Construction of Structurally Annotated Spoken Dialogue Corpus

Shingo Kato
Graduate School of Information Science,
Dept. of Information Engineering,
Nagoya University
Furo-cho, Chikusa-ku, Nagoya
gotyan@el.itc.nagoya-u.ac.jp

Shigeki Matsubara
Yukiko Yamaguchi
Nobuo Kawaguchi
Information Technology Center,
Nagoya University
Furo-cho, Chikusa-ku, Nagoya

Abstract

This paper describes the structural annotation of a spoken dialogue corpus. By statistically dealing with the corpus, the automatic acquisition of dialogue-structural rules is achieved. The dialogue structure is expressed as a binary tree and 789 dialogues consisting of 8150 utterances in the CIAIR speech corpus are annotated. To evaluate the scalability of the corpus for creating dialogue-structural rules, a dialogue parsing experiment was conducted.

1 Introduction

With the improvement of speech processing technologies, spoken dialogue systems that appropriately respond to a user's spontaneous utterances and cooperatively execute a dialogue are desired.

It is important for cooperative spoken dialogue systems to understand the intentions of a user's utterances, the purpose of the dialogue, and its achievement state (Litman, 1990). To solve this issue, several approaches have been so far proposed. One of them is an approach in which the system expresses the knowledge of the dialogue with a frame and executes the dialogue according to that frame (Goddeau, 1996; Niimi, 2001; Oku, 2004). However, it is difficult to make a frame that totally defines the content of the dialogue. Additionally, there is a tendency for the dialogue style to be greatly affected by the frame.

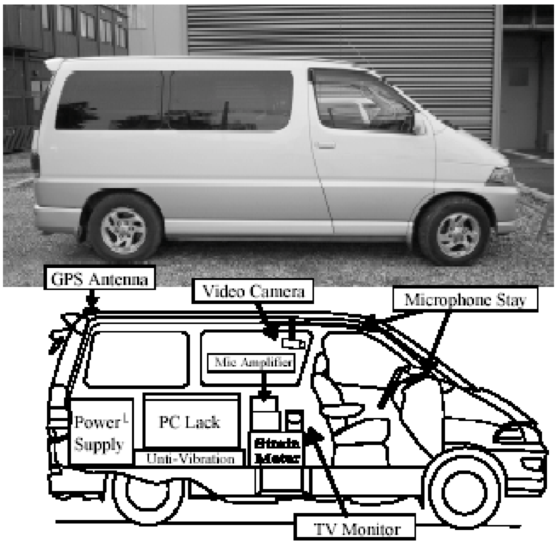


Figure 1: The data collection vehicle(DCV)

In this paper, we describe the construction of a structurally annotated spoken dialogue corpus. By statistically dealing with the corpus, we can achieve the automatic acquisition of dialogue-structural rules. We suppose that the system can figure out the state of the dialogue through the incremental building of the dialogue structure.

We use the CIAIR in-car spoken dialogue corpus (Kawaguchi, 2004; Kawaguchi, 2005), and describe the dialogue structure as a binary tree. The tree expresses the purpose of partial dialogues and the relations between utterances or partial dialogues. The speaker's intention tags were provided in the transcription of the corpus. We annotated 789 dialogues consisting of 8150 utterances. Due to the advantages of the dialogue-

```
0022 - 01:37:398-01:41:513 F:D:I:C:
(F えーっと)              [FILLER:well]  &(F エーッット)
おいしい                     [delicious]   &オイシー
おうどんの                   [Udon]        &オウドンノ
お店                         [restaurant]  &オミセ
行きたいんですが<SB> [want to go] &イキタインデスガ<SB>
0023 - 01:42:368-01:49:961 F:O:I:C:
はい                         [well]       &ハイ
この                         [this area]  &コノ
近くですと                   [near]       &チカクデスト
諏訪屋                       [SUWAYA] &スワヤ
千種豊月が                   [“CHIKUSA
                          HOUGETSU”]&チクサホーゲツガ
ございますが<SB>          [there are ] &ゴザイマスガ<SB>
```

