VERBMOBIL Dialogues: Multifaced Analysis

Akira Kurematsu[(1)], Youichi Akegami [(2)], Susanne Burger[(3)], Susanne Jekat [(4)],
Brigitte Lause[(4)] , Victoria L Maclaren[(3)], Daniela Oppermann [(5)], Tanja Schultz[(6)]
(1) Univ. of Electro-Communications, JAPAN, kure@apple.ee.uec.ac.jp
(2) ATR International, JAPAN, akegami@ctr.atr.co.jp
(3) Carnegie Mellon Univ. USA, {sburger,vlmst12+}@cs.cmu.edu
(4) Univ. of Hamburg, GERMANY, {jekat,etti}@nats.informatik.uni-hamburg.de
(5) Uni. of Muenchen, GERMANY, daniela@phonetik.uni-muenchen.de
(6) Uni. of Kahlsruhe, GERMANY, tanja@ira.uka.de

ABSTRACT

This paper describes the outline of collecting and transcribing spontaneous spoken dialogues for VERBMOBIL, the German research project on multilingual processing of spontaneous speech. The method and conditions of data collection performed using the same scenario and the transliteration convention of spontaneous speech were described. The characteristics of VERBMOBIL corpus were presented in terms of the size of dialogues, turns, sentences, words, perplexity based on the linguistic analysis.

1 INTRODUCTION

The corpora of person-to-person dialogues are especially needed for a speech translation system, which is human-to-human communication via a machine conducting language translation. Spontaneous dialogue speech corpora are essentially important to model relevant features of spontaneous speech, such as pauses, hesitations, turn-taking behaviors, etc and dialogue structures.

The German research project, VERBMOBIL(VM) aims to translate spontaneously-spoken dialogues robustly and bi-directionally for German/English and German/Japanese. The project was performed in two phases, VM-1 and VM2. The spontaneous spoken dialogues were collected under the constraints as performed in the scenarios.

To achieve a relatively broad coverage of cooperative colloquial speech and various accents, situational dialogues that are as close as possible to actual dialogues were recorded with a role-playing manner. The unified file structure and consistent transliteration were fulfilled to make the computational processing easy. The standard convention to transliterate spontaneous speech was defined and applied to describe the VM corpus.

In this paper, we describe the scenario and conditions of data collection in section 2 and then, outline of a convention of transliteration is presented in section 3. Next, the characteristics of VERBMOBIL corpus are described in section 4.

2 DIALOGUE SPEECH DATA COLLECTION

2.1 Monolingual dialogue speech data collection

Monolingual dialogue data collection was conducted in three languages of German, English and Japanese. In VM-1, dialogues were collected in a cooperative fashion using “Push-to-talk” equipment. The domain was a scheduling appointment. In VM-2, all dialogues were collected with face-to-face spontaneously spoken conversation. The common topic of dialogues was a travel arrangement and the scenario was prepared in order that the dialogues do not diverse too broadly. Speakers were instructed to utter with the formal manners in business life. For each speaker, the purpose of the trip was instructed before the start of recording.

(1) Scenario

In VM-1, the scenario was the negotiation of two persons concerning the schedule of appointment. In VM-2, the main scenario was the arrangement of business partners concerning a business trip to Hannover in Germany (Scenario A). The task was that of finding a date suitable to both, as well as using the best means of transportation, and discussing the modalities of booking a hotel. A further topic could be the visit of sights and cultural events. The calendar of three months for each speaker was used to arrange the dates of the trip together.

The information on the train or flight schedule changes according to the place of recording. The topic of the dialogue included greetings, schedule arrangement, train or flight schedule, hotel, and sight seeing. The following documents were prepared for the recordings: calendar sheets, train or flight timetable, hotel information. The visit of sights and events was only broadly discussed. There were a few modifications for different languages according to the different locations of recordings.

In German collection, the number of days for trip was one and half day and the city of the origination of the trip was from Munchen or Hannover. The transportation was by train or by plane. In German data collection, the other scenario of travel agency of was added (scenario B). The task was two people had to arrange a business trip to Hannover in a travel agency. In English collection, the total number of days for trips from the city in USA was extended to five days and at least one and half days of that for staying in Hannover. In Japanese collection, the total number of days for the trip from the city in Japan was extended to five days in the same way as in English.

