NAIST-SIC-Aligned: an Aligned English-Japanese Simultaneous Interpretation Corpus

Jinming Zhao[1,*,+], **Yuka Ko**[2,+], **Kosuke Doi**[2], **Ryo Fukuda**[2]
Katsuhito Sudoh[2], **Satoshi Nakamura**[2]
[1]Monash University, Australia
[2]Nara Institute of Science and Technology, Japan
jinming.zhao@monash.edu, {ko.yuka.kp2, sudoh, s-nakamura}@is.naist.jp

Abstract

It remains a question that how simultaneous interpretation (SI) data affects simultaneous machine translation (SiMT). Research has been limited due to the lack of a large-scale training corpus. In this work, we aim to fill in the gap by introducing *NAIST-SIC-Aligned*, which is an automatically-aligned parallel English-Japanese SI dataset. Starting with a non-aligned corpus NAIST-SIC, we propose a two-stage alignment approach to make the corpus parallel and thus suitable for model training. The first stage is coarse alignment where we perform a many-to-many mapping between source and target sentences, and the second stage is fine-grained alignment where we perform intra- and inter-sentence filtering to improve the quality of aligned pairs. To ensure the quality of the corpus, each step has been validated either quantitatively or qualitatively. This is the first open-sourced large-scale parallel SI dataset in the literature. We also manually curated a small test set for evaluation purposes. Our results show that models trained with SI data lead to significant improvement in translation quality and latency over baselines. We hope our work advances research on SI corpora construction and SiMT. Our data can be found at `https://github.com/mingzi151/AHC-SI`.

Keywords: simultaneous interpretation, interpreting corpus, corpus construction

1. Introduction

Simultaneous interpretation (SI) is a task where an utterance is translated in real-time. Simultaneous machine translation (SiMT) systems should produce reasonably good translations with low latency (Ma et al., 2019; Arthur et al., 2021). Due to the lack of a large-scale SI corpus, most SiMT systems are trained with offline machine translation (MT) corpora which are often abundant with easy access (Zheng et al., 2019). Yet, MT data is different from SI data because of the difference between offline and online translation in nature. Despite some efforts (Zhao et al., 2021), it has always remained a question on how and to what extent SI data affects an SiMT system.

Corpora	Language	Size
Toyama et al. (2004)	En↔Jp	182 hours
Shimizu et al. (2014)	En↔Jp	22 hours
Doi et al. (2021)	En↔Jp	304.5 hours
Pan (2019)	Zh↔En	6M tokens
Zhang et al. (2021) (dev/test)	Zh→En	3 hours
Paulik and Waibel (2009)	En↔Es	217 hours
Bernardini et al. (2016)	En, Fr, It	95K tokens
Kunz et al. (2021)	En↔De	83 hours
Zhao et al. (2021)	En↔De	1K sent.
Macháček et al. (2021)	En→De, Cs	10 hours
Wang et al. (2021)	15 lang.	17.3K hours
Przybyl et al. (2022)	En, De, Es	10K sent.

Table 1: Existing SI corpora.

Zhao et al. (2021) recently made a call for constructing SI corpora to learn interpreters' behaviours in modelling, but no corpus has been proposed as of now. This is mainly contributed by the fact that collecting and building parallel SI corpora is exceptionally costly and time-consuming (Shimizu et al., 2013). Most existing corpora are in a small scale, mainly for testing purposes (Zhao et al., 2021; Bernardini et al., 2016); see Table 1[1] for detailed statistics. In addition, aligning source and target sentences presents great challenges. Doi et al. (2021) presented a large-scale document-level SI corpus, NAIST-SIC, and conducted analyses on a small manually-aligned subset. However, sentences are not aligned within each document pair for most of the corpus, so it cannot be used for model training due to the lack of alignment.

* Work done while Jinming was a research intern at NAIST.
+These authors contributed equally to this work.

