Word-level Dependency-structure Annotation to Corpus of Spontaneous Japanese and Its Application

Kiyotaka Uchimoto* **and Yasuharu Den†**

* National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
uchimoto@nict.go.jp
† Department of Cognitive and Information Sciences, Faculty of Letters, Chiba University
1-33 Yayoicho, Inage-ku, Chiba 263-8522, Japan
den@cogsci.l.chiba-u.ac.jp

Abstract

In Japanese, the syntactic structure of a sentence is generally represented by the relationship between phrasal units, *bunsetsu*s in Japanese, based on a dependency grammar. In many cases, the syntactic structure of a *bunsetsu* is not considered in syntactic structure annotation. This paper gives the criteria and definitions of dependency relationships between words in a *bunsetsu* and their applications. The target corpus for the word-level dependency annotation is a large spontaneous Japanese-speech corpus, the *Corpus of Spontaneous Japanese* (CSJ). One application of word-level dependency relationships is to find basic units for constructing accent phrases.

1. Introduction

The *Corpus of Spontaneous Japanese* (CSJ) (Maekawa et al., 2000) — the largest spontaneous-speech corpus in the world available to the public — is a collection of monologues and dialogues. 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, prosodic information, and the like. In Japanese sentences, word order is flexible, and subjects or objects are often omitted. In Japanese, therefore, the syntactic structure of a sentence is generally represented by the relationship between phrasal units, *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 the dependency relationships between *bunsetsu*s (Uchimoto et al., 2006). However, in many cases, the syntactic structure of a *bunsetsu* is not considered in syntactic structure annotation.

This paper mainly gives the dependency structure within clauses and between *bunsetsu*s in the CSJ and discusses relevant applications.

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 analyzing because spontaneous speech basically contains no periods to mark sentence boundaries. Moreover, it is fundamentally difficult to find obvious sentence boundaries in spontaneous utterances, which usually contain utterance errors, utterance stops, and other characteristic phenomena. It is thus necessary to define and detect reasonable segmented units for processing as a "sentence" in spontaneous speech. In the CSJ, therefore, "sentences" are defined as "clause units". These "clause units" were originally defined as the basic processing units of spontaneous Japanese speech. They can be obtained by automatically detecting Japanese clause boundaries using the CBAP program (Maruyama et al., 2004) and then manually modifying them (Takanashi et al., 2003). Dependency relationships between *bunsetsu*s are annotated within "sentences" in the CSJ, and dependency relationships between words are annotated within *bunsetsu*s. Two types of word segments are defined in the CSJ: *short words* and *long words*. The term *short word* approximates an item found in an ordinary Japanese dictionary, and *long word* represents various compounds. In this project, the word segments in a word-level dependency structure are short words.

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

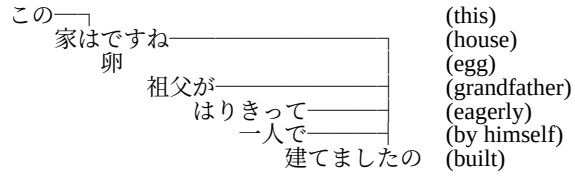
In producing spontaneous speech, speech plans constructed beforehand are sometimes changed during the utterance because of phonological, lexical, syntactic or ordering problems. In particular, long spontaneous monologues impose heavy linearization problems on speakers, such as deciding what to say first or next (Levelt, 1989). This causes various disfluencies, such as utterance stops, self-corrections, insertions, inversions, and distortions. For these disfluencies characteristic to spontaneous speech, dependency relationships between *bunsetsu*s and between short words in a *bunsetsu* are annotated in the following way.

- Utterance stop

  Utterance stops are basically detected as individual "sentences" in the CSJ, except if there is a depen-

dency relationship between *bunsetsu*s across an utterance stop. In that case, the utterance stop is defined to have no modifiee.

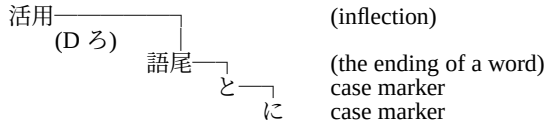
ex) "卵 (egg)" is an utterance stop and has no modifiee. In this example, each line represents a *bunsetsu*.



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."

An utterance stop is often preceded and followed by a short or longer pose. Therefore, an utterance stop holding its meaning usually consists of one or more *bunsetsu*s and is rarely found in a *bunsetsu*. When it is found, it is a word fragment and already marked with the label (D) in the CSJ. In this case, the utterance stop is also defined to have no modifiee.

