Data Collection and Transliteration of Japanese Spontaneous Database in the Travel Arrangement Task Domain

Akira Kurematsu[1], *Youichi Akegami*[2], *Tanja Schultz*[3], *Susanne Burger*[4]

[1] Univ. of Electro-Communications, JAPAN, e-mail: kure@apple.ee.uec.ac.jp
[2] ATR International, JAPAN, e-mail: akegami@ctr.atr.co.jp
[3] Univ. of Karlsruhe, Germany, e-mail: tanja@ira.uka.de
[4] Univ. of Munchen, German, e-mail: burger@phonetik.uni-muenchen.de

ABSTRACT

This paper describes the method to construct and transcribe Japanese spontaneous speech data for VERBMOBIL, the German research project of speech translation.. Spontaneous spoken dialogue database is the basis for developing speech and language processing for dialogue systems such as speech translation system. The extended data of human-to-human spoken dialogue in the scenario of travel arrangement has been initiated to be collected in German, English and Japanese in the travel arrangement task. Romanized transcription is used to develop acoustic model and language model in speech recognition system, and natural language translation system. In this paper, issues of transliteration method and several rules and conventions to transcribe Japanese spoken dialogue will be described.

1. Introduction

The corpora of person to person dialogue areneeded especially for a spoken language system and a speechtranslation system. The speechsituation with two people is more natural than a guided speech by themachine. The study of simultaneous speech from two speakers isimportant for research on dialogue or dialogue analysis, intentionanalysis, and spoken language understanding. In addition, speechrecognition system, which are now only able to deal with one speakerat a time, would also have to be able to deal with different speakerstalking simultaneously. Speech corpora containing dialogues will beused for the training and testing data for advanced recognizers. The issues of spoken language corpus design is found in detail in[1].

The spontaneous speech corpora has been collected under VERBMOBILproject, the German project of VERBMOBIL of face-to-face speechtranslation system. The project have supported to collectspontaneous spoken dialogue data in German, English and Japanese inthe same scenario. The corpora of VERBMOBIL2 in succession to VERBMOBIL1 is gathered in the travel arrangement task with the roleplaying manner. The purpose of the VERBMOBIL2 speech data collectionis to record situational dialogues as close as possible to actualdialogues between businessman. In the following chapters, issues ofdata collection of the travel arrangement task, and thetransliteration method and several rules and conventions to transcribe Japanese spoken dialogue will be described.

2. Data Collection

2.1. Scenario

In the VERBMOBIL data collection, there are several assumptions forconversation. Two speakers are businessmen from the same company. As for the scenario of the Japanese data collection, two speakers isassumed to trip together and the trip is planned to visit the officein Hannover in Germany from Japan and starting point of the trip isthe same place of the recording place. Theinformation on the schedule of the trip and flight schedule and hotelname change according to the selected trip document. Speakers look at separate calendars and flight schedule and hotel information. Theschedules are conceived that the date of the dialogue will be setproperly in reference to the calendar of the speakers. The calendar of three months are given to each speaker. We assume that the place where the conversation is to be held is thesame as the starting point for the trip, i.e., the place ofrecording. The information on the flight schedule changes according to the place of recording.

The whole business trip is to take five days. The business inHannover will require one and a half days, and the remaining time isfor sight-seeing.

Although the entire recording is to be performed over several days, a schedule is to be prepared every year. The speech data is made tocover all twelve months. Therefore, the day of conversation willdiffer from the day of recording.

Three set schedule sheets of the same month are prepared. The samespeaker presupposes that a different schedule is to be used. The total number of the dialogues is assumed to be approximately more than 200.Conversations by a calendar of the same month are taken as 14conversations.

As information on the stock of a hotel, information on three hotels,a place, some equipment, and a price list of 10 hotels are used. Thecombination of hotel selections changes each time. The visit of sights and events is only broadly conversed. No specialdocuments are prepared on this business trip. This is due to reducethe complexity of conversations.

2.2. Data Collection

Data sheet for recording As the data sheet for recording, following documents are prepared for the recordings.

- For experimental persons, a name-tag on which the name is written is placed. With this the speakers can easily address themselves with names.
- Calendar sheet containing three months. The schedule of the calendar is designed in rather general terms such as meeting, holiday, seminar etc.
- For speaker A, information on the timetable of flights from Kansai or Narita airport to Hannover. The schedule of the plane is designed in rather general terms of the plane company name and flight number and departure time and arrival time. The airport of the transit is also shown.
- For speaker B, information on hotels in Hannover.
- For each experimental person, consent documents on using the speech data for the research.