Figure 2: Transcription of in-car dialogue speech

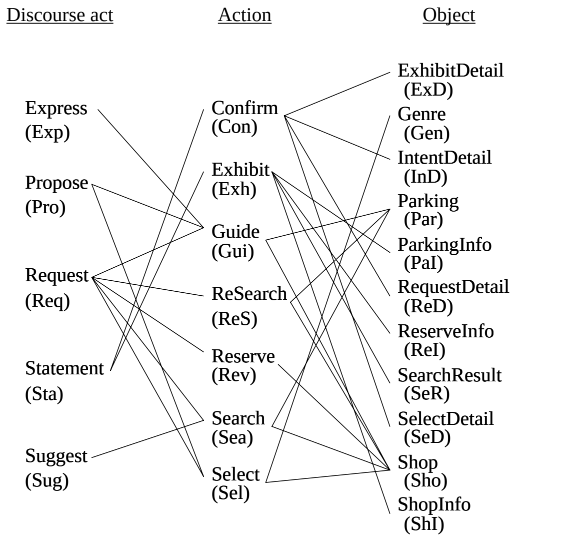


Figure 3: A part of the LIT

structural rules being represented by context free grammars, we were able to use an existing technique for natural language processing to reduce the annotation burden.

In section 2, we explain the CIAIR in-car spoken dialogue corpus and the speaker's intention tags. In sections 3 and 4, we discuss the design policy of a structurally annotated spoken dialogue corpus and the construction of the corpus. In section 5, we evaluate the corpus.

2 Spoken Dialogue Corpus with Layered Intention Tags

The Center for Integrated Acoustic Information Research (CIAIR), Nagoya University, has been compiling a database of in-car speech and dialogue since 1999, in order to achieve robust spoken dialogue systems in actual usage environments (Kawaguchi, 2004; Kawaguchi, 2005). This corpus has been recorded using more than 800 subjects. Each subject had conversations with three types of dialogue system: a human operator, the Wizard of OZ system, and the conversational system.

In this project, a system was specially built in a Data Collection Vehicle (DCV), shown in Figure 1, and was used for the synchronous recording of multi-channel audio data, multi-channel video data, and vehicle related data. All dialogue data were transcribed according to transcription standards in compliance with CSJ (Corpus of Spontaneous Japanese) (Maekawa, 2000) and were assigned discourse tags such as fillers, hesitations, and slips. An example of a transcript is shown in Figure 2. Utterances were divided into utterance units by a pause of 200 ms or more.

These dialogues are annotated by speech act tags called Layered Intention Tags (LIT) (Irie, 2004(a)), which indicate the intentions of the speaker's utterances. LIT consists of four layers: “Discourse act”, “Action”, “Object”, and “Argument”. Figure 3 shows a part of the organization of LIT. As Figure 3 shows, the lower layered intention tag depends on the upper layered one. In principle, one LIT is given to one utterance unit. 35,421 utterance units have been tagged by hand (Irie, 2004(a)).

In this research, we use parts of the restaurant guide dialogues between a driver and a human operator. An example of the dialogue corpus with LIT is shown in Table 1. In the column called *Speaker*, “D” means a driver's utterance and “O” means an operator's one. We used the Discourse act, Action, and Object layers and extended them with speaker symbols such as *“D+Request+Search+Shop”*. There are 41 types of extended LIT. Because the “Argument” layer is too detailed to express the dialogue structure, we omitted it.

