# SPIRAL CONSTRUCTION OF SYNTACTICALLY ANNOTATED SPOKEN LANGUAGE CORPUS

*Tomohiro Ohno*†, *Shigeki Matsubara*‡, *Nobuo Kawaguchi*‡ *and Yasuyoshi Inagaki*§

Graduate School of Information Science, Nagoya University †
Information Technology Center/CIAIR, Nagoya University ‡
Faculty of Information Science and Technology, Aichi Prefectural University §
Furo-cho, Chikusa-ku, Nagoya, 464-8603 Japan
ohno@inagaki.nuie.nagoya-u.ac.jp

## ABSTRACT

Spontaneous speech includes a broad range of linguistic phenomena characteristic of spoken language, and therefore a statistical approach would be effective for robust parsing of spoken language. Though a large-scale syntactically annotated corpus is required for the stochastic parsing, its construction requires a lot of human resources. This paper proposes a method of efficiently constructing a spoken language corpus for which the dependency analysis is provided. This method uses an existing spoken language corpus. A stochastic dependency parse is employed to tag spoken language sentences with the dependency structures, and the results are corrected manually. The tagged corpus is constructed in a spiral fashion where in the corrected data is utilized as the statistical information for automatic parsing of other data. Taking this spiral approach reduces the parsing errors, also allowing us to reduce the correction cost. An experiment using 10,995 Japanese utterances shows the spiral approach to be effective for efficient corpus construction.

**Keywords:** Stochastic parsing, Dependency parsing, Language database, Spoken dialogue corpus

## 1. INTRODUCTION

A large-scale text corpus for which the syntactic bracketing information is provided plays an important role in natural language processing. In fact, the various languages' parse-trees data of written language such as that used in newspapers and magazines, for instance, Penn Treebank [6], NEGRA Treebank [13], TIGER Treebank [8], Prague Dependency Treebank [3], Kyoto corpus [9], EDR corpus [2], etc., have been widely utilized not only for language parsing, but also for information retrieval, automatic summarization, machine translation, and so on. In these corpus, the EDR corpus and the Kyoto corpus are syntactically annotated corpora for Japanese language, and were built by sufficiently considering various kinds of syntactic features peculiar to Japanese language. On the other hand, turning our attention to those of spoken language, despite the fact that we can enumerate the Switchboard corpus [5], Verbmobil Treebanks [1], Spoken Dutch Corpus [11], etc., very few attempts have been made for Japanese spoken language so far.

Constructing a large-scale syntactically annotated corpus of spontaneously spoken language and utilizing it as the statistical information would be effective for developing a robust spoken language parsing. Since manually providing the annotation for a Japanese text corpus calls for several difficult tasks such as morphological analysis, bunsetsu segmentation, and dependency analysis[1], it therefore requires considerable human resources.

This paper describes spiral construction of a spoken language corpus in which a dependency structure is given to each utterance. A stochastic dependency parser is utilized for automatic annotation to construct the corpus at a lower cost; that is, our approach to corpus construction is to alternately provide the dependency analyses automatically and repair it manually. The key to this approach is parsing: the parser is based on statistical information, so the more the learning data there is, the more precise the parsing. It can be expected that the data would be corrected less in the spiral construction than that in the non-spiral one.

Stochastic dependency parsing was developed for

[1] A bunsetsu is one of the linguistic units in Japanese, and roughly corresponds to a basic phrase in English. A bunsetsu consists of one independent word and more than zero ancillary words. A dependency is a modification relation between two bunsetsus.


Figure 1: The data collection vehicle (DCV)

the purpose of spiral construction [10], and our approach was evaluated using over 10,000 spoken dialogue turns in a large-scale spoken dialogue corpus. The result has shown it to be effective for efficiently constructing a syntactically annotated spoken language corpus.

The paper is organized as follows: The next section explains the in-car speech dialogue corpus. Section 3 presents syntactically annotated spoken language corpus. Our corpus construction method and our support tool are described in Section 4 and 5 respectively. The evaluation of our method is reported in Section 6.

