Bilingual Speech Dialogue Corpus for Simultaneous Machine Interpretation Research

Koichiro Ryu[†], **Shigeki Matsubara**[††], **Nobuo Kawaguchi**[††]
and **Yasuyoshi Inagaki**[†††]
[†]Graduate School of Engineering, Nagoya University
[††]Information Technology Center/CIAIR, Nagoya University
[†††]Faculty of Information Science and Technology, Aichi Prefectural University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
E-mail:k_ryu@inagaki.nuie.nagoya-u.ac.jp

Abstract

Speech-to-speech translation has been an important research topic with the advance of technologies for speech processing and language processing. This paper describes a bilingual speech dialogue corpus which has been constructed for research on simultaneous machine interpretation at the Center for Integrated Acoustic Information Research (CIAIR), Nagoya University. The corpus has been implemented by collecting simulated cross-lingual conversations between English speech and Japanese speech through simultaneous interpretation, and by transcribing them manually with bilingual sentence alignment. In the year 2002, 216 spoken dialogues have been collected under a real environment, and transcribed into text files consisting of about 300,000 morphemes. In order to utilize the bilingual corpus effectively, every source utterance speech has been segmented into interpreting units according to its word-for-word translation and the word alignment of them. The interpreting unit means a linguistic chunk that could be interpreted separately and simultaneously. This paper has investigated linguistic characters of such the unit, and examined the feasibility of simultaneous machine interpretation.

1 Introduction

Speech-to-speech translation 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 et al., 1998) (Watanabe et al., 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, it is desired to develop a simultaneous interpretation system. Toward simultaneous machine interpretation, 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.

This paper describes a bilingual speech dialogue corpus, which has been collected as part of the project on the massive-scale speech database construction in the Center for Integrated Acoustic Information Research (CIAIR) (Kawaguchi et al., 2002). The bilingual spoken dialogues between an English native speaker and a Japanese native speaker through professional simultaneous interpreters have been collected and transcribed into ASCII text files. The exact beginning time and end time are provided for each utterance, and moreover, each source utterance is aligned with the corresponding target utterance. In this year, 216 spoken dialogues have been collected in total.

In order to utilize the corpus for the research on simultaneous machine interpretation, every speaker's utterance is segmented into "interpreting unit" The segmentation is executed based on the word-for-word translations for spoken source sentences and the word alignments of them. The interpreting unit means the smallest speech unit which can be translated separately and simultaneously. Therefore, it is very important for a simultaneous interpretation system to

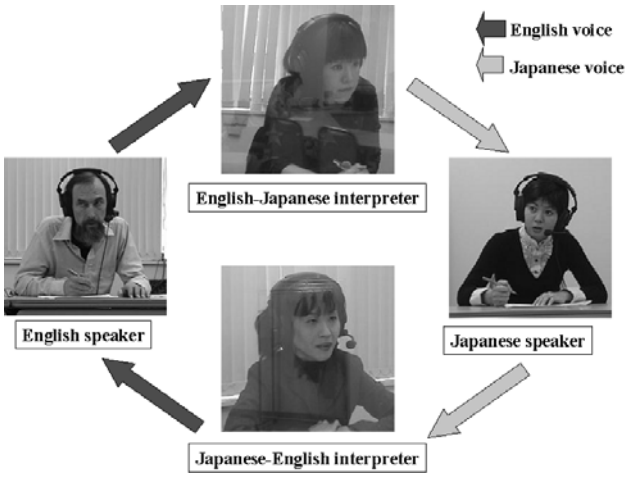

Figure 1: Recording scene (Upper left: English speaker, Upper right: Japanese speaker, Lower left: English-Japanese interpreter, and Lower right: Japanese-English interpreter)

recognize such a unit. In this paper, we make an investigation into a possibility of simultaneous machine interpretation by extracting such the units from our bilingual corpus.

This paper is organized as follows: Section 2 explains the design, construction and statistics of our corpus. Section 3 describes the segmentation of source utterances into interpreting units. Section 4 provides the observations of the interpreting unit.

