Construction of an Advanced In-Car Spoken Dialogue Corpus and its Characteristic Analysis

Itsuki Kishida[1], Yuki Irie[1], Yukiko Yamaguchi[2], Shigeki Matsubara[2, 4], Nobuo Kawaguchi[2, 4] and Yasuyoshi Inagaki[3]

[1] Graduate School of Information Science, Nagoya University,
[2] Information Technology Center, Nagoya University,
[3] Graduate School of Engineering, Nagoya University,
[4] Center for Integrated Acoustic Information Research, Nagoya University

Abstract

This paper describes an advanced spoken language corpus which has been constructed by enhancing an in-car speech database. The corpus has the following characteristic features: (1) **Advanced tag**: Not only linguistic phenomena tags but also advanced discourse tags such as sentential structures, and utterance intentions, have been provided for the transcribed texts. (2) **Large-scale**: The sentential structures and the intentions are currently provided for 45,053 phrases and 35,421 utterance units, respectively. (3) **Multi-layer**: The corpus consists of different levels of spoken language data such as speech signals, transcribed texts, sentential structures, intentional markers and dialogue structures, moreover, they are related with each other. It allows a very wide variety of analysis of spontaneous spoken dialogue to utilize the multi-layered corpus. This paper also reports the result of investigation of the corpus, especially, focusing on the relations between the syntactic style and the intentional style of spoken utterances.

1. Introduction

At the Center for Integrated Acoustic Information Research(CIAIR), Nagoya University, we have collected an in-car spoken dialogue corpus aiming at realization of a robust spoken dialogue system[2]. This corpus is the multi-modal corpus consisting of audio, videos, driving information and transcripts, and the huge scale corpus recording the dialogues between a driver and a navigator by about 800 subjects. We have analyzed the linguistic phenomena[6] and developed an example-based dialogue system [7] using the corpus. Large-scale corpora can become the important resources for promoting various researches, and will be expected to expand in application.

This paper describes the tagging of linguistic structure information and utterance intention information as an example of advancements of the corpus. By tagging these information, this corpus turned into a multi-layered corpus in which the analysis and use from various perspectives are possible. Moreover, this corpus is characterized by collecting both spoken dialogues with a human navigator and spoken dialogues with a WOZ system. In this paper, we have compared these different kinds of dialogues using the nature of multi-layer with the corpus.

2. In-car spoken dialogue corpus

The recording of the in-car spoken dialogue aims at collecting the data for the analysis, investigation and use[4]. In this recording, we set up three kinds of navigators described below as the dialogue partner in order to investigate the influence of the dialogue by the difference in a dialogue partner.

Figure 1: *The recording of the in-car spoken dialogue.*

- Human: He/she gets a workout as a navigator in advance and has the detailed information for the task achievement. However, in order to avoid a dialogue divergence, some restriction is put on the way he/she talks.
- WOZ system : It is a spoken dialogue system which has a touch-panel input by man, and speech synthesizer by the machine[7].
- ASR system : It performs a system-initiative dialogue, and the dialogue domain is the restaurant retrieval[6].

Fig.1 shows the recording of the in-car spoken dialogue. A subject (lower left) is located in a driver's seat and drives a car(upper left). An experimental auxiliary person (lower right) is located in a backseat, and operates WOZ(upper right) etc. using a touch panel[2]. In table1, the outline of the CIAIR in-car spoken dialogue database which was constructed for three years from 1999 to 2001, is shown, divided into a spoken dialogue with a human navigator(HUM), a spoken dialogue with a WOZ system(WOZ), and a spoken dialogue with an ASR system(SYS). In addition, in 1999 only human-human conversations were recorded, and in the last 2 years all kinds of the conversations were recorded. The total recording time is about 179 hours.

The recording of dialogue speech is simulated under the

Table 1: *The outline of in-car spoken dialogue database.*

	99HUM	00HUM	00WOZ	00SYS	01HUM	01WOZ	01SYS	Total
Rec. time(sec)	141822	94692	65746	77922	93465	93862	78169	645678
Sessions	209	294	199	288	295	294	287	1866
Speaking time(sec)	98100	69390	31672	54056	67635	47424	48877	417154
driver	44722	28085	12425	11515	26055	18127	11001	151930
operator	53328	41305	19247	42541	41580	29297	37876	265174
Utterance Unit	38760	25251	11992	24944	24178	19993	22904	168022
driver	20493	12555	6099	10567	11985	9245	10722	81666
operator	18267	12696	5893	14377	12193	10748	12182	86356
Sentence	36691	23892	10767	23088	22582	16172	21270	154462
driver	19007	11675	5628	9515	10983	8475	9722	75005
operator	17684	12217	5139	13573	11599	7697	11548	79457

