cmt**: CMT sentences from the NAIST English-to-Japanese Chunk-wise Monotonic Translation Evaluation Dataset
- **si_hum**: SI sentences from NAIST-SIC, manually aligned to the source speech
- **si_auto**: SI sentences from NAIST-SIC-Aligned, aligned automatically
- **offline**: offline translation sentences from the subtitles of TED talks.

Because `si_auto` was created by applying automatic alignment and filtering techniques to SI sentences in `si_hum`, it may contain alignment errors. In addition, SI sentences in `si_auto` tend to be shorter than those in `si_hum` (see Table 4). Therefore, we used the two SI-based datasets for our evaluation.

5.2 Speech Translation Models

We used three existing models (*i.e.*, one ST and two simulST models):

- **ST_offline**: an ST model trained on offline data (Fukuda et al., 2023)
- **simulST_offline**: a simulST model trained on offline data (Ko et al., 2023)
- **simulST_si_offline**: a simulST model trained on both offline and SI data (Ko et al., 2023).

All the models were built by connecting two pre-trained models, HuBERT-Large (Hsu et al., 2021) for their speech encoder and the decoder of mBART50 (Tang et al., 2020) for their text decoder. The encoder and decoder were connected by Inter-connection (Nishikawa and Nakamura, 2023) and a length adapter (Tsiamas et al., 2022). Both SimulST models used bilingual prefix pairs extracted using Bilingual Prefix Alignment (Kano et al., 2022) for the model training and employed a decoding policy called local agreement (Liu et al., 2020). For ST_offline, we used a model with checkpoint averaging (`Inter-connection + Ckpt Ave.` in Fukuda et al. (2023)). For simulST_offline and simulST_si_offline, we used the models that satisfy the task requirement of the simultaneous track in the IWSLT 2023 Evaluation Campaign[11], latency measured by Average Lagging (Ma et al., 2019) $\leq$ 2 seconds (`Offline FT` and `Mixed FT + Style` in Ko et al. (2023), respectively).

5.3 Metrics

We evaluated the translation quality of the output from the ST and simulST models (see Section 5.2)

[11] https://iwslt.org/2023/simultaneous

Model	BLEU				BLEURT				COMET			
	n-cmt	si_hum	si_auto	offline	n-cmt	si_hum	si_auto	offline	n-cmt	si_hum	si_auto	offline
ST_offline	14.487	8.856	8.637	17.775	0.553	0.447	0.414	**0.538**	**0.838**	**0.797**	**0.781**^{*1}	**0.833**
simulST_offline	15.406†	8.446†	7.773†	**17.907**	0.556	0.442	0.406	0.531	0.826	0.780	0.763	0.821
simulST_si_offline	**15.982**†	**12.031**†	**11.020**†	13.191†	**0.567**	**0.493**^{*1}	**0.460**^{*1}	0.519	0.807^{*2}	0.774^{*3}	0.761	0.789^{*2}

Model	BERTScore (Pre.)				BERTScore (Rec.)				BERTScore (F1)			
	n-cmt	si_hum	si_auto	offline	n-cmt	si_hum	si_auto	offline	n-cmt	si_hum	si_auto	offline
ST_offline	0.801	0.735	0.722	**0.789**	0.769	0.739	0.735	**0.788**	0.784	0.737	0.728	**0.788**
simulST_offline	0.799	0.730	0.717	0.783	0.770	0.738	0.734	0.786	0.783	0.734	0.725	0.784
simulST_si_offline	**0.817**^{*1}	**0.764**^{*1}	**0.746**^{*1}	0.759^{*2}	**0.784**^{*1}	**0.766**^{*1}	**0.760**^{*1}	0.757^{*2}	**0.800**^{*1}	**0.764**^{*1}	**0.752**^{*1}	0.757^{*2}

Table 8: Results of quality evaluation metrics across ST and simulST models. †: significantly different from ST_offline. $*1$: significantly higher than other two. $*2$ significantly lower than other two. $*3$ significantly lower than ST_offline. Significance threshold was set to $p < .05$ for all tests.

using BLEU[12], BLEURT[13] (Sellam et al., 2020), COMET[14] (Rei et al., 2020), and BERTScore[15] (Zhang et al., 2020). BERTScore was calculated using bert-base-multilingual-cased. We used the four datasets described in Section 5.1 as references.