Conditions for speakers As for the conditions for speakers, two speakers sit along a desk and have a face to face conversation. Although the dialog of the two may be superimposed, it is made to be as short as possible.Both speakers' native language is Japanese. Speakers are allowed to speak standard spoken Japanese, but they are not required to have had such special training in speaking. No strict rules are set on pronunciation.

Conditions for recording The recording system uses SennHeiser HMD414 microphones and records with a DAT deck. The DAT's sampling rate is 48KHz and the signal data is resampled at 16KHz by using workstations. Speech data are recorded to one channel per speaker.

Instructions for testconductor Data is collected so that the data are as close as possible torealistic dialogues of business men. The purpose of the datacollection of the dialogue is explained that the data will be used for the development of speech recognition and language translation of thespeech translation system. Corresponding to the pattern of thedialogue, the instructor explains the situation of the dialogue andthe stream of the topics of dialogue.

3. Transliteration

3.1. Transliteration of Spontaneous Speech

The standard method to transliterate spontaneous speech is defined in VERBMOBIL [2]. Japanese spontaneous speech data for VERBMOBIL istranscribed in Kanji-kana and Roman script with the segmentation intowords. By using an orthographic transliteration, the data containedin the dialogues are made accessible in written symbols to a widerange of research and usage. The user of the corpora listen to therecorded dialogues and transliterate these at the level of lexicalelements.
Transliteration conventions are used for various reasons:

- phenomena occur in spontaneous such as disruptions ofsentences, corrections and repetitions of utterances, reductions andhesitations,
- dialogues technical artifacts may occur, such as technically caused disruptions of recordings or noises resulting from themicrophone,
- situation in dialogues may lead to interference of speakersby the partner.

The outline of conventions to transliterate various phenomena ofspontaneous is described in [3].

3.2. Basic Requirements and Limitations

The basic requirements of the transliteration is given in two issues;a computational processibility and requirements regarding contents. As for the computational processibility, a unified file structure andconsistent transliteration is maintained. Transliteration conventions which can be parsed by a parser which will be a tool for filteringwith various filter options. As for the requirements regardingcontents, all audible dialogues are to be transliteration.Interference by noise or speaker are to be indicated. Certain wordcategories (names, numbers, foreign words) are to be indicated. There remains some limitations of transliteration. Audible eventswill not be described exactly. Orthography can represent dialogues in a word level and not a phonetic description of oral utterances. Forthe recording of noise and nonverbal oral sound production onlylimited categories may be used for identification.

3.3. Transliteration Conventions

The format of the transliteration file is defined. A file formatconsists of a header and turns. Each turn begins with anidentification of the turn or a turn name. The turn body contains all audible events, syntactical and semantic markers and comments. At the end of a line within a line there is a carriage return and the newline is indented by one white space. Transliteration rules of turn body for Japanese text are defined.Punctuation marks separate sentences, phrases and words. For Roman scriptsentences, a period is set after a sentence.

3.4. Transliteration of Turn Elements

Lexical elements are defined as words, interjections, regular reducedforms of words, classified words, words with articulatoryirregularities, and words with comments regarding theirpronunciation.
Words of a dictionary are words which may be listed in theirtransliterated form in the VERBMOBIL word list and can be translatedto different languages. They are also well intelligible and notdistorted.
Interjections are short sounds inserted for the purpose ofinvestigating surprise, affirmation, or doubt. Interjections a, ee, oo, uu, ara, ma, eeto, etto, etc., are handled as general words. Like e and ee,long duration interjections determined to be short durationinterjections are regarded as "other" interjections.
Compound words are multiple words in a series. In Kanji-kanasentences, compound words are simply "proper nouns" + "commonnouns" and are connected with a hyphen. Before each compound word, a marker (˜) is added.
There are several classified lexical elements; names, numbers and foreign words. They are attached with a specific symbols. Proper nouns have amarker b̃efore them. For numbers, marker # is used before them. (Days are included.)Evenwhen a number is used as an adverb, marker # is used. For compoundwords that include numbers, marker # is added in front. For foreign loan words, marker <*ENG> is used in front of them. Foreign words are non-Japanese words which do not exist in theJapanese dictionary.
Lexical elements with irregularities in pronunciation includeslengthening and poor comprehensibility of words and interruption oflexical elements. They are marked with symbols.

