Sara Schwartz
LIN 308 Final Paper
May 9th, 2023

Examining English-Japanese Simultaneous Interpretation
日英 同時 通訳 の 検討

Introduction

Simultaneous interpretation (SI) is the task of translating speech from a source language into a target language in real time (Doi, Sudoh, & Nakamura, 2021). In contrast to consecutive interpretation, where the interpretation is done after the speaker pauses, in SI the translation process starts while the speaker is still talking. The simultaneous nature of this work creates a major challenge concerning pacing. On one hand, the further behind the speaker an interpreter gets, the more they have to keep listening and processing the speaker's words in short-term memory. On the other hand, the closer to the speaker the interpreter gets, the more likely they are to miss the big picture and make grammar, syntax, or style errors.[1] No matter the pacing, simultaneous interpreters need to pick up the real topic of a long-winded sentence as quickly as possible and reformulate them without knowing where they are going. They also must divide their attention between what they are saying and what they are hearing the speaker say, which quickly becomes incredibly exhausting (Bellos, 2011). As these challenges are reflected in the types and patterns of errors observed in SI transcriptions, analyzing SI transcriptions can help gain a better understanding of the cognitive processes at work behind SI and promises valuable insights into improving SI quality.

Parallel to human translation and interpretation, the research in automatic speech processing can also be divided into two subfields. Automatic speech translation aims to convert speech signals in the source language to the target language, while automatic simultaneous interpretation aims to minimize the delay between the speaker and the translation. Thus, while corpora used to train speech translation systems are generated based on complete audio data or transcripts, SI corpora are generated from transcribing real-time human interpretation (Shimizu

[1] https://www.youtube.com/watch?v=twCpijr_GeQ

et al., 2014). Previous works have mainly focused on consecutive speech translation between English and Indo-European languages. As a result, many existing SI corpora are too small to train data-hungry neural models (Zhang et al., 2021).

In “Large-Scale English-Japanese Simultaneous Interpretation Corpus: Construction and Analyses with Sentence-Aligned Data”, authors Doi, Sudoh, & Nakamura addressed this scarcity of publicly available SI data by creating a large-scale English↔Japanese SI Corpus. In this paper, I review the construction and analysis of this corpus. I compare their methodology to several other related works and evaluate their proposed results in connection with the practice of professional SI. Based on these insights, I re-analyze a subset of their corpus and discuss the implications of these findings both for human interpreters and for future research on the development of automatic interpretation systems.

Review of Doi, Sudoh, & Nakamura (2021)

In the following section, I conduct a thorough review of Doi, Sudoh, & Nakamura's (2021) approach to the creation and analysis of their English↔Japanese SI Corpus.

Corpus Creation

To create their SI corpus, professional simultaneous interpreters were recruited to simultaneously interpret speech from TED talks.[2] Based on their years of experience, the interpreters were categorized into the following ranks: `S-rank` (15 years), `A-rank` (4 years), and `B-rank` (1 year). In a professional SI environment, interpreters utilize both visual information (e.g. expressions, gestures, and slides) and audio information (e.g. intonation) to inform their interpretations. To simulate this environment, Doi, Sudoh, & Nakamura had interpreters both watch a video alongside an audio recording of each talk, and interpreters were given documents related to each speech in advance. However, the authors did not specify what content was included in these documents. Existing research has reported that providing access to presentation slides can greatly improve the quality of interpretation (Shimizu et al. 2014). Furthermore, for speeches like TED talks that may involve technical terminology and proper names, materials that allow the interpreter to familiarize themselves with the subject matter can be invaluable. Ensuring that all of the necessary information was provided to these interpreters may improve the quality of their resulting interpretations.

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

After collecting the SI data, these transcripts were manually aligned to the source speech based on segments at the sentence level. The data was automatically divided into *bunsetsus*, basic units of dependency in Japanese that consist of one or more content words and any following function words. Each *bunsetsu* that appeared in the translation was considered a unit of ideas (Doi, Sudoh, & Nakamura, 2021).

Corpus Analysis

The analysis was conducted on English→Japanese SI data from a subset of 14 TED talks that have SI data from three interpreters. Additional translation data (e.g. Japanese subtitles) were also used to compare with the SI data, allowing the authors to examine the SI quality and word order.