5.4 Evaluation Results

Table 8 shows the results of the quality evaluation metrics across the ST and simulST models. For the BLEU scores, we conducted paired significance tests using paired bootstrap resampling (Koehn, 2004). We specified ST_offline as the baseline for the significance tests. For the other scores, we conducted a one-way ANOVA, followed by Tukey's multiple comparisons test.

When the translation quality was evaluated using BLEU with n-cmt as the reference, simulST_si_offline achieved the highest score. On the SI-based test sets (*i.e.*, si_hum and si_auto), simulST_si_offline also had the highest score. On the offline-based test set, in contrast, the models trained on only offline data achieved much higher scores than simulST_si_offline. The same tendencies were observed in BLEURT and BERTScore. These results suggest that the models trained on both SI and offline data generated more SI-like translations, and such models perhaps should be evaluated using a reference closer to SI sentences. In addition, using an offline-based test set might underestimate the performance of models trained on both SI and offline data.

Comparing n-cmt, si_hum, and si_auto, the scores were highest for n-cmt, followed by si_hum and si_auto on all the metrics and models. Because si_hum is based on SI sentences generated by human simultaneous interpreters, some content in the source speech might be omitted or inadequately translated (under-translation). SI sentences in si_auto, which were automatically created based on human SI sentences, might contain less source speech content than those in si_hum due to the alignment and filtering techniques applied (see Zhao et al., 2024). In fact, BERTScore precision was higher than recall on n-cmt, in which there were almost no omissions, while recall was higher than precision on si_auto and precision and recall were almost equal on si_hum. These results indicate the possibility that the existing SI-based test sets (Ko et al., 2023; Zhao et al., 2024) underestimate the model performance.

However, the COMET results were different from those on the other metrics (Table 8). On all four test sets, ST_offline achieved the highest score, followed by simulST_offline and simulST_si_offline. One possible reason is that COMET uses source sentences to calculate its scores.

To examine the impact of the source sentences, we also calculated a reference-free COMET-QE using wmt22-cometkiwi-da and got similar results (0.813, 0.798, and 0.766 for ST_offline, simulST_offline, and simulST_si_offline, respectively). We further calculated COMET-QE for n-cmt and offline, regarding them as oracle data, and found that n-cmt had a higher score than offline (n-cmt: 0.832, offline: 0.812). Because some translation sentences in offline are under-translated, these results suggest that the COMET scores tend to become high when more content in the source sentences is covered in the

[12] BLEU was calculated using sacreBLEU. (Post, 2018) https://github.com/mjpost/sacrebleu

[13] https://github.com/google-research/bleurt

[14] https://github.com/Unbabel/COMET

[15] https://github.com/Tiiiger/bert_score

target sentences. This feature does not fit the nature of SI, where human interpreters use sophisticated strategies (see He et al., 2016; Cai et al., 2020, for example). We need to carefully interpret COMET scores when we use them for evaluating simulST models.

6 Conclusion

This paper focused on monotonic translations in English-Japanese SI. Our analyses revealed some grammatical structures that make monotonic translations difficult and that human interpreters/translators address these challenges by repeating or deferring some phrases in source language in the subsequent chunks. The grammatical structures that might cause delays would be useful information for developing segmentation or decoding policies for simultaneous machine translation systems. One possible direction would be predicting whether a phrase in a chunk is the head of a phrase in subsequent chunks.