[1]The pSp corpus introduced in Paulik and Waibel (2009) replies heavily on time information. Its quality is unclear due to the lack of open-source accessibility.

LREC-COLING 2024, pages 12046–12052
20-25 May, 2024.

This work aims to fill the gap by building a large-scale parallel English-Japanese SI corpus named *NAIST-SIC-Aligned*. We start with the non-parallel NAIST-SIC and propose a two-stage pipeline alignment approach consisting of coarse and fine-grained alignment to make it parallel. The initial stage is coarse alignment which involves identifying minimal groups of source and target sentences that are considered translations of each other. The next stage is fine-grained alignment, where intra- and inter-sentence filtering techniques are applied over coarse-aligned pairs to improve data quality. We validate each step either manually or automatically, to ensure the quality of the data. Meanwhile, we compile a small-scale, manually curated SI test set for testing purposes. We additionally summarize alignment challenges and findings to guide future SI corpus construction for other language pairs. Lastly, we build SiMT systems based on our corpus and show significant improvement over baselines in both translation quality and latency.

2. Raw Corpus Construction

Raw SI data collection We used a portion of SI data from NAIST-SIC in this study, referred to as SI^{RAW}. This data comprises professional simultaneous interpreters' real-time interpretations of TED talks, which span a variety of topics from technology to entertainment.[2] Interpreters with varying experience levels (S-rank: 15 years, A-rank: 4 years, B-rank: 1 year) participated in the interpretations. Note that SI^{RAW} underwent manual transcription. We specifically used interpretations by S-rank and A-rank interpreters, denoted as SI^{Srank_RAW} and SI^{Arank_RAW}, respectively. Overall, the quality of SI^{Srank_RAW} surpasses that of SI^{Arank_RAW}.

Manual subset alignment Doi et al. (2021) conducted sentence-level alignment on 14 talks (a subset of talks interpreted by all three rank interpreters) for analysis. This process involved manual alignment of source sentences with target sentences, referred to as true-align. We refer readers to the original paper for more details.

Error analysis After a manual investigation on true-align, we categorise issues with the SI data into two groups: *under-translation* and *mis-translation*. *Under-translation* occurs when interpreters unintentionally omit content due to memory overload. Interpreters may also omit information intentionally, using tactics such as summarization, which is permissible. For instance, in the following interpretation, omitting "And if you drill into that" does not impede comprehension of the source sentence.

Source: And if you drill into that, it's especially the case for men.

Interpretation: 特に、男の人はそうなんです。 [This is especially true for men.]

Mis-translation is the result of interpreters mistakenly rendering source sentences, often due to cognitive overload. The interpretation below clearly misrepresents the source sentence.

Source: I think we should make this even more explicit.

Interpretation: 分かったと思います。 [I think I understand.]

3. Parallel Corpus Construction

In this section, we present how we perform alignment on SI^{RAW} (§3.1). Then, we detail how we split data to train/dev/test (§3.2), followed by describing manual compilation of the test set (§3.2).

3.1. Alignment

We propose a two-stage alignment method that the first stage involves coarse alignment, grouping minimal source and target sentences to establish translations, and the second stage refines these pairs through fine-grained alignment. Both stages are validated quantitatively and qualitatively, with insights and findings shared at each step.

3.1.1. Coarse Alignment

Alignment Each talk consists of M source sequences, $(e_1, e_2, \ldots, e_M)$, and N target sequences, $(f_1, f_2, \ldots, f_N)$, where each e is at the sentence level, each f is at the chunk level, and $N > M$. Note that there is no clear punctuation to group target sequences to sentence level, due to the nature of SI. The first step is to detect groups between these sequences that are translations of each other.[3]

Due to *under-translation* and *mis-translation*, some sequences lack corresponding translations, requiring deletion operations. We used the vecAlign (Thompson and Koehn, 2019) sentence aligner, which suits the above purposes. VecAlign creates source and target language sentences, compares sentence embeddings computed by

[2] https://www.ted.com/

[3] We tested various sentence and chunk combinations but found the current setup, to be the most effective. This aligns with our intuition, as source-side punctuation aids in semantically grouping target chunks.

