Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya,
Hirofumi Yamamoto, and Seiichi Yamamoto

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
{toshiyuki.takezawa, eiichiro.sumita, fumiaki.sugaya, hirofumi.yamamoto, seiichi.yamamoto}@atr.co.jp

Abstract

At ATR Spoken Language Translation Research Laboratories, we are building a broad-coverage bilingual corpus to study corpus-based speech translation technologies for the real world. There are three important points to consider in designing and constructing a corpus for future speech translation research. The first is to have a variety of speech samples, with a wide range of pronunciations and speakers. The second is to have data for a variety of situations. The third is to have a variety of expressions. This paper reports our trials and discusses the methodology. First, we introduce a bilingual travel conversation (TC) corpus of spoken languages and a broad-coverage bilingual basic expression (BE) corpus. TC and BE are designed to be complementary. TC is a collection of transcriptions of bilingual spoken dialogues, while BE is a collection of Japanese sentences and their English translations. Whereas TC covers a small domain, BE covers a wide variety of domains. We compare the characteristics of vocabulary and expressions between these two corpora and suggest that we need a much greater variety of expressions. One promising approach might be to collect paraphrases representing various different expressions generated by many people for similar concepts.

1. Introduction

The current speech translation research assumes a limited task such as hotel room reservations and carries out experiments with medium-sized vocabularies of about several tens of thousand words (Takezawa et al., 1998b). In order to enlarge the application area of speech translation systems, we must develop a large vocabulary speech translation system that can accept more expressions in not one but multiple situations. With this in mind, at ATR Spoken Language Translation Research Laboratories, we are building a broad-coverage bilingual corpus to study corpus-based speech translation technologies for the real world. This paper reports our trials and discusses how to build a broad-coverage bilingual corpus based on numerical data from various aspects.

Various kinds of corpora for analyzing linguistic phenomena and for gathering statistical information are now accessible via electronic media, and these can be used to study natural language processing. However, since these corpora generally contain written-language and monolingual data, they are not necessarily useful for speech translation. With this in mind, we built a bilingual travel conversation (TC) corpus of spoken languages (Morimoto et al., 1994; Takezawa et al., 1998a; Takezawa, 1999).

The TC corpus covers a limited task, i.e., hotel room reservations. All of the contents are transcriptions of spoken dialogues between Japanese and English speakers through human interpreters. The size of this corpus is more than 20,000 sentences.

In order to cover many situations for Japanese going abroad, we recently built a broad-coverage bilingual basic expression corpus (BE), which is a collection of Japanese sentences and their English translations usually found in phrase-books for foreign tourists. The size of this corpus is more than 200,000 sentences, roughly ten times larger than the TC corpus. The contents, however, are not spoken dialogues.

Despite one workshop on comparing corpora (Kilgarriff and Sardinha, 2000), there have not been enough studies on comparing corpora to quantify similarities and differences between them. Therefore, this paper reports characteristics of vocabulary and expressions between the above-mentioned two kinds of corpora and investigates the requirements for corpus construction methodologies. Our belief is that a practical and quantitative analysis is needed to develop useful guidelines for giving a speech translation system the ability to deal with a broad range of application areas.

Section **2** describes the corpora and their basic characteristics such as sentence length. Section **3** reports vocabulary characteristics as well as numerical data such as that for cross entropy (perplexity) from the viewpoint of information theory. Section **4** presents the characteristics of expressions and a related preliminary experiment. Section **5** discusses future research directions. Finally, section **6** offers our conclusions.

2. Corpora

2.1. A broad-coverage bilingual basic expression corpus

At ATR Spoken Language Translation Research Laboratories, we have built a broad-coverage bilingual basic expression (BE) corpus, which is a collection of Japanese sentences and their English translations usually found in phrase-books for foreign tourists. Many phrase-books are published for tourists going abroad, and they cover a number of situations. From such phrase-books, we collected Japanese/English parallel conversational sentences. Then additional important information was tagged to these sentences. For example, there are situation tags indicating airport, hotel, restaurant, shopping, and problems.