Doi, Sudoh, & Nakamura (2021) identified several patterns in interpretation style across the three ranks of interpreters. Some of these patterns intuitively align with the interpreters' years of experience: they found that more experienced interpreters tended to have a lower ratio of segments that were unintentionally omitted from interpretation (`drop`) and a higher ratio of segments that were intentionally omitted from interpretation (`skip`). Other findings were less obvious: `B-rank` interpreters produced the highest number of *Bunsetsu*, and frequently added segments not spoken by the original speaker (`en null`). This suggests that inexperienced interpreters are less comfortable with rephrasing and condensing the original speech to suit the fast-paced nature of SI.

Next, the authors compared three properties that are commonly investigated in translation studies: latency, quality, and word order (Doi, Sudoh, & Nakamura, 2021). They used Ear-Voice Span (EVS), a measure of the lag between the original utterances and the corresponding interpretation, to evaluate latency. `A-rank` interpreters had the largest `EVSend`, or difference between the end of the original speech in the source language and the end of the interpreted speech in the target language. In addition, approximately 57% of the segments following the segments with the largest `EVSend` were unintentionally omitted (`drop`) by `A-rank` interpreters. It appears that these moderately experienced interpreters more diligently attempt to thoroughly interpret each segment, while the less experienced (`B-rank`) and more experienced (`S-rank`) interpreters prefer to quickly wrap up each segment.

To evaluate the quality of interpretations, Doi, Sudoh, & Nakamura (2021) used two metrics. The *Bunsetsu*-level semantic preservation score (BSPS) was used to evaluate the faithfulness of the SIs against their corresponding reference translations by measuring how many

ideas from the original speech were covered in each interpreted sentence. The results were as expected, where higher-ranked interpreters achieved a higher BSPS. The second metric, BERTScore, is based on contextualized subword embeddings and was used to compare aspects of meaning between SIs and their translations (Zhang et al., 2019). Mirroring the idea that interpreters sometimes synthesize the content of the original speech to handle latency, precision was higher than recall for all three ranks of interpreters. However, BERTScore was not an appropriate metric in all cases, especially when interpreters used a strategy. Table 1 below describes an example of such a case. The F1 score of this example was approximately 0.55, demonstrating that BERTScore did not successfully capture the interpreter's strategy. Here, the strategy appears to be synthesizing the original utterance while still conveying its core ideas.

Table 1: Synthesis Strategy Example

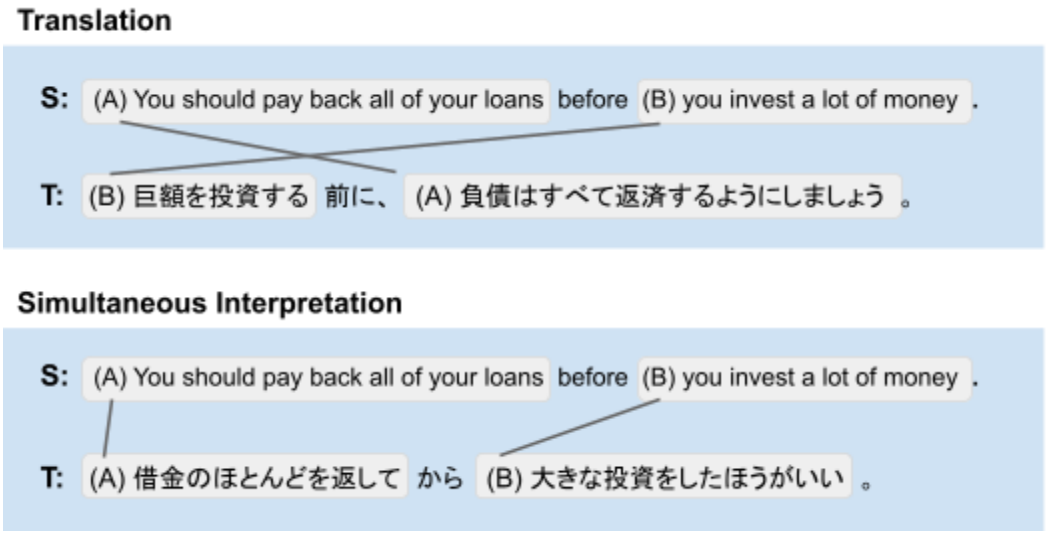
Original Sentence (English)	We can all think of some examples, right?
Reference Translation (Japanese)	例を挙げる事ができると思います。
S-Rank Simultaneous Interpretation (Japanese)	例えば、[For example]