## 2. CIAIR IN-CAR SPEECH DIALOGUE CORPUS

Human-machine speech interface in a moving car is one of the important applications of spoken language systems because the conventional models of data input/output such as video display, keyboard and mouse are not convenient to use while driving a car. The development of an in-car spoken language interface has to deal with fatal problems such as noise robustness and distortion of distant speech. Since the background noise in a moving car is not stationary and consists of a variety of sounds, a large corpus is required for training acoustic models in the presence of different background noise conditions. Another important issue is that the in-car speech communication for information

0003 – 00:09:382-00:13:652 F:D:I:I:

| | | |
|---|---|---|
| 今日 | [today] | & キョー |
| 朝 | [morning] | & アサ |
| パン | [bread] | & パン |
| 食べて | [ate] | & タベテ |
| お昼は | [lunchtime] | & オヒルワ |
| おそばを | [soba] | & オソバオ |
| 食べたんですよ<H><SB> | [ate] | & タベタンデスヨ<H><SB> |

0004 – 00:13:995-00:17:328 F:D:I:I:

| | | |
|---|---|---|
| 今晩は | [tonight] | & コンバンワ |
| 何 | [what] | & ナニ |
| 食べよっかな<SB> | [want to eat] | & タベヨッカナ<SB> |

Figure 2: Transcript of in-car dialogue speech

access by the driver has to deal with the continuously changing environment depending on the factors such as traffic condition and the distance to the destination. For a system to understand the environmental condition, it may be helpful to use audio data along with other types of data such as video images of the persons involved in dialogue, images of the road in front of the car and vehicle related data such as the angle of the steering wheel, status of the accelerator and speed of the car.

The Center for Integrated Acoustic Information Research (CIAIR), Nagoya University has been collecting large-scale in-car speech dialogues. The main objectives of this data collection are as follows: 1) training acoustic models for the in-car speech data under various driving conditions, 2) training language models of spoken dialogue for different task domains related to information access while driving a car, and 3) modeling communication by analyzing the interaction among different types of multimedia data. In an ongoing project, a system specially built in a Data Collection Vehicle (DCV), which is shown in Figure 1, has been used for synchronous recording of multi-channel audio data, multichannel video data and the vehicle related data. About 400 GB of data has been collected by recording three sessions of spoken dialogue in about a 60-minute drive by each of 200 drivers.

A spontaneously spoken language corpus has been constructed by transcribing the collected speech data into ASCII text files by hand in accordance with the rule of the corpus of spoken Japanese (CSJ) [4]. The corpus is composed of in-car dialogues between drivers and navigators about shop retrieval, driving directions, and so on. An example of a transcript is shown in Figure 2. For advanced analysis, discourse tags are assigned to fillers, hesitations, slips, and so on. Furthermore, each speech is segmented into utterance units by a pause, and their exact start and end times are provided. Other relevant environmental information regarding speaker's