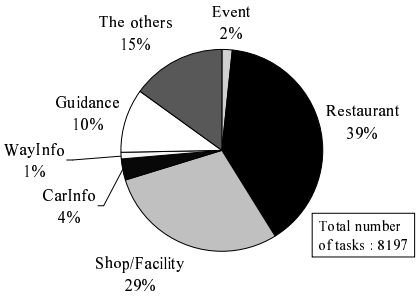

Figure 2: *The distribution of dialogue tasks.*

running car environment in which some tasks to be performed in the car such as store search and guidance are set. The distribution of the recorded dialogue task is shown in Fig. 2. Also the transcription of dialogue speech was based on the transcription criteria for the Corpus of Spontaneous Japanese(CSJ)[5]. In addition to the transcript data, time information, speaker information, and linguistic phenomenon tags, and so on, are marked. An example of the transcripts is shown in Fig. 3.

3. Advancement of the Corpus

Generally, a spoken dialogue system can be developed by the combination of the different level of components, such as speech processing, language processing, dialogue processing, and so on. In order to use the collected dialogue data for upgrading a system, not only a simple recording and transcription of speech but advanced information is needed. Then, we have advanced the dialogue corpus by giving various linguistic analysis on syntax and semantics to the text data of the corpus. Thereby, the multi-layered spoken dialogue corpus as shown in Fig. 6 could be constructed. Below, the corpus to which the dependency analysis and utterance intention analysis were given is described as the example.

3.1. Corpus with Dependency Tags

We gave the dependency analysis to the driver's utterances[8]. Dependency in Japanese is a dependency relation between the head of bunsetsu and the other bunsetsu. In addition, the bunsetsu, correponding to the basic phrase in English roughly, is a

0032 - 01:48:502-01:54:821 M:D:I:DI:
(F n) karai taiwanramen-ga tabe-tai-n-da-kedo dokka nai-kana<SB>
(Well, are there some places I can eat a spicy Taiwanese noodle?)

0033 - 01:55:975-01:58:093 M:O:I:DB:
hai niken ari-masu<SB>
(Yes, there're two.)

Figure 3: *The example of a transcribed text.*

minimum unit into which a sentence can be divided naturally in terms of meanings and pronunciations. The dependencies might be over two utterance units which are segmented by a pause. And such a dependency as a bunsetsu depends on a forward bunsetsu is also accepted. So we adopt the data specification accommodating to spontaneous utterances. The example of a corpus with dependency tags is shown in Fig. 4. The corpus includes not only the dependency between bunsetsus but morphological information, utterance unit information, dialogue turn information, and so on. Thus it has various levels of the linguistic information. This corpus is used for acquisition of the dependency probability for stochastic dependency parsing [8].

3.2. Corpus with Layered Intention Tags

We gave layered intention tags to both the driver's utterances and operator's utterances of the restaurant retrieval dialogues for either human-human conversations or human-WOZ conversations[1]. We designed the hierarchized system of the intention tags according to the degree of abstraction about intentions. The example of a corpus with layered intention tags is shown in Fig. 5. Because tags of speaker information, time information and linguistic phenomena are also recorded, it meets that specification suitable for quantitative discourse analysis. In addition, the data is utilized as an example database to predict utterance intentions on the basis of examples[9].

4. Characteristic Analysis of the Corpus

We gave the characteristic analysis to the advanced spoken dialogue corpus. This section describes the result of the analysis about the relation between an utterance intention and utterance length, and the relation between utterance intentions and linguistic phenomena. Especially, paying attention to the driver's utterances in human-human conversations and human-

```
(TIME 01:48:502-01:54:821)
((1   ( n (Well) filler ))
->    None )
((2   ( karai(spicy) adjective ))
->    (3   ( Taiwanramen-ga (a Taiwanese noodle) noun-particle )))
((3   ( Taiwanramen-ga (a Taiwanese noodle) noun-particle ))
->    (4   ( tabe-tai-n-da-kedo (I can eat) verb-auxiary-noun-auxiary-particle )))
((4   ( tabe-tai-n-da-kedo (I can eat) verb-auxiary-noun-auxiary-particle ))
->    (7   ( nai-ka-na (are there) adjective-auxiary-auxiary   )))
((5   ( dokka (some) noun ))
->    (7   ( nai-ka-na (are there) adjective-auxiary-auxiary   )))
((6   ( o-mise (places) prefix-noun ))
->    (7   ( nai-ka-na (are there) adjective-auxiary-auxiary   )))
((7   ( nai-ka-na (are there) adjective-auxiary-auxiary ))
->    None )
```

Figure 4: *An example of the corpus with dependency tags.*

```
0032    D    Request+Search+Shop
Well, are there some places I can eat a spicy Taiwanese noodle?
0033    O    Exhibit+SearchResult+NumOfShops
Yes, there're two.
```

Figure 5: *An example of the corpus with layered intention tags.*

WOZ conversations, we compare those utterances.