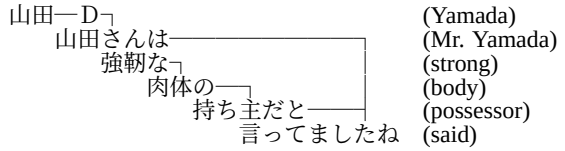
ex) "ろ" marked with (D) is an utterance stop and has no modifiee. In this example, each line represents a short word.



- Self-correction

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

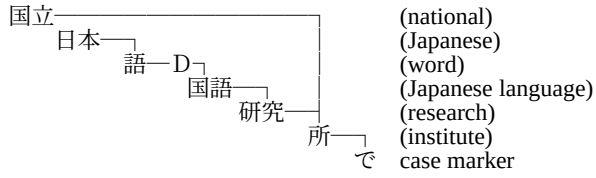
  ex) "山田 (Yamada)" is corrected to "山田さん (Mr. Yamada)" by the speaker. In this example, each line represents a *bunsetsu*.



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

  When self-corrections are found in a *bunsetsu*, they are represented as dependency relationships between short words and assigned with the label D in the same way.

  ex) "日本語 (Japanese)" is corrected to "国語 (Japanese language)" by the speaker. In this example, each line represents a short word.



- Inserted clause

  In spontaneous speech, speakers insert clauses in the middle of other clauses. This occurs when they 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 while detecting sentence boundaries.

  An inserted clause consists of one or more *bunsetsu*s and is not found in a *bunsetsu*.

- Inversion

  In the CSJ, inversions are represented as dependency relationships going from right to left. Inversions are not found in a *bunsetsu*.

- 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 that include topicalized expressions are often divided into different sentences. Distortions are not found in a *bunsetsu*.

Self-corrections differ from dependency relationships as well as from coordination and appositives. However, they are represented as dependency relationships between *bunsetsu*s, and the labels D, P, and A are assigned to self-corrections, coordination, and appositives, respectively. The definitions of coordination and appositives follow those of the Kyoto University text corpus (Kurohashi and Nagao, 1997). The definition of self-corrections and those of coordination and appositives for the word-level dependency structure were newly added to them[1].

3. Dependency-structure Annotation

Dependency relationships between *bunsetsu*s were manually annotated to 199 speeches, which included all standard "core" monologues, and a test set in the CSJ. Dependency relationships between short words in a *bunsetsu* were

[1]The detailed criteria for annotations to dependency relationships between *bunsetsu*s is down-loadable from the CSJ web page (CSJ, 2004). The detailed criteria for word-level dependency-structure annotation will also be available soon.

