# Dependency-structure Annotation to Corpus of Spontaneous Japanese

**Kiyotaka Uchimoto*, Ryoji Hamabe†, Takehiko Maruyama‡, Katsuya Takanashi†, Tatsuya Kawahara†, and Hitoshi Isahara***

* National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
{uchimoto, isahara}@nict.go.jp
† Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan
{hamabe, takanasi, kawahara}@ar.media.kyoto-u.ac.jp
‡ The National Institute for Japanese Language
3591-2 Midori-machi, Tachikawa, Tokyo 190-8561, Japan
maruyama@kokken.go.jp

**Abstract**

In Japanese, syntactic structure of a sentence is generally represented by the relationship between phrasal units, or *bunsetsu*s in Japanese, based on a dependency grammar. In the same way, the syntactic structure of a sentence in a large, spontaneous, Japanese-speech corpus, the *Corpus of Spontaneous Japanese* (CSJ), is represented by dependency relationships between *bunsetsu*s. This paper describes the criteria and definitions of dependency relationships between *bunsetsu*s in the CSJ. The dependency structure of the CSJ is investigated, and the difference in the dependency structures of written text and spontaneous speech is discussed in terms of the dependency accuracies obtained by using a corpus-based model. It is shown that the accuracy of automatic dependency-structure analysis can be improved if characteristic phenomena of spontaneous speech — such as self-corrections, basic utterance units in spontaneous speech, and *bunsetsu*s that have no modifiee — are detected and used for dependency-structure analysis.

## 1. Introduction

The "Spontaneous Speech: Corpus and Processing Technology" project sponsored the construction of a large, spontaneous, Japanese-speech corpus, the *Corpus of Spontaneous Japanese* (CSJ) (Maekawa et al., 2000). The CSJ — the biggest spontaneous-speech corpus in the world that is open to the public — is a collection of monologues and dialogues, the majority being monologues such as academic presentations. It includes transcriptions of speeches as well as audio recordings. Approximately one tenth of the CSJ has been manually annotated with information about morphemes, sentence boundaries, syntactic structures, discourse structures, and so on. In Japanese sentences, word order is rather free, and subjects or objects are often omitted. In Japanese, therefore, syntactic structure of a sentence is generally represented by the relationship between phrasal units, or *bunsetsu*s in Japanese, based on a dependency grammar, as represented in the Kyoto University text corpus (Kurohashi and Nagao, 1997). In the same way, the syntactic structure of a sentence in the CSJ is represented by dependency relationships between *bunsetsu*s.

This paper describes the dependency structure of the CSJ and discusses the difference in the dependency structures of written text and spontaneous speech.

## 2. Dependency Structure in the CSJ

In general, a sentence is a necessary standard unit for natural language processing, syntactic analysis in linguistics, and ordinary human-language activities. In dealing with spontaneous speech, however, a sentence is not necessarily appropriate for processing or analysis because spontaneous speech basically contains no periods to mark sentence boundaries. Moreover, it is fundamentally difficult to find obvious sentence boundaries from spontaneous utterances, which usually contain utterance errors, utterance stops, and other characteristic phenomena. It is thus necessary to define and detect some reasonable segmented units for processing as a "sentence" in spontaneous speech. In the CSJ, therefore, "sentences" are defined as "clause units". The "clause units" are originally defined as basic processing units of spontaneous Japanese speech. They can be obtained by automatically detecting Japanese clause boundaries using a program called CBAP (Maruyama et al., 2004) and manually modifying them (Takanashi et al., 2003). Dependency relationships between *bunsetsu*s are annotated within a "sentence" in the CSJ.

The criteria and definitions of dependency relationships between *bunsetsu*s in the CSJ basically follow those in the Kyoto University text corpus. However, the criteria and definitions in the Kyoto University text corpus do not cover all the linguistic phenomena observed in the CSJ because there are many differences between written text and spontaneous speech. In the CSJ, therefore, we added new criteria and definitions for dependency-structure annotation to those in the Kyoto University text corpus.