2 Simultaneous Interpreting Corpus

Several bilingual spoken dialogue corpora have been constructed, and played important roles in the advancement of the speech translation technology (Ehara et al., 1990; Morimoto and et al., 1994). Most of them, however, are limited to the collection of consecutive interpreting data, and therefore, simultaneous interpreting corpora will be getting valuable in the coming machine interpretation research.

The various types of large-scale multilingual speech database have been constructed at CIAIR from 1999 (Aizawa et al., 2000; Matsubara et al., 2002). The bilingual speech dialogues between an English speaker and a Japanese speaker through simultaneous interpreters were collected in the year 2002. This section explains the recording, transcription, and size of our spoken dialogue corpus.

2.1 Data Collection

We have set up "airport" and "hotel" as the typical situations of dialogue communications in an overseas travel. To put it concretely, the following topics were selected: "airport check-in", "hotel check-in/check-out", "booking of a room at a hotel", and "booking of a seat in an airplane", and so on. Since it is difficult for the subjects to talk naturally and smoothly with no motivation, we prepared the dialogue specification sheets and showed the subjects them in advance.

Because 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 native speakers and the interpreters were digitized by 16kHz of sampling frequencies and 16 bits, and recorded in stereo onto digital audio tapes (DAT). We have recorded 216 dialogues in total. Generally simultaneous interpreters use not only speaker's utterances but also speaker's facial expressions and conversational behaviors. So simultaneous interpreters enter a booth, where he or she can look speakers over glass window, when our recording was carried out(Figure 1). In order to enhance the quality of the interpretations for the English speaker and the Japanese speaker, each speaker was accompanied by one interpreter. To ensure all the participants the speakers can only listen to the output from the other speaker's interpreter, and the interpreter can only listen to

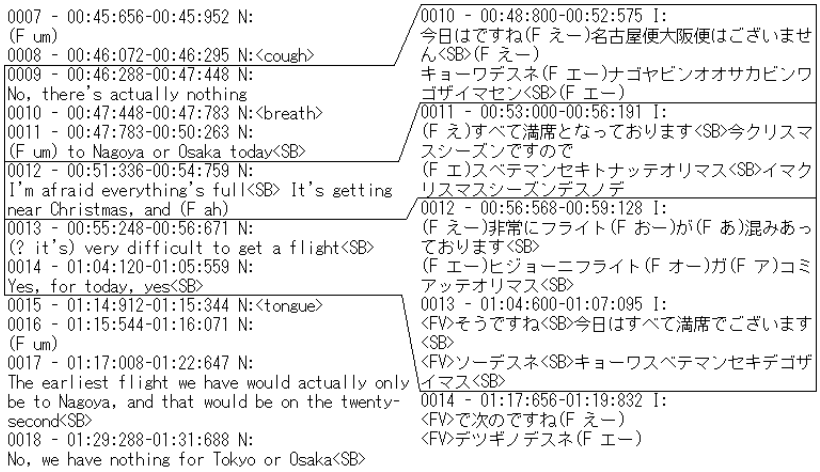

Figure 2: Sample of the transcript

Table 1: Basic statistics of the simultaneous interpreting corpus

item	English	E-J	Japanese	J-E
recording time(minute)	1082	1082	1082	1082
speaking time(sec)	17865	17495	18534	17128
morphemes(words)	73513	77790	75064	64306
kinds of morpheme(word)	2713	2855	2601	2615
utterance units	10858	11354	13176	10813
sentence breaks	7889	9595	8725	6991
fillers	2376	4396	4892	2712

the speech that they are assigned to interpret. Please note that this is a simulative dialogue, in which the contents of the utterances can be limited. Therefore, such background information as the speakers' roles and the dialogue tasks were informed to the speakers in advance, in attempt to collect as spontaneous an utterance as possible.

2.2 Speech Transcription

The collected speech data was transcribed into the text manually. The text transcription was done according to the manual for Corpus of Spoken Japanese(CSJ) produced by National Japanese Language Research Institute (Maekawa et al., 2000). Figure 2 shows the sample of the transcript. The dialogue tags were provided for the language phenomena characteristic of spoken language, such as fillers(F), corrections(X), hesitations(D), mistatements(W), repetitions(S), monologues(L) and sentence breaks ⟨SB⟩. Moreover, the transcript is broken up into utterance units by the pauses for 200ms or more, or those for 50ms after the sentence break, and the beginning time and end time have been provided for every utterance unit.