Table 1: Example of the dialogue corpus with LIT

Utterance Number	Speaker	Transcription	LIT		
			First layer (Discourse Act)	Second layer (Action)	Third layer (Object)
277	D	kono hen de tai ga tabera reru tokoro nai kana. (I'd like to eat some sea bream.)	Request	Search	Shop
278	O	hai. (Let me see.)	Statement	Exhibit	IntentDetail
279	O	o ryori wa donna o ryouri ga yorosi katta desuka. (Which kind do you like?)	Request	Select	Genre
280	D	nama kei ga ii kana. (Fresh and roe.)	Statement	Select	Genre
281	D	Nabe ga tabe tai desu. (I want to have a Hotpot.)	Statement	Select	Genre
282	O	hai kono tikaku desu to tyankonabe to oden kaiseki ato syabusyabu nado ga gozai masu ga. (Well, there are restaurants near here that serve sumo wrestler's stew, Japanese hot-pot, and sliced beef boiled with vegetables.)	Statement	Exhibit	SearchResult
283	D	oden kaiseki ga ii. (I love Japanese Hotpot.)	Statement	Select	Genre
284	O	hai sou simasu to "MARU" to iu omise ni nari masu ga. ("MARU" restaurant is suitable.)	Statement	Exhibit	SearchResult
285	O	yorosi katta de syou ka. (How about this?)	Request	Exhibit	IntentDetail
286	D	yoyaku wa hituyou ari masu ka. (Should I make a reservation?)	Request	Exhibit	ShopInfo
287	O	a yoyaku no hou wa yoyoku sare naku temo o mise ni wa hairu koto ga deki masu ga. (No, a reservation is not necessary.)	Statement	Exhibit	ShopInfo
288	D	a zya soko made annai onegai si masu. (I see. Please guide me there.)	Request	Guide	Shop
289	O	kasikomari masi ta. (Sure.)	Statement	Exhibit	IntentDetail
290	O	sore dewa "MARU" made go annnai itasi masu. (Now, I'm navigating to "MARU")	Express	Guide	Shop
291	D	hai. (Thanks.)	Statement	Exhibit	IntentDetail

3 Description of Dialogue Structure

3.1 Dialogue structure

In this research, we assume that the fundamental unit of a dialogue is an utterance to which one LIT is given. To make the structural analysis of the dialogue more efficient, we express the dialogue structure as a binary tree. We defined a category called POD (Part-Of-Dialogue), according to the observations of the restaurant guide task, that was especially focused on what subject was dealt with. As a result of this, 11 types of POD were built (Table 2). Each node of a structural tree is labeled with a POD or LIT. The dialogue structural tree of Table 1 is shown in Figure 4.

3.2 The design policy of dialogue structure

To consider a dialogue as an LIT sequence, LIT providing process (Irie, 2004(b)) usually should be done. Furthermore, repairs and corrections are eliminated because they do not provide LIT. In this research, we used an LIT sequence provided in the corpus. After that, the annotation of the dialogue structure was done in the following way.

Merging utterances: When two adjoining utterances such as request and answer, they seem to be able to pair up and merge with an appropriate POD. In Table 1, for example, the utterance “Should I make a reservation?” (#286) is a request and the answer to #286 is “No, a reservation is not necessary”(#287). In this way, utterances are combined with the POD “S_INFO”.

When the LIT’s of two adjacent utterances are corresponding, these utterances are supposed to be paired and merged with the same LIT. Utterance “Fresh and roe” (#280) and “I want to have Hotpot” (#281) are related to choosing the style of restaurant and are provided with the same LIT. Therefore they are combined with the LIT “*D+Statement+Select+Genre*”.

Merging partial dialogues: When two adjoining partial dialogues (i.e. a partial tree) are composing another partial dialogue, they are merged with a proper POD. In Table 1, for example, a search dialogue (from #277 to #285, SRCH) and a shop information dialogue helping search (from #286 to #287, S_INFO) are combined and labeled as the POD “SLCT”.

When the POD’s of two adjacent partial dialogues are corresponding, these dialogues are merged with the same POD. Two search dialogues (one is from #277 to #282, other is from #283 to #285) are combined with the same POD “SRCH”.