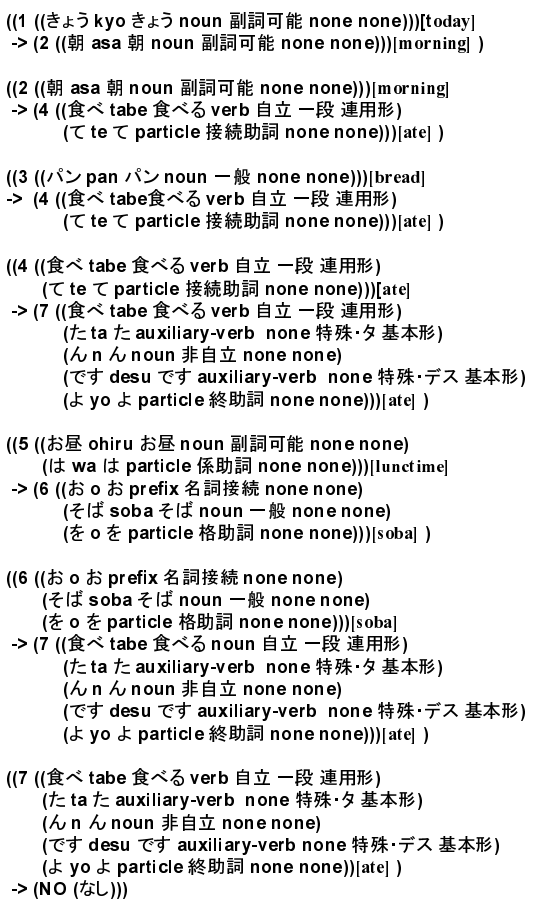


Figure 3: Spoken Japanese sentence annotated by dependency structure

sex (male/female), role (driver/navigator), dialogue task (navigation/information retrieval/...), and noise conditions (noisy/clear) is provided for each utterance unit.

## 3. SYNTACTICALLY ANNOTATED SPOKEN LANGUAGE CORPUS

Our syntactically annotated spoken language corpus has been constructed by providing the following information for each of the driver's utterances in CIAIR In-car Speech Dialogue Corpus.

- Morphological information
  - Boundaries between words
  - Pronunciation, basic form, part-of-speech, conjugation type, conjugated form of each word
- Syntactic information
  - Boundaries between bunsetsus
  - Dependencies between bunsetsus

Here, the specification of the parts-of-speech is in accordance with that of IPA parts-of-speech in a morphological analyzer called ChaSen [12], the rules of the bunsetsu segmentation with those of CSJ [4], and the dependency grammar with that of the Kyoto Corpus [9]. We have provided the following criteria for the linguistic phenomena peculiar to spoken language:

- There is no bunsetsu on which fillers and hesitations depend. They form dependency structures independently.
- A bunsetsu whose head bunsetsu is omitted does not depend on any bunsetsu.
- The specification for parts-of-speech has been provided for phrases peculiar to spoken language by adding lexical entries to the dictionary.
- We define one conversational turn as a unit of dependency parsing. The dependencies might be over two utterance units, but rarely over two conversational turns.

Figure 3 shows an example of spoken Japanese sentences annotated by dependency structure. It illustrates a sequence of dependency relations, each of which consists of a dependency bunsetsu and a head bunsetsu. Each bunsetsu is listed with it's number and its constituent morphemes.

## 4. SPIRAL CONSTRUCTION OF THE CORPUS

In this study, the dependency analyses are provided for a spoken language sentences by correcting the results of the stochastic dependency parsing. To correct them efficiently, it is desirable to execute dependency parsing with high precision. Generally speaking, that can be attained by increasing the corpus data that is used for acquiring the statistical information.

This section explains the method of spirally constructing spoken language corpus with dependency structures by splitting the corpus into several sets and carrying out the annotation incrementally. That is, the data constructed by manually correcting the parsed data is added to the annotated corpus and utilized as statistical information for the dependency parsing of another data set. Since the statistical information can