Table 1 shows the basic characteristics of the corpus. The Japanese words in the corpus were obtained automat-

	Japanese	English
Number of utterances	200,241	200,241
Number of sentences	220,244	223,482
Total number of words	1,689,442	1,230,650
Number of word entries	21,329	17,076
Average number of words per sentence	7.67	5.51

Table 1: Basic characteristics of broad-coverage bilingual basic expression corpus

Number of collected dialogues	618
Speaker participants	71
Interpreter participants	23

Table 2: Overview of bilingual travel conversation corpus of spoken languages

ically by a Japanese morphological analyzer at ATR, and the English words were automatically tagged by an English morphological analyzer at ATR. Interjections such as "thank you" and "may I help you" were tagged as one word. This corpus covers a number of situations for Japanese going abroad.

2.2. A bilingual travel conversation corpus of spoken languages

At ATR Interpreting Telecommunications Research Laboratories, we have built a bilingual travel conversation (TC) corpus of spoken languages for speech translation research (Morimoto et al., 1994; Takezawa et al., 1998a; Takezawa, 1999). An interpreter is assigned to each translation direction (J to E or E to J) when a dialogue is collected in order to gather good quality data. The task of the corpus involves travel conversations between a tourist and a front desk clerk at a hotel. This task was selected because of its familiarity to people and its expected use in future speech translation systems. The interpreters speak English and Japanese in all of the dialogues and serve as a speech translation system. Human interpreters successively interpreted each utterance, enabling us to gather basic data for developing a speech translation system.

Table 2 shows an overview of the bilingual travel conversation corpus of spoken languages. **Table 3** shows the basic characteristics of the corpus.

	Japanese	English
Number of utterances	16,084	16,084
Number of sentences	21,769	22,928
Total number of words	236,066	181,263
Number of word entries	5,298	4,320
Average number of words per sentence	10.84	7.91

Table 3: Basic characteristics of TC corpus

	Japanese	English
Total number of words	94%	97%
Number of word entries	73%	75%

Table 4: Coverage of BE corpus relative to TC corpus

	Japanese	English
Total number of words	73%	85%
Number of word entries	17%	17%

Table 5: Coverage of TC corpus relative to BE corpus

2.3. Comparison of basic characteristics

The BE corpus is expected to cover many situations for Japanese going abroad. The contents, however, are not spoken dialogues. The size of the corpus is more than 200,000 sentences, which is about ten times larger than the TC corpus.

The TC corpus covers a limited task, i.e., hotel room reservations. All of the contents are transcriptions of spoken dialogues. The size of this corpus is more than 20,000 sentences. Although it is about one tenth the size of the BE corpus, the average number of words per sentence is longer than that of the BE corpus.

Figure 1 shows sentence length distribution of the BE corpus (upper) and the TC corpus (lower). The horizontal line indicates the number of phrases per sentence, and the vertical line indicates frequencies. The average number of phrases per sentence is 2.88 for the BE corpus and 4.37 for the TC corpus. According to the study by the National Language Research Institute (Maekawa, 2001), the average number of phrases per sentence for Japanese daily conversations is 3.81, and its sentence length distribution is quite similar to that of the TC corpus. This is because the TC corpus contains transcriptions of spoken dialogues but the BE corpus contains edited colloquial sentences.

The common characteristics between them are as follows. The number of English sentences is slightly more than that of Japanese sentences. However, the average number of English words per English sentence is less than that of Japanese sentences. These characteristics do not depend on the translation direction because the TC corpus contains both translation directions (J to E and E to J), although the BE corpus may contain translations only from Japanese to English.

3. Characteristics of Vocabulary

3.1. Coverage

Table 4 shows the coverage of the BE corpus relative to the TC corpus. A high coverage (Japanese side: 94%; English side: 97%) was obtained because the size of the BE corpus was much larger than that of the TC corpus. Most of the words not covered were nouns.

Table 5 shows the coverage of the TC corpus relative to the BE corpus. A decent coverage (Japanese side: 73%; English side: 85%) was obtained because the domain was the same, i.e., the travel domain. Again, most of the words not covered were nouns.

