A Syntactically Annotated Corpus of Japanese Spoken Monologue

Tomohiro Ohno[*] [a)], **Shigeki Matsubara**[†‡],
Hideki Kashioka[‡], **Naoto Kato**[‡], **Yasuyoshi Inagaki**[§]

[*]Graduate School of Information Science, Nagoya University
[†]Information Technology Center, Nagoya University
[‡]ATR Spoken Language Translation Research Laboratories
[§]Faculty of Information Science and Technology, Aichi Prefectural University
Furo-cho, Chikusa-ku, Nagoya, 464-8601 Japan
[a)]ohno@el.itc.nagoya-u.ac.jp

Abstract

Recently, monologue data such as lecture and commentary by professionals have been considered as valuable intellectual resources, and have been gathering attention. On the other hand, in order to use these monologue data effectively and efficiently, it is necessary for the monologue data not only just to be accumulated but also to be structured. This paper describes the construction of a Japanese spoken monologue corpus in which dependency structure is given to each utterance. Spontaneous monologue includes a lot of very long sentences composed of two or more clauses. In these sentences, there may exist the subject or the adverb common to multi-clauses, and it may be considered that the subject or adverb depend on multi-predicates. In order to give the dependency information in a real fashion, our research allows that a bunsetsu depends on multiple bunsetsus.

1. Introduction

Recently, monologue data such as lecture and commentary by professionals have been considered as valuable intellectual resources, and have been gathering attention. Actually, monologue data have been accumulated by ELRA [1], and LDC [2], etc.

On the other hand, in order to use these monologue data effectively and efficiently, it is necessary for the monologue data not only just to be accumulated but also to be structured. The various parse-trees data of written language or spoken dialogue have been constructed (e.g. (Kurohashi and Nagao, 1997; Marcus et al., 1993; Ohno et al., 2003)) and widely utilized not only for language parsing, but also for information retrieval, automatic summarization, machine translation, and so on. However, compared with these corpora, the syntactically annotated corpus of spoken monologue is not necessarily promoted enough.

This paper describes the construction of a Japanese spoken monologue corpus in which dependency structure is given to each utterance. Spontaneous monologue includes a lot of very long sentences composed of two or more clauses. In these sentences, there may exist the subject or the adverb common to multi-clauses, and it may be considered that the subject or adverb depend on multi-predicates. Even if these cases happen, the conventional dependency-annotated corpus does not allow that a *bunsetsu*[3] depends on multiple different bunsetsus (Kurohashi and Nagao, 1997; Maekawa et al., 2000). In order to give the dependency information in a real fashion, our research allows that a bunsetsu depends on multiple bunsetsus. In our research, the dependency relations of which the dependent bunsetsu depends on a nearer head bunsetsu than the conventional research are also provided. Therefore, the corpus is suitable to the incremental dependency parsing, which should decide the dependency relations simultaneously with the monologue speech input.

We have constructed a syntactically annotated corpus of spoken monologue by providing dependency information for "Asu-wo-yomu" transcription data, which is a collection of transcriptions of a TV commentary program, according to our specification. Moreover, we conducted the incremental dependency parsing experiment by using the constructed corpus, and confirmed the availability of this corpus.

The paper is organized as follows: The next section explains the "Asu-wo-yomu" transcription data. Section 3 presents our syntactically annotated corpus of spoken monologue. The application of our corpus to the incremental dependency parsing is described in Section 4. The related works of our research is reported in Section 5.

2. "Asu-wo-yomu" Corpus

Transcription data of 327 programs of "Asu-wo-yomu," which is a TV commentary program of the Japan Broadcasting Corporation (NHK), have been collected [4]. In this program, a commentator, which is different every time, speaks on some current social issue for 10 minutes. For advanced analysis, discourse tags are assigned to fillers, hesitations, slips, sentence breaks, and so on. Furthermore, each speech is segmented into utterance units by a pause, and the exact start and end times are provided.

[1]European Language Resources Association, http://www.elra.info/

[2]Linguistic Data Consortium, http://www.ldc.upenn.edu/

[3]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 in which a dependent bunsetsu depends on a head bunsetsu. That is, the dependent bunsetsu and the head bunsetsu work as modifier and modifyee, respectively.