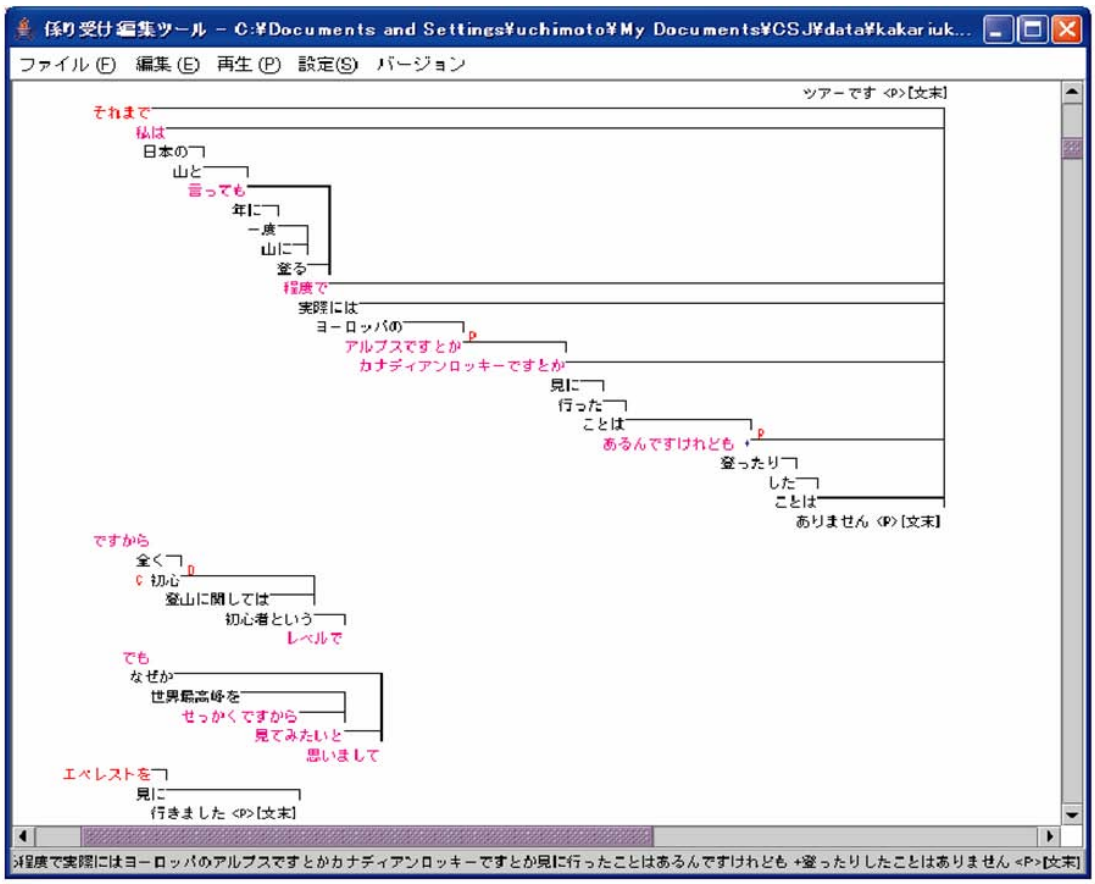


Figure 1: Dependency-structure annotation tool.

also manually annotated to 50 speeches, which included the main parts of the standard monologues. The definition of a *bunsetsu* followed that defined by the National Institute for Japanese Language (Nishikawa et al., 2004). The annotation tools shown in Figures 1 and 2 were used to assist human annotators. Each line represents a *bunsetsu* in Figure 1 and a short word in Figure 2, and each dependency can be modified by a mouse drag-and-drop. Self-corrections, coordination, and appositives can be annotated with the labels D, P, and A by right-clicking the mouse. Initial dependencies were annotated so that every *bunsetsu* and every short word depends on the *bunsetsu* or short word. In the first annotation step, two annotators examined each dependency and modified it if it was inappropriate. In the second step, a checker examined all the dependencies annotated in the first step. The annotators referred to audio recordings as well as transcriptions during these steps.

4. Word-level Dependency-structure Analysis

We investigated the word-level dependency accuracy in a *bunsetsu*. Analyzing word-level dependency can be reduced to the problem of finding a modifiee for each word in a *bunsetsu*. In Japanese, the rightmost word in a *bunsetsu* is the head of the *bunsetsu* and, in our experiments, was assumed to have no modifiee. The input for a dependency analysis is the sequence of words in a *bunsetsu* and their POS tags. Each word's modifiee is determined by using a dependency model. Therefore, the output is the set of dependencies between each word and its modifiee. Existing methods, such as shift-reduce (Nivre and Scholz, 2004), were applied for the analysis[2]. In the shift-reduce method, for example, four classes (left, right, shift, and reduce) should be discriminated for each pair of a target word and its modifiee candidate. Therefore, an SVM multi-class classifier based on pairwise coupling was used to discriminate the four classes. Two words on the left of the target word and two words on the right of the modifiee candidate and the modifiers of the target word and those of the modifiee candidate were features. The distance between the target word and its modifiee candidate was also used as a feature. The 50 speeches annotated with word-level dependency-structures were used for training and testing the model. The dependency accuracies were calculated by using 10-fold cross validation. Utterance stops were eliminated before the training and testing. The rightmost word in a *bunsetsu* is the root. The second rightmost word in a *bunsetsu* always depends on the next word. Therefore, every rightmost dependency in a *bunsetsu* was not counted in the evaluation. The total number of dependencies was 33,429. The dependency accuracies are shown in Table 1. The base-

[2]For the MSTParser (http://sourceforge.net/projects/mstparser) and CaboCha (http://chasen.org/˜taku/software/cabocha/), default parameter settings were used. The default unit for CaboCha is a *bunsetsu*. Therefore, each word was assumed to be a *bunsetsu* when CaboCha was applied for the analysis.