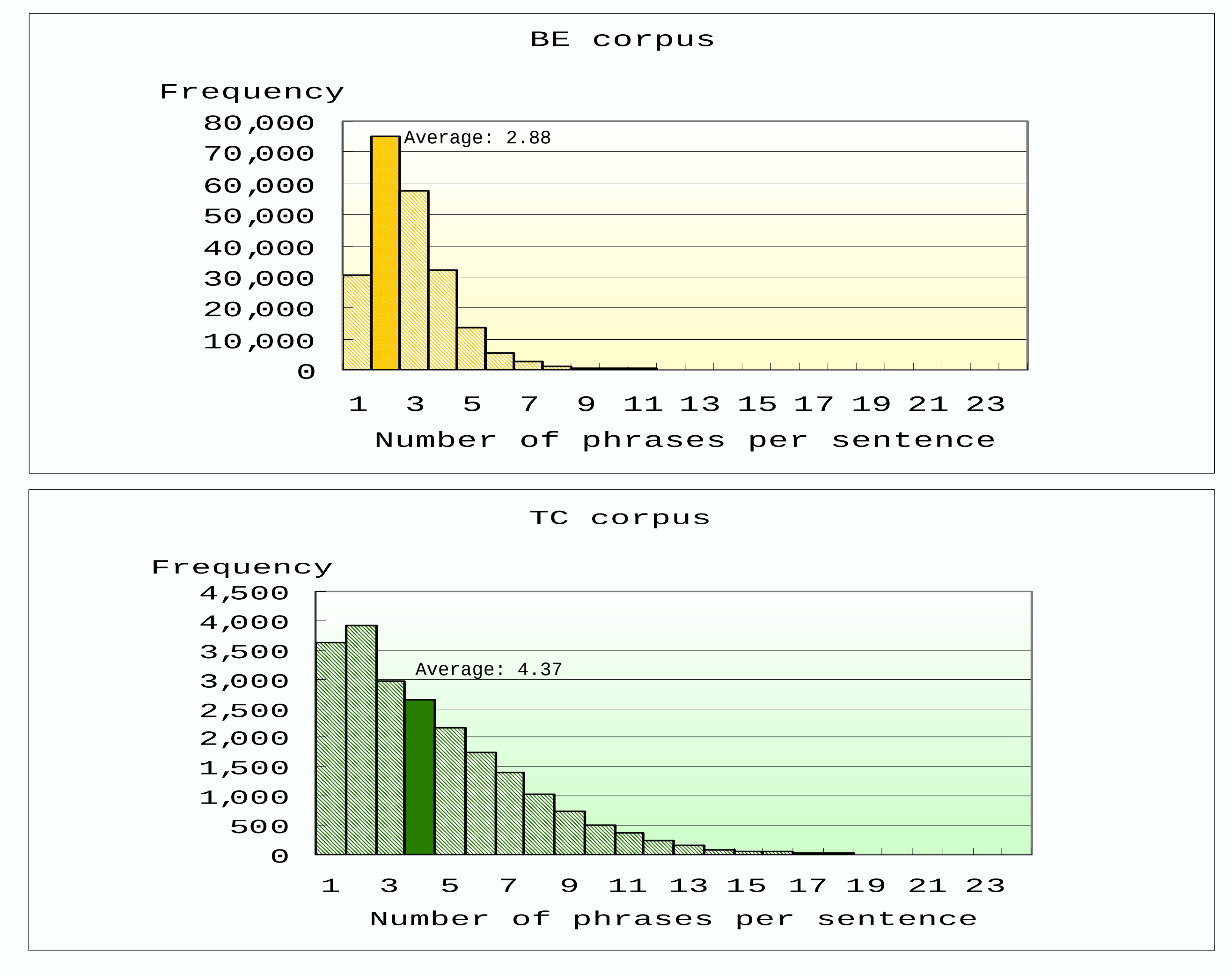

Figure 1: Sentence length distribution

Test set	Language model		
	BE	TC	BE+TC
BE	16.4	58.3	17.6
TC	72.3	16.3	17.1

Table 6: Cross perplexity

3.2. Cross perplexity

Table 6 shows the cross perplexity. Three kinds of language models were built. BE in Table 6 indicates a language model trained by the BE corpus. Approximately 5% (10,000 utterances) of the BE corpus was reserved for future usages such as test sets and development test sets. The remaining 95% was used to train a language model based on a multi-class composite N-gram model (Yamamoto et al., 2001). TC in Table 6 indicates a language model trained by the TC corpus. Not only the bilingual TC corpus but also Japanese monolingual spoken dialogue corpora (Nakamura et al., 1996; Takezawa et al., 1998a) were used, so the size of the training data for TC was comparable to that of BE. BE+TC in Table 6 indicates a mixed language model, i.e., trained by both corpora.