[4]This data is used in a collaborative research of ATR and NHK.

```
ID |230-1-2| |2005/6/15| |14:18:6| 大野誠寛

1 13D            /主題ハ(topicalized element)/ LAST          (the supreme court)
最高裁判所 サイコウサイバンショ 最高裁判所 名詞-固有名詞-組織(org-proper-noun)
は ハ は 助詞-係助詞(kakari-particle)

2 13D            /テ節(compound clause-te)/ NONLAST          (today)
今日 キョウ 今日 名詞-副詞可能(adverbial-noun)

3 7D 5S          /テ節(compound clause-te)/ NONLAST          (the prosecution)
検察 ケンサツ 検察 名詞-サ変接続(sahen-conjunctive-noun)
側 ガワ 側 名詞-接尾-一般(general-suffix-noune)
が ガ が 助詞-格助詞-一般(general-kaku-particle)

4 5D             /テ節(compound clause-te)/ NONLAST          (the death penalty)
死刑 シケイ 死刑 名詞-一般(general-noun)
を ヲ を 助詞-格助詞-一般(general-kaku-particle)

5 7D             /テ節(compound clause-te)/ LAST             (demanded)
求め モトメ 求める 動詞-自立(self-sufficient-verb) 一段(ichidan type) 連用形(adverbial form)
て テ て 助詞-接続助詞(conjunctive-particle)

6 7D             /連体節(adnominal clause)/ NONLAST          (a final appeal)
上告 ジョウコク 上告 名詞-サ変接続(sahen-conjunctive-noun)
を ヲ を 助詞-格助詞-一般(general-kaku-particle)

7 8D             /連体節(adnominal clause)/ LAST             (made)
し シ する 動詞-自立(self-sufficient-verb) サ変・スル(sahen-suru type) 連用形(adverbial form)
て テ て 助詞-接続助詞(self-sufficient-verb)
おり オリ おる 動詞-非自立(ancillary-verb) 五段・ラ行(godan-ragyou type) 連用形(adverbial form)
まし マシ ます 助動詞(auxiliary verb) 特殊・マス(tokusyu-masu type) 連用形(adverbial form)
た タ た 助動詞(auxiliary verb) 特殊・タ(tokusyu-ta type) 基本形(basic form)

(The rest is omitted. )
```

Figure 1: Sample of syntactically annotated corpus of Japanese spoken monologue

3. Dependency Structure Annotation of "Asu-wo-yomu" Corpus

Our syntactically annotated corpus of spoken monologue has been constructed by providing the following information for the above-mentioned "Asu-wo-yomu" transcription data.

- Morphological information
  - Boundary between morphemes
  - Pronunciation, basic form, part-of-speech, conjugation type, conjugated form of each morpheme
- Bunsetsu information
  - Boundary between bunsetsus
- Clause information
  - Boundary between clauses
  - Type of clause
- Dependency information
  - Dependency relation between bunsetsus
  - Type of dependency relation

Here, the specification of the parts-of-speech is in accordance with that of IPA parts-of-speech in a morphological analyzer called ChaSen (Matsumoto et al., 1999), the rules of the bunsetsu segmentation with those of CSJ (Maekawa et al., 2000), the rules of the clause boundary analysis with those of Maruyama et al. (2004). The specification of dependency grammar is detailed in the following section.

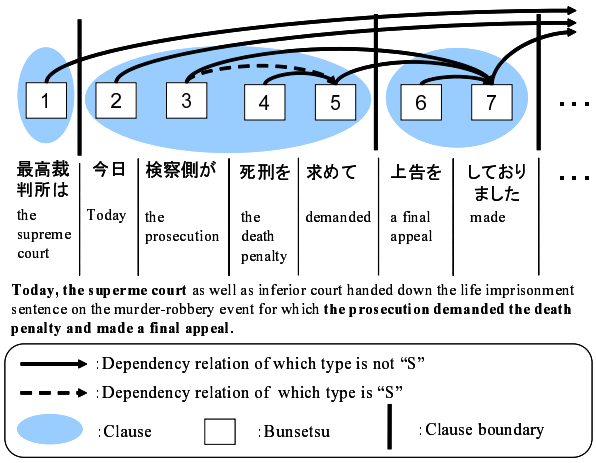

Figure 2: Dependency structure of the sample in Figure 1

3.1. The Specification of Dependency Grammar

The dependency grammar is in accordance with that of the Kyoto Text Corpus (Kurohashi and Nagao, 1997) in principle, but our grammar relaxes the conventional syntactic constraint: each bunsetsu, except the last one, depends on only one bunsetsu. This is because in Japanese there exist the following cases in which a bunsetsu can be thought to depend on two or more bunsetsus.