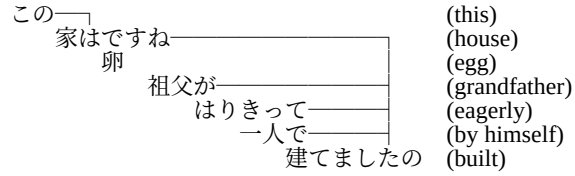
In the production of spontaneous speech, speech plans constructed beforehand are sometimes changed during the utterance because of phonological, lexical, syntactic or ordering problems. In particular, long spontaneous monologues impose heavy linearization problems on speakers, such as deciding what to say first and what to say next (Levelt, 1989). This causes various disfluencies such as utterance stops, self-correction, insertions, inversions, and distortions. For these disfluencies characteristic to spon-

taneous speech, dependency relationships are annotated in the following way.

- Utterance stop

  Utterance stops are basically detected as individual "sentences" in the CSJ, except for the case that there is a dependency relationship between *bunsetsu*s across an utterance stop. In that case, the utterance stop is defined to have no modifiee.

  **ex)** "卵 (egg)" is an utterance stop, and it has no modifiee.

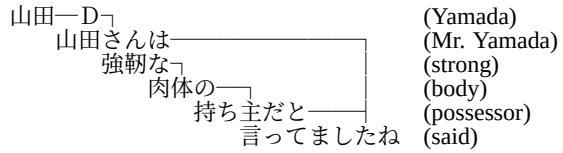
  
  

  In this example, the speaker wanted to say "This house, my grandfather eagerly built it by himself." However, the word "egg" was inserted into the utterance to form "This house is an egg, my grandfather eagerly built it by himself."

- Self-correction

  In the CSJ, self-corrections are represented as dependency relationships between *bunsetsu*s, and label D is assigned to them. We established new criteria for the annotation of the self-corrections. Although there are various types of self-corrections, all of them were labeled with D, because we focus not on classifying the self-corrections into fine-grained types but on discriminating them from ordinary dependency relationships.

  **ex)** "山田 (Yamada)" is corrected as "山田さん (Mr. Yamada)" by the speaker.

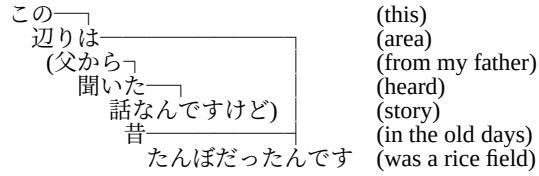
  
  

  This example can be translated as "Yamada, Mr. Yamada said that he had a strong body."

- Inserted clauses

  In spontaneous speech, it can be observed that speakers insert clauses in the middle of other clauses. This occurs when speakers change their speech plans while producing utterances, which results in supplements, annotations, or paraphrases of main clauses. In the CSJ, inserted clauses are manually detected and bracketed with (. . .). Dependency relationships within an inserted clause are closed. And the boundaries of the inserted clause are detected in the process of detecting sentence boundaries.

  **ex)** "父から聞いた話なんですけど (which is a story that I heard from my father)" is an inserted clause.

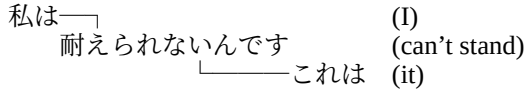
  
  

  This example means "This area was a rice field in the old days, which is a story that I heard from my father."

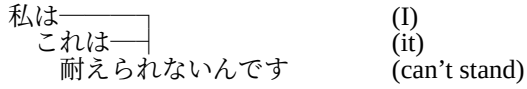
- Inversion

  In the CSJ, inversions are represented as dependency relationships going from right to left.

  **ex)** "これは (it)" is an inversion.

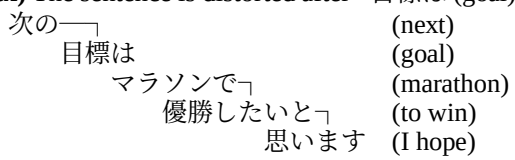
  The canonical word order in Japanese is as follows.