LASER (Artetxe and Schwenk, 2019) using dynamic programming, and yields candidate pairs $< E, F >$ with the lowest cost.[4] Pairs with a cost higher than 1 or with empty E or F indicating no corresponding translation were removed.

Validation The outputs of this stage are validated both qualitatively and quantitatively. Manual inspection reveals a relatively high alignment quality, particularly for SI^{Srank_RAW}. To perform quantitative evaluation, we use true-align to automatically generate coarse alignment from talks, denoted as auto-align, and then measure the extent to which true-align can be recovered. For a given source sentence, e, we measure the similarity between the automatically aligned target sentence, F^a, and the manually aligned sentence, F^m, as follows:

$$s = \frac{LCS(F^a, F^m)}{|F^m|} \tag{1}$$

where $LCS(A, B)$ stands for the longest common substring between strings A and B.

We define f^a as correctly aligned to e when the similarity measure s exceeds a threshold ϵ. Higher ϵ values result in lower accuracy. Notably, with ϵ set to 0.8, the overall accuracy for the S-rank subset in true-align remains at 80%, underscoring vecAlign's effectiveness in addressing the issue. We name the output at this stage COARSE, and we also prepare S-rank data as COARSE^{Srank}.

Finding After the manual check of the outputs, we summarise our main findings as follows.

- The quality of alignment results largely depends on the quality of interpretation data. SI^{Srank_RAW} has a better quality of data than SI^{Arank_RAW}, and thus the aligned pairs of the former have higher quality.
- The quality of interpretation data is influenced by talk difficulty, which can be assessed from two perspectives: the speaker and the interpreter. A talk becomes challenging when the speaker speaks rapidly, employs jargon, has an accent, etc. It is also contingent on the interpreter's skills and background knowledge, etc.
- Interpreters may exhibit varying performance levels for different talks. For easy talks, their performance remains consistent, contrasting the typical phenomenon of performance decline due to cognitive load over time. We attribute this to the interpreters working in a less stressful, simulated environment. In the case of difficult talks, the expected performance decline is observed.

[4]Even when substituting LASER with other advanced encoders, it consistently outperforms them.

Human Evaluation We further analysed half of a talk from SI^{Srank_RAW}, denoted as SAMPLE (comprising of 78 pairs). Our findings indicate that 82% of the pairs are well-aligned, 11% are reasonably aligned, and the remainder are poorly aligned. Within the 11% subset, misalignment often occurs at the beginning/end of the target sentence, and removing it would enhance alignment quality.[5] While these statistics may vary across talks, these observations guide our alignment design.

3.1.2. Fine-grained Alignment

The second stage is to perform fine-grained alignment consisting of two filtering stages. The purpose is to improve the accuracy of aligned pairs.

Intra-sentence filtering Based on our observations, we decide to filter out the beginning and last chunk of a target sentence if it does not carry any substantial meaning. For example, じゃあ (meaning "then" in English) often appears as a filer in Japanese and should be removed. We consider Japanese words conveying important information (i.e., content words): "NOUN", "PROPN", "PRON", "VERB" or "NUM". Manual inspection of the output from this step on SAMPLE shows that this simple heuristic accurately detects and removes chunks that contain no content words in most of the time. We call the resulting data together the S-rank subset INTRA and INTRA^{Srank}.

