In this paper, we describe the design and development of a new version of the Corpus of Spontaneous Japanese (CSJ–RDB), which is a large-scale spoken corpus released in 2004. CSJ contains various annotations that are represented in XML format (CSJ–XML). CSJ–XML, however, is very complicated and suffers from some problems. To overcome this problem, we have developed and released, in 2013, a relational database version of CSJ (CSJ–RDB). CSJ–RDB is based on an extension of the segment and link-based annotation scheme, which we adapted to handle multi-channel and multi-modal streams. Because this scheme adopts a stand-off framework, CSJ–RDB can represent three hierarchical structures at the same time: inter-pausal-unit-top, clause-top, and intonational-phrase-top. CSJ–RDB consists of five different types of tables: segment, unaligned-segment, link, relation, and meta-information tables. The database was automatically constructed from annotation files extracted from CSJ–XML by using general-purpose corpus construction tools. CSJ–RDB enables us to easily and efficiently conduct complex searches required for corpus-based studies of spoken language.

Keywords: spoken corpus, relational database, segment and link-based annotation scheme

1. Introduction

The Corpus of Spontaneous Japanese (CSJ) is a large-scale annotated spoken corpus that was collaboratively developed by the National Institute of Japanese Language and Linguistics, National Institute of Information and Communications Technology, and Tokyo Institute of Technology between 1999 and 2003 (Maekawa, 2003). Since its release in 2004, CSJ has been used in various fields such as linguistics, speech and language technologies, psychology, and education. This release of CSJ contains various annotations that are represented in XML format (Maekawa et al., 2004). The XML documents of CSJ, however, are very complicated, which hinders general use of CSJ in research fields such as humanities. To overcome this problem, we have developed and released, in 2013, a relational database (RDB) version of CSJ (CSJ–RDB), which enable a wider range of researchers to access various annotations more easily and efficiently. This paper describes the design and development of this new version of CSJ.

2. Corpus of Spontaneous Japanese

The CSJ contains about 625 hours of monolog speech that is mainly sourced from academic presentation speech and general speech on everyday topic. It also contains about 15 hours of dialog speech and about 20 hours of read speech. The speech material is transcribed, and part-of-speech (POS) analysis is applied based on two different types of words: short-unit word (SUW), which approximates dictionary item of ordinary Japanese dictionary, and long-unit word (LUW), which represents various compounds.

The core is a fully annotated subset of CSJ that contains around 44 hours of speech (Figure 1). In addition to transcriptions and POS information, the following annotations are included in it (NINJAL, 2006):

- Discourse structure label: The dependency structure between bunsetsu phrases, which are basic phrasal units in Japanese, is labeled.
- Impressionistic rating: Impression rating scores about talks, such as fluency, politeness, and spontaneity, are given by twenty raters.
- Segmental label: Most labels are basically phonemic (e.g., ‘a’, ‘o’, ‘k’), but some are phonetic (e.g., ‘k’).
- Clause boundary label: The syntactic clause boundary, its type, and miscellaneous annotations such as inversion and parenthetical clause are labeled.
- Dependency structure label: The dependency structure between bunsetsu phrases, which are basic phrasal units in Japanese, is labeled.

The whole CSJ (7.5 million words, 660 hours)

CORE (44 hours)

POS info (manual)
Intonation label
Segmental label

Clause boundary (manual)
Dependency Structure
Impressionistic rating (multiplex)
Discourse segment (excluding dialog in the core)

Speaker info

Monolog part of the core (excluding dialog and reading)

Manual POS analysis
Total=One million words

Figure 1: Layered structure of CSJ

Intonation label: The prosodic grouping of an utterance (e.g., accentual and intonational phrases) and its tonal events (e.g., word accents, phrase initial tones, and boundary tones) are annotated based on X–JToBI scheme.

Segmental label: Most labels are basically phonemic (e.g., ‘a’, ‘o’, ‘k’), but some are phonetic (e.g., ‘k’).

