Computational Linguistics and Chinese Language Processing
Vol. 12, No. 3, September 2007, pp. 303-324

Multilingual Spoken Language Corpus Development for Communication Research

Toshiyuki Takezawa[*], Genichiro Kikui[*+], Masahide Mizushima[*+], and

Eiichiro Sumita[#*]

Abstract

Multilingual spoken language corpora are indispensable for research on areas of spoken language communication, such as speech-to-speech translation. The speech and natural language processing essential to multilingual spoken language research requires unified structure and annotation, such as tagging. In this study, we describe an experience with multilingual spoken language corpus development at our research institution, focusing in particular on speech recognition and natural language processing for speech translation of travel conversations. An integrated speech and language database, Spoken Language DataBase (SLDB) was planned and constructed. Basic Travel Expression Corpus (BTEC) was planned and constructed to cover a variety of situations and expressions. BTEC and SLDB are designed to be complementary. BTEC is a collection of Japanese sentences and their translations, and SLDB is a collection of transcriptions of bilingual spoken dialogs. Whereas BTEC covers a wide variety of travel domains, SLDB covers a limited domain, *i.e.*, hotel situations. BTEC contains approximately 588k utterance-style expressions, while SLDB contains about 16k utterances. Machine-aided Dialogs (MAD) was developed as a development corpus, and both BTEC and SLDB can be used to handle MAD-type tasks. Field Experiment Data (FED) was developed as the evaluation corpus. We conducted an experiment, and based on analysis of our follow-up questionnaire, roughly half the subjects of the

[*] ATR Spoken Language Communication Research Laboratories, 2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan Telephone: +81 774 95 1301 Fax: +81 774 95 1308
E-mail: toshiyuki.takezawa@atr.jp

[+] currently with NTT Cyberspace Laboratories, Japan
E-mail: {kikui.genichiro, mizushima.masahide}@lab.ntt.co.jp

[#] National Institute of Information and Communications Technology, Japan
E-mail: eiichiro.sumita@{nict.go.jp; atr.jp}

[Received April 6, 2007; Revised August 3, 2007; Accepted August 6, 2007]

experiment felt they could understand and make themselves understood by their partners.

Keywords: Multilingual Corpus, Spoken Language, Speech Translation, Dialog, Communication.

1. Introduction

Various kinds of corpora developed for analysis of linguistic phenomena and statistical information gathering are now accessible via electronic media and can be utilized for the study of natural language processing. Since these include written-language and monolingual corpora, however, they are not necessarily useful for research and development of multilingual spoken language processing. A multilingual spoken language corpus is indispensable for research on areas of spoken language communication such as speech-to-speech translation.

Research on speech translation began in the 1980s. NEC demonstrated a prototype speech translation system at the Telecom ’83 exhibition. ATR Interpreting Telephony Research Laboratories was established in 1986 for the research of basic speech translation technologies and produced ASURA [Morimoto *et al*. 1993]. This system can recognize well-formed Japanese utterances in a limited domain, translate them into both English and German, and output synthesized speech. The ASURA system was used for the International Joint Experiment of Interpreting Telephony with participants from Kyoto, Japan (ATR), Pittsburgh, USA (Carnegie Mellon University [Lavie *et al*. 1997]) and Munich, Germany (Siemens and University of Karlsruhe) in January 1993 [Morimoto *et al*. 1993].

Many projects on speech-to-speech translation began at that time [Rayner *et al*. 1993; Roe *et al*. 1992; Wahlster *et al*. 2000]. SRI International and Swedish Telecom developed a prototype speech translation system that could translate queries from spoken English to spoken Swedish in the domain of air travel information systems [Rayner *et al*. 1993]. AT&T Bell Laboratories and Telefónica Investigación y Desarrollo developed a restricted domain spoken language translation system called Voice English/Spanish Translator (VEST) [Roe *et al*. 1992]. In Germany, *Verbmobil* [Wahlster 2000], was created as a major speech-to-speech translation research project. The *Verbmobil* scenario assumes native speakers of German and of Japanese who both possess at least a basic knowledge of English. The *Verbmobil* system supports them by translating from their mother tongue, *i.e.* Japanese or German, into English.

In the 1990s, speech recognition and synthesis research shifted from a rule-based to a corpus-based approach such as HMM and N-gram. However, machine translation research still depended mainly on a rule-based or knowledge-based approach. In the 2000s, wholly corpus-based projects such as European TC-STAR [Höge 2002; Lazzari 2006] and DARPA GALE [Roukos 2006] began to deal with monologue speeches such as broadcast news and

European Parliament plenary speeches. In this paper, we report corpus construction activities for translation of spoken dialogs of travel conversations.

There are a variety of requirements for every component technology, such as speech recognition and language processing. A variety of speakers and pronunciations may be important for speech recognition, and a variety of expressions and information on parts of speech may be important for natural language processing. The speech and natural language processing essential to multilingual spoken language research requires unified structure and annotation, such as tagging.