If the test set is selected from the same corpus as the language model, the value of the perplexity is approximately 16 in the both cases. However, the value of the cross perplexity is much larger than 16 in both cases (58.3 and 72.3, respectively) if the test set is from a different corpus than the language model. Therefore, the statistics may be different between them. By using the mixed language model, a similar value of about 17 could be obtained in both cases.

3.3. Word distribution

According to the experimental results of the cross perplexity, we found that the statistics of the BE corpus may be different from that of the TC corpus. From the viewpoint of vocabulary characteristics, we examined the differences of word unigrams between them. **Figure 2** shows an example of an English word distribution. The horizontal line indicates values of English word unigrams in the BE corpus. For example, the occurrence of the word “the” is 42,889 and its frequency is 3.49%. The vertical line indicates relative frequencies of word occurrences in the TC corpus. For example, the frequency of the word “the” in the TC corpus is 4.24%, so its relative frequency is $4.24/3.49 = 1.21$. According to Figure 2, we can find that words such as “front desk,” “Suzuki,” and “thank you very much” frequently oc-

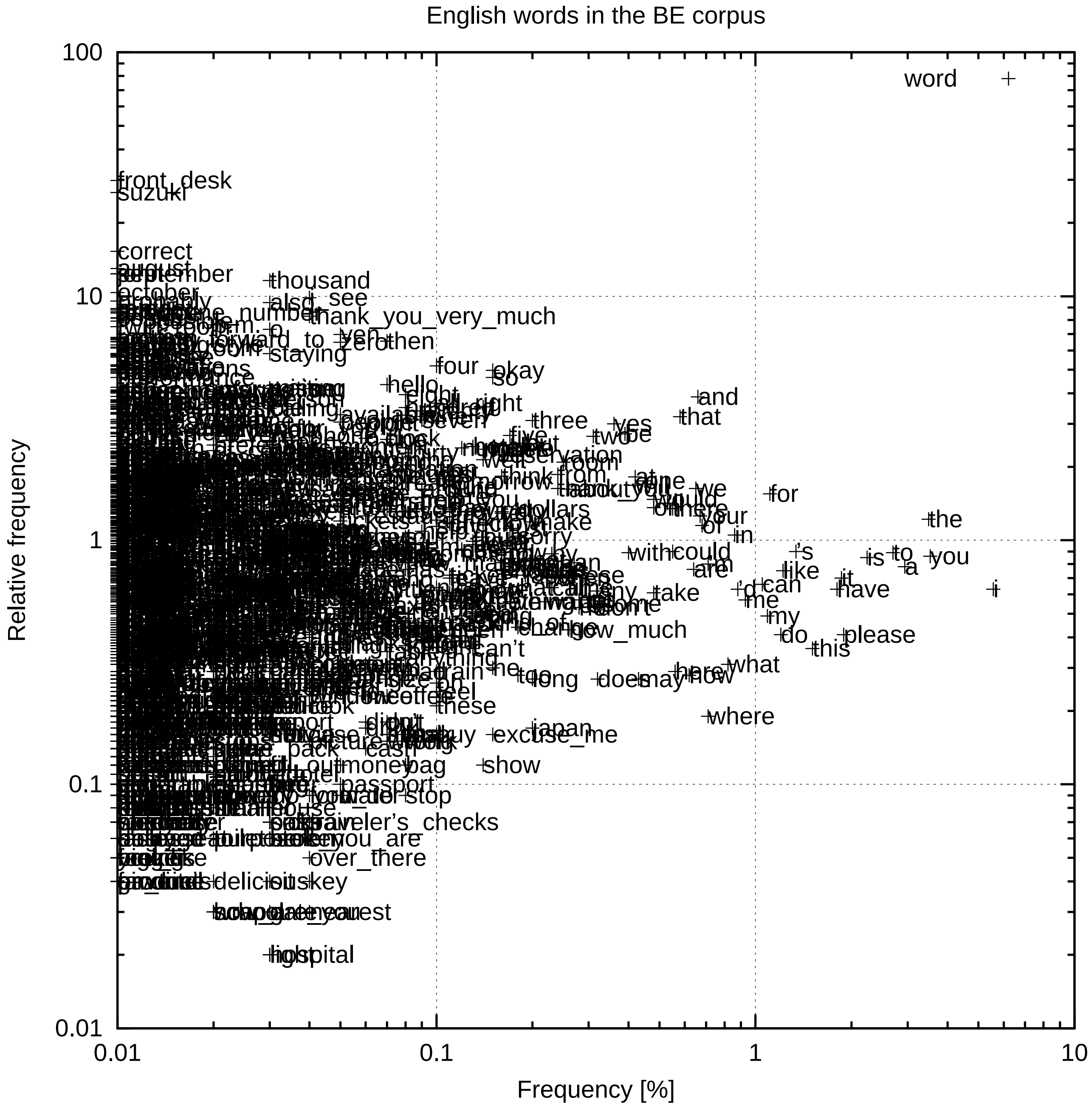

Figure 2: English word distribution

cur in the TC corpus, but other words such as “hospital,” “over there,” and “traveler’s checks” rarely occur.