The root of the tree: The POD of the root of the tree is “GUIDE”, because the domain of the corpus is restaurant guide task.

4 Construction of Structurally Annotated Spoken Dialogue Corpus

4.1 Work environment and procedures

We made a dialogue parser as a supportive environment for annotating dialogue structures.

Applying the dialogue-structural rules, which are obtained from annotated structural trees (like Figure 4.), the parser analyzes the inputs of the LIT sequences and the outputs off all available dialogue-structural trees. An annotator then chooses the correct tree from the outputs. When the outputs don’t include the correct tree, the annotator should rectify the wrong tree rewriting the list form of the tree. In this way, we make the annotation more efficient.

Table 2: Type and substance of POD’s

POD	Substance
GENRE	choosing style of cuisine.
GUIDE	guidance to restaurant or parking.
P_INFO	extracting parking information such as vacant space, neighborhood.
P_SRCH	searching for a parking space.
S_INFO	extracting shop information such as price, reservation, menu, area, fixed holiday.
SLCT	selecting a restaurant or parking space.
SRCH	searching for a restaurant.
SRCH_RQST	requesting a search.
RSRV	making a reservation.
RSRV_DTL	extracting reservation information such as time, number of people, etc.
RSRV_RQST	requesting a reservation.

The dialogue parser was implemented using the bottom-up chart parsing (Kay, 1980). The structural rules were extracted from all annotated dialogues. In the environment outlined above, we have worked at bootstrap building. That is, we

1. outputed the dialogue structures through the parser.

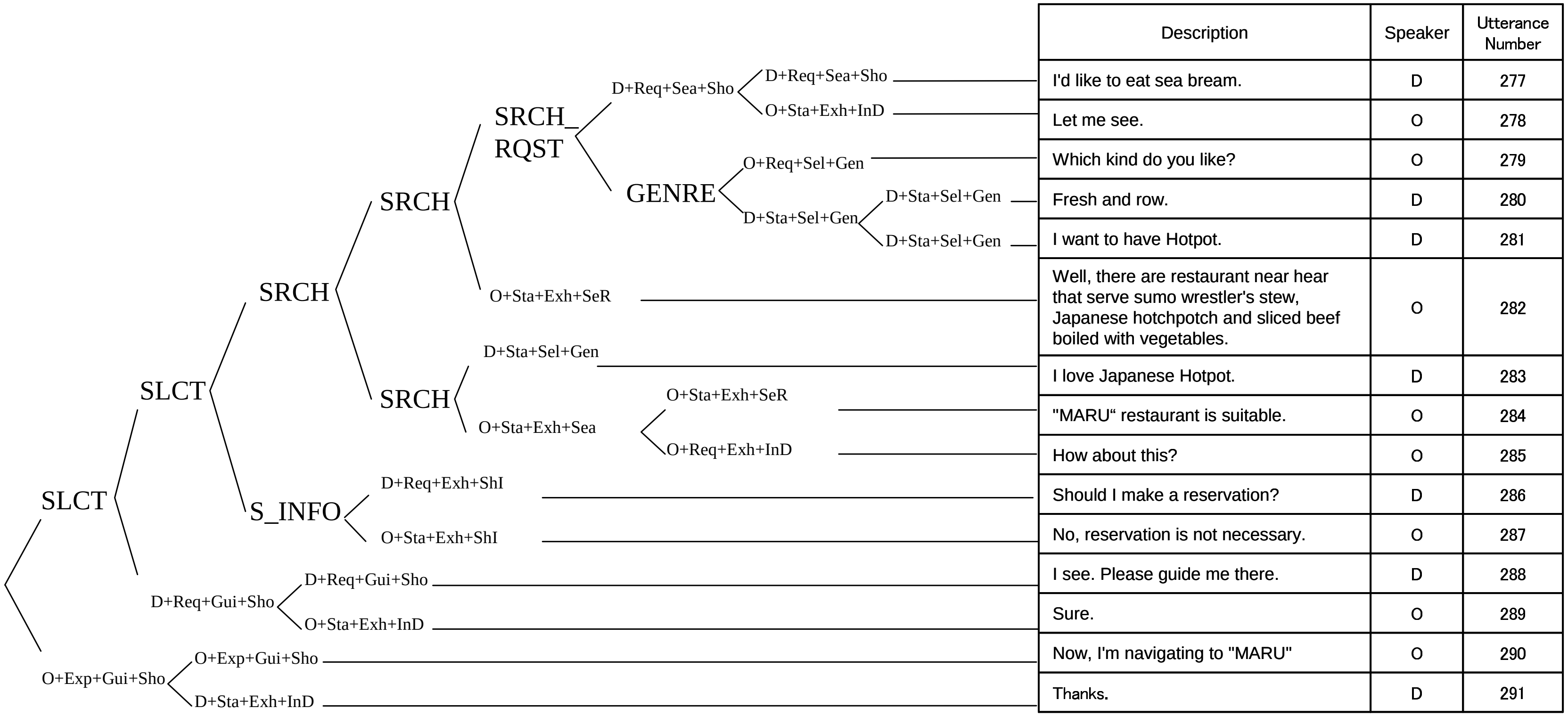
2. chose and rectified the dialogue structure using an annotator.