Clause boundary label: The syntactic clause boundary, its type, and miscellaneous annotations such as inversion and parenthetical clause are labeled.

Dependency structure label: The dependency structure between bunsetsu phrases, which are basic phrasal units in Japanese, is labeled.

Impressionistic rating: Impression rating scores about talks, such as fluency, politeness, and spontaneity, are given by twenty raters.

Discourse structure label: The middle level of discourse structure (sub-story) is labeled based on a committee-based decision of discourse purposes.
These annotations are represented in XML format (CSJ–XML). The core part of CSJ–XML has a hierarchical data structure composed of six elements, as shown in Figure 2. Each element has its own attributes; for example, an SUW-type element has POS information, a phone-type element contains a phone class such as “vowel” and “consonant,” and accentual phrase (AP)-type element includes a boundary tone category such as “falling” and “rising.” Although CSJ–XML represents a large number of information in a systematic way, it suffers from some problems. Most seriously, there are two more hierarchies in addition to the inter-pausal-unit (IPU) layer at the top—one with a clause and the other with an intonational phrase (IP) at the top—and features related to elements of these two hierarchies are necessarily crowded into the IPU layer. Searching across two or more hierarchies is also very difficult. Moreover, the relation between IPUs and LUWs is not strictly hierarchical; some LUWs extend across more than one IPU, which is implemented in an ad-hoc way in CSJ–XML. These problems, caused by the hierarchical design, make CSJ–XML complex and inconvenient, and lead to occasional failure in extracting some features from it.
3. Design of CSJ–RDB

To represent the multiple hierarchies of the core, we developed a relational database (RDB) version of CSJ (CSJ–RDB) based on a segment and link-based annotation scheme (Noguchi et al., 2008; Kaplan et al., 2010). This scheme was originally developed for written corpora. We adapted it to spoken corpora, in particular, handling multi-channel and multi-modal streams (Den and Koiso, in press). In the modified scheme, a specific type of segment, such as an SUW-type segment, is defined as a region identified by the starting and ending times on a certain channel, and a link is defined as a relationship between the source and the destination segments. Because this scheme adopts a stand-off framework, the hierarchical relation between two segments can be extracted based on temporal inclusion. For instance, when a phoneme-type segment ‘k’ is temporally included in an SUW-type segment ‘kore,’ there is a hierarchical relation between the two segments. For further convenience, however, the hierarchical relations among segments are explicitly described in relation tables in CSJ–RDB, as mentioned below. By virtue of the stand-off framework, CSJ–RDB can represent the three hierarchical structures at the same time: IPU-top, clause-top, and IP-top.

Figure 3 shows the basic data structure of CSJ–RDB. CSJ–RDB consists of five different types of tables: segment, unaligned-segment, link, relation, and meta-information tables.

3.1. Segment tables

Segment tables are prepared for all morpho-syntactic, phonetic, and prosodic segments, as shown in Figure 3. Each segment table contains, at least, five attributes, which uniquely identify every segment:

(1) talk ID
(2) segment ID
(3) channel label
(4) starting time
(5) ending time

In addition, attributes specific to each segment table are included (Table 1). There is another type of table called a point table. A point table is a special type of segment table with no duration, i.e., the starting and ending times are identical. A table named pointTone in Table 1 is the only point table in CSJ–RDB.

3.2. Unaligned-segment tables

In spontaneous speech, multiple words are occasionally assimilated, resulting in a contracted or reduced form such as “didja” (did you). In this case, it is impossible to identify the boundary between the component words. To deal with this problem, an assimilated form, whose starting and ending times can be specified, is treated as a segment, while original and unassimilated components of the segment are treated as temporally unaligned segment (Figure 4). The hierarchical relationship between (aligned) segments and unaligned segments is represented in the relation table to be described below.

Each unaligned-segment table contains, at least, three attributes:

Figure 4: Example of unaligned segment
3.3. Relation tables

The relation table represents the hierarchical relationship between two segments. As mentioned above, the hierarchical relation between two segments can be automatically extracted based on temporal inclusion. In spoken language, however, segments can be discontinuous due to the presence of intervening pauses, meaning that the ending time of the preceding segment and the starting time of the following segment are not always coincident. In such cases, an SQL query for extracting adjacent segments may become complicated.