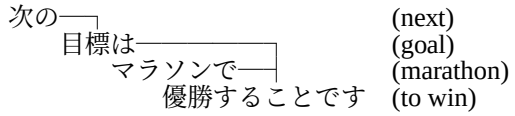
- Distortion

  Distortions are basically defined to have no modifiee because the change of the speech plans causes a distortion, and the distorted sentence has an unnatural syntactic structure. Distorted sentences are often divided into different sentences when topicalized expressions are included in the distorted sentence.

  **ex)** The sentence is distorted after "目標は (goal)".

  This utterance means "My next goal is, I hope, to win a marathon." The speaker should have said as follows to keep a natural syntactic structure.

  This is a normal sentence saying "My next goal is to win a marathon."

Self-corrections differ from dependency relationships as well as from coordination and appositives. However, they are represented as dependency relationships between *bunsetsu*s, and labels D, P, and A are assigned to self-corrections, coordination, and appositives, respectively.

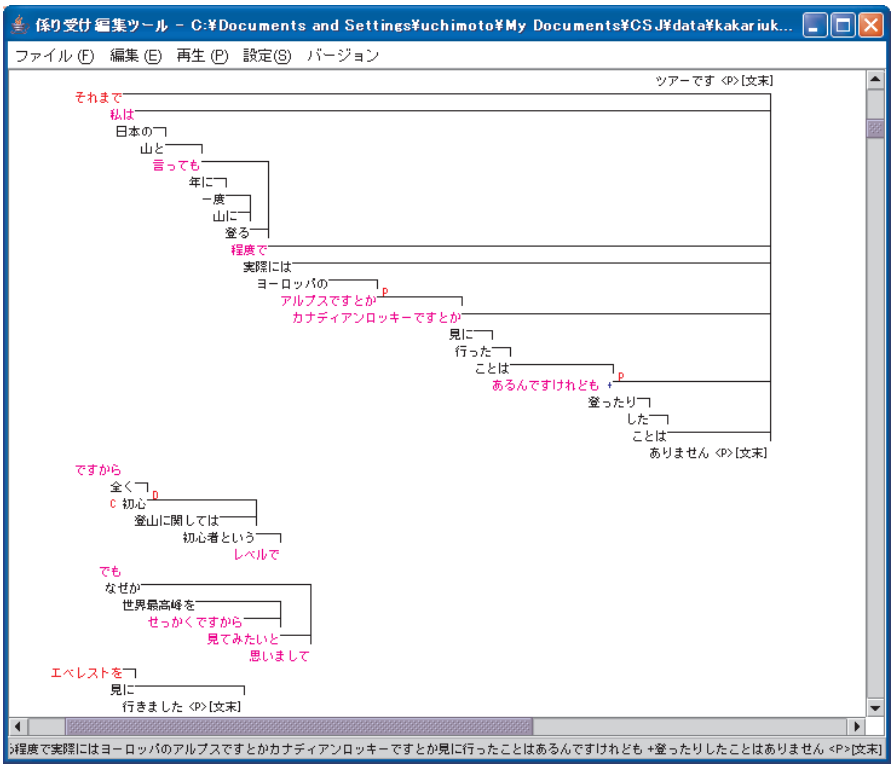



Figure 1: Dependency-structure annotation tool.

The definitions of coordination and appositives follow those of the Kyoto University text corpus (Kurohashi and Nagao, 1997). The definition of self-corrections was newly added to them [1].

## 3. Dependency-structure Annotation

Dependency structures were manually annotated to 199 speeches, which include all standard monologues, called "core", and a test set in the CSJ. The definition of a *bunsetsu* followed that defined by the National Institute for Japanese Language (Nishikawa et al., 2004). The annotation tool shown in Figure 1 was used to assist human annotators. In this figure, each line represents a *bunsetsu*, and each dependency is modified by mouse drag-and-drop. Self-corrections, coordination, and appositives can be annotated with labels D, P, and A by right-clicking the mouse. Initial dependencies were annotated so that every *bunsetsu* depends on the next. In the first step, two annotators examined each dependency and modified it if it was inappropriate. In the second step, a checker examined all dependencies annotated in the first step. The annotators referred to audio recordings as well as transcriptions during the annotation steps. For the dependency-structure annotation to the remaining parts of the CSJ, initial dependencies can be automatically annotated by using a corpus-based parser (Fujio and Matsumoto, 1998; Haruno et al., 1998; Uchimoto et al., 1999; Uchimoto et al., 2000; Kudo and Matsumoto, 2000; Matsubara et al., 2002; Shitaoka et al., 2004).