Inter-sentence filtering To address potential quality disparities in COARSE alignment for talks other than SAMPLE, we introduce rigorous surface-level and semantic-level filtering rules. i) For an aligned pair <E, F>, we calculate α, representing the percentage of content words in E covered by F. Content words include "NOUN," "PROPN," and "NUM,"; "VERB" is excluded due to its perceived lesser importance during interpretation (Seeber, 2001). This accounts for interpreters' techniques aimed at managing cognitive load, allowing for certain word omissions. ii) We compute the length ratio γ between F and its corresponding offline translation T, generated via Google Translate from E. Large or small γ indicate potential *mis-translation* and *under-translation* issues. iii) We measure semantic similarity, η, between F and T with BLEURT (Sellam et al., 2020), a semantic-based MT metric. As T is the correct translation of E, high η implies high coverage of E by F. We denote the entire data and S-rank data as INTER and INTER^{Srank}, respectively.

[5]Occasionally, the misaligned part should belong to the previous or the next sentence. We leave the potential improvement to future work.

Data	Subset	# Talks	# Pairs
COARSE	train	831	67,079
COARSE^{Srank}	train	472	41,597
INTRA	train	831	66,834
INTRA^{Srank}	train	472	41,436
INTER	train	831	50,096
INTER^{Srank}	train	472	32,039
AUTO-DEV	dev	10	732
AUTO-TEST	test	15	1,176
SI^{DEV}	dev	4	238
SI^{TEST}	test	5	383

Table 2: Dataset statistics

Validation Each of the above steps has been manually validated by one of the authors. That said, the optimal values for α, γ and η vary by talk, and we leave automatically learning optimal values per talk for future work.

3.2. Data Split

Both our data and MuST-C (Di Gangi et al., 2019), a common dataset in SiMT (both speech and text), contain TED Talks. To prevent potential data contamination, we ensured our train set (831 talks) is present in Must-C train data. We created training set variations for experimentation. For evaluation, we selected 10 and 15 talks from INTRA as our dev and test sets, referred to as AUTO-DEV and AUTO-TEST. Importantly, these sets do not intersect with MuST-C training data. Additionally, we curated manual dev and test sets for testing.

3.3. Dev & Test Set Curation

To guarantee the quality of the dev and test sets, we randomly selected four and five talks interpreted by S-rank interpreters for annotation for the subsets.[6] Data curation involves a two-step process.

Label aligning We observed that a correctly aligned pair may not be suitable for evaluation due to issues such as *under-translation* or *mistranslation*. For instance, in the following pair, the target SI sentence aligns correctly with the source sentence but is not a correct translation.

Source: They haven't given up on government.

Interpretation: 政府は、諦めていないのです。 [The government hasn't given up.]

[6]The annotators are two PhD students with backgrounds in interpreting and language analysis. Agreement on the annotations was reached between them in cases of ambiguity.

Another instance involves words with similar pronunciations, especially for proper nouns. While hearing them in a speech may be appropriate, using them in written text could lead to errors. Hence, we instructed the annotators to assign two labels: *good_align* for well-aligned pairs, and *good_mt* for suitable for evaluation. Assigning the *good_align* label is generally straightforward and less ambiguous. However, assigning the *good_mt* label requires a judgement call; annotators collaborated on this task to reach a consensus and minimize bias. It is unrealistic to expect perfect interpretations; therefore, the principle for annotation dictates that omitting important content (e.g., numbers and proper nouns) is not allowed, but omitting less important content is permissible.

For the above example, the labels would be True and False, respectively.

Sentence editing We further asked annotators to perform two sequential tasks on the target sentence. (1) Rectify alignment issues from automatic alignment. This is mainly done by introducing "true COARSE" for the test set if chunks are removed from or re-introduced to the target sentence as a result of auto alignment. We note that the need of merging sentences is rare; even in the case of merging, the resulting sentences do not lead to good pairs. This can be explained by human's behavior that when interpreters have to merge sentences, they make mistakes more easily. (2) Perform minimal manual post-editing to ensure faithfulness and reliability without altering the data distribution significantly. Examples include, but not limited to, replacing pronouns with referred entities, removing self-correction to improve fluency, and correcting numbers.

In the example below, the annotator would add the text highlighted in red to the target sentence.