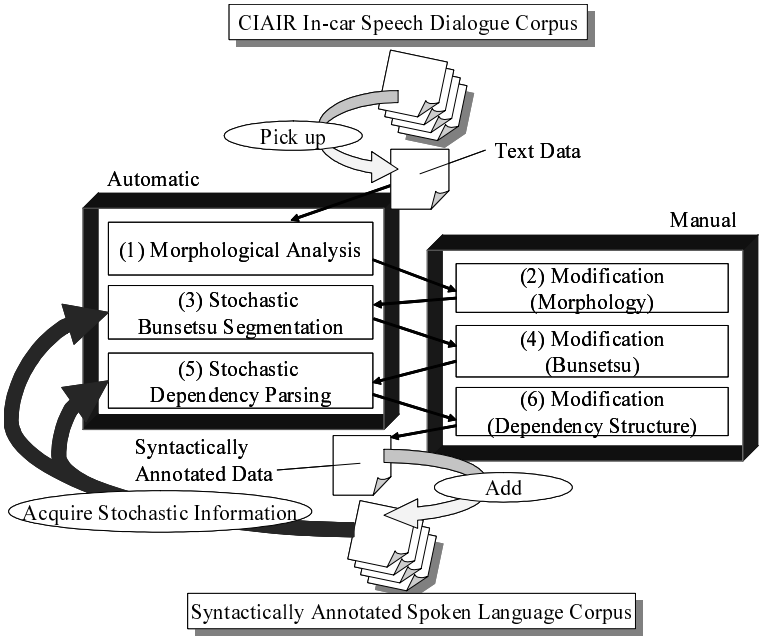


Figure 4: Flow of corpus construction

be automatically acquired from the corrected data, no human resources are necessary except to correct the parsed data. The larger the corpus' scale, the higher the precision of the stochastic dependency parsing becomes.

### 4.1. Flow of Corpus Construction

Figure 4 shows the flow of the corpus construction. The transcripts of spoken language are split into several sets, and the following tasks are executed in order for each set:

**(1)** The transcribed text is segmented into morphemes, and automatically tagged with respect to parts-of-speech by ChaSen.

**(2)** Morpheme segmentation and part-of-speech tagging are corrected manually. Any new expressions unique to spontaneously spoken language or proper nouns are added to the ChaSen dictionary.

**(3)** The corrected text is automatically segmented into bunsetsus based on stochastic bunsetsu segmentation.

**(4)** The bunsetsu segmentation is corrected manually.

**(5)** The dependency analysis is automatically provided for the bunsetsu-segmented text based on stochastic dependency parsing [10].

**(6)** The dependency analysis is corrected manually.

The result of (6) is newly added to the spoken language corpus with dependency analyses. Detailed accounts of the bunsetsu segmentation and dependency parsing are given below.

### 4.2. Stochastic Bunsetsu Segmentation

Bunsetsu segmentation is used to decide whether to insert a bunsetsu boundary between adjoining morphemes. For example, a sequence of fifteen morphemes: "kyo-asa-pan-tabe-te-ohiru-wa-o-soba-o-tabe-ta-n-desu-yo (I had some bread in this morning and soba at lunchtime.)" is segmented into the following seven bunsetsus:

kyo(today)/asa(morning)/pan(bread)/tabe␣te(ate)/
ohiru␣wa(lunchtime)/o␣soba␣o/tabe␣ta␣n␣desu␣yo(ate)

The method acquires the stochastic information on bunsetsu boundaries from a bunsetsu-segmented corpus, and utilizes it to segment a sequence from left to right. When considering the bunsetsu segmentation between adjoining morphemes $m_i$ and $m_{i+1}$, the

method uses the following attributes as stochastic information:

- The basic forms of $m_i$ and $m_{i+1}$: $h_i$, $h_{i+1}$
- The parts-of-speech of $m_i$ and $m_{i+1}$: $t_i$, $t_{i+1}$
- The conjugated forms or the detailed parts-of-speech of $m_i$ and $m_{i+1}$: $s_i$, $s_{i+1}$

The probability that the boundary between $m_i$ and $m_{i+1}$ is the bunsetsu boundary, i.e., that the morphemes are not constituents of the same bunsetsu, is calculated as follows:

$$P(m_i/m_{i+1}|m_i, m_{i+1}) \quad (1)$$
$$= \frac{C(m_i/m_{i+1}, h_i, h_{i+1}, t_i, t_{i+1}, s_i, s_{i+1})}{C(h_i, h_{i+1}, t_i, t_{i+1}, s_i, s_{i+1})}.$$

Here, $m_i/m_{i+1}$ means that there is a bunsetsu boundary between $m_i$ and $m_{i+1}$, and $C$ is a cooccurrence frequency function. If $P(m_i/m_{i+1}|m_i, m_{i+1}) \geq 0.5$, the boundary between $m_i$ and $m_{i+1}$ can be regarded as the bunsetsu boundary.