Moreover, the alignment between Japanese speaker's utterances and Japanese-English interpreter's utterances, and that between English speaker's utterances and English-Japanese interpreter's utterance have been provided. Figure 2 shows the sample of the alignment.

2.3 Statistics of the Corpus

The basic statistics of our corpus, such as recording time, speaking time, number of morphemes (words), different morphemes, utterance units, sentences, and fillers, are shown in Table 1. E-J and J-E in the table indicate English-Japanese interpreter's and Japanese-English interpreter's, 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 and et al., 2000). The number of kinds of morpheme in English is the

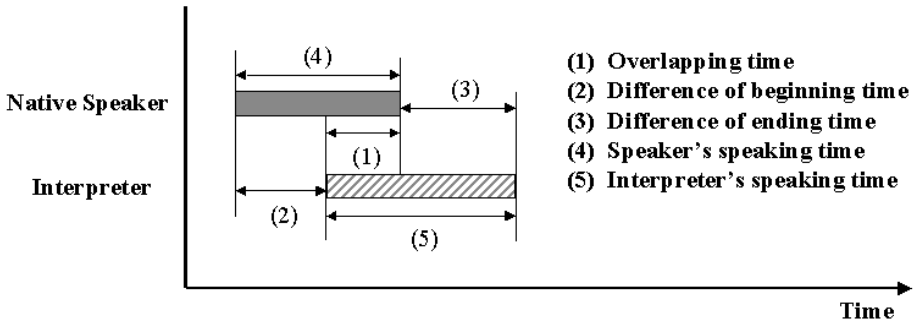

Figure 3: Pattern diagram of simultaneous interpretation

Table 2: Statistics of the aligned corpus

item	E-J	J-E
overlapping time(sec) (1)	0.63	0.65
difference of beginning time(sec) (2)	1.68	1.88
difference of end time(sec) (3)	1.81	1.78
speaker's speaking time(sec) (4)	2.19	2.21
interpreter's speaking time(sec) (5)	2.20	2.15
speaker's morphemes(words)	9.47	9.33
interpreter's morphemes(words)	10.29	8.49

number of words whose notations differ, and that in Japanese is the number of morphemes whose basic forms differ. The recording time is about 18 hours. The total of the speaking time is about 20 hours data. The reason for that the total of the speaking time is longer than that of the recording time is that the utterance of the interpreter overlaps with the speaker's utterance.

As the result of the parallel corpus alignment, the average time of several indices shown in Figure 3 has been measured.

(1) The time overlapping with the speaker's utterance in the interpreter's utterance.
(2) The difference between the beginning time of the speaker's utterance and that of the interpreter's one.
(3) The difference between the end time of the speaker's utterance and that of the interpreter's one.
(4) The speaking time by speakers.
(5) The speaking time by interpreters.

In addition, the average number of the morphemes of speaker's utterances and interpreter's utterances has been also counted. Table 2 shows the statistics of the aligned corpus.

3 Speech Segmentation into Interpreting Units

The biggest difference between consecutive interpretation and simultaneous interpretation would be the beginning time of the interpretation. The simultaneous interpreter begins to speak before the source utterance finishes, and that enables the simultaneous interpreter to reduce listener's waiting time and make a contribution to natural bilingual conversations. In general, simultaneous interpreters break up the utterance by speakers into several meaningful segments, and translate them incrementally. We call such the segment (simultaneous) interpreting unit. If a simultaneous interpretation system can recognize such the unit, it can begin to make a translation without waiting for the completion of the speaker's utterance. Moreover, there is an advantage that the output staring timing can be decided inevitably. As one of the techniques for acquiring the interpreting units, a group of the meaning is formed by referring to the correspondence of the words of the original and the interpreted sentence. That is, the group becomes an interpreting

unit. However, it is difficult to use simultaneous interpreter's transcript data to acquire the interpreting unit. Since simultaneous interpreters use advanced peculiar techniques such as the free translation, the omissions, and restating to overcome a time restriction, it does not reflect the sentence structure of the speaker's utterance in precision. Therefore, in our research, the word-for-word translations are given to the transcript of the speaker's utterance manually without putting time pressure. These are the interpretations based on the structure of the speaker's utterance. That is, they are interpretations with little free translation and omission. And, the group of the meaning is found to the speaker's utterance by refering the word correspondence between the word-for-word translation and the speaker's utterance. Interpreting unit can be acquired by breaking up the speaker's utterance into the unit.