Finally, to investigate differences in word order, the authors used Kendall's K distance to measure the degree of reordering between SIs and their corresponding reference translations. Given the grammatical difference between English (SVO and head-initial) and Japanese (SOV and head-final), it is unsurprising that they found large differences between the word order of the interpretations and the translations. By choosing not to reverse the original sequence of *bunsetsu* whenever possible, interpreters do not need to wait for the end of the sentence to begin their interpretation, thus decreasing the amount of information they must hold in their memory. However, this strategy sometimes requires additional efforts to "repair" or "join" segmented sentences to make sense in the target language. This poses a tradeoff between accuracy and efficiency: throughout SI, interpreters must decide whether to wait for more information and properly inverse the SVO source language into the SOV target language, or start interpreting current items to reduce the delay and working memory load.

In Figure 1 below, I illustrated an example case where word order was reversed between the original sentence and the translation but maintained between the original sentence and the SI. For both the translation and the SI of the original sentence in the source (S) language, the resulting translation into the target (T) language is below. The main clause is word string A, and the adverbial clause is word string B. The Japanese translation of this sentence inverts these two

clauses, standard to Japanese grammatical structure. In contrast, the Japanese SI of this sentence maintains the original order of the two clauses. To compensate for this, the original sentence `“A before B”` is interpreted more along the lines of `“A [after which] B”.`

Figure 1: Comparing an original sentence, its translation, and its SI.



However, no clear pattern emerged concerning interpreter ranks. I wonder if an interpreter’s familiarity with other SOV languages, bias towards their L1, and other factors related to their background may have influenced these results. In a paper by Shimizu et al. (2014), which focused on the creation of a smaller English↔Japanese simultaneous translation corpus, the authors note that all of their interpreters were native Japanese speakers and worked as professional interpreters in both directions (English→Japanese and Japanese→English). If the same is true of Doi, Sudoh, & Nakamura's (2021) work, it may be the case that the interpreters demonstrated a preference towards the syntax of their L1, and therefore were less inclined to reverse the word order of the source language. Additionally, the “Successive Language Acquisition" chapter of the textbook “The Psycholinguistics of Bilingualism” examined the competition and interaction that occurs between successive bilinguals’ L1 and L2, and how these relations differ based on successive bilinguals’ phonological and lexical knowledge of their L2. The chapter reviewed several studies demonstrating that simultaneous bilinguals have a different structural representation of their L2 compared to successive bilinguals and that the organization of the L2 is dependent on the successive bilinguals' age of acquisition of their L2. Therefore, I am curious as to whether a simultaneous interpreter’s age of acquisition of English as their L2 may also influence their performance. Including this sort of background information on the participating interpreters would be a valuable addition to this paper.

Despite these limitations, Doi, Sudoh, & Nakamura (2021) do align with findings from previous works. In particular, Cai et al. (2020) conducted a statistical comparison of English↔Japanese translation and SI and found that simultaneous interpreters often preferred maintaining the word order in the original speech. Furthermore, they investigated multiple factors that may influence an interpreter's choice of strategy. Looking at syntactic factors, or factors due to grammatical differences between the source and target language, one interesting insight they gleaned was on the length of the post-modifier. In traditional English↔Japanese translation, the post-modifier is primarily translated first. In SI, however, if the post-modifier was longer than three words, interpreters tended to avoid waiting until the end of the post-modifier to begin interpreting, thus maintaining the original word order. Similarly, if two chunks are related by a grammatical dependency and are far away from each other, an interpreter has to retain a massive quantity of information in their working memory before the posterior chunk is input and the relation can be decided. Again, interpreters tended to interpret in the original word order to avoid overload.