Interruptions of lexical elements Articulatory abortion or termination is indicated with a mark to alocation within that word. For the restating of words afterthey had been completely uttered markers are placed so that there isno unnaturalness even if the enclosed part is omitted.
If a pause is inserted between words and then the remaining pronouncementcontinues, the marker is used.
When the pronunciation is colloquial or is not correct in the case ofspoken language, the correct expression is added.Colloquial pronunciation is transcribed as it is heard in bracket andthe normal expression is shown in front of the colloquialpronunciation. The dictionary of terms has a base of Japanesedictionary such as 'DAIJIRIN" dictionary[4]. When a word is not found in the dictionary, it isregarded as an "incorrect expression".

3.5. Syntactic Semantic Structure

Syntactic semantic structures are markers for the structuring of thesentence flow. The transliterations tries to mark regular sentencesby a subset of punctuation marks. Non grammatical phenomena likecorrection of sentence abortions are marked so that they may beremoved by specific text filters and correct grammatical structures in a syntactic and semantic sense are left.

Nonverbal articulatory productions and sounds For noise not expressed with words uttered by a person, an appropriate mark is placed (depending on the type). A "< >" mark is used fornoise heard but with completely no meaning (not words).
A respiration is a sound made when breathing. The symbol "<B>" is used.
For completely difficult utterances to hear, a marker is added. Poor comprehensibility of words, symbol "<%>" is used.
Filled pause or hesitations are when long duration sounds are made midway through saying a word, stopping, or inserting a vowel. Hesitation i.e. human noises and articulating human noises i.e. human non-phonemic noises are enclosed by angle brackets. Typical examples of filled pause are as follows; <uh>, <ah>, <uhm>, <ahm>,<e>, <eeto>, <ma>, <ano>. All other articulation which may not be classified among the other listed above is indicated as <hes>.
Nonverbal Articulatory Sounds
For noise not expressed with words uttered by a person, an appropriate mark is placed depending on the type. Predetermined limited symbolsare used for nonverbal articulatory sounds. Those symbols are: <Smack>, <Throat>, <Cough>, <Laugh>,<Swallow>, and <Noise>.

{subsubsectionNoise and technical artifact Nonverbal Articulatory Sounds
For machine noise from recording equipment or peripherals, severallimited numbers of these sounds are marked. They are <#Click>,<#Ring>, <#Knock>, <#Mtouch>(A microphone that is touched),<#Mwind>(Blowing on a microphone), <#Rustle>,<#Squeak>, and <#>(Others). When a word is cut into parts by a noise expression, XX_<marking>_XX is used.

Acoustic interference Interference By Dialog Partners
Overlapping utterances with two speakers are written down with numbers attached in the order of the utterances. At the points of overlapping utterances, application is done with the utterances of the two. Markers are added to the descriptions of both speakers. The markers are n@ and @n. n is the serial number used for overlapping voices; it starts from 1. For passive interference, n@ is put at the starting position and n@ is put at the end position; corresponding to the active interference. For active interference, @n is put at the starting position and @n is put at the end position. It is possible for many small utterances to be made between multiple sentences over the topics of conversation.

Interference by noise There may be one or more lexical elements in an utterance, so when noise covers all of those parts, markers are added. Comprising the interruption noise are the noise voiced by the speakers, thebackground noise, and the surrounding noise. The symbol, "< >" denotes interruption noise. In a part of a language, when noise covers the whole body, the marker "<: :>" is used.

4. Romanized Transcription and Segmentation

4.1. Romanization

The collected dialogues are transcribed in kanji-kana sentences and Roman script sentences. In Roman transcription, there are two types of transliteration. One is segmented into words and the other is without segmentation.

1. Orthographic representation
   In Kanji-kana sentence, the romanization depending on Japanese orthography [5] is used. Three kinds of particles, i.e. "ha", "wo", "he", which are written in the orthogonal way in the kanji-kana text are transcribed in romanization in the way to consistent to the pronunciation as "wa", "o", "e".
2. Long Vowels
   In the Japanese colloquial speech, some words are pronounced as if they are long vowels. Long vowels are transcribed by duplicating vowels. For example, "iu" as "yuu", etc. A long vowel is used in the interjections and the foreign word. It is indicated by duplicated vowel . (Example: aa, @sukejuuru)
3. Double Consonant
   The double consonant is transcribed by duplicating consonant in Romanization. (Example: jikkeN, chotto).