In this paper, we introduce an interpreter-aided spoken dialog corpus and discuss corpus configuration. Next, we introduce the basic travel expression corpus developed to train machine translation of spoken language among Japanese, English, and Chinese speakers. Finally, we discuss the Japanese, English, and Chinese multilingual spoken dialog corpus that we created using speech-to-speech translation systems.

2. Overview of Approach

We first planned and constructed an integrated speech and language database called Spoken Language DataBase (SLDB) [Morimoto *et al.* 1994; Takezawa *et al.* 1998]. The task involved travel conversations between a foreign tourist and a front desk clerk at a hotel; this task was selected because people are familiar with it and because we expect it to be included in future speech translation systems. All of the conversations for this database take place in English and Japanese through interpreters because the research at that time concentrated on Japanese and English. The interpreters serve as the speech translation system. One remarkable characteristic of the database is its integration of speech and linguistic data. Each conversation includes data on recorded speech, transcribed utterances, and their correspondences. This kind of data is very useful because it contains transcriptions of spoken dialogs between speakers who speak different mother tongues. However, the cost of collecting spoken languages is too high to expand the size.

There are three important points to consider in designing and constructing a corpus for dialog-style speech communication such as speech-to-speech translation. The first is to have a variety of speech samples with a wide range of pronunciations, speaking styles, and speakers. The second point is to have data for a variety of situations. A “situation” means one of various limited circumstances in which the system’s user finds him- or herself, such as an airport, a hotel, a restaurant, a shop, or in transit during travel; it also involves various speakers’ roles, such as communication with a middle-aged stranger, a stranger wearing jeans, a waiter or waitress, or a hotel clerk. The third point is to have a variety of expressions.

According to our previous study [Takezawa *et al.* 2000], human-to-machine conversational speech data shared characteristics with human-to-human indirect communication speech data such as spoken dialogs between Japanese and English speakers through human interpreters. Moreover, human-to-human indirect communication data had an intermediate characteristic, *i.e.*, it was positioned somewhere between direct communication data, that is, Japanese monolingual conversations, and speech data from conversational text. If we assume that a speaker would accept a machine-friendly speaking style, we could take a great step forward: a clear separation of speech data collection and multilingual data collection. In the following, we focus on multilingual data collection. In order, Basic Travel Expression Corpus (BTEC) [Takezawa *et al*. 2002; Kikui *et al.* 2003] was planned to cover the varieties of situations and expressions.

Machine-aided Dialogs (MAD) was planned as a development corpus to handle the differences between the target utterance with which speech translation systems must deal and the following two corpora.

SLDB contains no recognition/translation errors because the translations between people speaking different languages are done by professional human interpreters. However, even a state-of-the-art speech translation system cannot avoid recognition/translation errors.

BTEC contains edited colloquial travel expressions, which are not transcriptions, so some people might not express things in the same way, and the frequency distribution of expressions might be different from actual dialogs.

Field Experiment Data (FED) was planned as the evaluation corpus. Table 1 shows an overview of the corpora. In the table, S2ST stands for speech-to-speech translation, MT stands for machine translation, J, E, and C stand for Japanese, English, and Chinese, respectively.

Table 1. Overview of corpora

Name	SLDB	BTEC	MAD	FED
	Spoken Language DataBase	Basic Travel Expression Corpus	Machine-Aided Dialogs	Field Experiment Data
Purpose	Developing S2ST	Training MT	Developing S2ST	Evaluation of S2ST
Domain	Hotel	Travel	Travel	Travel
Languages	J E (C)	J E C	J E (C)	J E C
Speaker Participants	71 (+23 Interpreters)	Not spoken	45	84
Size	16k	588k	13k	2k

3. Interpreter-Aided Spoken Dialog Corpus (SLDB)

SLDB contains data from dialog spoken between English and Japanese speakers through human interpreters [Morimoto *et al.* 1994; Takezawa *et al.* 1998]. All utterances in SLDB have been translated into Chinese. The content is entirely travel conversations between a foreign tourist and a front desk clerk at a hotel. Human interpreters serve as the speech translation system.

Table 2 is an overview of the corpus, and Table 3 shows its basic characteristics.

Table 2. Overview of SLDB

Number of collected dialogs	618
Speaker participants	71
Interpreter participants	23

Table 3. Basic characteristics of SLDB

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

One remarkable characteristic of SLDB is its integration of speech and linguistic data. Each conversation includes recorded speech data, transcribed utterances, and the correspondences between them.

The transcribed Japanese and English utterances are tagged with morphological information. This kind of tagged information is crucial for natural language processing as well as for speech recognition language modeling. The recorded speech signals and transcribed utterances in the database provide both examples of various phenomena in bilingual conversations, and input data for speech recognition and machine translation evaluation purposes.

Data can be classified into the following three major categories.

1. Transcribed data
2. Tagged data
3. Speech data

Transcribed data consists of the following.

(a) Bilingual text

(b) Japanese text

(c) English text

The recorded bilingual conversations are transcribed into a text file. The bilingual text contains descriptions of the situations in which a speech translation system is used.

J: Arigatou gozaimasu. Kyoto Kankou Hotel de gozaimasu.