(2) Recording conditions

As for the conditions of data collection, two speakers sat along a desk and had face to face conversations. The utterances of the two speakers were asked to speak freely where the interferences by dialogue partner were allowed. Speakers' languages were their own native languages. No strict rules were set on pronunciation. In order to keep control over the speech material, the speaker had to talk about was determined beforehand by explaining the scenario. The signals were digitized with a stereo DAT recorder with 48kHz sampling frequency per channel. Later the speech signals were low-pass filtered and downsampled at 16 kHz. The starting end ending time of each turn were identified and recorded.

Most German monolingual dialogues were recorded in three different channels (close microphones, room microphones placed on the table and telephone (standard telephones and mobile phones). All recordings were done in a quiet room, with the dialogue partners sitting at one table together.. In our investigation we extracted the data separately for each scenario. Recordings in Bonn to record in scenario of travel agency were made by close microphone or room microphone.

Recordings in English were done at CMU (Pittsburgh). Close microphones were used in all recordings. The English scenario sheets were changed for the last part in the hopes of adding more and new vocabulary.

In Japanese recordings, each speaker was allowed to appear in two dialogues at most. The speakers were recruited from the company and their occupations were officer workers. Recordings were conducted at Osaka and Tokyo. All recordings were made using close microphone. The speech data were recorded in the office room with air-conditioned and carpet and windows. Various noises from outdoor such as car noises including an ambulance car were input from the microphone.

2.2 Multi-lingual Data Collection

The multi-language dialogue speech data collection consistsed of face-to-face conversations between one native speaker of Japanese or American English and one native speaker of German. Both speakers do not know the language of the other. A human interpreter or VERBMOBIL translates into both directions. The course of the dialogue waas predictable in that both speakers had the same interest in reaching the communicative goal and in that the dialogues covered a limited domain (travel planning). The speakers reacted cooperatively, none of them had to convince the other to do such that (s)he did not want to.

(1) Scenario

To test possible VERBMOBIL applications and to list requirements of the user, different recording situations within different domains were designed. These situations were called scenarios and characterized the context in which the dialogues take place.

The basic design for the recording of interpreted VM-dialogues was a dialog with the goal to arrange an appointment for a business trip [6]. The dialogue was carried out either directly between two business partners of different native tongues or between two business partners of different native tongues with the support of a travel agent. Each of the subjects used one calendar, in which different dates were marked as occupied. Therefore, the subjects could only choose from a restricted set of dates that were convenient for both of them. Thus negotiations about a suitable date, hotel and means of transport were required. Accordingly, in the course of the dialogue several proposals were made and rejected.

(2) Recording conditions

The multilingual VM-dialogs were recorded in a speechlab by a hard disk recorder which delivers digitalized speech data with high acoustic quality and channel separation for different speakers. All speakers were recorded by close microphones but we also use room microphones to generate a variation of the acoustic quality and to take safety copies

3 TRANSLITERATION

3.1 Overview of transliteration convention

A great amount of dialogues and multiparty conversations between various speakers were recorded in the VM corpora. The standard convention to transliterate spontaneous speech was defined in VERBMOBIL group [1][2] and was revised as needed. Transliteration of spontaneous speech is needed to transcribe various situations. Typical one is phenomena that occur in spontaneous speech such as the disruption of sentences, the correction and repetition of utterances, reductions and hesitations. Other is the situation in dialogues such as interference of speakers by the partner.

By using an orthographic transliteration recorded dialogues were transliterated at the level of lexical elements. Symbolized objects of transliteration of elements contain lexical elements, syntactical structures, nonverbal articulatory productions, noises, pauses, acoustic interference, comments and special comments.

Regarding contents in the convention, there are following characteristics: (1) All audible dialogues were recorded in the transliteration. (2) Syntactic structures were marked. (3) Interference by noise or speaker were indicated. (4) Certain word categories (names, numbers, foreign words) were indicated.

Transliteration conventions which can be parsed by a parser function as a tool for filtering with various filter options.