We also evaluated the output from the existing ST and simulST models on the NAIST English-to-Japanese Chunk-wise Monotonic Translation Evaluation Dataset as well as on existing SI-based and offline-based test sets. The BLEU, BLEURT, and BERTScore results supported using CMT sentences for evaluating simulST models trained using SI data, although the results with COMET were different. Further analysis across various evaluation metrics is necessary. Analyzing how the source and target sentences are aligned monotonically on different types of translations (*e.g.*, Han et al., 2021) would also be useful.

This paper investigated the impact of using CMT sentences for evaluation purposes. A future study would involve using monotonic translation sentences for developing simulST models (Sakai et al., 2024)[16]. It could potentially address the problem that simulST models trained using SI sentences suffered from under-translation (Ko et al., 2023). However, CMT sentences tend to be long. Investigating the trade-offs between longer CMT sentences and the potential cognitive load on listeners/readers might provide further insights.

Acknowledgments

A part of this work was supported by JSPS KAKENHI Grant Numbers JP21H05054 and by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2137.

References

Zhongxi Cai, Koichiro Ryu, and Shigeki Matsubara. 2018. Statistical analysis of missing translation in simultaneous interpretation using a large-scale bilingual speech corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Zhongxi Cai, Koichiro Ryu, and Shigeki Matsubara. 2020. What affects the word order of target language in simultaneous interpretation. In *Proceedings of 2020 International Conference on Asian Language Processing (IALP)*, pages 135–140.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. 2021. Large-scale English-Japanese simultaneous interpretation corpus: Construction and analyses with sentence-aligned data. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 226–235, Bangkok, Thailand (online). Association for Computational Linguistics.

Ryo Fukuda, Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. 2024. 原発話に忠実な英日同時機械翻訳の実現に向けた順送り訳評価データ作成 [Creation of Evaluation Data for Monotonic Translation toward the Realization of Simultaneous English-Japanese Machine Translation Faithful to the Source Speech]. In *Proceedings of the 259th meeting of Special Interest Group of Natural Language Processing (IPSJ-SIGNL), 2024-NL-259(14)*, pages 1–6. (in Japanese).

Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Yuka Ko, Tomoya Yanagita, Kosuke Doi, Mana Makinae, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. 2023. NAIST simultaneous speech-to-speech translation system for IWSLT 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 330–340, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Kosuke Futamata, Katsuhito Sudoh, and Satoshi Nakamura. 2020. 英日同時通訳システムのための疑似同時通訳コーパス自動生成手法の提案[Proposal

[16] Published around the same time as the submission of this paper.

of a Method for Automatically Generating Pseudo-simultaneous Interpretation Corpora for English-Japanese Simultaneous Interpretation Systems]. In *Proceedings of the 26th Annual Meeting of the Association for Natural Language Processing*, pages 1281–1284. (in Japanese).

HyoJung Han, Seokchan Ahn, Yoonjung Choi, Insoo Chung, Sangha Kim, and Kyunghyun Cho. 2021. Monotonic simultaneous translation with chunk-wise reordering and refinement. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1110–1123, Online. Association for Computational Linguistics.

He He, Jordan Boyd-Graber, and Hal Daumé III. 2016. Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 971–976, San Diego, California. Association for Computational Linguistics.

He He, Alvin Grissom II, John Morgan, Jordan Boyd-Graber, and Hal Daumé III. 2015. Syntax-based rewriting for simultaneous machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 55–64, Lisbon, Portugal. Association for Computational Linguistics.

Shohei Higashiyama, Kenji Imamura, Masao Utiyama, and Eiichiro. Sumita. 2023. GCP 同時通訳コーパスの構築[Construction of GCP Simultaneous Interpretation Corpus]. In *Proceedings of the 29th Annual Meeting of the Association for Natural Language Processing*, pages 1405–1410. (in Japanese).