Cai et al. (2020) also investigated non-syntactic factors, or factors imposed by the complicated circumstances of simultaneous interpretation. Similarly to Doi, Sudoh, & Nakamura (2021), they hypothesized that an experienced interpreter may be more skilled in the strategy of maintaining the original word order. However, Cai et al. (2020) conducted their analysis of 17 interpreters divided into two groups, less than 10 years and greater than 10 years of experience, and did not find a difference at the 0.05 significance level between the two groups. Although they presented compelling evidence that syntactic factors (e.g. dependencies of chunk pairs and length of post-modifier) affect word order strategy more than non-syntactic factors, the correlations presented between these factors and word order were generally weak. It may be beneficial to redo this analysis while combining Cai et al.'s (2020) larger participant pool (17 interpreters) with Doi, Sudoh, & Nakamura's (2021) finer partitioning of groups (1 year, 4 years, and 15 years). Investigating the effect of multiple factors combined also may help to arrive at significant results.

Returning to Doi, Sudoh, & Nakamura's (2021) analysis, the authors then examined the trade-off relationship between latency and quality. They found that as latency increased, the number of *bunsetsu* in interpretations decreased, aligning with previous studies that have shown that higher latency damages quality. They also found that as the gap between the start time of the original speech in the source language and the start time of the interpretation in the target language (`EVSstart`) increased, the interpretation's BERTScore and BSPS decreased. Considering these findings alongside the previously mentioned negative effect of a large

EVSend on the following sentence, these results further emphasize why interpreters often have to prioritize latency at the expense of quality.

As a final method of analysis, Doi, Sudoh, & Nakamura (2021) asked three professional translators to subjectively evaluate the faithfulness of each sentence on a 4-point scale. The resulting human evaluation scores were low, most often less than 2. The authors reason that one possible reason for these low scores is that the translators were strict about the sentence structure in the source language. Previous works have used similar approaches to evaluating SI data. For example, following their creation of a large-scale Chinese↔English speech translation dataset, Zhang et al. (2021) asked human translators to evaluate the transcribed interpretations of their human interpreters based on adequacy, fluency, and correctness. Compared to Doi, Sudoh, & Nakamura's (2021) 4-point scale, Zhang et al. (2021) used a more descriptive ranking system. They defined "acceptable" translations as those that either contained no obvious errors, or those that were comprehensible but may have had "minor errors such as incorrect function words and less fluent phrases" (Zhang et al., 2021). This design, which prioritizes measuring the faithfulness of the translations rather than their exact word-for-word accuracy, seems better equipped to capture the nuances of SI. Nevertheless, due to the unique strategies that simultaneous interpreters employ to work around time and working memory constraints, it seems unreasonable to expect that the resulting interpretations would match those expected by professional translators. This raises the question: how can we evaluate the semantic similarity between sentences that are structurally different from their corresponding reference translations? Regardless of the exact measure that would be best suited for this task, it seems to me like future work should at the very least involve human evaluation with other simultaneous interpreters that understand these constraints.

Overall, Doi, Sudoh, & Nakamura (2021) conducted a deep analysis of their proposed English↔Japanese SI corpus that led them to highlight common SI errors and strategies, including the influence of the post-modifier length on word order and the inverse relationship between SI experience and the number of produced bunsetsu. These findings have potential applications in improving SI practices for human interpreters, decreasing latency and increasing quality, and possibly even finding methods to decrease the working memory load on these interpreters.

Added Analysis: Word Substitution

Background

The Japanese language uses three different writing systems: Hiragana, Katakana, and Kanji. Hiragana and Katakana are phonetic symbols, each representing one syllable. Typically, native Japanese words are written using Hiragana, and words borrowed from other languages are written using Katakana. Kanji are symbols that represent parts or all of individual words or concepts.