JE: Thank you for calling Kyoto Kanko Hotel. |How may I help you?

E: Good evening. |I’d like to make a reservation, please.

EJ: Konbanwa. |Yoyaku wo shi tai n desu keredomo.

J: Hai,[e-]go yoyaku no hou wa itsu desho u ka?

JE: Yes, when do you plan to stay?

E: I’d like to stay from August tenth through the twelfth, for two nights.|
If possible, I’d like a single room, please.

EJ: Hachigatsu no tooka kara juuni-nichi made, ni-haku shi tai n desu.|
Dekire ba, single room de onegaishimasu.

J: Kashikomarimashita. |Shoushou o-machi kudasai mase.

JE: All right, please wait a moment.

J: O-mata se itashimashita.|
Osoreiri masu ga, single room wa manshitsu to nat te orimasu.

JE: I am very sorry our single rooms are all booked.

J: [e]Washitsu ka twin room no o-hitori sama shiyou deshi tara o-tori dekimasu ga.

JE: But, Japanese style rooms and twin rooms for single use are available.

E: [Oh] what are the rates on those types of rooms?

EJ: Sono o-heya no ryoukin wo oshie te kudasai.

Figure 1. Conversation between an American tourist and a Japanese front desk clerk.

Figure 1 shows an example of transcribed conversations. The Japanese text in Figure 1 has been transcribed into Romanized Japanese for the convenience of readers who do not understand Japanese *hiragana*, *katakana*, and *kanji* (Chinese characters). The original text was transcribed in Japanese characters *hiragana*, *katakana*, and *kanji*. Interjections are bracketed. J, E, JE, or EJ at the beginning of a line denotes a Japanese speaker, an English speaker, a Japanese-to-English interpreter, or an English-to-Japanese interpreter, respectively. "｜" denotes a sentence boundary. A blank line between utterances shows that the utterance's right was transferred.

The Japanese text is produced by extracting the utterances of a Japanese speaker and an English-to-Japanese interpreter, while the English text is produced by extracting the utterances of an English speaker and a Japanese-to-English interpreter. These two kinds of data are utilized for such monolingual investigations as morphological analysis.

The tagged data consists of the following.

(d) Japanese morphological data

(e) English morphological data

SLDB is available to outside research institutions and can be accessed at the following URL: http://www.atr.jp.

4. Basic Travel Expression Corpus (BTEC)

The Basic Travel Expression Corpus (BTEC) [Takezawa *et al*. 2002; Kikui *et al*. 2003] was designed to cover utterances for possible travel conversations topic and their translations. Since it is practically impossible to collect them by transcribing actual conversations or simulated dialogs, we decided to use sentences provided by bilingual travel experts based on their experience. We started by looking at phrasebooks that contain bilingual sentence pairs (in this case Japanese/English) that the editors consider useful for tourists traveling abroad. Such sentence pairs were collected and rewritten to make translation as context-independent as possible and to comply with the speech transcription style of our research institution. Sentences that were outside of the travel domain or have very special meanings were removed.

Table 4 lists the basic statistics of the BTEC collections, called BTEC1, 2, 3, 4, and 5. Each collection was created using the same procedure in a different time period or using a different translation direction from the source language to target languages. Strictly speaking, morphemes are used as the basic linguistic unit for Japanese (instead of words), since morpheme units are more stable than word units.

Table 4. Overview of BTEC

	BTEC1	BTEC2	BTEC3	BTEC4	BTEC5
Number of utterance-style expressions	172k	46k	198k	74k	98k
Number of Japanese word tokens	1,174k	341k	1,434k	548k	1,046k
Number of Japanese word types	28k	20k	43k	22k	28k
Languages (Source:Targets)	J:EC	J:EC	J:EC	E:JC	E:JC

The aims of the BTEC corpus are for translation and language modeling for automatic speech recognition. For translation, one of the key points to cover is the translation direction from the source language to target languages. For automatic speech recognition in the travel domain, one of the key points to cover is multiple sub-domains such as airport-related dialogs, hotel-related dialogs, and so on.

For translation, the BTEC collections cover both translation directions. BTEC1, BTEC2, and BTEC3 contain expressions for Japanese tourists visiting the USA, UK, or Australia. The translation direction is from Japanese to English and Chinese. BTEC4 mainly contains expressions for American tourists who visit Japan. The translation direction is from English to Japanese and Chinese. BTEC5 contains various expressions, such as those for American tourists who go to Korea. The translation direction is from English to Japanese and Chinese.

For automatic speech recognition, BTEC covers multiple domains. Domain information is given for BTEC1, BTEC2, and BTEC3. Table 5 shows an overview.

BTEC sentences, as described above, did not come from actual conversations but were generated by experts as reference materials. This approach enabled us to efficiently create a broad corpus; however, it may have two problems. First, this corpus may lack utterances that occur in real conversation. For example, when people ask the way to a bus stop, they often use a sentence like (1). However, in BTEC this is expressed more directly, as in (2).

(1) I’d like to go downtown. Where can I catch a bus?

