Bilingual Spoken Monologue Corpus for Simultaneous Machine Interpretation Research

Shigeki Matsubara[*], **Akira Takagi**[†], **Nobuo Kawaguchi**[*], and **Yasuyoshi Inagaki**[†]

[*]Information Technology Center/CIAIR, Nagoya University
[†]Graduate School of Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
{matu, atakagi, kawaguti, inagaki}@inagaki.nuie.nagoya-u.ac.jp

Abstract

This paper describes a large-scale bilingual corpus of spoken monologues and their simultaneous interpretation, which has been constructed at CIAIR. The corpus has the following characteristics: (1) English and Japanese speeches are recorded in parallel, (2) the data contains monologue speeches such as lecture and self-introduction, and (3) the exact beginning and ending times are provided for each utterance. We have collected a total of about 70 hours of speech data and transcribed them into ASCII text files. The corpus will be made publicly available in the near future. This paper also provides an analysis of the professional interpreter's speeches using the bilingual corpus. The following points have been investigated: (1) the interpreting unit of simultaneous interpretation, (2) the difference between the beginning time of the lecturer's utterance and that of the interpreter's utterance, and (3) the interpreter's speaking speed. The characteristic features about the timing at which simultaneous interpreters start to speak is presented. The analysis will be available for the development of a simultaneous machine interpreting system.

1. Introduction

Machine interpreting has become one of the important research topics with the advance of technologies for speech processing and language translation. Several experimental systems of spoken dialogue translation for specific task domains have been developed so far (Takezawa, 1998; Watanabe, 2000). The interpreting style of them is within so-called consecutive interpretation. In order to provide an environment for supporting natural and smooth cross-lingual communication, to develop a technique for simultaneous machine interpretation is desired and has been tried recently (Mima, 1998; Amtrup, 1999; Matsubara, 1999).

Towards a simultaneous interpreting system, not only the quality of the interpretation but its output timing is also important, and it would be effective to investigate and analyze the interpreting process of professional simultaneous interpreters. The Center for Integrated Acoustic Information Research, Nagoya University (CIAIR), is constructing and maintaining the various types of speech and language database for the purpose of the advancement of robust speech information processing technology (Kawaguchi, 2001). Moreover, a bilingual database of simultaneous interpretation is also constructed as part of the project (Kawaguchi, 2002).

This paper describes a large-scale bilingual corpus of spoken monologues and their simultaneous interpretation, which has been constructed at CIAIR. The corpus, among other things, has the following characteristics: (1) English and Japanese speeches are recorded in parallel, (2) the data contains monologue speeches such as lecture and self-introduction, and (3) the exact beginning and ending times are provided for each utterance. We have collected a total of about 70 hours of speech data and transcribed them into ASCII text files. Furthermore, we have developed the software tools for the investigation of the corpus.

This paper provides also an analysis of interpreter's speeches using the simultaneous interpreting corpus. Using the time information in the transcript of this corpus, we have aligned the interpreter's utterances with the lecturer's utterances, and investigated the following points: (1) the interpreting units in simultaneous interpretation, (2) the difference between the beginning time of the interpreter's utterance and that of the lecturer's utterance, (3) interpreter's speaking speed, and so on. The analysis about the timing of the interpreting process of professional simultaneous interpreters will be available for the development of simultaneous machine interpretation.

2. Simultaneous Interpreting Corpus

The project of multilingual data collection in CIAIR is collecting speeches of the simultaneous interpretation in both Japanese-English and English-Japanese, and constructing a large-scale bilingual speech corpus (Aizawa, 2000). In 2001, the monologue speeches of English and Japanese and their simultaneous interpretation speeches have been collected.

2.1. Corpus Design

Several bilingual spoken dialogue corpora have been constructed so far and played an effective role to advance speech translation technologies. It is expected that those of spoken monologues would be also important.