We, thus, create relation tables that explicitly represent the hierarchical relationship between the ancestor and descendant segments. Each relation table contains five attributes:

(1) talk ID
(2) segment ID
(3) segment ID of the parent segment to which the segment belongs

In addition, attributes specific to each unaligned-segment table are included (Table 2).

### Table 1: Attributes specific to each segment table

| table name   | column name          | synopsis                                      | possible values / examples                                             |
|--------------|----------------------|-----------------------------------------------|-----------------------------------------------------------------------|
| segClause    | Text                 | orthographic transcription                     | (F. ま) 私生活もぼろぼろですから (well, because my personal life is also miserable) |
|              | ClauseBoundaryLabel  | clause type                                    | /causal clause-kara/                                                   |
|              | CU, ObligateComment  | comment on special cases                       | parenthetical clause                                                   |
| segBunsetsu  | Text                 | orthographic transcription                     | 私生活も (my personal life is also)                                    |
| segLUW       | Text                 | orthographic transcription                     | 私生活 (personal life)                                                |
| segSUW       | Text                 | orthographic transcription                     | 私 (personal)                                                         |
|              | PhonLabel            | phoneme string                                 | si                                                                    |
| segIP        | Text                 | orthographic transcription                     | (F. ま) 私生活もぼろぼろですから (well, because my personal life is also miserable) |
|              | fbt                  | final boundary tone                            | L%, LH%, HL%, or HLH%                                                 |
| segAP        | Text                 | orthographic transcription                     | 私生活も (my personal life is also)                                    |
|              | break                | break index                                    | 2, 2+b, 2+bp, 3, etc.                                                |
|              | fbt                  | final boundary tone                            | L%, LH%, HL%, or HLH%                                                 |
|              | prm                  | prominence                                     | PNLP, EUAP, FR, or HR                                                 |
|              | misc                 | miscellaneous info                             | AYOR, QQ, or HRP                                                      |
| segMora      | MoraEntity           | mora entity                                    | り (absent) or 1 (present)                                           |
|              | PerceivedAcc         | presence of accent nucleus                     |                                                                      |
| segPhone     | PhoneEntity          | phone entity                                   | #j                                                                    |
|              | PhoneClass           | phone class                                    | consonant, vowel, or special                                         |
|              | Devoiced             | presence of devoicing                          | 0 (absent) or 1 (present)                                            |
|              | StartTimeUncertain   | uncertainty of start time                      | 0 (certain) or 1 (uncertain)                                         |
|              | EndTimeUncertain     | uncertainty of end time                        | 0 (certain) or 1 (uncertain)                                         |
| pointTone    | tone                 | tone label                                     | %L, H-, A, H%, etc.                                                 |
|              | F0Uncertain          | uncertainty of F0 value                        | 0 (certain) or 1 (uncertain)                                         |
|              | CategoryUncertain    | uncertainty of tone category                   | 0 (certain) or 1 (uncertain)                                         |
|              | PositionUncertain    | uncertainty of tone position                   | 0 (certain) or 1 (uncertain)                                         |

(1) talk ID
(2) segment ID
(3) segment ID of the parent segment to which the segment belongs

3.4. Link tables

The link table represents the relationship between two segments. CSJ–RDB has two link tables, one describing the dependency structure between the modifier bunsetsu phrases and the modified bunsetsu phrases, and the other representing the relationship between tonal events and accentual phrases to which they belong.

Each link table contains, at least, three attributes:

(1) talk ID
(2) segment ID of the source segment
(3) segment ID of the destination segment

In addition, attributes specific to each link table, such as the type of dependency (“parallel,” “appositive,” “reversed,” etc.), may be included.