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. 2022. Simultaneous neural machine translation with prefix alignment. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 22–31, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Yuka Ko, Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. 2023. Tagged end-to-end simultaneous speech translation training using simultaneous interpretation data. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 363–375, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Danni Liu, Gerasimos Spanakis, and Jan Niehues. 2020. Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis Selection. In *Proceedings of Interspeech 2020*, pages 3620–3624.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Kayo Matsushita, Masaru Yamada, and Hiroyuki Ishizuka. 2020. An overview of the Japan National Press Club (JNPC) Interpreting Corpus. *Invitation to Interpreting and Translation Studies*, 22:87–94. (in Japanese).

Masaki Murata, Tomohiro Ohno, Shigeki Matsubara, and Yasuyoshi Inagaki. 2010. Construction of chunk-aligned bilingual lecture corpus for simultaneous machine translation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Akiko Nakabayashi and Tsuneaki Kato. 2021. 同時機械翻訳のための文脈を考慮したセグメントコーパス[Context-Aware Segment Corpus for Simultaneous Machine Translation]. In *Proceedings of the 27th Annual Meeting of the Association for Natural Language Processing*, pages 1659–1663. (in Japanese).

Yuta Nishikawa and Satoshi Nakamura. 2023. Interconnection: Effective Connection between Pretrained Encoder and Decoder for Speech Translation. In *Proceedings of INTERSPEECH 2023*, pages 2193–2197.

Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Optimizing segmentation strategies for simultaneous speech translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 551–556, Baltimore, Maryland. Association for Computational Linguistics.

Yuki Okamura and Masaru Yamada. 2023. 「順送り訳」の規範と模範 同時通訳を模範とした教育論の

試論 [Norms and Canon of Progressive Translation - An Exploratory Study on Educational Theories Using Simultaneous Interpretation as a Canon]. In Hiroyuki Ishizuka, editor, *Word Order in English-Japanese Interpreting and Translation: The History, Theory and Practice of Progressive Translation*, pages 217–250. Hitsuji Syobo. (in Japanese).

Takahiro Ono, Hitomi Tohyama, and Shigeki Matsubara. 2008. Construction and analysis of word-level time-aligned simultaneous interpretation corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Sara Papi, Marco Turchi, and Matteo Negri. 2023. AlignAtt: Using Attention-based Audio-Translation Alignments as a Guide for Simultaneous Speech Translation. In *Proceedings of INTERSPEECH 2023*, pages 3974–3978.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Koichiro Ryu, Atsushi Mizuno, Shigeki Matsubara, and Yasuyoshi Inagaki. 2004. Incremental japanese spoken language generation in simultaneous machine interpretation. In *Proceedings of Asian Symposium on Natural Language Processing to Overcome language Barriers*, pages 91–95.

Yusuke Sakai, Mana Makinae, Hidetaka Kamigaito, and Taro Watanabe. 2024. Simultaneous Interpretation Corpus Construction by Large Language Models in Distant Language Pair. *arXiv*. ArXiv:2404.12299.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2013. Constructing a speech translation system using simultaneous interpretation data. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*, Heidelberg, Germany.

Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Collection of a simultaneous translation corpus for comparative analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 670–673, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv*, arXiv:2008.00401.

Hitomi Tohyama and Shigeki Matsubara. 2006. Collection of simultaneous interpreting patterns by using bilingual spoken monologue corpus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Hitomi Toyama, Shigeki Matsubara, Koichiro Ryu, Nobuo Kawaguchi, and Yasuyoshi Inagaki. 2004. CIAIR Simultaneous Interpretation Corpus. In *Proceedings of Oriental COCOSDA*.

Ioannis Tsiamas, Gerard I. Gállego, Carlos Escolano, José Fonollosa, and Marta R. Costa-jussà. 2022. Pretrained speech encoders and efficient fine-tuning methods for speech translation: UPC at IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 265–276, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Jinming Zhao, Yuka Ko, Kosuke Doi, Ryo Fukuda, Katsuhito Sudoh, and Satoshi Nakamura. 2024. NAIST-SIC-Aligned: Automatically-Aligned English-Japanese Simultaneous Interpretation Corpus. *arXiv*. ArXiv:2304.11766, version 4.