4. Characteristics of Expressions

The BE corpus contains edited colloquial sentences that are not transcriptions of spoken utterances; as a result, the characteristics of some expressions may be different from those of daily conversations. Therefore, we carried out a preliminary experiment. We selected 50 sentences for this experiment. An example is shown as follows.

(Japanese) *Watashi no nimotsu wo sagashite kudasai.*

(English) Please look for my luggage.

We showed only the English part to three Japanese native subjects (Subjects A, B, and C) who were able to understand English sufficiently and then collected one Japanese utterance per subject that was freely uttered as the subject imagined the situation. The Japanese utterances are shown as follows.

(Subject A) *Kaban wo sagashite itadake masen ka?*
(Could you please look for my baggage?)

(Subject B) *Watashi no nimotsu wo sagashite kudasai.*
(Please look for my luggage.)

(Subject C) *Kaban sagashite iru n desu ga.*
(I’m looking for my baggage.)

Test set	Perplexity
Original 50 sentences	18.0
Subject A	23.0
Subject B	19.9
Subject C	35.5

Table 7: Test set perplexity of transcriptions of spoken utterances

	BE	TC
Expressions	edited colloquial	spoken
Size	large	medium
Sentence length	short	medium
Situations	multiple	limited

Table 8: Summary of corpora characteristics

Table 7 shows the test set perplexity of the original 50 sentences and the corresponding transcriptions of the utterances spoken by the three subjects. The test set perplexity of the original 50 sentences was 18.0, and the test set perplexity values of the corresponding transcriptions of spoken utterances by the three subjects (A, B and C) were 23.0, 19.9, and 35.5, respectively. One of them (Subject B) uttered similar expressions enabling, us to obtain a similar perplexity to the original one. However, another subject (Subject C) yielded a higher perplexity that was approximately double the original one. This suggested that some speakers might utter similar expressions to those in the BE corpus while others might not use similar expressions in the same situations.

5. Discussions

At ATR Spoken Language Translation Research Laboratories, we are building a broad-coverage bilingual corpus to study corpus-based speech translation technologies for the real world. As the first step, we created a broad-coverage bilingual basic expression (BE) corpus. The corpus is expected to cover many situations for Japanese going abroad, but the contents are not spoken dialogues. Considering this, we carried out a comparison between the BE corpus and a bilingual travel conversation (TC) corpus of spoken languages. The characteristics of both corpora are summarized in **Table 8**.

The cost of collecting spoken dialogues is higher than that of collecting edited colloquial sentences. Accordingly, it is difficult to enlarge the sizes of spoken language corpora. A large corpus, which is about ten times larger than a spoken language corpus, could be built by collecting Japanese/English parallel conversational sentences, which are usually found in phrase-books for foreign tourists. Such a BE corpus would be adequate for example-based machine translation research (Sumita, 2001). The corpus can be expected to cover a number of situations such as conversations about airports, hotels, restaurants, shopping, and problems.

However, this is insufficient for speech translation research in the real world because some people might not use similar expressions, according to the preliminary experiment shown in Table 7. Therefore, we are now collecting paraphrases by using some parts of the corpora to build a speech translation system that can accept more expressions by many speakers.

As the first trial, we collected English paraphrased translations against each Japanese example by five Americans with a sufficient Japanese understanding. Each American made three paraphrased expressions for each Japanese example. Some of the expressions from the different Americans were duplicate sentences. We selected 330 utterances from the TC corpus for this experiment. The average number of paraphrased expressions was 14.4 for each Japanese example, indicating that the Americans gave diverse sentences under the condition of three-sentence generation per person.

The example shown below is the smallest case, in which only five English paraphrased sentences could be obtained.

(Japanese) *Futari desu.*

(Original English) We're a party of two.