### 4.3. Stochastic Dependency Parsing

Our dependency parsing method can robustly parse grammatically ill-formed linguistic phenomena unique to spoken language, e.g. inversion and no head bunsetsu [10]. For a sequence of bunsetsus, $B$ $(= b_1 \cdots b_n)$, the method identifies the dependency structure $S$.

The conventional methods of dependency parsing for a written language have assumed the following three syntactic constraints: dependencies don't cross each other, no dependencies are directed from right to left, and each bunsetsu except the last one depends on only one bunsetsu. Considering that there are frequent inversions, fillers, hesitations and slips, we established that a dependency structure only fulfill one constraint: dependencies don't cross each other. However, we consider the other two constraints by reflecting the stochastic information.

Assuming that each dependency is independent, the $P(S|B)$ can be calculated as follows:

$$P(S|B) = \prod_{i=1}^{n} P(b_i \overset{rel}{\rightarrow} b_j|B), \quad (2)$$

where $P(b_i \overset{rel}{\rightarrow} b_j|B)$ is the probability that a bunsetsu $b_i$ depends on a bunsetsu $b_j$ when the sequence of bunsetsus $B$ is provided. The parameter $S$, which maximizes the conditional probability $P(S|B)$ is regarded as the dependency structure of $B$ and identified by DP.

Figure 5: Support tool for corpus correction

Next, we explain the calculation of $P(b_i \overset{rel}{\rightarrow} b_j|B)$. First, the basic form of independent words in a dependent bunsetsu is represented by $h_i$, its parts-of-speech $t_i$, type of dependency $r_i$, and the basic form of the independent word in a head bunsetsu $h_j$, its parts-of-speech $t_j$. Furthermore, the distance between bunsetsus is depicted as $d_{ij}$, the number of pauses between them $p_{ij}$, and the location of the dependent bunsetsu $l_i$. Here, if a dependent bunsetsu has an ancillary word, the type of dependency is the lexicon, part-of-speech and conjugated form of that ancillary word, and if not so, it is the part-of-speech and conjugated form of the last morpheme. Moreover, the dependent bunsetsu's location indicates whether it is the last one of the turn.

By using the above attributes, the conditional probability $P(b_i \overset{rel}{\rightarrow} b_j|B)$ is calculated as follows:

$$P(b_i \overset{rel}{\rightarrow} b_j|B) \quad (3)$$
$$\cong P(b_i \overset{rel}{\rightarrow} b_j|h_i, h_j, t_i, t_j, r_i, d_{ij}, p_{ij}, l_i)$$
$$= \frac{C(b_i \overset{rel}{\rightarrow} b_j, h_i, h_j, t_i, t_j, r_i, d_{ij}, p_{ij}, l_i)}{C(h_i, h_j, t_i, t_j, r_i, d_{ij}, p_{ij}, l_i)}.$$

Note that $C$ is a cooccurrence frequency function. The probability of a bunsetsu not having a head bunsetsu can also be calculated in formula (3) by considering that such a bunsetsu depends on itself ( i.e. $i = j$).

## 5. SUPPORT TOOL FOR CORPUS CORRECTION

To reduce the human resources needed for correcting the parsing errors, we created a graphical user interface. Figure 5 shows an example of the interface.

The left-hand part of the interface window is used to correct the parsed morpheme and the bunsetsu boundary. One morpheme is represented by one row, each of which contains buttons that display more information about the morpheme. Part-of-speech, detailed part-of-speech, conjugation type and conjugated form are displayed by menu buttons, and the user can correct the data by selecting the appropriate one from the menu bar. The bunsetsu boundary is modified by clicking the button in the left hand corner and changing the color of that row.

The right-hand side of the interface window is used for correcting the dependency structure. Here, the part not only indicates the bunsetsu and its head bunsetsu, but also visually displays the dependency relations. If the head bunsetsu number is "NO," it implies that the bunsetsu does not have any head bunsetsu.