- The subject, adverb, or conjunction etc. in sentences consisting of subordinate clauses and a main clause may depend on two or more bunsetsus.

  e.g.) 今朝 (this morning) 東京へ (to Tokyo) 行く (going) 途中で (when) 考えた (thought)
  "I thought when going to Tokyo this morning."

  In the conventional dependency grammar of Kyoto Text Corpus, the bunsetsu "今朝 (this morning)" depends on only one bunsetsu "考えた (thought)." However, the idea that the bunsetsu "今朝 (this morning)" depends on also another bunsetsu "行く (going)" is also true.

- The subject common to compound clauses and main clause may depend on two or more bunsetsus.

  e.g.) 小泉首相が (The prime minister Koizumi) 二倍の (about twice) 差を (lead) つけて (had) 圧勝した (and won)
  "The prime minister Koizumi had about twice lead and won."

  In the conventional dependency grammar of Kyoto Text Corpus, the bunsetsu "小泉首相が (The prime minister Koizumi)" depends on only one bunsetsu "圧勝した (and won)." However, the bunsetsu "小泉首相が (The prime minister Koizumi)" depend on also another bunsetsu "つけて (had)."

Therefore, our research allows that a bunsetsu depends on two or more bunsetsus if these dependency relations are based on grammatical functions about the inflected forms or the ancillary words, and are not necessarily incorrect semantically. Meanwhile, for comparison, the dependency

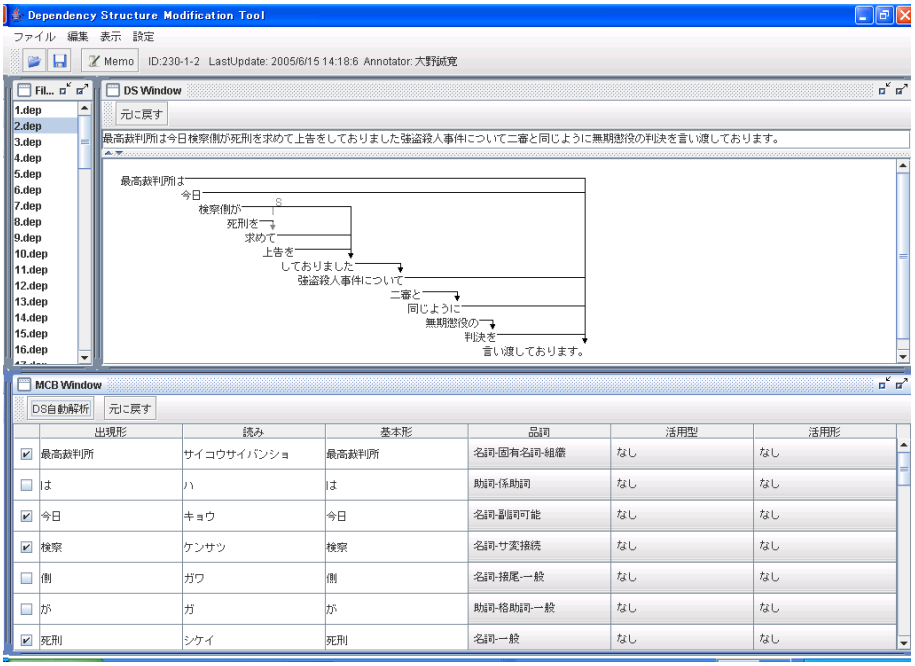

Figure 3: Support tool for corpus correction

information is provided distinguishing dependency relations based on Kyoto Text Corpus's grammar from other dependency relations which can be thought not necessarily incorrect.

Figure 1 shows an sample of the constructed dependency corpus. It illustrates a sequence of bunsetsus. Each of bunsetsu is listed with its number, its head bunsetsus, type of clause to which the bunsetsu belongs, information whether the bunsetsu is last of a clause or not, and its constituent morphemes. Here, a alphabet next to a number of a head bunsetsu indicates the type of the dependency relations. "S" means the dependency relation which is not annotated by the specification of Kyoto Text Corpus but is annotated by our specification. Figure 2 shows the dependency structure of sample of Figure 1. In this figure, the dotted line indicates the dependency relation of which type is "S."