3.1 Word-for-Word Translation

The word-for-word translations are given to the transcript of the speaker's utterance manually. To collect data with good quality that aimed at the construction of the simultaneous machine interpretation, the person who had simultaneous interpreter's expertise did this work. The method of making the word-for-word translation is described in detail as follows.

Rules of word-for-word translation:

1. Each utterance is given the word-for-word translation. (Exceptionally, it is likely to extend two or more utterances only when not interpretable by one utterance alone.)
2. The interpreter's utterances in the simultaneous interpreter corpus are not referred to.
3. The listener can understand the semantic content of the original utterance exactly.
4. The system can translate the source utterance incrementally.
5. To avoid the dependence on the context, the free translation and the omission are avoided.

Here, the example of the word-for-word translation is shown (Figure 4). The beginning utterance ID and the end utterance ID of the corresponding speaker's utterance are given to each translation. In Figure 2, the interpreter doesn't translate "I'm afraid" that exists in the original. (It is guessed that there was no translated time.) Moreover, the interpreter has translated the part of "very difficult" in "very difficult to get a flight" freely with "komi atte iru (It is crowded)." In word-for-word translation "I'm afraid" in the speaker's sentence is interpreted like "moushiwake gozaimasenn (I'm sorry)." And the part of "very difficult" reflects speaker's sentence like "sore-ha muzukashi-i (It's very difficult)." The omission and the free translation interferes to the translation making when using it as study data of the simultaneous machine interpretation of the example base. Therefore, making the word-for-word translation without omissions and free translations is profitable.

3.2 Acquirement of Interpreting Units

This section describes the technique for acquiring the interpreting unit by using the word-for-word translation made with section 3.1. We explain the procedure of the unit division by using Figure 5.

Step 1 We break up English utterance/interpretation into words (Japanese utterance/interpretation into morphemes).

Step 2 The word correspondence is taken between speaker's utterances and interpretation. In Figure 5 an arrow show the correspondence.

Step 3 Then, sometimes, the phenomenon of arrows intersecting occurs because of the difference of the grammar structure between English and Japanese. In Figure 5 the intersection of the arrow exists. It is caused by the difference of grammar that a verb is located just behind the subject in English and it is located in the end of sentence in Japanese. The part where the arrow intersects settles the two as one interpreting unit. Therefore, it is possible to translate natural interpretation without disregarding the grammar.

```
0007 - 00:45:656-00:45:952 N:
(F um)
0008 - 00:46:072-00:46:295 N:<cough>
0009 - 00:46:288-00:47:448 N:
No, there's actually nothing
0010 - 00:47:448-00:47:783 N:<breath>
0011 - 00:47:783-00:50:263 N:
(F um) to Nagoya or Osaka today<SB>
0012 - 00:51:336-00:54:759 N:
I'm afraid everything's full<SB> It's getting
 near Christmas, and (F ah)
0013 - 00:55:248-00:56:671 N:
it's very difficult to get a flight<SB>
0014 - 01:04:120-01:05:559 N:
Yes, for today, yes<SB>
0015 - 01:14:912-01:15:344 N:<tongue>
0016 - 01:15:544-01:16:071 N:
(F um)
0017 - 01:17:008-01:22:647 N:
The earliest flight we have would actually on
ly be to Nagoya, and that would be on the twe
nty-second<SB>
0018 - 01:29:288-01:31:688 N:
No, we have nothing for Tokyo or Osaka<SB>
```

(a) English native speaker

```
0009 - 0009-0009 T:
いいえ、ございません。
0010 - 0010-0010 T:

0011 - 0011-0011 T:
本日の名古屋か大阪行きですね。
0012 - 0012-0012 T:
申し訳ございませんが、全て満席となっておりま
す。クリスマスも近いですので、
0013 - 0013-0013 T:
予約は大変難しくなっております。
0014 - 0014-0014 T:
はい、本日ですね、はい。
0015 - 0015-0015 T:

0016 - 0016-0016 T:

0017 - 0017-0017 T:
お取りできる１番早い便は、名古屋行きのみで、
２２日の発となります。
0018 - 0018-0018 T:
いいえ、東京行き、大阪行きもございません。
0019 - 0019-0019 T:
一席もございません。
```