One line of investigation that does not seem to have been explored by Doi, Sudoh, & Nakamura (2021) or related works is word substitution between English and Japanese. During SI, it seems reasonable to believe that interpreters may occasionally substitute the Japanese translation of an English word for the original English word, particularly for acronyms, proper names, and technical terms. I was curious to see if this strategy extends to the use of Katakana. Although these alphabets are technically writing systems, spoken words in Katakana can be distinguished from spoken words in Hiragana or Kanji due to their pronunciation, where Katakana uses Japanese phones that are similar to the pronunciation of the language the word is originally borrowed from. For example, the Japanese word for *star* is 星「ほし」. However, it is not uncommon to see *star* written using Katakana: スター, which is pronounced similarly to the pronunciation of *star*: *sutā*. This led me to ask the question: is there a relationship between an interpreter's experience and the frequency of English or Katakana substitution of a Japanese word traditionally written in Hiragana or Kanji? Originally, I hypothesized that less experienced interpreters would be more likely to rely on both English and Katakana word substitution.

Distribution of Written Systems

To begin exploring this question, I downloaded the subset of the English↔Japanese SI Corpus created by Doi, Sudoh, & Nakamura (2021) that was publicly available for download.[3] This data included the 14 TED talks that had been interpreted by all three ranks of interpreters. For each of these talks, the corpus included the English and Japanese subtitles of the talk in JSON format and SI transcripts from each interpreter given by a tab-separated text file in UTF-8 encoding. After extracting the relevant data from these files, I conducted a general analysis of the Japanese SI transcripts.

---

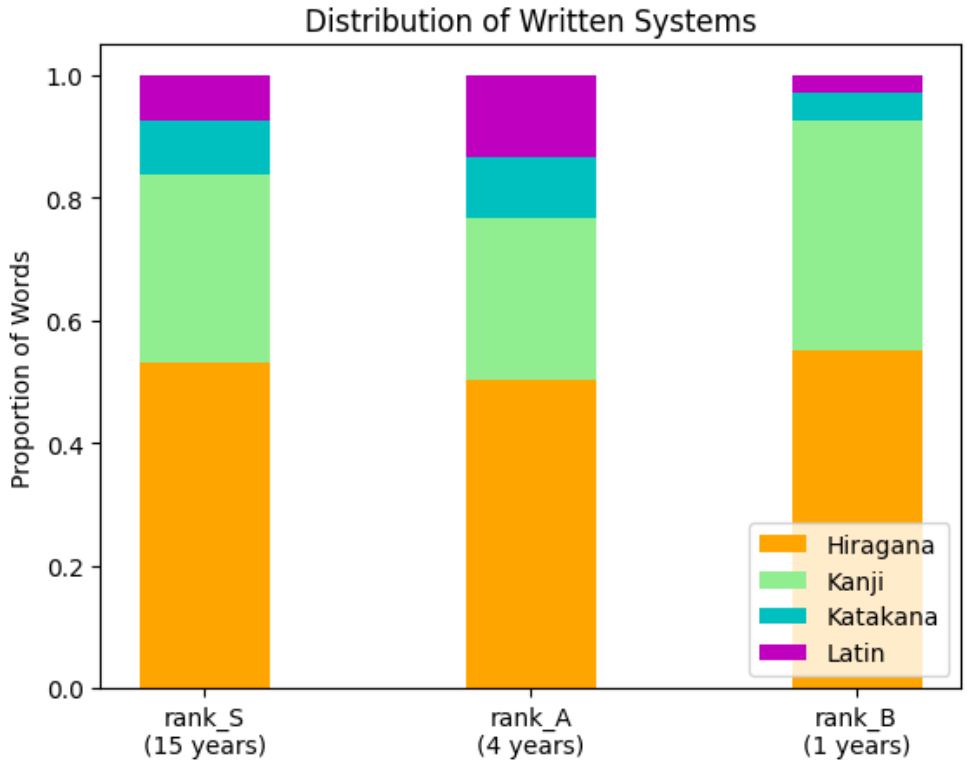
[3] https://dsc-nlp.naist.jp/data/NAIST-SIC/

The top 10 most frequently used English words are listed below. As expected, many of these words are the names of products, schools, and acronyms specific to the United States:

`AI` (42), `PTSD` (21), `NASA` (12), `CO` (7), `iPhone` (5), `MIT` (5), `CEO` (4), `NRA` (3), `IT` (3), `GDP` (3)

Next, I analyzed the number of words in each written system. For each of the three interpreter ranks, I averaged the number of words in Hiragana, Katakana, Kanji, and Latin (English) across all 14 talks, and calculated the proportion of total words in each of the 4 writing systems. The stacked bar chart in Figure 2 below illustrates this distribution. In direct contrast to my hypothesis, we can see that `B-rank` made the smallest number of both English and Katakana word substitutions, and `A-rank` made the most.