3.2. Construction of Syntactically Annotated Corpus of Spoken Monologue

Here, we constructed the syntactically annotated corpus by modifying each information automatically analyzed by ChaSen (Matsumoto et al., 1999), CaboCha (Kudo and Matsumoto, 2002), and CBAP (Maruyama et al., 2004). As annotation tool for reducing the human resources needed for correcting the parsing errors, we created a graphical user interface as shown in Figure 3. The interface allows us to provide syntactic information that a bunsetsu depends on multiple different bunsetsus. Table 1 shows the size of our constructed corpus.

4. Application of Our Corpus to Dependency Parsing of Monologue

We have proposed a technique for stochastic dependency parsing using our proposed corpus for dependency probability calculation (Ohno et al., 2005). Because the parsing

Table 1: Size of syntactically annotated corpus of Japanese spoken monologue

programs	102
sentences	6,026
clauses	28,512
bunsetsus	71,037
morphemes	178,254

works incrementally on a clause-by-clause basis, it could be used as a basic technique of the applications such as simultaneous speech translation and real-time caption generation. The technique identifies the clauses, analyzes the dependency structures, and decides the dependency relations with another clauses, simultaneously with the monologue speech input. In this technique, we assume that a bunsetsu depends on another bunsetsu in the same clause. However, in the conventional dependency corpus, dependency relations that do not satisfy such assumption tended to exist frequently. This is because even if there exists the subject common to multi-clauses. The subject can be considered to depend on multi-predicates, the conventional dependency grammar tends to make the subject depend on only the furthest bunsetsu. On the other hand, our corpus allows a bunsetsu to depend on multiple bunsetsus, which include not only the furthest but also the nearest. This implies that our grammar can get the dependency structure which is closed in the clause more easily than the conventional grammar. Therefore, our syntactically annotated corpus can be considered to be suitable for the incremental dependency parsing.

In the following, we first outline a dependency parsing based on clause boundaries and next describe the parsing

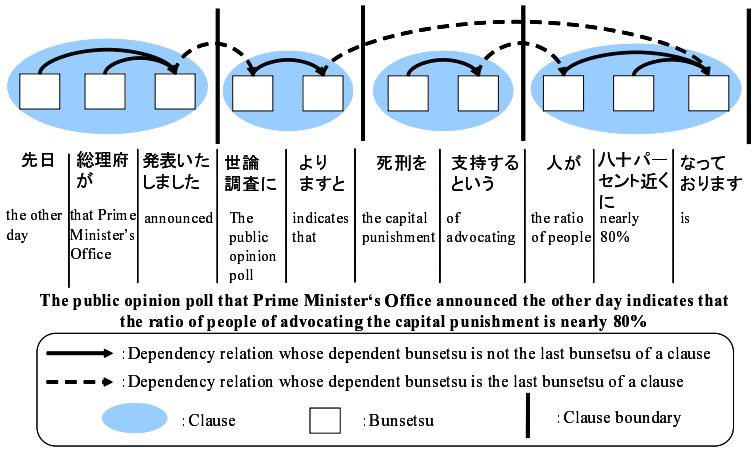

Figure 4: Relation between clause boundary and dependency structure

experiment.

4.1. Incremental Dependency Parsing Based on Clause Boundaries

In our research, we adopt a clause as a parsing unit and work out the incremental dependency parsing system which can output the dependency structure of a clause simultaneously with the monologue speech input.

In Japanese, a clause basically contains one verb phrase. Therefore, a complex sentence or a compound sentence contains one or more clauses. Moreover, since a clause constitutes a syntactically sufficient and semantically meaningful language unit, it can be used as an alternative parsing unit to a sentence. Our proposed method assumes that a monologue is a sequence of one or more clauses, and every bunsetsu in a clause, except the last bunsetsu, depends on another bunsetsu in the same clause. As an example, the dependency structure of a part of a Japanese spoken monologue:

> "先日総理府が発表いたしました世論調査によりますと死刑を支持するという人が八十パーセント近くになっております（The public opinion poll that Prime Minister's Office announced the other day indicates that the ratio of people of advocating the capital punishment is nearly 80%.)"

is presented in Figure 4. Here, although it is essentially difficult to divide a monologue into clauses on one dimension, a monologue can be approximately segmented into clauses by a clause boundary annotation program (Maruyama et al., 2004). In our research, we call the unit sandwiched between two clause boundaries detected by the clause boundary analysis ***clause boundary unit*** and adopt it as an alternative paring unit.