## 6. EVALUATION

We have evaluated our method through an experiment on corpus construction, noting the amount of manually corrected automatic parsing results.

### 6.1. Outline of the Experiment

We performed an experiment on syntactical annotation of a large-scale spoken language corpus using 221 spoken dialogues in the CIAIR speech database [7]. The data consists of 10,995 dialogue turns, and the length of a dialogue turn is 4.1 bunsetsus on average. We used 10 turns as the basic learning data, while the remaining 10,985 turns were used as the test data. To examine the effectiveness of the spiral construction, we equally split the test data into 110 sets (approx. 100 turns each), and for practical convenience, we allocated numbers from 1 to 110 to each set.

We constructed a corpus by the following two methods:

- **Spiral Construction:** We provided the dependency structures from the data 1 to 110 in order, according to the method described in Section 4. The completed data with the dependency structures was added to the learning data for the rest of the parsing.

- **Non-Spiral Construction:** We used only the basic learning data to parse the entire test data, after which they were corrected manually.

The results were evaluated by comparing the precision of bunsetsu segmentation and dependency parsing between the above two construction methods.

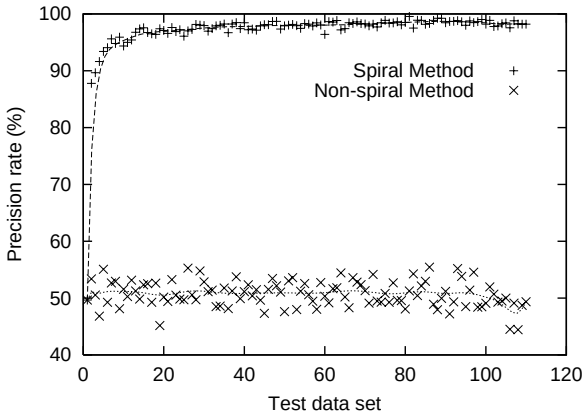


Figure 6: The result of stochastic bunsetsu segmentation

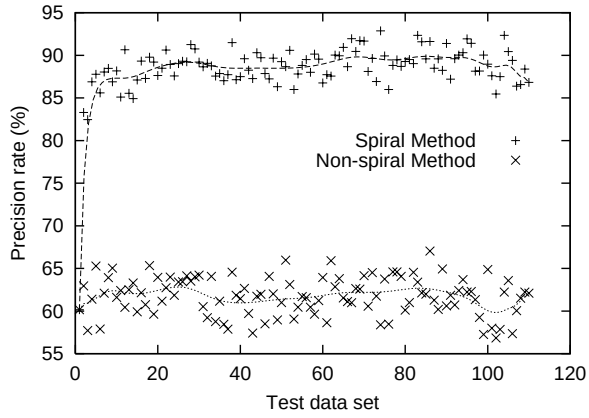


Figure 7: The result of stochastic dependency parsing

### 6.2. Experimental Results

Figure 6 shows the experimental result for bunsetsu segmentation. The segmentation conducted with the spiral method attained a level of precision about 46.4% higher than the non-spiral one. There were 96,675 boundaries between the morphemes included in the text data, among which 48,969 were correctly judged by the non-spiral method, and 93,880 by the spiral method. In other words, this means that adopting the spiral method reduced the number of corrections by 44,911 and that the corrections made by the spiral method should be 94.1% fewer than by the non-spiral one.

A comparison of the dependency parsing is shown in Figure 7. The probability of a correct judgment made with the spiral method is 88.4%, whereas that with the non-spiral method is 61.7%. As it was

with bunsetsu segmentation, the spiral method showed higher precision in ascribing dependency data. Among the 45,012 dependencies in the test data, 27,767 were correctly ascribed by the non-spiral method, whereas 39,804 were correctly ascribed by the spiral method. In other words, it means that by adopting the spiral method, the number of corrections were reduced by 12,037, a reduction of roughly 69.8% compared with the non-spiral method.