Our database supplies the data of spoken monologues by native speakers and the interpretations of them by simultaneous interpreters. Socially major topics such as politics, economy, history, technology, and environment have been

Table 1: Outline of simultaneous interpreting corpus

type of speech	monologue(lecture)
type of interpretation	simultaneous
language	Japanese, English
data type	speech, text
topics	computer science, politics, economy, history, etc.
number of interpreters	21
number of English lectures	24 (239 min. in total)
number of Japanese lectures	15 (146 min. in total)

```
0001 - 00:05:120-00:08:024 N:
For my final topic I will speak about
0002 - 00:08:616-00:12:592 N:
the little differences between Tokyo and Kansai<SB>
0003 - 00:13:512-00:14:095 N:
(F umm)
0004 - 00:14:696-00:17:407 N:
Before I begin speaking about Tokyo and Kansai,
0005 - 00:17:920-00:20:636 N:
I'd like to speak about an analogy that
0006 - 00:21:032-00:21:296 N:
a
0007 - 00:21:704-00:22:903 N:
professor of (F mines;mine) (S had)
0008 - 00:23:215-00:24:015 N:
had said (X to the)
0009 - 00:24:216-00:25:471 N:
to his class once<SB>
```

Figure 1: Sample of the transcript (English native speech)

selected as the themes. Since simultaneous interpreting is one of the extremely advanced human language technologies, the interpreting styles may differ from one interpreter to another. To investigate the relation between the level of the skill and the interpreting result, a spoken monologue by one native speaker is interpreted by two or more interpreters whose skill is different from each other.

2.2. Data Collection

CIAIR has a research purpose of collecting large-scale speech data under real environments, and therefore our recording was carried out in a classroom. The sounds of lecturers and the interpreters were digitized by 16kHz of sampling frequencies and 16 bits, and recorded in stereo onto digital audio tapes (DAT). 4,196 minutes have already been recorded in total. Although lectures utter toward audiences, they cannot listen the interpreter's sound. It enables them to speak at the pace of themselves, seeing the audience. As English speeches, the lectures about politics, economy, history, etc., and, as Japanese speeches, those about the themes relevant to computer science have been adopted. Moreover, each monologue speech is interpreted by two or four professional interpreters of which the degree of experience differs one another (5 years over or not). Simultaneous interpreters go into a booth, and interpret the lecturer's sounds from the headphone, seeing the lecturer. The sound for about 10 minutes is recorded per lecture. In recording of interpretation, the interpreter can grasp the contents of the lecture in advance by the manuscript note of the speech. In a Japanese lecture, since the lecturer uses the presentation slide, and the interpreter can also see the slide. And, the interpreter can use the glossary of technical terms which is prepared by the lecturer.

2.3. Database Construction

The collected speech data was transcribed into the text manually. It is necessary to take into consideration about which kinds of discourse tags to be assigned. In consideration of both sides of the quality and quantity, we created two kinds of data: that by advanced tags attachment and that to which only the simple tags were given. The text transcription was done according to the manual for Corpus of Spoken Japanese (CSJ) produced by National Japanese Language Research Institute (Maekawa, 2000). Figure 1 and Figure 2 show the sample of the transcript. The discourse tags were provided for the language phenomena characteristic of spoken language, such as fillers, corrections, and misstatements. Moreover, the transcript is segmented into utterance units by the pause for 200ms or more, or that for 50ms after the sentence break, and the starting time and ending time have been provided for every utterance unit. Table 1 shows the outline of the simultaneous interpreting corpus. The database consists of wave files, transcription files and environment data files and contain about 569,000 morphemes in 26,000 utterance sentences.

```
0001 - 00:06:232-00:07:856 I:
私の最後のトピック
0002 - 00:08:062-00:08:887 I:
と(W したし;いたし)まして
0003 - 00:10:320-00:10:968 I:
(F え)ちょっと
0004 - 00:11:360-00:12:055 I:
東京と
0005 - 00:12:584-00:14:927 I:
関西が違うという話をしてみたいと思います<SB>
0006 - 00:16:328-00:18:887 I:
その東京と関西の話をする前に
0007 - 00:19:232-00:19:727 I:
まず
0008 - 00:20:144-00:22:064 I:
アナロジーを話したいと思います<SB>
0009 - 00:22:856-00:24:767 I:
私の教授の一人が
0010 - 00:25:440-00:27:919 I:
(F あー)一度授業で話をしてくれた
```

Figure 2: Sample of the transcript (English-Japanese interpreting)

2.4. Development of Corpus Use Environment

In order to make the corpus use convenient, several software tools which work on the internet have been developed using CGI script.