3.2 Format of a Transliteration File

The file consists of header and transliteration. The transliteration is subdivided in turns. Turns consists of the turn identification and the turn body containing the turn elements, that is all audible events, syntactical and semantic markers and comments. Followings are formats of the turn body:

(1) As for the coding scheme, for German, English, and Romanized Japanese script sentences, the ASCII code was used, while for Kanji-kana sentences, the JIS code was used.

(2) Turn elements were separated by one white space. At the end of a turn, one additional white space and a carriage return were inserted.

(3) Words were not syllabified.

(4) Japanese spontaneous speech data was transcribed in Kanji-Kana and Roman script with segmentation into words.

3.3 Transliteration of Turn Elements

(1) Lexical elements: Lexical elements consist of words as they may be found in dictionaries of the respective language , interjections, regular reduced forms of words, compound words, classified words such as names and numbers, words with articulatory irregularities (reductions, disruptions, lengthening), and words with comments regarding their pronunciation.

As for the orthographic convention, depending on the language of concern dialogues were transliterated according to the relevant orthographic rules of the specific language. Lexical item conventions cover the following items; capitalization, foreign word, aborted articulation, letter spelling, undefinable or hard to identify pronunciation, interjection, names, numbers, neologisms, pronunciation comments, articulatory interruption of lexical items.

Dialected pronunciations, slips/erratic pronunciations and other divergences were transliterated in standard language, e.g. in German according to the Duden rules and in Japanese according to Daijirin [4] rules. A written version of the diverging pronunciation was indicated in the orthographic way.

(2) Sentence structure: Sentence structures are non time intensive markers for the structuring of sentence flow. Non grammatical phenomena like corrections or sentences were marked so that they might be removed by specific text filters and more or less grammatical structures in a syntactic and semantic sense were left. Rules of components of sentence structure were defined in the following items; punctuation, question mark, comma, period, repetition or correction, false start.

(3) Prosodic definition: Prosodic definition covers the items of empty pause, filled pause such as hesitation, lengthening, and breathing. A hesitation, or filled pause is an articulation by the speaker that may be encountered between utterances.

(4) Noise conventions: Noises and technical artifact are audible elements of the signal which were not produced by articulation by a human speaker. These noises arise partially due to the recording process (e.g. touching of a microphone) or to other things happening in the background (Knocking, ringing, etc.). Noise conventions included human noise, technical noises that were not attributable to any speaker. To keep the number of symbols as little as possible limited noise categories were transliterated, the entire rest of possible noise was subsumed under the general <#> symbol.

(5) Uncategorized items: Conventions of interference by dialog partner and interference by noise were defined. Speaker interference occur when speakers speak at the same time. An interference was referred either passive or active, depending on the situation of the utterance interruption. Interference by noise is the interruption by various noises, which may be either human noise or background noise.

3.4 Word Segmentation

In Japanese, Kanji-Kana and Romanized transliteration were not segmented into words. Words were basically segmented into morphological unit by use of Japanese morphological analyzer CHASEN [5]. Then the segmentation was modified by human checking. The definition of the item of the word was followed CHASEN segmentation therein. In Japanese verbs and adjectives take the form of the combination of the stem and the ending. In the analysis of CHASEN, the inflation of the ending of verbs and adjectives was included in one word. Words appearing in the "Daijirin" dictionary were all treated as one word. In order to apply the segmented words to speech recognition and language processing, some modifications were conducted to form compound words according to the rules.

4 .OVERVIEW ABOUT THE CORPUS

Table 1 gives the summary of characteristics of VERBMOBIL corpus in terms of numbers of dialogues, turns, sentences, words, unique words and perplexities and out of vocabulary rate. Monolingual data in VM-2 contains altogether German 584 dialogues consisting of 39,014 turns, English 127 dialogues consisting of 10566 turns, Japanese 220 dialogues consisting of 12897 turns and multilingual (English-German, Japanese-German) 166 dialogues consisting of 9480 turns. Average turns per one dialogue in travel arrangement scenario of VM-2 were 68.7 for German, 46.4 for English and 58.6 for Japanese. Not so large variances of average number of turns per dialogue among different languages was derived from the use of the common scenario. It is to be noted that the number of words and unique words were counted for the filtered text which was obtained from the transliteration parser and the all noises and disfluencies were filtered out. The perplexity and out of vocabulary rate for the filtered text were calculated by use of CMU toolkit. Perplexity is a testset perplexity using bigram as a language model. Low perplexity of the Japanese language was observed. On the one hand this is a result from a segmentation giving short vocabulary units but on the other hand that Japanese speakers acted in speaking dialogues in a more disciplined way in this data collection. It is interesting that the number of unique words is rather limited by the reason of the use of the scenario and guidance for conversations.