Mod_Interpretation: 彼らは政府を諦めていないのです。 [They haven't given up on government.]

The entire process, consuming approximately 20 hours, resulted in SI^{TEST}. Additionally, we manually annotated three talks for the dev set, creating SI^{DEV}. Data statistics are provided in Table 2.

4. Experiments

4.1. Experimental settings

Datasets In addition to our proposed data shown in Table 2, we also complemented it with the Must-C v2 En-Ja dataset, which is an offline dataset with 328,639 training instances.

baselines We applied test-time wait-k (Ma et al., 2019) on offline MT models trained on offline (i.e.,

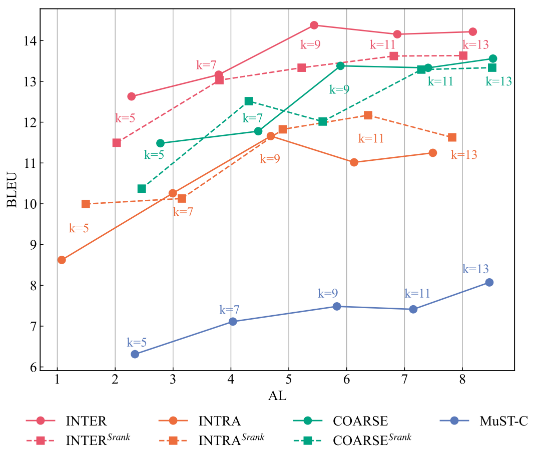

Figure 1: Translation quality and latency for wait-k systems trained on Must-C and various SI data.

Must-C) and various online datasets, including COARSE, COARSE^{Srank}, INTRA and INTRA^{Srank}, as our baseline models.

Implementation details For the wait-k systems, we primarily followed the fairseq toolkit (Ott et al., 2019) instructions with some distinctions. We used separate vocabularies for English and Japanese with SentencePiece (Kudo and Richardson, 2018) which was trained on Must-C training data. We set the vocabulary size to 8,000, as larger or smaller sizes yielded worse results. We used a batch size of 7,168 with the update frequency of 4, and set the dropout rate on 0.3. Early-stopping occurred after 16 epochs if validation loss stagnated. We selected the best checkpoint based on the loss on SI^{DEV} for all SI experiments. We evaluated model performance on SI^{TEST} with the SimulEval toolkit (Ma et al., 2020) on SI^{TEST} where translation quality is measured in BLEU[7] and latency in average lagging (AL) (Ma et al., 2019). We also measured translation quality with BLEURT. All models start from an offline MT system trained with MuST-C.

4.2. Results

Figure 1 shows BLEU and AL scores in a set of k values (i.e., 5, 7, 9, 11, 13). SiMT systems trained INTER and INTER^{Srank} outperform models trained on Must-C significantly, by an average of 6.43 and 5.74 BLEU scores, respectively, across all latency settings. This demonstrates the effectiveness of SI data despite Must-C En-Ja having significantly more data than our corpus. BLEURT scores confirm that both INTER and INTER^{Srank} offer the highest translation quality.[8]

[7] https://github.com/mjpost/sacrebleu

[8] The performance order of other systems differs with BLEU and BLEURT. We attribute this to imperfect measurement metrics for SiMT, which is beyond the scope of this work.

5. Conclusion

The question of how simultaneous interpretation (SI) data impacts simultaneous machine translation (SiMT) remains unresolved. This is due to the lack of large-scale SI training data, constructing which imposes great challenges. In this work, we filled in the gap by introducing NAIST-SIC-Aligned, an automatically aligned English-Japanese SI corpus, together with a small-scale human annotated SI test set. We proposed a two-stage alignment approach to align source data with target SI data. Our results show that systems trained on our proposed data surpassed baselines by a large margin. We share findings and insights with SI corpus, hoping to offer guidance on future research on SI corpus construction.

Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number JP21H05054.