(Paraphrased English 1) Two.
[five persons]

(Paraphrased English 2) Just two.
[four persons]

(Paraphrased English 3) There will be two of us.
[three persons]

(Paraphrased English 4) There'll be two of us.
[one person]

(Paraphrased English 5) There are two of us.
[one person]

[] indicates the number of persons who generated each paraphrase. This trial confirmed the feasibility of obtaining paraphrases and their frequencies. Such paraphrased data can be efficiently used for automatic evaluation of speech translation performance (Sugaya et al., 2001) and as well as for language modeling.

6. Conclusions

At ATR Spoken Language Translation Research Laboratories, we are building a broad-coverage bilingual corpus to study corpus-based speech translation technologies for the real world. This paper reports our trials and discusses how to build a broad-coverage bilingual corpus based on numerical data from various aspects. In order to build a broad-coverage bilingual corpus for future speech translation research, there are three important points. The first is to have a variety of speech data, with a wide range of pronunciations, speaking styles, and speakers. We could conceivably cover such a variety by collecting not only bilingual spoken dialogues but also monolingual spoken dialogues. The second is to have data for a variety of situations. We could conceivably cover such a variety by collecting bilingual basic expressions usually found in phrase-books. The third is to have a variety of expressions. One approach might be to collect paraphrases representing various different expressions generated by many people for similar concepts.

7. Acknowledgments

The authors wish to thank members of ATR Spoken Language Translation Research Laboratories for their contributions to building the broad-coverage bilingual corpus for speech translation of travel conversations in the real world.

The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, “A study of speech dialogue translation technology based on a large corpus.”

8. References

Adam Kilgarriff and Tony Berber Sardinha, editors. 2000. *Proceedings of the Workshop on Comparing Corpora*, Hong Kong. Held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics.

Kikuo Maekawa. 2001. Compiling the corpus of spontaneous Japanese. In *Proceedings of the 1st Spontaneous Speech Science and Technology Workshop*, pages 7–12. *(in Japanese).*

Tsuyoshi Morimoto, Noriyoshi Uratani, Toshiyuki Takezawa, Osamu Furuse, Yasuhiro Sobashima, Hitoshi Iida, Atsushi Nakamura, Yoshinori Sagisaka, Norio Higuchi, and Yasuhiro Yamazaki. 1994. A speech and language database for speech translation research. In *Proceedings of the 3rd International Conference on Spoken Language Processing*, pages 1791–1794.

Atsushi Nakamura, Shoichi Matsunaga, Tohru Shimizu, Masahiro Tonomura, and Yoshinori Sagisaka. 1996. Japanese speech databases for robust speech recognition. In *Proceedings of the 4th International Conference on Spoken Language Processing*, pages 2199–2202.

Fumiaki Sugaya, Keiji Yasuda, Toshiyuki Takezawa, and Seiichi Yamamoto. 2001. Precise measurement method of a speech translation system’s capability with a paired comparison method between the system and humans. In *Proceedings of the Machine Translation Summit VIII*, pages 345–350.

Eiichiro Sumita. 2001. Example-based machine translation using DP-matching between word sequences. In *Proceedings of the ACL-2001 Workshop on Data-Driven Methods in Machine Translation*, pages 1–8.

Toshiyuki Takezawa, Tsuyoshi Morimoto, and Yoshinori Sagisaka. 1998a. Speech and language databases for speech translation research in ATR. In *Proceedings of the 1st International Workshop on East-Asian Language Resources and Evaluation — Oriental COCOSDA Workshop ’98 —*, pages 148–155.

Toshiyuki Takezawa, Tsuyoshi Morimoto, Yoshinori Sagisaka, Nick Campbell, Hitoshi Iida, Fumiaki Sugaya, Akio Yokoo, and Seiichi Yamamoto. 1998b. A Japanese-to-English speech translation system: ATR-MATRIX. In *Proceedings of the 5th International Conference on Spoken Language Processing*, pages 2779–2782.

Toshiyuki Takezawa. 1999. Building a bilingual travel conversation database for speech translation research. In *Proceedings of the 2nd International Workshop on East-Asian Language Resources and Evaluation — Oriental COCOSDA Workshop ’99 —*, pages 17–20.

Hirofumi Yamamoto, Shuntaro Isogai, and Yoshinori Sagisaka. 2001. Multi-class composite N-gram language model for spoken language processing using multiple word clusters. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 531–538.