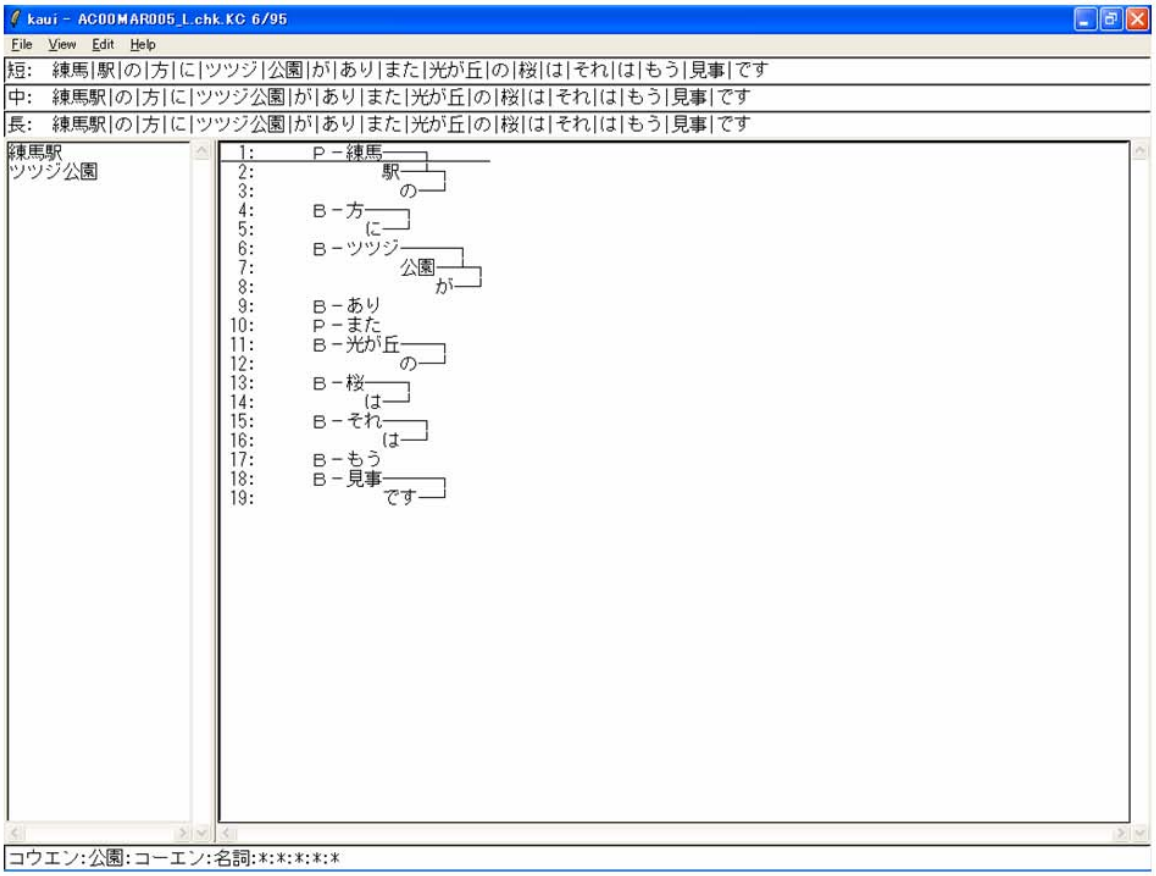


Figure 2: Word-level dependency-structure annotation tool.

line in the table is the accuracy obtained when we assumed that every bunsetsu depended on the next one. There is no significant difference between the baseline accuracy and the accuracy obtained by the existing methods. Improving dependency accuracy is our future work. For this, semantic relationships between words should be considered.

Method	Dependency accuracy
Baseline	98.6%
Shift-reduce (Nivre and Scholz, 2004)	99.1%
MSTParser (McDonald et al., 2005)	99.1%
CaboCha (Kudo and Matsumoto, 2000)	99.1%

Table 1: Word-level dependency accuracy.