(2) Where is a bus stop (to go downtown)?

We will discuss this issue in the section on MAD.

The second problem is that the frequency distribution of this corpus may be different from the actual distribution. In this corpus, the frequency of an utterance most likely reflects the best trade-off between usefulness in real situations and compactness of the collection. Therefore, it is possible to think of this frequency distribution as a first approximation of reality, but this is an open question.

A part of BTEC was distributed to the participants in the International Workshop on Spoken Language Translation (IWSLT) [IWSLT 2006].

Table 5. Domain information of BTEC

Domain	Frequency
Communication	20.4%
Basic	19.1%
Trouble	8.7%
Shopping	7.9%
Stay	6.9%
Sightseeing	6.6%
Transfer	6.6%
Restaurant	5.9%
Business	3.8%
Airport	3.6%
Contact	3.3%
Airplane	2.3%
Drink	1.0%
Home stay	1.0%
Exchange	0.8%
Snack	0.8%
Beauty	0.5%
Study overseas	0.5%
Go home	0.3%
Total	100.0%

5. Machine Translation-Aided Spoken Dialog Corpus (MAD)

The approach exemplified by BTEC focuses on maximizing the coverage of the corpus rather than creating an accurate sample of reality. Users may use different wording when they speak to the system. In addition, there may be differences between the target utterance with which speech translation systems must deal and the following two corpora.

SLDB contains no recognition/translation errors because the translations between people speaking different languages are done by professional human interpreters. However, even a state-of-the-art speech translation system cannot avoid recognition/translation errors.

BTEC contains edited colloquial travel expressions, which are not transcriptions, so some people might not express things in the same way and the frequency distribution of expressions might be different from actual dialogs.

Therefore, MAD is intended to collect representative utterances that people will input into S2ST systems. For this purpose, simulated dialogs (*i.e.*, role play) were carried out between two native speakers of different mother tongues with a Japanese/English bi-directional S2ST system, instead of using human interpreters.

During the first half of the research program, human typists were used instead of speech recognizers to ensure that we collected good quality data. During the second half of the research program, the S2ST system between English and Japanese was used.

5.1 Collecting Spoken Dialog Data Using Typists

Figure 2 is an overview of the data collection environment. An English typist transcribes an English utterance and inputs it into a machine translation system from English to Japanese. The translated Japanese text and its synthesized speech are sent to a Japanese speaker. Likewise, a Japanese typist transcribes a Japanese utterance and inputs it into a machine translation system from Japanese to English. The translated English text and its synthesized speech are sent to an English speaker. By repeating this process, an MT-aided bilingual dialog continues. Speech waves, transcriptions, and translated texts are stored in log files.

Five sets of simulated dialogs (MAD1 through MAD5) have so far been developed, changing parameters such as system configurations, complexity of dialog tasks, instructions to speakers, and so on. Table 6 shows a summary of the five experiments, MAD1-MAD5. In this table, the number of utterances includes both Japanese and English.

The first set of dialogs (MAD1) was collected to see whether conversation through a machine translation system is feasible. The second set (MAD2) focused on task achievement by assigning complex tasks to participants. The third set (MAD3) contains carefully recorded speech data of medium complexity. MAD4 and MAD5 aim to investigate how utterances change based on a change in setting.

It is very likely that people would speak differently to a spoken language system based on the instructions given to them. Instructions were conveyed to subjects for all sets other than MAD1 using instructional movies to ensure that the same instructions were given to each subject. Before starting the experiments, subjects were asked to watch these movies and then try the system with test dialogs. Instructions and practice took about 30 minutes. We gave different types of instructions for the fourth set (MAD4).