This method detects a clause boundary for speech input as needed and whenever the clause boundary unit is identified, it executes dependency parsing for a sequence of bunsetsus which was provided up to that point. The detection of a clause boundary is parsed by CBAP (Maruyama et al., 2004). In dependency parsing, our method constructs the dependency structure within a clause boundary unit and decides the head bunsetsus of the last bunsetsus of clause boundary units which was previously provided if possible.

In this method, the transcribed sentence for which a morphological analysis, clause boundary detection, and bunsetsu segmentation are provided is considered as an input. In addition, in the above both procedures, our method assumes the following three syntactic constraints:

1. No dependency is directed from right to left.
2. Dependencies don't cross each other.
3. Each bunsetsu, except the last one in a sentence, depends on only one bunsetsu.

These assumptions are usually used for Japanese dependency parsing.

In what follows in the section, we describe the following processing.

1. The dependency relations of a clause boundary unit inside are identified for every clause boundary unit in a monologue. **(clause-level parsing)**
2. The dependency relations of which the dependent bunsetsus is the last bunsetsus of a clause boundary units in a monologue are identified. **(monologue-level parsing)**

In those two parsing, the structure, which maximizes the likelihood which is calculated by using the dependency probability provided from the corpus, is regarded as the dependency structure and calculated by dynamic programming (DP). Refer to the literature (Ohno et al., 2005) for details of the statistical model.

In monologue-level dependency parsing, since it is not clear when their head bunsetsus are provided, the timing on which the dependency relation is decided is important. In our research, by taking into consideration that a dependency relation which crosses over a sentence boundary does not exist and that the length between dependent bunsetsu and head one is not long, we thought of deciding the head bunsetsu when the analysis advances to some degree after the last bunsetsu of a clause boundary unit is provided.

Table 2: Size of experimental data set

	test data	learning data
programs	7	95
sentences	470	5,556
clause boundary units	2,126	26,386
bunsetsus	5,054	65,983
morphemes	12,748	165,506

Concretely speaking, in our method, whenever a clause boundary unit is provided, the maximum likelihood dependency structure of that point is parsed and if a dependency relation for the last bunsetsu of a clause boundary unit does not change during a fixed input time (hereinafter referred to as a fixed value), the dependency relation is decided as having the head bunsetsu. Refer to the literature (Ohno et al., 2005) for details of the algorithm of incremental parsing.

4.2. Parsing Experiment

To evaluate the availability of our syntactically annotated corpus of Japanese spoken monologue, we conducted an experiment on the incremental dependency parsing.

4.2.1. Outline of Experiment

Table 2 shows the monologue data used in the experiment. We used 7 programs (470 sentences) as the test data and 95 programs (5,532 sentences) as the learning data. Here, in order to evaluate the appropriateness for the incremental parsing between our dependency grammar and the conventional dependency grammar, we conducted the following two experiments by changing the grammar of the dependency information, which was used as the correct data or the learning data.

- *Nearest dependency*
  In this experiment, among dependency relations provided according to our specification, only the dependency relations of which the dependent bunsetsu depended on the nearest head bunsetsu were used as dependency information.

- *Conventional dependency*
  In this experiment, among dependency relations provided according to our specification, only the dependency relations which were based on also Kyoto Text Corpus's grammar were used as dependency information.

We implemented our parsing method in GNU Common Lisp on Linux PC (Pentium4 2.4GHz, with main memory 2GB). In addition, we performed each experiment 12 times by changing the fixed value described in Section 4.1 from 1 to 12 and estimated the dependency accuracy respectively.

4.2.2. Experimental Result

Figure 5 shows the dependency accuracy when the nearest dependency was used and when the conventional dependency was used. In this figure, a fixed value and the dependency accuracy are respectively shown in a horizontal axis and vertical axis. We can confirm that, in all fixed values,

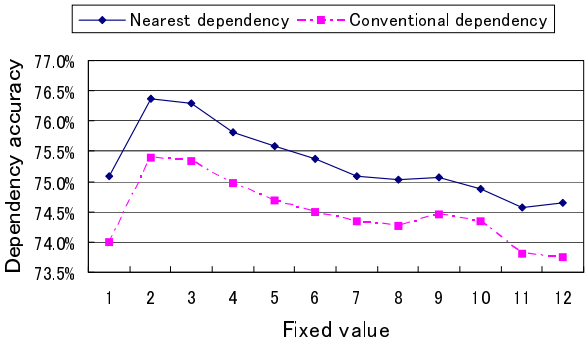

Figure 5: Dependency accuracy for each fixed vale (total)