Table 2: Statistics of the simultaneous interpreting corpus

item	English	E-to-J	Japanese	J-to-E
recoding time (sec)	14359	57300	8767	35078
speaking time (sec)	11458	40473	5690	22289
morphemes	35474	165127	21708	63872
kinds of morpheme	3506	6115	1457	3130
utterance units	5573	20496	4505	12468
sentence breaks	1263	7423	651	3315
fillers	1251	8067	1482	3261

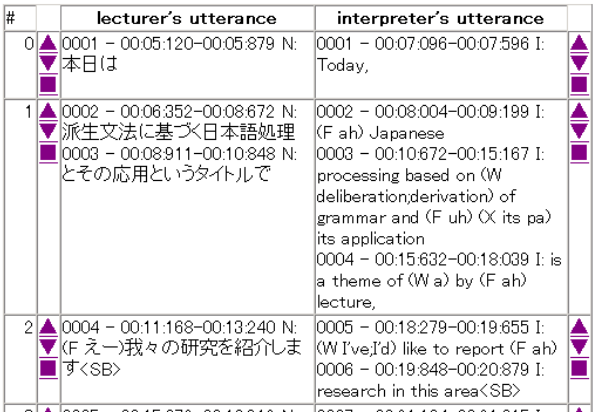

Figure 3: Alignment support tool

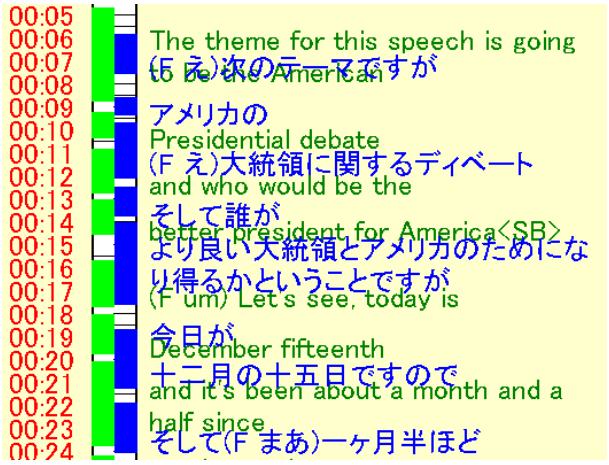

Figure 4: Timing information visualization tool

1. **Data perusal tool** (Figure 1 and 2): By GUI operation on a browser, the users attain the perusal of the data put on a WWW server computer. They can select the data types such as field data, tag removal data, and Japanese morphological analysis data.

2. **Alignment support tool** (Figure 4): Users can perform alignment work by carrying out the mouse click on the bilingual text displays. The alignment data can be used for analysis of interpreting units and timing.

3. **Timing information visualization tool** (Figure **??**): This displays the speech timing of lecturers and interpreters visually. The speaking time of English lecturers, English-Japanese interpreters, Japanese lecturers and Japanese-English interpreters and their overlap relation are displayed.

3. Statistics of the Corpus

As fundamental statistics of the simultaneous monologue interpreting corpus, using the transcript date with advanced discourse tags, we have examined the recording time, the speaking time, the number of morphemes(words), the number of different morphemes(words), the number of utterance units, the number of sentences, and the number of discourse tags. The result is shown in Table 2. Eng, E-J, Jap and J-E in the table mean English lecture, English-Japanese simultaneous interpretation, Japanese lecture, and Japanese-English simultaneous interpretation, respectively. In this paper, a morpheme in English means a word, and the number of morphemes in Japanese was calculated on the basis of the result of a Japanese morphological analyzer called ChaSen (Matsumoto, 2000). The number of kinds of morphemes in English is the number of words whose notations differ, and that in Japanese is the number of morphemes whose basic forms differ.

The interpreter's speaking time is about 4 times as much as the lecturer's time because the simultaneous interpretations by four interpreters were collected for one lecturer's speech. On the other hand, the number of utterance sentences of English-Japanese interpreters is about 6 times as much as that of the English lecturers and that of Japanese-English interpreters is about 5 times as much as that of the Japanese lecturers. This fact substantiates the empirical knowledge supported by an interpreting theory that a professional interpreter may segment a lecturer's utterance sentence into 2 or more sentences, and interpret it (Mizuno, 1995).