(b) English-Japanese interpretation

Figure 4: Sample of the transcript (word-for-word translation)

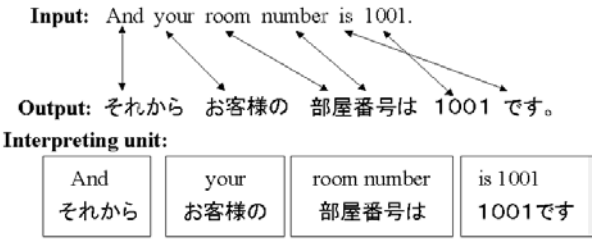

Figure 5: Example of segmentation of interpreting unit

4 Investigation of Interpreting Unit

4.1 Statistics of Interpreting Unit

As the basic statistics of word-for-word translation data, the number of morphemes (words), utterance units, sentences, and turns. are shown in Table 3. The turn is a group of one or more utterances until a speaker alternates. And Table 4 shows the number of interpreting units per a utterance, number of interpreting units per a sentence, number of interpreting units per a turn, and number of morphemes (words) per a unit. According to Table 4, it is possible to segment a turn into the unit of 3.67 in English and 3.73 in Japanese on average. It means that the chance to output the interpretation increases compared with before speaker's utterances are segmented.

4.2 Utilization of Interpreting Unit

It is explained that the process of acquiring the interpreting unit from speaker's utterances in section 3.2. Acquiring this interpreting units may give the possibility that the simultaneous machine interpretation can imitate simultaneous interpreter's behavior. Figure 6 shows the distribution of the number of morphemes in the first unit of a turn. The units composed less than four words accounts for 90% or more in English and less than six morphemes accounts for 90% or more in Japanese. That is, if the interpretation is output incrementally in each interpreting unit, the system begins the interpretation before the speaker finishes speaking. Therefore it can reduce listener's waiting time, and a conversation between different languages be achieved

Table 3: Basic statistics of the word-for-word translation

item	English	E-J	Japanese	J-E
morphemes(words)	69424	68675	70225	62605
utterance units (translations)	10858	10387	13176	12516
sentence breaks	7889	8108	8725	8231
turns	4451	4451	4422	4422

Table 4: Basic statistics of interpreting unit

item	E-J	J-E
interpreting units	16338	16516
average units per a utterance	1.50	1.25
average units per a sentence	2.07	1.89
average units per a turn	3.67	3.73
average morphemes per a unit	4.25	4.25

smoothly. Figure 7 and 8 show the position where the unit boundary in the sentence appeared at the rate. This indicates the possibility that the translations can be output incrementally along a speaker's utterance if an interpreting unit is used. Figure 9 shows the distribution of the number of morphemes per a unit. About 80% of all units is composed less than five words in English and less than 7 morphemes in Japanese. Thus, being possible to segment into a short unit is that there is a chance to output interpretation. If these chances are used, listener's waiting time is reduced and the interpreting system can output interpretation incrementally.

5 Concluding Remarks

This paper has described the design of a bilingual speech dialogue corpus, which has been constructed at the CIAIR, and its investigation. Our corpus contains spoken dialogues between an English native speaker and a Japanese native speaker through professional simultaneous interpreters, and each source utterance has been aligned with the corresponding target utterance. In order to utilize the corpus for the research on simultaneous machine interpretation, every speaker's utterance is segmented into interpreting units. The segmentation is executed according to word-for-word translation for spoken source sentences and alignment between the corresponding words. The results of the investigation of interpreting units are as follows:

- It is possible to segment into the unit of 3.67 in English and 3.73 in Japanese averages a turn.
- The first unit of turn composed less than four words accounts for 90% or more in English and less than six morphemes accounts for 90% or more in Japanese.
- About 80% of all units is composed less than five words in English and less than 7 morphemes in Japanese.