Figure 2: Proportion of total words in each of the 4 writing systems.



Considering Doi, Sudoh, & Nakamura's (2021) findings that `A-rank` interpreters had the largest difference between the end of the original speech in Japanese and the end of the interpreted speech in English (`EVSend`), I wonder if this trend points to a shift in preferred strategies as interpreters gain experience. Less experienced interpreters (`B-rank`) may be focused on fully interpreting the English speech in Japanese, sacrificing clarity and fluidity in

speech. Then, as interpreters gain experience (A-rank), they may begin utilizing both English and Katakana word substitution as a tool to save time and memory load. However, I can imagine that more frequent switching between English and Japanese in SI may increase the likelihood of language interference and general confusion, and therefore may have contributed to Doi, Sudoh, & Nakamura's (2021) findings that A-rank interpreters have a higher frequency of unintentional omissions during sentences with larger EVSend as compared to other ranks. Therefore, as interpreters continue to gain experience (S-rank), they may decrease English word substitution while maintaining Katakana word substitution, allowing them to fill in blanks without switching languages. Although this line of thinking provides an interesting way to contextualize Doi, Sudoh, & Nakamura's (2021) findings, in order to validate this theory I would need to perform much more rigorous data collection and analysis.

Intersection of Latin and Katakana Word Sets

To examine which words were commonly substituted in Japanese interpretation, I compiled a complete set of unique Latin and Katakana words and compared the intersection of sets for each interpreter rank. The Venn diagram in Figure 3 illustrates the set intersections for Latin words, and the one in Figure 4 illustrates the set intersections for Katakana words.

Figure 3

Intersection of Latin Sets for Each Rank

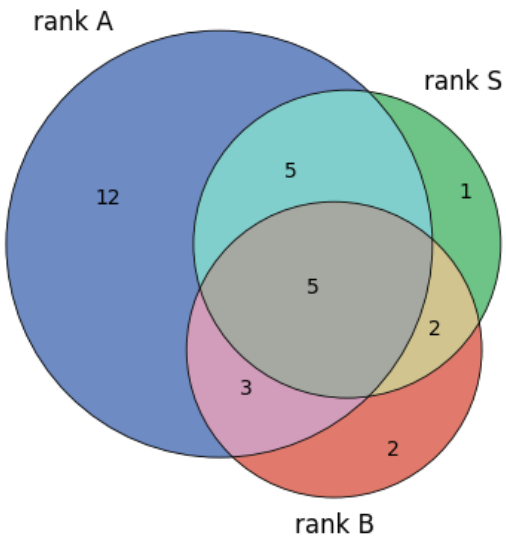
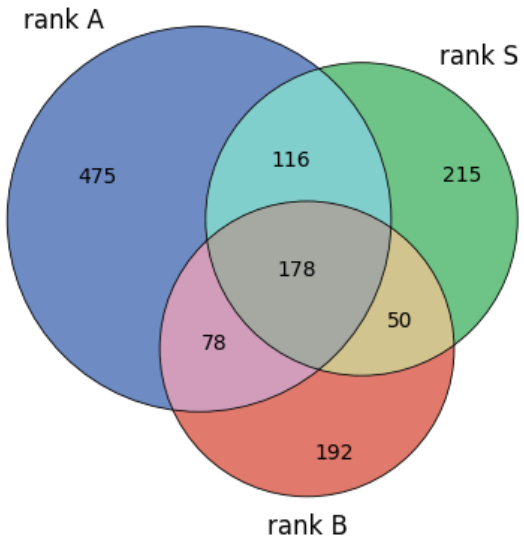


Figure 4

Intersection of Katakana Sets for Each Rank



Immediately, we can see that in both cases, rank-A's set covers the majority of both Latin and Katakana words used by all three ranks. Also in both cases, the intersection between rank-A and rank-S is greater than the intersection between rank-A and rank-B.

