

COLLECTING OF SIMULTANEOUS INTERPRETATION PATTERNS USING BILINGUAL SPOKEN LANGUAGE CORPUS

Hitomi Tohyama[†] *Shigeki Matsubara*^{*}

[†]Graduate School of Information Science, Nagoya University

^{*}Information Technology Center, Nagoya University

ABSTRACT

This paper provides an investigation of simultaneous interpreting patterns using a bilingual spoken monologue corpus. 4,578 pairs of English-Japanese aligned utterances in CIAIR simultaneous interpretation database were used. This investigation is the largest scale as the observation of simultaneous interpreting speech. As a result, we collected 203 cases out of aligned utterances in which simultaneous interpreters' strategies for raising simultaneity were observed. These 203 cases could be categorized into 12 types of interpreting pattern. These interpreting patterns can be expected to be used as interpreting rules of simultaneous machine interpretation.

1. INTRODUCTION

To achieve the supporting environment for the natural and smooth multilingual communication, the studies on the simultaneous machine interpreting has been conducted at several institutions [1]. In order to develop the simultaneous interpretation system with high performance, it is important to observe how professional interpreters generate simultaneous interpretation and to make use of their techniques. Their skills are acquired by tremendous amount of training over many years. Since there existed no large-scale data on the English-Japanese simultaneous interpretation, the study which theorized the English-Japanese simultaneous interpretation based on the actual data is yet to be done. Therefore, it is actually indispensable for us to collect various interpreting patterns by observing thoroughly a large-scale interpretation data.

2. INVESTIGATED DATA

In this research, we used English-Japanese monologue data in CIAIR simultaneous interpretation corpus [2]. The contents of speeches are economics, history, and, culture, etc. Moreover, each monologue speech is interpreted by two or four professional interpreters. The flexibility of a corpus is raised by using four interpreters. Therefore, it becomes possible to compare two or more interpretation examples in

Table 1. Statistics of investigated data

	Speaker (English)	Interpreter (Japanese)
No.of lectures	26	26
recording time (min)	15,627	15,627
utterance time (min)	12,780	11,001
No.of words/morpheme	38,437	44,399
No.of utterance units	1,494	1,977
No.of alligening papirs		4,578

a sample of utterances. The speech was recorded for about 10 minutes per lecture. In this research, 4,578 pairs of English-Japanese aligned utterances in the corpus were used. This investigation is the largest scale as the manual observation of English-Japanese simultaneous interpreting data. Table1 shows the statistics of the investigated data.

3. TRANSLATION TECHNIQUES EXAMINED

In this research, the complex sentence and the passive construction were focused. The reason why those sentences were focused was that complex sentences are usually long and the passive constructions in the original English sentences have to be translated as the active sentences in Japanese. And this linguistic difference between English and Japanese is the obstacle for the simultaneous machine translation because it is hard for the system to determine translation units and translation timing. Therefore, we paid attention to interpreter's technique such as when and how they translate the speaker's utterances and how they omit unnecessary words and phrases.

4. COLLECTING THE INTERPRETING PATTERNS

The table2 shows the examples of interpreters' technique to translate concerning complex sentences with conjunction, relative clauses, and passive construction sentences, which were discovered by observing the corpus. The following alphabets and numbers are the symbols used in order to explain interpreting patterns. They show what each symbol stands for. "X", "Y" are clause. "Z" is words. One example is shown as follows.