In Table 2, the summary of numbers of speakers and age group are shown. It was observed that aspeakers were mainly centered in the age group of 21-30.

5 CONCLUSION

In this paper we described the outline of collection of spontaneous spoken dialogues which were recorded under the same scenario in a travel arrangement in German, English and Japanese and multilingual dialogues. The method and conditions of data collection were presented. The transliteration convention was developed and applied for the implementation of VERBMOBIL corpus. The summary of characteristics of VERBMOBIL monolingual and multilingual languages was described in terms of the size of dialogues, turns, sentences, words, perplexity based on the linguistic analysis These large amount of spontaneous dialogue corpus will be used for research and development in the field of speech recognition, speech synthesis and speech translation.

Some limitations on transliteration procedure remain. Audible events are not described exactly in nature. Orthography can represent dialogues in a word level and not a phonetic description of oral utterances. In order to apply to many other languages and various spontaneous speech dialogues, improvement in transliteration procedure will be further study.

6 ACKNOWLEDGEMENT

The authors gratefully acknowledge support of the German Ministry of Education, Science, Research and Technology. We wish to thank all members of Verbmobil groups.

7 REFERENCES

[1] S. Burger, "Transliteration of spontaneous speech data Manual of transliteration conventions VERBMOBIL 2", English version Release 2, August 1997.

[2] K. Weilhammer, S. Burger, "File names, Formats and Structures in VERBMOBIL •", VERBMOBIL MEMO-131

[3] A. Kurematsu, Y. Akegami, T. Schultz, S. Burger "Data collection and transliteration of Japanese spontaneous database in the travel arrangement task domain", Oriental COCOSDA'99, pp.125-128,(1999)

[4] A. Matsumura, "Daijirin Dictionary", Sanseido Publishing, 1985.

[5] Y. Matsumoto, "Japanese Morphological Analysis System: CHASEN", Information Science Technical Report NAIST-IS-TR97007, Nara Institute of Science and Technology, 1997.

[6] S. Jekat, B.Lause, "VMII Szenario A und B: Instruktionen f"ur alle Sprachstellungen", Universit"at Hamburg, Friedrich-Alexander Universit"at Erlangen-N"urnberg, VM-Techdoc 71, August 1999.

Table 1. Figures characterizing VERBMOBIL corpus

	#Dialogue	#Turn	#Sentence	#Words	#Unique Words	Perplexity	Out of Vocabulary
German (A)	536	36,789	41,877	168,695	5,645	109.60	2.37%
German (B)	48	2,225	3,909	26,692	2,072	101.91	4.56%
English	127	10,566	16,525	118,356	2,557	38.18	0.81%
Japanese	220	12,897	21,078	165,755	4,447	30.24	1.74%
Multi-language	166 {105(E-G) 61(J-E)}	4,505(G) 2,749(E) 1,776(J)	11,080	55,989			
German-VM1	793	30,750	13,940	308,028	7,120	69.02	1.62
English-VM1	488	9,897	4,081	96,487	1,932	35.18	2.12
Japanese-VM1	150 800	2,196 11,358	3,880 20,596	24,205	1,253	22.49	0.39

Table 2 Speakers characteristics

	#Speaker	Age group -20	21-30	31-40	41-	Unknown
German (A)	173 (M:79, F: 94)	6	106	45	11	5
German (B)	60 (M:31, F:29)	0	22	4	0	0
English	60 (M:31, F:29)	5	39	12	4	0
Japanese	120 (M:58, F:62)	12	75	26	7	0
Multi-languag e	29 (M:12, F:17)	3	19	7	0	0
German-VM1	717(M:393, F:334)	0	226	131	40	396
English-VM1	257(M:183, F:74)	1	132	49	11	64
Japanese-VM 1	60(M:3, F:57) 324(M:148, F:176)	0 29	21 223	32 51	7 21	0 0