4.1. Relation between Utterance Intention and Utterance Length

It can be expected that the amount of information for conveying the intention depends on the type of the intention. Then, we investigated the relation between utterance intention and utterance length. The utterances were classified using the corpus with layered intention tags, and the number of bunsetsus in an utterance was investigated by using the corpus with dependency tags.

Fig. 7 shows the average number of bunsetsus for each layered intention tag which is one of the top ten intentions in frequency of appearance. The ten intentions account for 91.4% of the total. This graph means that the utterance with the intention relevant to "request" tends to become long. In an in-car dialogue, when a driver requests something of a navigator, there is a tendency to explain why a driver gives the request, and that makes utterances become long owing to that. Moreover, regardless of a kind of the intention, the utterance in a dialogue with a human is longer than in a dialogue with a WOZ system. The driver's utterance actually consists of 3.0 bunsetsus in HUM data on average, and 3.4 bunsetsus in the WOZ data. The reason is that the driver speaks briefly in consideration of the dialogue ability of a system.

4.2. Utterance Intention and Linguistic Phenomena

The frequency and position of appearance of linguistic phenomena such as filler, hesitation has been studied using various data so far[4]. In this research, we analyze the relation between the frequency of appearances of linguistic phenomena and utterance intentions. By classifing utterances using the corpus with layered intention tags and counting up the number of the linguistic phenomenon tags. We focus on the top ten intention

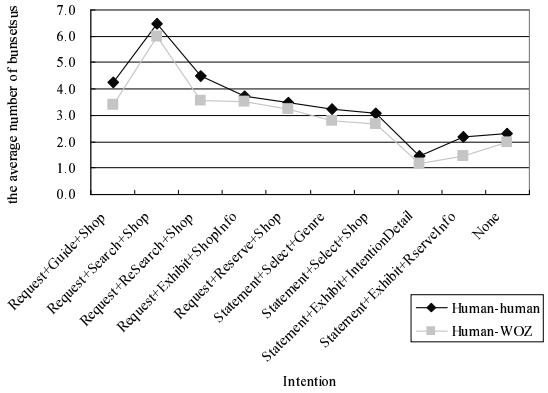

Figure 7: *The relation between utterance intention and number of average bunsetsus*

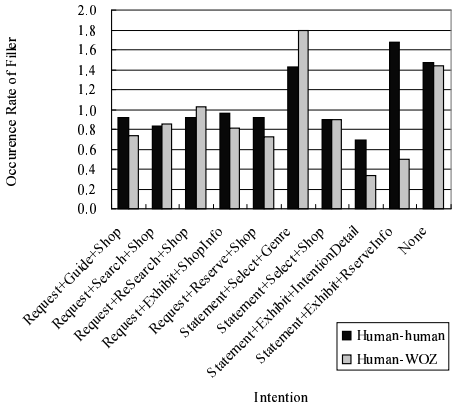

Figure 8: *The relation between utterance intention and filler frequency of appearance.*

tags in frequency of appearances as the previous section. Fig. 8 shows the result of investigation on the average frequency of appearances of fillers per bunsetsu. In addition, the value in this figure represents the ratio to the average appearances of fillers per bunsetsu in the whole corpus (0.15 in HUM dialogue and 0.12 in WOZ dialogue). In the utterance relevant to "request", each of the appearance densities indicates the average value, but in the other utterances the values are dependent on the intention. Moreover, in the utterance with the intention relevant to "exhibit", it turned out that the frequency of appearances of fillers in human-human conversations is lower than in human-WOZ conversations. The reason is that a driver might respond to it briefly, hearing a synthesized speech in a spoken dialogue with a WOZ system. By the way, it is known that the number of fillers in human-WOZ conversations is generally smaller than that in human-human conversations as shown in Table 2[4]. The above phenomena might be the cause in part.

5. Conclusion

This paper has described the advancement of the in-car spoken dialogue corpus which has been collected at CIAIR, Nagoya University. The multi-layered spoken dialogue corpus by tag-