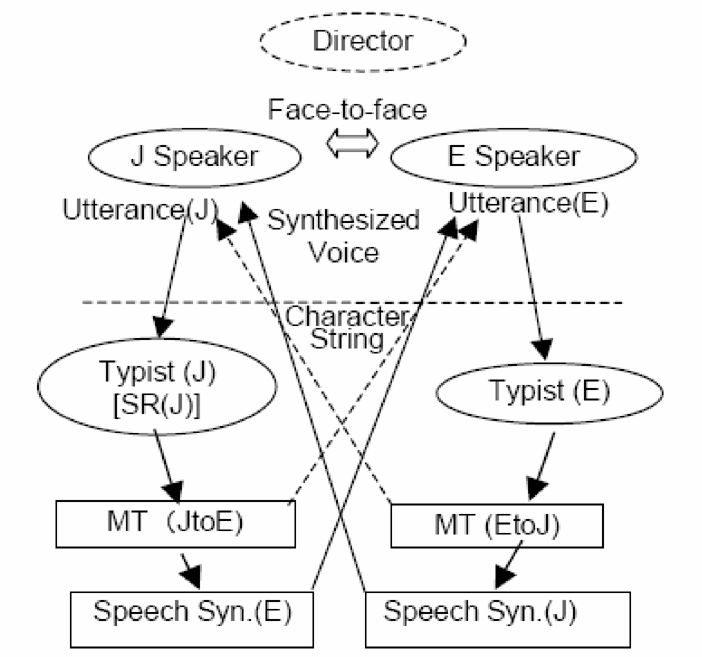

Figure 2. Data collection environment of MAD

Table 6. Statistics of MAD corpora

Subset ID	MAD1	MAD2	MAD3	MAD4	MAD5
Reference	[Takezawa and Kikui 2003]	[Takezawa and Kikui 2003]	[Takezawa *et al*. 2003]	[Takezawa and Kikui 2004]	[Mizushima *et al*. 2004]
Number of utterances	3,022	1,696	2,180	1,872	1,437
Number of words per utterance	10.0	12.6	11.1	9.82	8.47
Number of utterances per dialog	7.8	49.3	18.8	22.0	27.0
Task complexity	Simple	Complex	Medium	Medium	Medium

Average numbers depend on experimental conditions.

S2ST presupposes that each user understands the translated utterances of the other. However, the dialog environment described so far allows the user to access other information, such as translated text displayed on a PDA. We tried to control the extra information in MAD5 to see how utterances would be affected.

Part of the MAD corpus has been translated into Chinese.

5.2 Collecting Spoken Dialog Data Using Speech Translation Systems

Spoken dialog data was collected using the S2ST system for English and Japanese. This data collection experiment is called MAD6 because five data collection experiments were carried out using typists. The system was configured as follows.

- Acoustic model for Japanese speech recognition: Speaker-adapted models.
- Language model for Japanese speech recognition: Vocabulary size 52,000 words.
- Acoustic model for English speech recognition: Speaker-adapted models.
- Language model for English speech recognition: Vocabulary size 15,000 words.

Table 7 is an overview of MAD6. Data collected by typists (MAD1 through MAD5) contains some translation errors but very few recognition errors. However, MAD6 data contains both recognition errors and translation errors. We found that translation errors caused by recognition errors sometimes caused great confusion. That is, users need many more turns to recover from translation errors caused by recognition errors than to recover from mere translation errors. Moreover, we found that the user's speaking style changed similar to read speech when using speech recognizers. This was because users could confirm their recognition results using a PC display. Experienced users soon understood that they were confused by translation errors caused by recognition errors and adopted strategies to avoid recognition errors. As a result, their speaking style seemed to change from a natural dialog style to a read speech style.

Table 7. Overview of MAD6

	MAD6
Purpose	Spoken dialog data collection using S2ST system
Task	Simple as in MAD1
Number of utterances	2,507
Number of dialogs	139

5.3 Comparative Analysis and Discussion

BTEC and SLDB are designed to be complementary. BTEC is a collection of Japanese sentences and their translations, and SLDB is a collection of transcriptions of bilingual spoken dialogs. Whereas BTEC covers a wide variety of travel domains, SLDB covers a limited domain, *i.e.*, hotel situations. BTEC contains approximately 588k utterance-style expressions, while SLDB contains about 16k utterances. Thus, we can hypothesize that BTEC and SLDB together cover the same content as MAD. This hypothesis is partly validated by the cross-perplexity shown in Table 8. In this table, BTEC1＋SLDB combines two language models trained on BTEC1 and SLDB with linear interpolation. Similarly, BTEC1＋Extra combines BTEC1 and a corpus called Extra, which is a sample of a BTEC-type extra corpus of about the same size as SLDB. This clearly shows that both BTEC1 and SLDB are required for handling MAD-type tasks. Further discussion is available in [Kikui *et al*. 2006].

Table 8. Cross-perplexity for MAD (Japanese)

	Training corpus			
	BTEC1	SLDB	BTEC1＋SLDB	BTEC1＋Extra
Size (Number of utterances)	162k	12k	174k	174k
Cross-perplexity	38.2	94.9	30.7	35.7

6. Field Experiment Data (FED)

An ideal approach to applying a system to real utterances is to let people use the system in real world settings to achieve real conversational goals (*e.g.*, booking a package tour). This approach, however, has at least two problems. First, it is difficult to back up the system when it makes errors because current technology is not perfect. Second, it is difficult to control tasks and conditions to do meaningful analysis of the collected data.

The new experiment reported here was still in the role-play style but its dialog situations were designed to be more natural. The S2ST system for travel conversation was set up at tourist information centers in an airport and a train station, and non-Japanese-speaking people were asked to talk with the Japanese staff at information centers using the S2ST system.

6.1 Experimental System for Data Collection

Figure 3 is a diagram of the overall experimental system. The system includes two PDAs, one for each language, and several PC servers. The PC servers are controlled by a special controller called the gateway for component engines, consisting of automatic speech recognition (ASR) [Itoh *et al*. 2004], machine translation (MT) [Sumita *et al*. 2004], and speech synthesis (SS) [Kawai *et al*. 2004] PCs for each language and each language-pair. The gateway is responsible for controlling information flow between PDAs and engines. It is also responsible for mediating messages from the ASR and MT engines to PDAs. Each PDA is connected to the gateway with a wireless LAN. The gateway and component engines are wired. Headset microphones were used in the FED experiment.