The results would provide the possibility that a simultaneous interpreting system can interpret source languages incrementally and play a role for supporting natural and smooth cross-lingual communication.

Acknowledgements

Inter Group Corporation has cooperated in the recording of the data. Many thanks to Mr. Masafumi Yokoo for his contribution. The authors would like to express their gratitude to Ms. Yayoi Banno and Ms. Nozomi Yamaguchi for their contribution in constructing the interpreting unit data. This research is partially aided by the Grant-in-Aid for COE Research for Young Scientists of the Ministry of Education, Science, Sports, and Culture, Japan and the Nakajima Foundation.

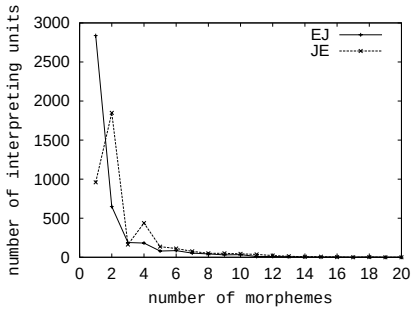

Figure 6: Distribution of number of morpheme in first unit

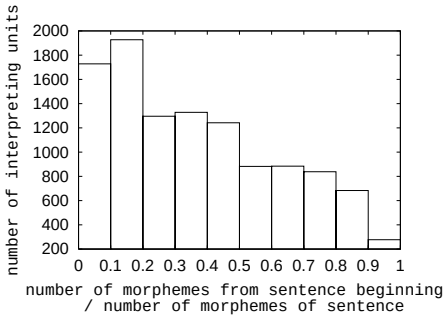

Figure 7: Distribution of interpreting units (English-Japanese)

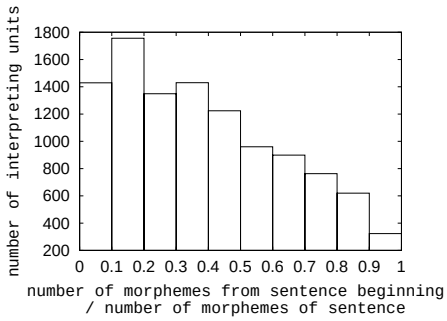

Figure 8: Distribution of interpreting units (Japanese-English)

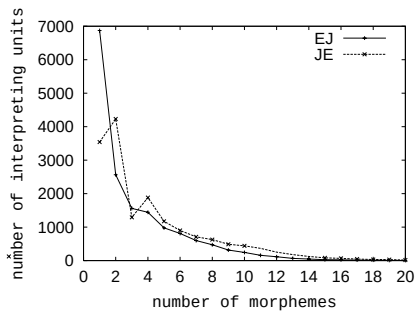

Figure 9: Distribution of number of morphemes per an interpreting unit

References

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

Ehara, T., K. Ogura, and T. Morimoto. 1990. ATR Dialogue Database. *Proceedings of 1st International Conference of on Spoken Language Processing*, pages 1093–1096.

Kawaguchi, N., S. Matsubara, K. Takeda, and F. Itakura. 2002. Multi-Dimensional Data Acquisition for Integrated Acoustic Information Research. *Proceedings of International Language Resources and Evaluation Conference*, pages 2043–2046.

Maekawa, K., H. Koiso, S. Furui, and H.Isahara. 2000. Spontaneous Speech Corpus of Japanese. *Proceedings of 2nd International Conference of Language Resources and Evaluation*, pages 945–952.

Matsubara, S., A. Takagi, N. Kawaguchi, and Y. Inagaki. 2002. Bilingual Spoken Monologue Corpus for Simultaneous Machine Interpretation Research. *Proceedings of International Language Resources and Evaluation Conference*, I:153–159.

Matsumoto, Y. and et al., 2000. *Morphological Analysis System ChaSen version 2.2.1 Manual*.

Morimoto, T. and et al. 1994. Speech and Language Database for Speech Translation Research. *Proceedings of 3rd International Conference of on Spoken Language Processing*, pages 1791–1794.

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

Watanabe, T., A. Okumura, S. Sakai, K. Yamabana, S. Doi, and K. Hanazawa. 2000. An Automatic Interpretation System for Travel Conversation. *Proceedings of 6th International Conference on Spoken Language Processing*, pages IV:444–447.