3. extracted some structural rules from some dialogue-structural trees.

We repeated these procedures and increased the structural rules incrementally, so that the dialogue parser improved it’s operational performance.

4.2 Structurally annotated dialogue corpus

We built a structurally annotated dialogue corpus in the environment described in Section 4.1, using the restaurant guide dialogues in the CIAIR corpus. The corpus includes 789 dialogues consisting of 8150 utterances. One dialogue is composed of 11.61 utterances. Table 3 shows them in detail.



Description	Speaker	Utterance Number
I'd like to eat sea bream.	D	277
Let me see.	O	278
Which kind do you like?	O	279
Fresh and row.	D	280
I want to have Hotpot.	D	281
Well, there are restaurant near hear that serve sumo wrestler's stew, Japanese hotchpotch and sliced beef boiled with vegetables.	O	282
I love Japanese Hotpot.	D	283
"MARU“ restaurant is suitable.	O	284
How about this?	O	285
Should I make a reservation?	D	286
No, reservation is not necessary.	O	287
I see. Please guide me there.	D	288
Sure.	O	289
Now, I'm navigating to "MARU"	O	290
Thanks.	D	291

GUIDE→SLCT O+Exp+Gui+Sho
SLCT→SLCT D+Req+Gui+Sho
SLCT→SRCH S_INFO
SRCH→SRCH SRCH
SRCH→SRCH_RQST O+Sta+Exh+SeR
SRCH_RQST→D+Req+Sea+Sho GENRE
D+Sta+Exh+SeR→O+Sta+Exh+SeR O+Re+Exh+InD

GENRE→O+Req+Sel+Gen D+Sta+Sel+Gen
S_INFO→D+Req+Exh+ShI O+Sta+Exh+ShI
D+Sta+Sel+Gen→D+Sta+Sel+Gen D+Sta+Sel+Gen
D+Req+Gui+Sho→D+Req+Gui+Sho O+Sta+Exh+InD
D+Req+Sea+Sho→D+Req+Sea+Sho O+Sta+Exh+InD
O+Exp+Gui+Sho→O+Exp+Gui+Sho D+Sta+Exh+InD

Figure 4: Dialogue-structural tree and rules for Table 1

Table 3: Corpus statistics

number of dialogues	789
number of utterances	8150
number of structural rules	297
utterances per one dialogue	11.61
number of dialogue-structural tree types	659
number of LIT sequence types	657

5 Evaluation of Structurally Annotated Dialogue Corpus

To evaluate the scalability of the corpus for creating dialogue-structural rules, a dialogue parsing experiment was conducted. In the experiment, all 789 dialogues were divided into two data sets. One of them is the test data consists of 100 dialogues and the other is the training data consists of 689 dialogues. Furthermore, the training data were divided into 10 training sets.