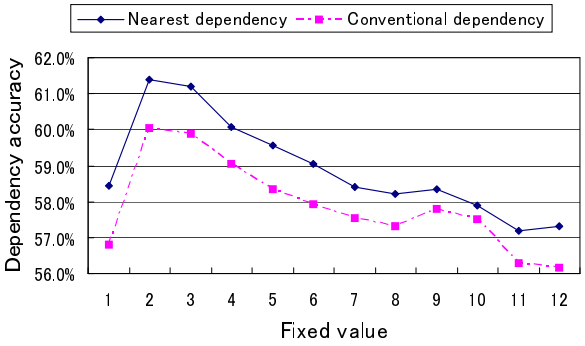

Figure 6: Dependency accuracy for each fixed vale (monologue-level parsing)

the dependency accuracy when the nearest dependency was used is higher than when the conventional dependency was used.

In the incremental dependency parsing based on clause boundaries, the parsing is performed by two stages: clause-level parsing and monologue-level parsing. Then, to analyze the above-mentioned experiment result, we evaluated the dependency accuracy in each stage respectively. In the clause-level parsing, the dependency accuracy when the nearest dependency was used exceeded about 1% in all fixed values compared with when the conventional dependency was used. This is because the nearest dependency tends to be the dependency structure closed in a clause, which fulfills the assumption described in Section 4.1, compared with the conventional dependency. Actually, the number of dependency relations which did not fulfill this assumption was 141 in case of the nearest dependency and 162 in case of the conventional dependency respectively.

Next, the dependency accuracy in the monologue-level parsing is shown in Figure 6. In the monologue-level parsing as well as the clause-level parsing, the dependency accuracy when the nearest dependency was used is the highest in all fixed values. This is because although, when the conventional dependency was used, the incorrect dependency relation was decided before inputting the correct head bunsetsu, it can be prevented when the nearest dependency was used.

5. Related Works

First, in CSJ (Maekawa et al., 2000), dependency information has been given to Japanese monologue data. The specific new criteria is provided for the linguistic phenomena peculiar to spoken language. The difference from our corpus is that we relax the above-mentioned conventional syntactic constraint.
Next, in Kyoto Text Corpus (Kurohashi and Nagao, 1997), dependency information has been provided for newspaper articles. Ellipsis information, which indicates the omitted case of a predicate and its referent, are additionally annotated for about 5,000 sentences. This ellipsis information can be thought to have a deep relation with the dependency information that a bunsetsu depends on multiple bunsetsus in our corpus.

6. Conclusions

In this paper, we have described the construction of the syntactically annotated monologue corpus. Our constructed corpus allows a bunsetsu to depend on multiple bunsetsus and is suitable to the incremental dependency parsing.
We plan to continue corpus construction and report a detailed analysis using a larger-scale corpus in another paper.

Acknowledgements The authors would like to thank graduate students of Nagoya University for their helpful support in correcting the monologue dependency corpus. The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, "A study of speech dialogue translation technology based on a large corpus."

7. References

T. Kudo and Y. Matsumoto. 2002. Japanese dependency analyisis using cascaded chunking. In *Proc. of 6th CoNLL*, pages 63–69.

S. Kurohashi and M. Nagao. 1997. Building a Japanese parsed corpus while improving the parsing system. In *Proc. of 4th NLPRS*, pages 451–456.

K. Maekawa, H. Koiso, S. Furui, and H. Isahara. 2000. Spontaneous speech corpus of Japanese. In *Proc. of 2nd LREC*, pages 947–952.

MP. Marcus, B. Santorini, and MA. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:310–330.

T. Maruyama, H. Kashioka, T. Kumano, and H. Tanaka. 2004. Development and evaluation of Japanese clause boundaries annotation program. *Journal of Natural Language Processing*, 11(3):39–68. (In Japanese).

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

T. Ohno, S. Matsubara, N. Kawaguchi, and Y. Inagaki. 2003. Spiral construction of syntactically annotated spoken language corpus. In *Proc. of IEEE NLPKE-2003*, pages 477–483.

T. Ohno, S. Matsubara, H. Kashioka, N. Kato, and Y. Inagaki. 2005. Incremental dependency parsing of Japanese spoken monologue based on clause boundaries. In *Proc. of 9th EUROSPEECH*, pages 3449–3452.