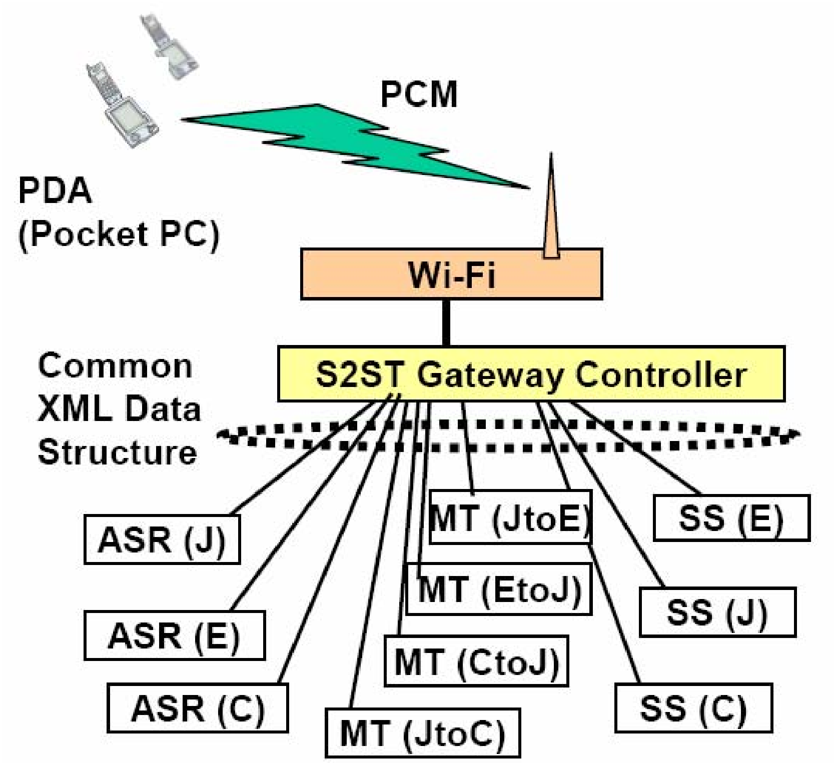

Figure 3. Overview of experimental system

An utterance spoken into a PDA is sent to the gateway server, which calls the ASR, MT, and SS engines in this order to have the utterance translated. Finally, the gateway sends the translated utterance to the other PDA.

Speaker-adapted acoustic models were used for Japanese speech recognition because only a few Japanese staff at the tourist office agreed to participate in the FED experiment. A few proper names that were deemed necessary to carry out the planned conversations were added to the lexicon. These included names such as those of stations near the locations of the experiment.

6.2 Locations

Data collection was conducted near two tourist information centers. One was in Kansai International Airport (hereafter, KIX), and the other was at Osaka City Air Terminal (hereafter, OCAT) in the center of Osaka. The former is in the main arrival lobby of the airport, which many tourists pass as they emerge from customs. The latter is a semi-enclosed area of about 40 m^2 enclosed by glass walls (but with two open doors).

Environmental noise was 60-65 dBA in both places but rose to 70 dBA when the public address system was in use.

The language pairs were English-Japanese/Japanese-English and Chinese-Japanese/Japanese-Chinese.

6.3 Scenario

A good method of collecting real utterances is to just let subjects talk freely without using predetermined scenarios. Analyzing uncontrolled dialog, however, is very difficult. In the FED experiment, eight dialog scenarios were prepared. These scenarios, listed below, are categorized by expected number of turns for each speaker into three levels of complexity.

Level-1 : Requires one or two turns per speaker plus greetings.
E.g., “Please ask where the bus stop for Kyoto station is.”

Level-2 : Requires three or four turns per speaker plus greetings.
E.g., “Please ask the way to Kyoto station.”

Level-3 : Free discussion.
E.g., “Please ask anything related to traveling in the Osaka area.”

Real dialogs included many clarification sub-dialogs necessitated by incomprehensible output from the system. This means that the number of turns was actually larger than we expected or planned.

6.4 Speakers

6.4.1 Japanese Speakers

We asked staff at the tourist information centers to participate in the experiments, and six people at KIX and three at OCAT agreed to take part.

6.4.2 Chinese Speakers

Since the Chinese speech recognizer was trained on Mandarin speech, we needed to recruit subjects from the Beijing region of China. It was, however, difficult to find tourists from China who had time to participate in the experiment because most of them came to Osaka as members of tightly scheduled group tours. Therefore, we relied on 36 subjects gathered by the Osaka prefectural government. These subjects are college students from China majoring in non-technical areas such as foreign studies and tourism.

6.4.3 English Speakers

The English speech recognizer was trained on North American English. Again, however, it was difficult to find volunteer subjects who speak North American English. We expected to recruit many individual tourists, and most of the English-speaking volunteer subjects were indeed tourists arriving at or leaving the airport during the experiment. In addition to these volunteers, Osaka prefecture provided nine subjects who were working in Japan as English teachers. The resulting 39 subjects were not all North Americans, as shown in Table 9.