[1] The detailed criteria for dependency annotation is downloadable from the web page of the CSJ (CSJ, 2004).

## 4. Dependency-structure Analysis of the CSJ

This section discusses the difference in the dependency structures of written text and that of spontaneous speech in terms of the dependency accuracies obtained by using a corpus-based model.

The 199 manually annotated speeches consist of 20,202 sentences, 176,870 *bunsetsu*s, and 457,399 morphemes. The number of types of morphemes is 14,029. The average number of *bunsetsu*s in a sentence and the average number of morphemes in a *bunsetsu* in the CSJ are the same as those in the Kyoto University text corpus. Although the number of types of morphemes in the CSJ is less than that in the Kyoto University text corpus, the dependency accuracy obtained for the CSJ is much worse than that obtained for the Kyoto University text corpus (Shitaoka et al., 2004) under the condition that the same size of training data was used for training a model.

We investigated the expected improvements obtained by eliminating the effect of characteristic phenomena on spontaneous speech. For training a model and testing it, the same model described in (Shitaoka et al., 2004) was used. We used 168 talks for training and 20 talks for testing.

The biggest problem with dependency-structure analysis is that sentence boundaries are ambiguous, and it has been reported that the dependency accuracy is improved by approximately 3% when correct sentence boundaries are given. Other problems are assumed to be self-corrections, inserted clauses, quotations, and *bunsetsu*s that have no modifiee. By giving the correct sentence boundaries, we investigated the accuracies obtained when a model was

trained and tested after eliminating self-corrections, and giving the correct boundaries of inserted clauses and quotations, and eliminating the *bunsetsu*s that have no modifiee. Table 1 lists the accuracies. According to the table, the accuracies were improved by 0.4% by eliminating self-corrections, by 2.3% by giving the correct boundaries of inserted clauses and quotations, and by 0.4% by eliminating the *bunsetsu*s that have no modifiee. The total improvement was thus 3.1%.

| Baseline (After giving correct sentence boundaries) | | |
|---|---|---|
| closed | 88.3% | (14,632 / 16,566) |
| open | 78.3% | (12,251 / 15,650) |
| After eliminating self-corrections | | |
| closed | 88.6% | (14,434 / 16,293) |
| open | 78.7% | (12,095 / 15,362) |
| After using clause information | | |
| closed | 90.2% | (14,692 / 16,293) |
| open | 81.0% | (12,444 / 15,362) |
| After eliminating the *bunsetsu*s that have no modifiee | | |
| closed | 90.4% | (13,346 / 14,771) |
| open | 81.4% | (11,427 / 14,033) |

Table 1: Effect of characteristic phenomena on spontaneous speech.

These results show that the accuracy of automatic dependency analysis can be improved when self-corrections, clause boundaries, and *bunsetsu*s that have no modifiee are detected. We are currently investigating the improvement of dependency analysis obtained by using automatically detected self-corrections, clause boundaries, and *bunsetsu*s that have no modifiee. However, there is still a big difference between dependency accuracies of written text and spontaneous speech; namely, the accuracies obtained for closed- and open-test data extracted from the Kyoto University text corpus were approximately 98% and 89%, respectively, when the same size of training data to that used in the above experiment for spontaneous speech was used. We are also investigating other problems that contribute to the difference between dependency accuracies of written text and spontaneous speech.