Table 2. The result of the investigation

Strategy	Syntax type of sentence	English sentence structure (speaker)	Standard Japanese sentence structure	Interpretation technique with high simultaneity (interpreter)	Occurrence count in the data	Occurrence count (the proportion of the technique (%))
Interpretation according to English structure (A-1)	sub-conjunction	<X because Y.>	<Y <i>nano-de</i> X.>	a. <X, <i>naze-nara-ba</i> Y.> b. <X, Y <i>nano-de-su</i> .> c. <X, Y.>	27	a. 23(85.2) b. 1(3.7) c. 1(3.7)
		<X before Y.>	<Y <i>su-ru mae-ni</i> X.>	a. <X <i>shi-te kara</i> Y.> b. <X <i>shita ato-ni</i> Y.>	4	a. 2(50.0) b. 1(25.0)
		<X even if Y.>	< <i>tatoe</i> Y <i>de-aru to-shite-mo</i> X.>	<X, Y <i>de-aru-to-shi-te-mo</i> .>	5	4(80.0)
	relative	<X which Y. >	<Y <i>de-aru</i> X.>	a. <X, <i>kore-ni kann-shi-te-wa</i> Y.> b. <X, Y, X>	30	a. 20(66.7) b. 4(13.3)
		<X where Y.>	<Y <i>to-iu</i> X.>	<X, Y.>	17	8(47.0)
Omission (A-2)	subject + perception verb	<Z think Y.>	<Z <i>wa</i> Y <i>to omoi-masu</i> .>	a. <Y <i>to omoi-masu</i> .> b. <Y <i>de-aru</i> .> c. <Y <i>de-syou</i> .>	101	a. 58(57.4) b. 35(34.7) c. 2(1.98)
	passive construction	<Z be+past participle.>	<Z <i>wa sa-reru</i> .>	<Z <i>wo su-ru</i> .>	79	65(82.3)

“before” Clause

The reverse clause order also can be observed in the sentences with “before.”

(1): X before Y.

(2): Y *su-ru mae-ni* X.

One interesting technique was observed. In the technique, the simultaneous interpreters generate the phrases such as “*shi-te kara*” and “*shi-ta ato-ni*” which have opposite meaning of “*su-ru mae-ni*” seen in (2) after they generate X, and, after that, they translate Y. The following two interpreting patterns were collected in this research.

<Interpreter's Techniques >

(3)-a: X *shi-te kara* Y. (Refer to Figure 2)

(3)-b: X *shi-ta ato-ni* Y.

Corpus Example (3)-a

N: Okay, you should pay back ahh most or all of your loans **before** you invest a lot of money.

I: *Mina-sann-no syakkinn-no hotonndo-wo kaeshi-te-kara toushi-wo suru-to-iu koto-ni shita-hou-ga ii-to omoi-masu.*

Corpus Example (3)-b

N: You should pay back ahh most or all of your loans

I: *kurezittokaado-de-no kariire-wo*

N: **before** you invest a lot of money.

I: *Hennsai-shita ato toushi-wo kanngae-nasai to-iu -koto-de-ari-masu.*

5. CONCLUSION

It is one of the effective techniques to use the translation strategies collected by observing interpreter's behavior for

the development of the simultaneous machine interpreting. In this research, 4,578 pairs of English-Japanese aligned utterances were examined. This investigation was large-scale as the observation of simultaneous interpretation data manually. As a result, we observed 203 interpreting patterns. These were able to be classified into 12 interpreting pattern. It was clarified that 4.5 percent of the English-Japanese monologue data were fitted in those interpreting patterns and those patterns can be expected to be used as translating rules for developing simultaneous machine interpretation. It will also be necessary to collect more simultaneous interpreters' techniques, such as how they translate other English sentence structures in addition to complex sentences. As for developing the simultaneous interpreting system, the further research concerning how to apply those interpreting patterns to the simultaneous interpreting system has to be done.

Acknowledgment: The authors would like to thank Prof. Toyohide Watanabe, Assoc. Prof. Nobuo Kawaguchi, and all members of Matsubara Laboratory at the Nagoya University for their precious advices.

6. REFERENCES

- [1] H. Mima, H. Iida, and O. Furuse., “Simultaneous Interpretation Utilizing Example-based Incremental Transfer,” *Proc. 17th COLING and 36th ACL*, pp.855-861. 1998.
- [2] H.Tohyama, S.Matsubara, N. Kawaguchi, and Y. Inagaki, “Construction and Utilization of Bilingual Speech Corpus for Simultaneous Machine Interpretation Research,” *Proceedings of 9th European Conference on Speech Communication and Technology*, pp. 1585-1588, Portugal ,2005