4. Bilingual Alignment and Its Analysis

4.1. Corpus Alignment

In order to analyze the process of simultaneous interpreting in detail, it is required to align utterances by the interpreters with those of the lecturers in a possibly small unit. We have aligned the corpus by hand according to the following conditions:

#	lecturer's utterance	interpreter's utterace
0	0001 – 00:05:120–00:05:879 N: 本日は	0001 – 00:07:096–00:07:596 I: Today,
1	0002 – 00:06:352–00:08:672 N: 派生文法に基づく日本語処理 0003 – 00:08:911–00:10:848 N: とその応用というタイトルで	0002 – 00:08:004–00:09:199 I: (F ah) Japanese 0003 – 00:10:672–00:15:167 I: processing based on (W deliberation;derivation) of grammar and (F uh) (X its pa) its application 0004 – 00:15:632–00:18:039 I: is a theme of (W a) by (F ah) lecture,
2	0004 – 00:11:168–00:13:240 N: (F えー)我々の研究を紹介します <SB>	0005 – 00:18:279–00:19:655 I: (W I've;I'd) like to report (F ah) 0006 – 00:19:848–00:20:879 I: research in this area<SB>
3	0005 – 00:15:872–00:16:912 N: (F えー)まず(F えー)	0007 – 00:21:184–00:21:615 I: First,
4	0006 – 00:17:200–00:19:919 N: コンピューターによる日本語処理についてですが	0008 – 00:22:567–00:26:472 I: (F ah) Japanese processing by computer is the first topic<SB>

Figure 5: Sample of the aligned corpus

Table 3: Statistics of the aligned corpus

	E-J: 1717 pairs		J-E: 1441 pairs	
item	Sum	Ave	Sum	Ave
overlapping time (sec)	1845	1.07	1232	0.86
difference of beginning time (sec)	5482	3.19	5955	4.13
difference of ending time (sec)	5383	3.13	5829	4.04
lecturer's speaking time (sec)	6593	3.84	5077	3.52
interpreter's speaking time (sec)	6214	3.62	5388	3.74
lecturer's morphemes	22360	13.02	18572	12.89
interpreter's morphemes	25926	15.10	15554	10.79

- The smallest unit of alignment is a utterance unit.
- The corpus is aligned as small as possible.
- There is no counterpart for the lecturer's utterance unit such as a filler, an abbreviation and a supplement, etc.

The alignment data is made by hand for 16 lectures presently in both Japanese-English and English-Japanese. Figure 5 shows a part of an aligned corpus. The left-hand side expresses Japanese lecturer's utterance units, and right-hand side expresses Japanese-English interpreter's utterance units. Each line means an alignment pair.

4.2. Statistics of the Aligned Corpus

For each alignment pair in 16 sets of the aligned corpus, we examined the following items:

- The time overlapping with the lecturer's utterance in the interpreter's utterance
- The difference between the beginning time of the lecturer's utterance and that of the interpreter's utterance.
- The speaking time of lecturers and interpreters.
- The numbers of the morphemes of lecturer's utterances and interpreter's utterances.

Table 3 shows the result. Sum and Ave in the table mean the total of the time for 16 lectures and the average of the time per alignment pair, respectively.

4.3. Interpreter's Speaking Timing

Simultaneous interpretation may overlap with the corresponding native speech. It is expected that an interpreter recognizes a part of a lecturer's utterance as an interpreting unit, and interprets it at an early stage. We have investigated the interpreting units of professional interpreter by analyzing the aligned corpus.

4.3.1. Japanese-English Simultaneous Interpretation

16 sets of the aligned data are used for the investigation. An interpreting unit may become a key phrase for determining at which timing the system starts to generate the

result. In order to clarify such a unit, we have analyzed the feature of each alignment pair. Figure 6 shows the distribution of the number of morphemes in lecturer's utterances by the solid line. The number of the average morphemes is 12.89, and the range of 20 morphemes occupies 83.1% of the whole.

Among these, the utterance consisting of a few morphemes could be regarded as an interpreting unit. Because the system can interpret it immediately when such utterance is detected.

We have extracted the alignment pairs in which lecturer's utterances consist of four or less morphemes. There exist 231 pairs equivalent to 16.0% of the whole. The main features are listed below: The frequency is shown in a parenthesis.