By increasing the training data sets, we extracted the probabilistic structural-rules from each data. We then parsed the test data using the rules and ranked their results by probability.

In the evaluation, the coverage rate, the correct rate, and the N-best correct rate were used.

$$\text{Coverage rate} = \frac{D_{parsed}}{D_{test}}$$

$$\text{Correct rate} = \frac{D_{correct}}{D_{test}}$$

$$\text{N-best correct rate} = \frac{D_{nbest}}{D_{test}}$$

D_{parsed} = Number of the dialogues which can be parsed

$D_{correct}$ = Number of the dialogues which include the correct tree in their parse trees

D_{nbest} = Number of the dialogues which include the correct tree in their n-best parse trees

D_{test} = Number of the dialogues in the test data

The results of the evaluation of the coverage rate and the correct rate are shown in Figure 5. The correct rates for each of the training sets, ranked from 1-best to 10-best, are shown in Figure 6.

In Figure 5, both the coverage rate and the correct rate improved as the training data was increased. The coverage rate of the training set consisting of 689 dialogues was 92%. This means that the rules that were from the training set enabled the parsing of a wide variety dialogues. The fact the correct rate was 86% shows that, using the rules, the correct structures can be built for a large number of dialogues.

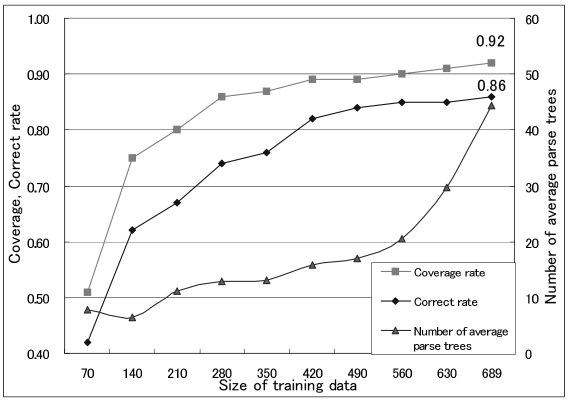


Figure 5: The relation between the size of training data and coverage and correct rate.

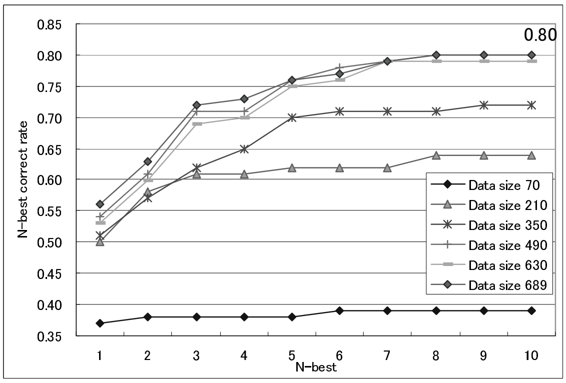


Figure 6: The relation between the size of training data and the n-best correct rate.

Three in eight failure dialogues had continued after a guidance for a restaurant. Therefore, we assume that offering guidance to a restaurant is a termination of the dialogue, in which case they couldn't be analyzed. Another three dialogues couldn't be analyzed because they included some LIT which rarely appeared in the training data. The cause of failure in the other two dialogues is that an utterance that should be combined with its adjoining utterance is abbreviated.

Figure 6 shows that the 10-best correct rate for the training set consisting of 689 dialogues was 80%. Therefore the correct rate is 86%, and approximately 93% (80/86) of the dialogues that can be correctly analyzed include the correct tree in their top-10. According to Figure 5, the number of average parse trees increased with the growth of the training data. However, most of the dialogues that can be analyzed correctly are supposed to include the correct tree in their top-10. Therefore, it is enough to refer to the top-10 in a situation where the correct one should be chosen from the set of candidates, such as in the speech prediction and the dialogue control. As a result, the high-speed processing is achieved.