6. Bibliographical References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Philip Arthur, Trevor Cohn, and Gholamreza Haffari. 2021. Learning coupled policies for simultaneous machine translation using imitation learning. In *Proceedings of EACL*.

Silvia Bernardini, Adriano Ferraresi, and Maja Miličević. 2016. From EPIC to EPTIC – exploring simplification in interpreting and translation from an intermodal perspective. *Target. International Journal of Translation Studies*, 28(1):61–86.

Erik Camayd-Freixas. 2011. Cognitive theory of simultaneous interpreting and training. In *Proceedings of AMTA*.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

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. 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.

Yanping Dong, Yinghui Li, and Nan Zhao. 2019. Acquisition of interpreting strategies by student interpreters. *The Interpreter and Translator Trainer*, 13(4):408–425.

Maha Elbayad, Laurent Besacier, and Jakob Verbeek. 2020. Efficient Wait-k Models for Simultaneous Machine Translation. In *Proc. Interspeech 2020*, pages 1461–1465.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.

Tomoki Fujita, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2013. Simple, Lexicalized Choice of Translation Timing for Simultaneous Speech Translation. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH-2013)*, pages 3487–3491.

Ryo Fukuda, Yui Oka, Yasumasa Kano, Yuki Yano, Yuka Ko, Hirotaka Tokuyama, Kosuke Doi, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. 2021. Naist english-to-japanese simultaneous translation system for iwslt 2021 simultaneous text-to-text task. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 39–45.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

He He, Jordan L. Boyd-Graber, and Hal Daumé III. 2016. Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. In *Proceedings of NAACL HLT*.

He He, Alvin Grissom II, John Morgan, Jordan L. 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, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 55–64. The Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Kerstin Kunz, Christoph Stoll, and Eva Klüber. 2021. Heicic: A simultaneous interpreting corpus combining product and pre-process data. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 8–14.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, et al. 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.

Xutai Ma, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. 2020. Simuleval: An evaluation toolkit for simultaneous translation. *EMNLP 2020*, page 144.

Dominik Macháček, Matúš Žilinec, and Ondřej Bojar. 2021. Lost in Interpreting: Speech Translation from Source or Interpreter? In *Proc. Interspeech 2021*, pages 2376–2380.

Shigeki Matsubara and Yasuyoshi Inagaki. 1997. Incremental transfer in english-japanese machine translation. *IEICE TRANSACTIONS on Information and Systems*, 80(11):1122–1130.

Akira Mizuno. 2017. Simultaneous interpreting and cognitive constraints. *Bull. Coll. Lit*, 58:1–28.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. Fairseq: A fast, extensible toolkit for sequence modeling. *NAACL HLT 2019*, page 48.

Jun Pan. 2019. The chinese/english political interpreting corpus (cepic): A new electronic resource for translators and interpreters. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 82–88.

Matthias Paulik and Alex Waibel. 2009. Automatic translation from parallel speech: Simultaneous interpretation as mt training data. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 496–501.

Heike Przybyl, Ekaterina Lapshinova-Koltunski, Katrin Menzel, Stefan Fischer, and Elke Teich. 2022. Epic uds-creation and applications of a simultaneous interpreting corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1193–1200.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643.

Kilian Seeber. 2001. Intonation and anticipation in simultaneous interpreting. *Cahiers de linguistique française*, 23:61–97.

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.

Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2013. Constructing a speech translation system using simultaneous interpretation data. In *Proceedings of IWSLT*.

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

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003.

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. In *Proceedings of the Second Workshop on Automatic Simultaneous Translation*, pages 28–35.

Jinming Zhao, Philip Arthur, Gholamreza Haffari, Trevor Cohn, and Ehsan Shareghi. 2021. It is not as good as you think! evaluating simultaneous machine translation on interpretation data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6707–6715.

Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019. Simultaneous translation with flexible policy via restricted imitation learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5816–5822.