5. Towards Construction of Middle Words Based on Word-level Dependency-structure

In text-to-speech synthesis, pronunciation and accent information are required to naturally read out a given text written in Japanese kana and kanji. To indicate the appropriate pronunciation and accent, we need a basic unit where a sound may change at the beginning or the ending of a word and/or an accent may change. For example, sequential voicing occurs by putting "段々(*dandan*)" (layered) and "畑 (*hatake*)" (fields) together to form a compound. Here, the *ha* of *hatake* is voiced as *ba*, and the compound word is pronounced as *dandanbatake* (layered fields). This phenomenon is called "rendaku" (Weijer et al., 2005). As for accent, the middle-accented word "*dandanbaˆtake*" (layered fields) is formed by putting the initial-accented word "*daˆndan*" and the unaccented word "*hatake*" together. On the other hand, sequential voicing does not occur in "仮名 (*kana*)" (kana) when "外来語 (*gairaigo*)" (foreign word) and "仮名表記 (*kanahyouki*)" (kana orthography) are put together to form a compound, although it occurs when "万葉 (*manyo*)"(myriad) and "仮名 (*kana*)" (kana) are put together; the *ka* of *kana* is voiced as *ga*, and the compound word is pronounced as *manyogana* (manyogana). In the above examples, "段々畑 (*dandanbatake*)", "外来語 (*gairaigo*)", "仮名表記 (*kanahyouki*)", and "万葉仮名 (*manyogana*)" are suitable units to indicate the appropriate pronunciation and accent. However, they are different from the existing two types of word segments, short words and long words. The above examples are segmented as "段々 (*dandan*)", "畑 (*hatake*)", "外来 (*gairai*)", "語 (*go*)", "仮名 (*kana*)", "表記 (*hyouki*)", "万葉 (*manyo*)", and "仮名 (*kana*)" for short words, and as "段々畑 (*dandanbatake*)", "外来語仮名表記 (*gairaigokanahyouki*)", and "万葉仮名 (*manyogana*)" for long words. Therefore, we propose constructing new basic units, *middle words*, that are defined as units between short and long words and that will be useful as constituents of accent phrases.

When putting words together to form a compound, a sound

change at the beginning or the ending of a word or an accent change are blocked by right branched tree structures (Kubozono, 1995). For example, "外来語仮名表記 (*gairaigokanahyouki*)" has a [[外来語 (*gairaigo*)] [仮名表記 (*kanahyouki*)]] structure where [仮名表記 (*kanahyouki*)] is a right branch. Therefore, we define middle words as units constructed by combining adjacent short words that have dependency relationships within a long word.
Middle words are constructed on the basis of the following rule.

- Combining adjacent short words that have dependency relationships under the condition that a middle word is not longer than a long word.

Morphological information is acquired in the following way.

- If a middle word corresponds to a long word, morphological information of the middle word is extracted from the long word.
- If the middle word does not correspond to a long word, morphological information of the middle word is extracted from the rightmost short word in the middle word.

After eliminating utterance stops and self-corrections, we automatically constructed middle words for 50 speeches in the CSJ based on the word-level dependency structure mentioned in Section 2. We found the following relationships between middle words and accent phrases (BI=2, 2+p, 2+b, 2+bp, 3) in the CSJ.

- The total numbers of short, middle, and long words were 117,145, 97,644, and 97,167, respectively.
- The number of middle words that were constituents of accent phrases in the CSJ was 94,702. Therefore, 2,942 middle words crossed the accent phrase boundaries in the CSJ.
- The number of long words that were constituents of accent phrases in the CSJ was 94,038. Of these, middle words corresponded to 93,797. It is desirable that the remaining 241 long words correspond to middle words.

We plan to revise the rules for constructing middle words so that the number of middle words that cross the accent phrase boundaries decreases and as many middle words as possible correspond to long words.

6. Conclusion

In this paper we described the dependency structure of a large spontaneous Japanese-speech corpus, *Corpus of Spontaneous Japanese* (CSJ), and discussed the application of a word-level dependency-structure. We proposed constructing new basic units, *middle words*, that are defined as units between short and long words and that will be useful as constituents of accent phrases. The word-level dependency structure and middle words annotated to the CSJ will be available to the public. We plan to use the data to annotate word-level dependency structures and middle words to the *Balanced Corpus of Contemporary Written Japanese* (BCCWJ), the compilation and analysis of which is supported by the priority area program 'Japanese Corpus', a five-year (2006-2010) project.

7. References

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

Haruo Kubozono. 1995. *Gokeisei to Onin-kouzou. [Word formation and phonological structure]*. Kurosio Publisher (in Japanese).

Taku Kudo and Yuji Matsumoto. 2000. Japanese Dependency Analysis Based on Support Vector Machines. In *Proceedings of the EMNLP/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).

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online Large-Margin Training of Dependency Parsers. In *Proceedings of the ACL*, pages 91–98.

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).

Joakim Nivre and Mario Scholz. 2004. Deterministic Dependency Parsing of English Text. In *Proceedings of the COLING*, pages 23–27.

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, Ryoji Hamabe, Takehiko Maruyama, Katsuya Takanashi, Tatsuya Kawahara, and Hitoshi Isahara. 2006. Dependency-structure Annotation to Corpus of Spontaneous Japanese. In *Proceedings of the LREC 2006*, pages 635–638.

Jeroen van de Weijer, Kensuke Nanjo, and Tetsuo Nishihara 2005. *Voicing in Japanese (Studies in Generative Grammar)*. Walter De Gruyter Inc.