Table 9. Origin of English-speaking subjects

Origin	Number of subjects
USA	15
UK	6
Australia	5
Canada	4
New Zealand	2
Denmark	2
Other	5

6.5 Collecting Data

First, we set up the S2ST system and asked the Japanese subjects (*i.e.*, service personnel at the tourist information centers) to stand by at the experimental sites.

When an English or Chinese speaking subject visited a center, he or she was asked to fill out the registration form. Then, the staff explained for 2-3 minutes how to use the S2ST system and asked the subject to try very simple utterances like “hello” or “thank you.” After the trial utterances, we had the subject try two dialogs: one dialog for practice using a level 1 scenario, and the other for the “main” dialog, which was a scenario chosen randomly from level 1 through level 3. Finally, the subject was asked to answer a questionnaire.

The average time from registration to filling out the questionnaire was 15-20 minutes. Since we conducted 4-5 hours of experiments each day, excluding system setup, we were able to obtain dialog data for 15 subjects per day.

Table 10 is an overview of FED data.

Table 10. Overview of FED

	J (to E)	E (to J)	J (to C)	C (to J)
Number of utterances	608	660	344	484
Number of speakers	7	39	6	36
Number of word tokens	3,851	4,306	2,017	422
Number of word types	727	668	436	382

6.6 Performance Evaluation

We collected questionnaires from all subjects. As mentioned above, all of the Chinese-speaking subjects were college students. Therefore, they had at least a basic understanding of Japanese because they attend lectures given in Japanese. Therefore, in the following, we will focus on the English side.

First, overall performance is measured in terms of subjective scores from A to D, defined as follows.

(A) Perfect: no problems in either information or grammar.

(B) Good: easy to understand, with either some unimportant information missing or flawed grammar.

(C) Fair: broken but understandable with effort.

(D) Nonsense or no-output: (including ASR errors).

Table 11 shows results of English-Japanese translation. About 50% of the utterances were translated into the target language with their original meaning preserved (*e.g.*, at rank B or above).

Table 11. Results of English-Japanese translation

Rank	J to E (%)	E to J (%)
A	37.1	36.2
B	10.2	18.2
C	10.9	5.7
D	41.4	24.5

The ultimate goal of speech translation systems is to help users achieve their conversational goals. Instead of evaluating “goal achievement,” we asked them to subjectively evaluate during the course of conversations to what extent 1) they could understand their partner’s utterances, and 2) they felt that their utterances were correctly understood. Table 12 shows the questionnaire results on these issues.

Table 12. Results of questionnaires on understanding a partner's utterances (English side)

	Make the hearer understood (%)	Understood what the partner said (%)
Complete	8.3	22.2
Almost	41.6	50.0
Half	33.3	22.2
Little	16.7	5.6

Note that, although the number of subjects (*i.e.*, samples) is limited, the table does show that roughly half the subjects felt they could almost understand and make themselves understood by their partners. The result seems to coincide with the overall performance shown in Table 11.

7. Conclusion

This paper described our experience with multilingual spoken language corpus development at our research institution, focusing in particular on speech recognition and natural language processing for speech translation of travel conversations.

First, we introduced an interpreter-aided spoken dialog corpus called SLDB, and mentioned corpus configuration. Next, we introduced BTEC, which was built to train machine translation of spoken language among Japanese, English, and Chinese speakers. BTEC and SLDB are designed to be complementary. BTEC is a collection of Japanese sentences and their translations, and SLDB is a collection of transcriptions of bilingual spoken dialogs. Whereas BTEC covers a wide variety of travel domains, SLDB covers a limited domain, *i.e.*, hotel situations. BTEC contains approximately 588k utterance-style expressions, and SLDB contains about 16k utterances.

Finally, we discussed a multilingual spoken dialog corpus between Japanese, English, and Chinese created using speech-to-speech translation systems. MAD was developed as a development corpus and we presented both BTEC and SLDB can be used to handle with MAD-type tasks. FED was planned as the evaluation corpus. According to analysis of the questionnaire, roughly half the subjects felt they could understand and make themselves understood by their partners.

In the future, we plan to expand our activities to multilingual spoken language communication research and development involving both verbal and nonverbal communication. Information is available at the following URL: http://www.atr.jp.

Acknowledgments

The work reported here was mainly conducted at ATR Spoken Language Communication Research Laboratories. The authors are grateful to Prof. Seiichi Yamamoto, Dr. Satoshi Nakamura, and all other staff who helped construct corpora.

The FED corpus was collected within the framework of "Social experiments on supporting non-Japanese-speaking tourists using information technologies", which was carried out by the Osaka prefectural government and related offices in the winter of 2004. In these experiments, Osaka's prefectural government negotiated with management of facilities frequented by foreign tourists, such as airports and bus terminals, to provide the necessary assistance (*e.g.*, use of public space and electricity). The government also gathered the volunteer subjects.

References

Höge, H., "Project Proposal TC-STAR: Make Speech to Speech Translation Real," *Proc. of International Conference on Language Resources and Evaluation*, 2002, pp. 136–141.