Table 2 below lists the total number of unique Latin and Katakana words used by each interpreter rank. Here, we see that the set of Latin words is much larger for rank-A than the other two ranks, which helps explain why this set covers the majority of Latin and Katakana words used by all three ranks, and also aligns with the fact that rank-A interpreters used Latin and Katakana word substitution proportionally more frequently than the other two ranks. We can also see that the sets of Latin and Katakana words for rank-S and rank-B are fairly similar sizes. Combining these observations with those above, it appears that rank-A interpreters use Latin and Katakana substitution frequently for different words, while rank-S interpreters use Latin and Katakana substitution at a slightly lower frequency for a smaller set of words. This suggests that rank-S interpreters may be using word substitution as a more intentional strategy than rank-B interpreters.

Table 2: Number of Unique Latin and Katakana Words

Rank	Num. Unique Latin Words	Num. Unique Katakana Words
rank-S	13	559
rank-A	25	847
rank-B	12	498

To further test this hypothesis, I calculated the frequency of each Latin and Katakana word used in substitution for each rank by dividing the number of unique words by the total number of recorded word substitutions. The results are shown in Table 3 below. These results further contribute to my theory that although rank-A interpreters made the largest number of word substitutions, they also used these words with the lowest frequency. Meanwhile, rank-S interpreters used their smaller set of words at the highest frequency. Given this emerging pattern in an interpreters choice of word substitution strategy, I am even more curious to see whether future works that expand on the investigation of word order by Doi, Sudoh, & Nakamura (2021) and Cai et al. (2020) may observe a similar pattern, where moderately experienced interpreters begin exploiting more complicated strategies, and more experienced interpreters demonstrate a more refined use of the same strategies.

Table 3: Frequency of Latin and Katakana words used in substitution.

Rank	Latin Word Frequency	Katakana Word Frequency
rank-S	0.4	0.33
rank-A	0.31	0.19
rank-B	0.36	0.42

In sum, this additional layer of analysis on word substitution across SI from interpreters of varied experience levels provides a different perspective from which we can analyze SI corpora like the one created by Doi, Sudoh, & Nakamura (2021). Future research may be able to answer some of the remaining questions, including how we can best evaluate the semantic similarity of SI against reference translations and how an interpreter's prior experience may inform their choice of SI strategy. Continued investigation of common SI errors and strategies made by human interpreters has potential applications in improving SI training and identification of individuals who are more likely to excel at this challenging work. In addition, gaining a better understanding of effective SI strategies used by experienced interpreters can help inform our design of automatic SI systems. For example, other factors to probe include: What influences an interpreter to predict the direction of the sentence rather than waiting to hear the remaining content? How do interpreters decide when to start translating, and how do they maintain a pace that maximizes the tradeoff between speed and accuracy? By continuing to create more extensive and more inclusive SI corpora for a variety of language pairs annotated by a larger pool of interpreters and translators, we may continue to discover more conclusive answers to these questions.

References

Doi, K., Sudoh, K., & Nakamura, S. (2021, August). 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)* (pp. 226-235).

Zhang, R., Wang, X., Zhang, C., He, Z., Wu, H., Li, Z., ... & Li, Q. (2021). Bstc: A large-scale chinese-english speech translation dataset. *arXiv preprint arXiv:2104.03575*.

Shimizu, H., Neubig, G., Sakti, S., Toda, T., & Nakamura, S. (2014, May). Collection of a Simultaneous Translation Corpus for Comparative Analysis. In *LREC* (pp. 670-673).

Cai, Z., Ryu, K., & Matsubara, S. (2020, December). What affects the word order of target language in simultaneous interpretation. In *2020 International Conference on Asian Language Processing (IALP)* (pp. 135-140). IEEE.

Chiari, S. (2012). David Bellos, Is That a Fish in your Ear? Translation and the Meaning of Everything. London, Penguin Books, 2011, 390 p. ISBN: 978-1-846-14464-6. *E-rea. Revue électronique d'études sur le monde anglophone*, (10.1).

Grosjean, F., & Li, P. (2013). *The psycholinguistics of bilingualism*. John Wiley & Sons.