6 Conclusion

In this paper, we described the construction of a structurally annotated spoken dialogue corpus. From observating the restaurant guide dialogues, we designed the policy of the dialogue structure and annotated 789 dialogues consisting of 8150 utterances. Furthermore, we have evaluated the scalability of the corpus for creating dialogue-structural rules.

We now introduce the application field of the structurally annotated dialogue corpus.

Discourse analysis: Using a POD labeled information for each partial structure of the dialogue, we can obtain information such as the structure of the domain, the user's tasks, the dialogue formats, etc.

Speech prediction and dialogue control: A system builds the structure of an input up to date and extracts the dialogue example that is most similar to the structure of the input from the corpus. If the next utterance or LIT of the extracted dialogue is the user's, the system waits for the user's utterance and predicts its meaning and intention. If the system's utterance is next, the system uses the utterance or LIT to control the dialogue.

At the present time, we have run up the data of the corpus and built probabilistic dialogue-structural trees. Next, we will apply the trees to some components of the spoken dialogue systems such as speech prediction and dialogue control.

Acknowledgments

The authors would like to thank Ms. Yuki Irie for her valuable comments about the design of the dialogue structure. This research was partially supported by the Grant-in-Aid for Scientific Research (No. 15300045) of JSPS.

References

David Goddeau, Helen Meng, Joe Poliformi, Stephanie Seneff, and Senis Busayapongchai: A form-based dialogue manager for spoken language applications, Proc. of ICSLP'96, pp.701-704, 1996.

Diane J. Litman and James F. Allen : Discourse Processing and Commonsense Plans. Phillip R. Cohen, Jerry Morgan, Martha E. Pollack, editors. Intentions in Communication. pp.365-388, MIT Press, Cambridge, MA, 1990.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara: Spontaneous speech corpus of Japanese, LREC-2000, pp.947-952, 2000.

Martin Kay: Algorithm Schemata and Data Structures in Syntactic Processing, TR CSL-80-12, Xerox PARC, 1980.

Nobuo Kawaguchi, Kazuya Takeda, and Fumitada Itakura: Multimedia corpus of in-car speech communication. J. VLSI Signal Processing, vol.36, no.2, pp.153-159, 2004.

Nobuo Kawaguchi, Shigeki Matsubara, Kazuya Takeda, and Fumitada Itakura: CIAIR In-Car Speech Corpus -Influence of Driving States-. IEICE Trans. on Information and System, E88-D(3), pp.578-582, 2005.

Tomoki Oku, Takuya Nishimoto, Masahiro Araki, and Yasuhisa Niimi: A Task-Independent Control Method for Spoken Dialogs, Systems and Computers in Japan, Vol.35, No.14, 2004.

Yasuhisa Niimi, Tomoki Oku, Takuya Nishimoto, and Masahiro Araki: A rule based approach to extraction of topic and dialog acts in a spoken dialog system, Proc. of EUROSPEECH2001, vol.3, pp.2185-2188, 2001.

Yuki Irie, Shigeki Matsubara, Nobuo Kawaguchi, Yukiko Yamaguchi, and Yasuyoshi Inagaki: Design and Evaluation of Layered Intention Tag for In-Car Speech Corpus, Proc. of the INTERNATIONAL SYMPOSIUM ON SPEECH TECHNOLOGY AND PROCESSING SYSTEMS iSTEPS-2004, pp.82-86, 2004.

Yuki Irie, Shigeki Matsubara, Nobuo Kawaguchi, Yukiko Yamaguchi, and Yasuyoshi Inagaki: Speech Intention Understanding based on Decision Tree Learning, Proceedings of 8th International Conference on Spoken Language Processing, Cheju, Korea, 2004.