- Lecturer's utterances consisting of a conjunction.(84 pairs)

  lecturer: *soshite*

  interpreter: Then

- Lecturer's utterances consisting of a subject.(42 pairs)

  lecturer: *yuza-ga*

  interpreter: the user,

- Lecturer's utterances to be interpreted as a prepositional phrase.(65 pairs)

  lecturer: *kobunkaiseki-dewa*

  interpreter: in syntactic parsing,

Next, we have examined the difference between the beginning time of the interpreter's utterance and that of the lecturer's utterance. If lecturer's utterance can be interpreted by small difference, it is expected that an appropriate reason exists and therefore it may be able to be used as a technique for starting to make an interpretation at an early timing. The solid line of Figure 7 shows a distribution of the difference between the beginning time of the interpreter's utterance and that of the lecturer's utterance. The average of the difference of the beginning time is 4.1 seconds.

We have extracted the factors that the interpreter can follow the lecturer by picking up 30 alignment pairs of smaller difference of the beginning time. The main features observed in these alignment pairs are as follows:

- Predicting of the next phrase by referring the information on the slide of the lecture(6 pairs)

  lecturer: *haikei-desu-ga*

  interpreter: This slide shows the background of us.

  The lecturer gives a lecture using presentation slides. By seeing this slide, the interpreter can predict and follow the contents which the lecturer speaks.

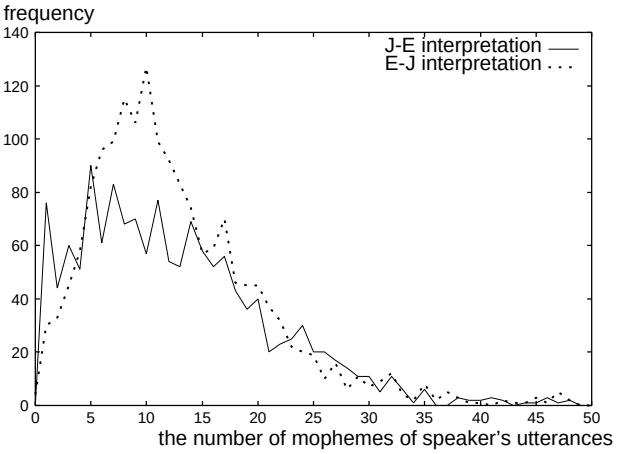

Figure 6: Distribution of the length of the lecturer's utterance

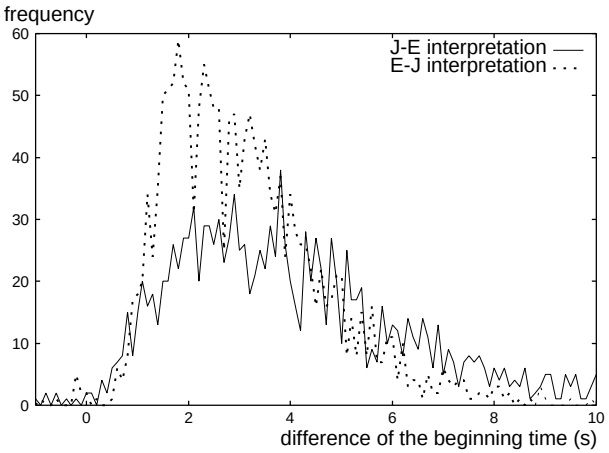

Figure 7: Distribution of difference between the beginning time of the interpreter's utterance and that of the lecturer's utterance

- Conjunction(11 pairs)

  lecturer: *shitagatte*

  interpreter: Therefore

  When a lecturer generates a conjunction, it can be interpreted without waiting for the next utterance.

- Interpolating a subject(4 pairs)

  lecturer: *bunrui-shi-ta-tokoro*

  interpreter: we analyzed such patterns

  A Japanese lecturer may omit a subject. At this time, an interpreter chooses the general subject "we" or "I" from the context, and generates it at an early stage.

- Interpreting a subject(3 pairs)

  lecturer: *yuza-ga*

  interpreter: the user,

  A subject appears at the beginning of a sentence in Japanese and in English. So the subject can be interpreted immediately.