Itoh, G., Ashikari, Y., Jitsuhiro, T., and Nakamura, S., "Summary and Evaluation of Speech Recognition Integrated Environment ATRASR," *Proc. of Autumn Meeting of the Acoustical Society of Japan*, 1-P-30, 2004, pp. 221–222.

IWSLT, *Proc. of International Workshop on Spoken Language Translation,* Kyoto, Japan, 2006.

Kawai, H., Toda, T., Ni, J., Tsuzaki, M., and Tokuda, K., "XIMERA: a New TTS from ATR Based on Corpus-based Technologies," *Proc. of 5th ISCA Speech Synthesis Workshop,* 2004, pp. 179–184.

Kikui, G., Sumita, E., Takezawa, T., and Yamamoto, S., "Creating Corpora for Speech-to-Speech Translation," *Proc. of European Conference on Speech Communication and Technology*, 2003, pp. 381–382.

Kikui, G., Yamamoto, S., Takezawa, T., and Sumita, E., "Comparative Study on Corpora for Speech Translation," *IEEE Trans. on Audio, Speech, and Language Processing*, 14 (5), 2006, pp. 1674–1682.

Lavie, A., A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavaldà, T. Zeppenfeld, and P. Zhan, "JANUS-III: Speech-to-speech Translation in Multiple Language," *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997, pp. 99–102.

Lazzari, G., "TC-STAR: a Speech to Speech Translation Project," *Proc. of International Workshop on Spoken Language Translation*, 2006, pp. xiv–xv.

Mizushima, M., T. Takezawa, and G. Kikui, "Effects of Audibility of Partner's Voice and Visibility of Translated Text in Machine-Translation-Aided Bilingual Spoken

Dialogues," *IPSJ SIG Technical Reports*, 2004 (74), 2004-HI-109-19/2004-SLP-52-19, 2004, pp. 99–106.

Morimoto, T., T. Takezawa, F. Yato, S. Sagayama, T. Tashiro, M. Nagata, and A. Kurematsu, "ATR's Speech Translation System: ASURA," *Proc. of European Conference on Speech Communication and Technology*, 1993, pp. 1291–1294.

Morimoto, T., N. Uratani, T. Takezawa, O. Furuse, Y. Sobashima, H. Iida, A. Nakamura, Y. Sagisaka, N. Higuchi, and Y. Yamazaki, "A Speech and Language Database for Speech Translation Research," *Proc. of International Conference on Spoken Language Processing*, 1994, pp. 1791–1794.

Rayner, M., I. Bretan, D. Carter, M. Collins, V. Digalakis, B. Gambäck, J. Kaja, J. Karlgren, B. Lyberg, S. Pulman, P. Price, and C. Samuelsson, "Spoken Language Translation with Mid-90's Technology: a Case Study," *Proc. of European Conference on Speech Communication and Technology*, 1993, pp. 1299–1302.

Roe, D. B., P. J. Moreno, R. W. Sproat, F. C. N. Pereira, M.D. Riley, and A. Macarrón, "A Spoken Language Translator for Restricted-domain Context-free Languages," *Speech Communication*, 11, 1992, pp. 311–319.

Roukos, S., "Recent Results on MT Evaluation in the GALE Program," *Proc. of International Workshop on Spoken Language Translation*, 2006, pp. xvi–xvii.

Sumita, E., H. Nakaiwa, and S. Yamamoto, "Corpus-Based Translation Technology for Multi-lingual Speech-to-Speech Translation," *Proc. of Spring Meeting of the Acoustical Society of Japan*, 1-8-26, 2004, pp. 57–58.

Takezawa, T., T. Morimoto, and Y. Sagisaka, "Speech and Language Databases for Speech Translation Research in ATR," *Proc. of Oriental COCOSDA Workshop*, 1998, pp. 148–155.

Takezawa, T., F. Sugaya, M. Naito, and S. Yamamoto, "A Comparative Study on Acoustic and Linguistic Characteristics Using Speech from Human-to-Human and Human-to-Machine Conversations," *Proc. of International Conference on Spoken Language Processing*, III, 2000, pp. 522–525.

Takezawa, T., E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Toward a Broad-Coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World," *Proc. of International Conference on Language Resources and Evaluation*, 2002, pp. 147–152.

Takezawa, T. and G. Kikui, "Collecting Machine-Translation-Aided Bilingual Dialogues for Corpus-Based Speech Translation," *Proc. of European Conference on Speech Communication and Technology*, 2003, pp. 2757–2760.

Takezawa, T., A. Nishino, K. Takashima, T. Matsui, and G. Kikui,, "An Experimental System for Collecting Machine-Translation-Aided Dialogues," *Proc. of Forum on Information Technology*, E-036, 2003, pp. 161–162.

Takezawa, T. and G. Kikui, “A Comparative Study on Human Communication Behaviors and Linguistics Characteristics for Speech-to-Speech Translation,” *Proc. of International Conference on Language Resources and Evaluation*, 2004, pp. 1589–1592.

Wahlster, W. (Ed.), “Verbmobil: Foundations of Speech-to-Speech Translation,” Springer, Germany, 2000.