4.2. Segmentation

Word separation is undertaken based on the morphological unit of Japanese sentence. The separation of word unit is based on Japanese morphological program CHASEN [6]. In order to apply the segmented words for speech recognition and language analysis, some modification is undertaken to form compound words according to the rules on common word in the dictionary , compound words, and collapsed forms.

- Common word
  Common word which is written in the dictionary is treated as one word.
- Compound word
  If the compound word is used common as one word in the dictionary, itwill be treated as a word. If the suffix or prefix to a common word is listed in the DAIJIRIN dictionary, it will be treated as a word.
- Collapsed form
  Collapsed form of inflated verb and particle connection is treated as compound word of suffix.
- Particles
  Almost all particles are segmented into words. When a particle such as "de" or "te" is connected to verb or adjective or auxiliary verb in a "Renyou" inflection, the connected words is treated as a compound word. (Example: susuN-de, itashi-te).
- Auxiliary verb
  An auxiliary verb has an inflection forms. The following auxiliary verbs in the list will be listed as suffixes and treated as a compound form. "masu" , "nai" , "rareru" , "reru" , "saseru" , "seru" , "ta" , "tai" , "you" (Example: ari-mase-N, de-rareru)When an auxiliary verb such as "ta" or the inflated form of the auxiliary verb is connected to adjective in a "Mizen" inflection, the connected words is treated as a compound word. (Example:atsukat-ta). When a concatenation of inflated auxiliary verb of "Renyo" form such as "mashi", and an inflated auxiliary verb such as such as "ta", the connected words is treated as a compound word. (Example: ari-mase-N).

4.3. Example of transliteration

Example of transliteration is shown in Figure 1.

Figure 1: Example of transliteration

Kanji-Kana transcrip-
tion j002ach1_007_BAB_150000: <> #-
płA <P> ZB Roman transcrip-
tion with phrase segmenta-
tion j002ach1_007_BAB_150000: <ah> #ku-
gatsuno kouhaNwa potsupotsuto youjiga hait-
teirunode chotto <P> isogashiiNdesukere-
domo . Roman transcription with word segmen-
tationj002ach1_007_BAB_150000: <ah> #ku-
gatsu no kouhaN wa pot-
supotsu to youji ga hai-tte iru node chotto <P> iso-
gashiiN desu keredomo . Parsed text by trpparserj002ach1_007_BAB_150000: ku-
gatsu no kouhaN wa potsupotsu to youji ga hai-
tte iru node chotto isogashiiN desu keredomo

5. Conclusion

We have described the method to construct and transcribe Japanese spontaneous speech data under VERBMOBIL project. The design method of spontaneous corpora in the travel arrangement task in a role playing situation has been explained in detail. Standard way of transliteration method and several rules and conventions to transcribe Japanese spoken dialogue have also described. Up to now, most of the spontaneous speech corpora have been constructed in the specific task domain in the feasible scale. When it is required to design the spoken language system in the different task domain, the problem of data collection arises and it takes much time to collect the data. The future issues will be how to make the spontaneous data corpora universal in the sense of dialog expressions and coverage of the usual usage of spoken language. It is also required to collect the unified data for the research of spoken language processing technology. There are many cases to be identified in the spoken dialogue in the pronunciation of variants association with the same spelling but different meaning. The tagging of other information than the word category, prosodic information and dialog tagging will also necessary to be used for the dialogue research.

6. Acknowledgement

The authors gratefully acknowledge the opportunity of German Ministry for Scientific Research and ATR International. I wish to thank Prof. Waibel@and colleagues of VERBMOBIL Japanese database and language processing group for their helpful suggestions and discussions on data collection of Japanese spontaneous dialogues.

References

1. D. Gibbon, et. al. ed. ”Spoken Language Systemand Corpus Design”, Mouton de Gruyter, (1998)
2. S. Burger, ”Transliteration of spontaneousspeech data: Manual of transliteration conventions VERBMOBIL 2English version Release 2”, (August 1997)
3. A. Kurematsu, ”Roman Transcription Of Japanese Soken Dialog In VERBMOBIL Project”, Proc. of First Int. Workshop on East-AsianLanguage resources and Evaluation,@pp.83-86, (May. 1998)
4. A. Matsumura, ”Daijirin Dictionary”, SanseidoPublishing, (1985)
5. K. Masuda, ”New Japanese-English Dictionary”, Kenkyuusha, (1988)
6. Y. Matsumoto, ”Japanese Morphological Analysis System: CHASEN”, Information Science Technical Repord NAIST-IS-TR97007, Nara Institute of Science and Technology, (1997)