- Insertion of a filler(1 pair)

  lecturer: *ijo-de owari-masu*

  interpreter: Ah this completes my lecture.

  While waiting for the next lecturer's utterance, the interpreter speaks a filler.

4.3.2. English-Japanese Simultaneous Interpretation

16 sets of the aligned English-Japanese simultaneous interpreting data are used for the investigation. Figure 6 shows the distribution of the number of morphemes in lecturer's utterances by the dotted line.

We have extracted the alignment pairs in which a lecturer's utterance consists of four or less morphemes. There exist 170 pairs equivalent to 9.90% of the whole. As a result, the same features of a Japanese-English interpretation were observed as follows:

- Lecturer's utterances consisting of a conjunction.(36pairs)

  lecturer: For example,

  interpreter: *tatoeba*

- Lecturer's utterances consisting of a subject.(32pairs)

  lecturer: some of the machines

  interpreter: *sono uchi-no kikai-no ikutsuka-ga*

- Lecturer's utterances of short clauses.(48pairs)

  lecturer: Someone once said that

  interpreter: *aru hito-ga ii-mashi-ta*

Next, we have examined the difference between the beginning time of the interpreter's utterance and that of lecturer's utterances, The dotted line of Figure 7 shows a distribution of the difference between the beginning time of the interpreter's utterance and of the lecturer's utterance. The average of difference of the beginning time is 3.2 seconds.

The following factors can be observed from the investigation of 30 alignment pairs with small difference of the beginning time:

- Interpreting a subject(7pairs)

  lecturer: I arrived in Japan

  interpreter: *e watashi nihon-ni tochaku-shi-mashi-ta-no-ga*

  A subject appears at the beginning of a sentence in Japanese and in English. So the subject can be interpreted immediately.

- Insertion of a filler(4pairs)

  lecturer: these are supporting countries.

  interpreter: *aa koyu-yona kuni-ga shien-wo shi-te-iru wake-de ari-masu*

  While waiting for the next lecturer's utterance, the interpreter speaks a filler.

- Conjunction(4pairs)

  lecturer: but after walking up and down

  interpreter: *e shikashi ee michi-wo arui-te iru uchi-ni*

  When a lecturer generates a conjunction, it can be interpreted without waiting for the next utterance.

- An adverbial phrase and a prepositional phrase(2pairs)

  lecturer: Here in Kansai

  interpreter: *mm kansai-dewa*

  The word order of Japanese is flexible in comparison with that of English. If the adverbial phrase and prepositional phrase appear at the beginning of the sentence, the phrase can be interpreted immediately.

4.3.3. Comparison

The delay of the English-Japanese interpretations is smaller than that of the Japanese-English interpretation about 1.0 seconds. The following reasons can be considered as the factors: (1)The high simultaneity is attained by interpreting according to the order of an appearance of an English word because the word order of Japanese is flexible in comparison with that of English. (2)It is possible to predict the next utterance by an interpreter's background knowledge because the topic of an English lecture was common. (3)Since a verb appears at the end of sentences in Japanese, the sentence structure of the interpretation in a Japanese-English at an early stage.

4.3.4. Analysis of Speaking Speed of Interpreter

We have investigated the speaking speed of the interpreter. The speed was calculated with the number of moras per utterance unit. We assigned one mora to a short vowel and a consonant behind vowel, and two moras to long vowel in English. And we calculated the number of moras by using the result of ChaSen (Matsumoto, 2000) in Japanese. Figure 8 shows the result.

Although the interpretation begins at an early stage in order that an interpreter may raise simultaneity, however, interpreter's speaking speed is slow because the role of the

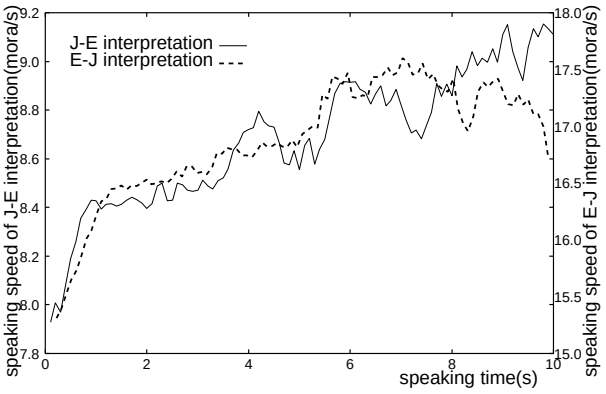

Figure 8: Interpreter's speaking speed

interpretation is not fixed at the beginning of the utterance. As the lecturer's utterance is progressing, the interpreter can grasp a lecturer's intentions gradually. Therefore, it is considered that the interpreter gathers the speed gradually.