## 7. CONCLUDING REMARKS

In this paper,we have proposed a method of spirally constructing a syntactically annotated spoken language corpus based on stochastic bunsetsu segmentation and dependency parsing. The method utilizes the corpus, which is constructed by manually correcting the automatic parsing result, as statistical information to provide the dependency structure for each utterance. The evaluation using the CIAIR spoken dialogue corpus has shown our approach to be effective in reducing the human resources necessary to correct the parsing result.

We expect this approaches effect to increase as the scale of the utilized spoken language corpus grows. We plan to continue corpus construction and report a detailed analysis using a larger-scale corpus in another paper.

## 8. ACKNOWLEDGEMENT

The authors would like to thank Ms. Hitomi Toyama of the Graduate School of Languages and Cultures, Nagoya University for her helpful support in correcting the dependency corpus. This work is partially supported by the Grant-in-Aid for COE Research and for Young Scientists of the Ministry of Education, Science, Sports and Culture, Japan and The Asahi Glass Foundation.

## References

[1] E.W. Hinrichs, J. Bartels, Y. Kawata, V. Kordoni and H. Telljohann, "The VERBMOBIL Treebanks," In Zuehlke, W. and Schukat-Talamazzini, E. G. (eds.), *Proc. of KONVENS-2000 Sprachkommunikation*, ITG-Fachbericht 161, pp. 107-112, 2000.

[2] Japan Electronic Dictionary Research Institute, Ltd., "The EDR electronic dictionary technical guide (second edition)," Technical Report TR-045, Japan Electronic Dictionary Research Institute, 1995.

[3] J. Hajic, "Building a Syntactically Annotated Corpus: The Prague Dependency Treebank," Issues of Valency and Meaning, pp. 106-132, 1998.

[4] K. Maekawa, H. Koiso, S. Furui and H. Isahara, "Spontaneous Speech Corpus of Japanese," *Proc. of LREC-2000*, No.262, pp. 947-952, 2000.

[5] M. Meteer, et al., "Dysfluency Annotation Stylebook for the Switch-board Corpus. Linguistic Data Consortium," revised by Ann Taylor, 1995, ftp://ftp.cis.upenn.edu/pub/treebank/swbd/doc/DFL-book.ps.

[6] M.P. Marcus, B. Santorini and M.A. Marcinkiewicz, "Building a Large Annotated Corpus of English: The penn treebank," Computational Linguistics, Vol.19, No.2, pp. 313-330, 1993.

[7] N. Kawaguchi, S. Matsubara, K. Takeda and F. Itakura, "Multimedia Data Collection of In-Car Speech Communication," *Proc. of Eurospeech-2001*, pp. 2027-2030, 2001.

[8] S. Brants, S. Dipper, S. Hansen, W. Lezius and G. Smith, "The TIGER Treebank," *Proc. of the Workshop on Treebanks and Linguistic Theories*, pp. 24-41, 2002.

[9] S. Kurohashi and M. Nagao, "Building a Japanese Parsed Corpus while Improving the Parsing System," *Proc. of NLPRS-97*, pp. 451-456, 1997.

[10] 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.

[11] T. van der Wouden, H. Hoekstra, M. Moortgat, B. Renmans and I. Schuurman, "Syntactic Analysis in The Spoken Dutch Corpus (CGN)," *Proc. of LREC-2002*, pp. 768-773, 2002.

[12] Y. Matsumoto, A. Kitauchi, T. Yamashita and Y. Hirano, "Japanese Morphological Analysis System ChaSen version 2.0 Manual," NAIST *Technical Report*, NAIST-IS-TR99009, 1999.

[13] W. Skut, T. Brants, B. Krenn and H. Uszkoreit, " A Linguistically Interpreted Corpus of German Newspaper Text," *Proc. of LREC-98*, pp. 705-711, 1998.