## 5. Conclusion

This paper described the dependency structure of a large, spontaneous, Japanese-speech corpus, *Corpus of Spontaneous Japanese (CSJ)*, and discussed the difference in the dependency structures of written text and spontaneous speech in terms of the dependency accuracies obtained by using a corpus-based model.

The biggest problem with dependency-structure analysis of spontaneous speech is that sentence boundaries are ambiguous. Other problems are assumed to be self-corrections, inserted clauses, quotations, and *bunsetsu*s that have no modifiee. We investigated the expected improvements in the accuracy of dependency analysis obtained by eliminating the effect of the characteristic phenomena on spontaneous speech, and found that the accuracy of dependency analysis could be improved if the characteristic phenomena were detected. We are therefore currently investigating the improvement of the accuracy of dependency analysis obtained by using automatically detected self-corrections, clause boundaries, and *bunsetsu*s that have no modifiee. We are also investigating other problems that cause a gap in the dependency-analysis accuracies for written text and spontaneous speech.

## 6. References


CSJ. 2004. Release information of the Corpus of Spontaneous Japanese. http://www.kokken.go.jp/katsudo/kenkyu_jyo/corpus/index.html.

Masakazu Fujio and Yuji Matsumoto. 1998. Japanese Dependency Structure Analysis based on Lexicalized Statistics. In *Proceedings of the Third Conference on EMNLP*, pages 87–96.

Masahiko Haruno, Satoshi Shirai, and Yoshifumi Ooyama. 1998. Using Decision Trees to Construct a Practical Parser. In *Proceedings of the COLING-ACL '98*, pages 505–511.

Taku Kudo and Yuji Matsumoto. 2000. Japanese Dependency Structure Analysis Based on Support Vector Machines. In *Proceedings of the 2000 Joint SIGDAT Conference on EMNLP and VLC*, pages 18–25.

Sadao Kurohashi and Makoto Nagao. 1997. Building a Japanese Parsed Corpus while Improving the Parsing System. In *Proceedings of the NLPRS*, pages 451–456.

Willem J. M. Levelt. 1989. *Speaking: From Intention to Articulation*. The MIT Press.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous Speech Corpus of Japanese. In *Proceedings of LREC2000*, pages 947–952.

Takehiko Maruyama, Hideki Kashioka, Tadashi Kumano, and Hideki Tanaka. 2004. Development and evaluation of Japanese Clause Boundaries Annotation Program. *Journal of Natural Language Processing*, 11(3):39–68. (in Japanese).

Shigeki Matsubara, Takahisa Murase, Nobuo Kawaguchi, and Yasuyoshi Inagaki. 2002. Stochastic Dependency Parsing of Spontaneous Japanese Spoken Language. In *Proceedings of the 19th COLING*, pages 640–645.

Ken'ya Nishikawa, Hideki Ogura, Satsuki Souma, Hanae Koiso, Yoko Mabuchi, Naoko Tsuchiya, and Miki Saito. 2004. Annotation Manual for *Bunsetsu*. http://www2.kokken.go.jp/~csj/public/members_only/manuals/bunsetsu_2004MAR24.pdf. (in Japanese).

Kazuya Shitaoka, Kiyotaka Uchimoto, Tatsuya Kawahara, and Hitoshi Isahara. 2004. Dependency Structure Analysis and Sentence Boundary Detection in Spontaneous Japanese. In *Proceedings of the 20th COLING*, pages 1107–1113.

Katsuya Takanashi, Takehiko Maruyama, Kiyotaka Uchimoto, and Hitoshi Isahara. 2003. Identification of "Sentences" in Spontaneous Japanese — Detection and Modification of Clause Boundaries —. In *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 183–186.

Kiyotaka Uchimoto, Satoshi Sekine, and Hitoshi Isahara. 1999. Japanese Dependency Structure Analysis Based on Maximum Entropy Models. In *Proceedings of the Ninth Conference of the EACL*, pages 196–203.

Kiyotaka Uchimoto, Masaki Murata, Satoshi Sekine, and Hitoshi Isahara. 2000. Dependency Model Using Posterior Context. In *Proceedings of the Sixth IWPT*, pages 321–322.