5. Conclusion

The design and construction of simultaneous interpreting corpus at CIAIR has been described. The corpus consists of the data of monologue speech and the interpreting speech through simultaneous interpreters, and could be utilized as the language data for interpretation research. While collection of data will be also advanced succeedingly, the corpus stated in this paper, after the further maintenance, will become publicly available as early a stage as possible.

This paper has also described an investigation of simultaneous interpreting corpus. The results of the investigation are as follows:

1. When a lecturer generates a conjunction, it can be interpreted immediately without waiting for the next lecturer's utterance.

2. Since a subject appears at the beginning of a sentence in both Japanese and English, the subject can be interpreted immediately.

3. While waiting for the next lecturer's utterance, the interpreter speaks a filler.

4. By controlling the speaking speed based on the quantity of the input utterance, the interpreter reduces the difference between the beginning time of the interpreter's utterance and that of the lecturer's utterance.

These results will be available for the development of a simultaneous interpreting system.

Acknowledgement: The collection and transcription of the speech data have been carried out cooperatively with Inter Group Corporation. The authors with to thank specially Mr. Masafumi Yokoo for his contribution. This work is partially supported by the Grand-in-Aid for COE Research of the Ministry of Education, Science, Sports and Culture, Japan.

6. References

Y. Aizawa, S. Matsubara, N. Kawaguchi, K. Toyama, Y. Inagaki, 2000. Spoken Language Corpus for Machine Interpretation Research, *Proceedings of 6th International Conference on Spoken Language Processing*, III:389–401.

J. Amtrup, 1999. Incremental Speech Translation, *Lecture Notes in Artificial Intelligence*, 1735.

N. Kawaguchi, S. Matsubara, K. Takeda, F. Itakura, 2001. Construction of Speech Corpus in Moving Car Environment, *Proceedings of 7th European Conference on Speech Communication and Technology*, 2:2027-2030.

N. Kawaguchi, S. Matsubara, K. Takeda, F. Itakura, 2002. Multi-Dimensional Data Acquisition for Integrated Acoustic Information Research, *Proceedings of 3rd International Conference on Language Resources and Development*.

K. Maekawa, T. Kagomiya, H. Koiso, H. Ogura, H. Kikuchi, 2000. Design of the Corpus of Spontaneous Japanese, *Journal of the Phonetic Society of Japan*, 4(2):51–61. (in Japanese)

S. Matsubara, K. Toyama, Y. Inagaki, 1999. Sync/Trans: Simultaneous Machine Interpretation between English and Japanese, N. Foo (Ed), *Advanced Topics in Artificial Intelligence, Lecture Note in Artificial Intelligence*, 1747:134-143.

Y. Matsumoto, et al., 2000. Morphological Analysis System ChaSen version 2.2.1 Manual, http://chasen.aist-nara.ac.jp/ .

H. Mima, H.Iida, O. Furuse, 1998. Simultaneous Interpretation Utilizing Example-based Incremental Transfer, *Proceedings of 15th International Conference on Computational Linguistics*, 855–861.

A. Mizuno, 1995. On Simultaneous Interpretation from Japanese into English, *Journal of the Interpreting Research Association of Japan*, 5(2):4–21. (in Japanese)

T. Takezawa, T. Morimoto, Y. Sagisaka, N. Cambell, H. Iida, F. Sugaya, A. Yokoo, S. Yamamoto, T., 1998. Japanese-to-English Speech Translation System: ATR-MATRIX, *Proceedings of 5th International Conference on Spoken Language Processing*, III:957–960.

T. Watanabe, A. Okumura, S. Sakai, K. Yamabana, S. Doi, K. Takahashi, (2000). An Automatic Interpretation Software for Travel Conversation, *Proceedings of Workshop on Multi-Lingual Speech Communication*, 21–24.