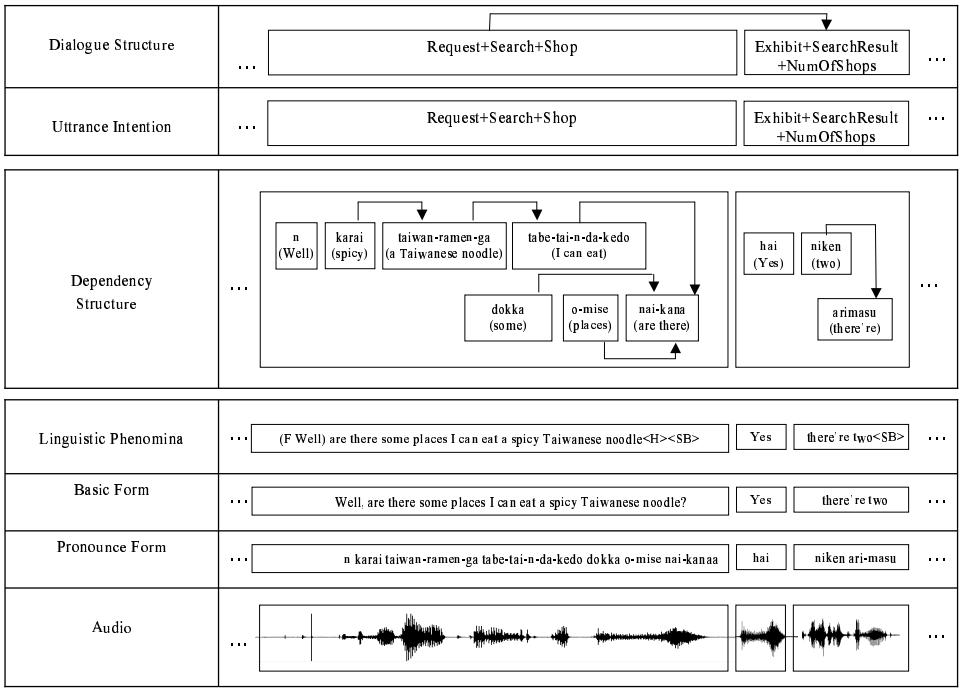

Figure 6: *The multi-layered spoken dialogue corpus.*

Table 2: *Appearance ratio of filler per utterance unit.*

	HUM	WOZ
filler(Driver)	46.8%	34.6%

ging the additional data such as linguistic structure information, utterance intentional information, and so on, to speech data and transcripts allows the analysis from various viewpoints. In this paper, we have also shown the result of the analysis on the relation between utterance intention and utterance length, and the relation between utterance intention and linguistic phenomena using the corpus with dependency tags and the corpus with layered intention tags.

6. Acknowledgements

The authors would like to thank all members of CIAIR, Nagoya University for their contribution to the construction of the in-car spoken dialogue corpus. They are grateful to Mr. Hiroya Murao of Sanyo Electric Co., Ltd. and Mr. Masahiro Ohno of our colleague for valuable comments on data analysis. This work is partially supported by the Grant-in-Aid for COE research(No. 11CE2005) of the Ministry of Education, Science, Sports and Culture, Japan and Nissan Science Foundation.

7. References

[1] N. Kawaguchi, S. Matsubara, I. Kishida, Y. Irie, Y. Yamaguchi, K. Takeda and F. Itakura, "Construction and Analysis of the Multi-Layered In-Car Spoken Dialog Corpus", Proc. of DSP in Mobile and Vehicular Systems, 2003.

[2] N. Kawaguchi, S. Matsubara, K. Takeda, F. Itakura and Y. Inagaki, "Design and Characterization of In-Car Speech Corpus", IPSJ SLP-39, pp.141-146, 2001(in Japanese).

[3] N. Kawaguchi, K. Takeda, S. Matsubara, I. Yokoo, T. Ito, K. Tatara, T. Shinde and F. Itakura, "CIAIR speech corpus for real world speech recognition", Proc. of SNLP-2002 & Oriental COCOSDA Workshop 2002, pp. 288-295, 2002.

[4] N. Kawaguchi, S. Matsubara K. Takeda and F. Itakura, "Multi-Dimensional Data Acquisition for Integrated Acoustic Information Research", Proc. of LREC-2002, pp. 2043-2046, 2002.

[5] K. Maekawa, H. Koiso, S. Furui, H. Isahara, "Spontaneous Speech Corpus of Japanese", Proc. of LREC-2000, No.262, 2000.

[6] T. Isobe, S. Hayakawa, H. Murao, K. Takeda and F. Itakura, "A Study on Domain Recognition of Spoken Dialog System", Proc. of Eurospeech-2003, 2003.

[7] H. Murao, N. Kawaguchi, S. Matsubara and Y. Inagaki, "Example-based Query Generation for Spontaneous Speech", Proc. of ASRU-2001, 2001.

[8] S. Matsubara, T. Murase, N. Kawaguchi and Y. Inagaki, "Stochastic Dependency Parsing of Spontaneous Japanese Spoken Language", Proc. of COLING-2002, Vol.1, pp.640-645, 2002.

[9] S. Matsubara, S. Kimura, N. Kawaguchi, Y. Yamaguchi and Y. Inagaki "Example-based Speech Intention Understanding and its Application to In-Car Spoken Dialogue System", Proc. of COLING-2002, Vol.1, pp.633-639, 2002.