3.5. Meta-information tables

CSJ–RDB includes four meta-information tables, one containing information about talks (e.g., talk ID, speaker ID, talk type, genre, topic), one including attributes about
Table 2: Attributes specific to each unaligned-segment table

| Table name  | Column name               | Synopsis                | Possible values / examples         |
|-------------|---------------------------|-------------------------|------------------------------------|
| usegLUW     | LUWDictionaryForm         | dictionary form         | イク (go)                          |
|             | LUWLemma                 | lemma                   | 行く                               |
|             | LUWPPOS                   | POS                     | verb                               |
|             | LUWConjugateType         | conjugate type          | 5-dan conjugation-ka               |
|             | LUWConjugateForm         | conjugate form          | adverbial form                     |
|             | LUWMiscPOSInfo1          | miscellaneous POS info 1| case particle                      |
|             | LUWMiscPOSInfo2          | miscellaneous POS info 2| consonant geminaton               |
|             | LUWMiscPOSInfo3          | miscellaneous POS info 3| collocation                        |
| usegSUW     | PlainOrthographicTranscription | orthographic transcription w/o tags | 行き |
|             | PhoneticTranscription    | phonetic transcription  | イク                               |
|             | SUWDictionaryForm         | dictionary form         | 行く                               |
|             | SUWLemma                 | lemma                   | verb                               |
|             | SUWPPOS                   | POS                     | 5-dan conjugation-ka               |
|             | SUWConjugateType         | conjugate type          | adverbial form                     |
|             | SUWConjugateForm         | conjugate form          | sentence-final particle            |
|             | SUWMiscPOSInfo1          | miscellaneous POS info 1| ellipsis                           |
|             | SUWMiscPOSInfo2          | miscellaneous POS info 2| reduced form                       |
|             | SUWMiscPOSInfo3          | miscellaneous POS info 3| <conditional clause-to>            |
|             | ClauseBoundaryLabel      | clause boundary label   | >>                                 |
|             | CU_preBracket            | open bracket for the scope of inversion, quotation, etc. | + |
|             | CU_postBracket           | close bracket for the case above | parenthesis clause               |
|             | CU_OperationSign         | unit operation sign     |                                    |
|             | CU_ObligateComment       | comment on special cases|                                    |

Clause segment table

| TalkID  | BunsetsuID | StartTime | EndTime  | Channel | OrthographicTranscription | ClauseBoundaryLabel |
|---------|------------|-----------|----------|---------|---------------------------|---------------------|
| A01F0067 | 00262895L  | 262.895042 | 264.895345 | L       | 次の三つの課題を行いました | [sentence boundary] |

Bunsetsu segment table

| TalkID  | BunsetsuID | StartTime | EndTime  | Channel | OrthographicTranscription |
|---------|------------|-----------|----------|---------|---------------------------|
| A01F0067 | 00262895L  | 262.895042 | 263.240038 | L       | 次の三つの課題を行いました   |
| A01F0067 | 00263240L  | 263.240038 | 263.769447 | L       | three-GEN                 |
| A01F0067 | 00263769L  | 263.769447 | 264.153538 | L       | 課題を                 |
| A01F0067 | 00264154L  | 264.153538 | 264.895345 | L       | conduct-POL-PAST         |

Bunsetsu-to-Clause relation table

| TalkID  | BunsetsuID | ClauseID | mh | len |
|---------|------------|----------|----|-----|
| A01F0067 | 00262895L  | 00262895L | 1  | 4   |
| A01F0067 | 00263240L  | 00262895L | 2  | 4   |
| A01F0067 | 00263769L  | 00262895L | 3  | 4   |
| A01F0067 | 00264154L  | 00262895L | 4  | 4   |

Figure 5: Examples of segment tables and a relation table

4. Development of CSJ–RDB

CSJ–RDB was constructed by the following steps (Figure 6).

Step 1: The following three types of annotation files, which can be used for existing annotation tools, were extracted from CSJ–XML (version 3, released in 2011): (1) syntactic information including clause boundary labels, bunsetsu boundaries, and dependency structures, (2) morphological information including POS information of both SUW and LUW, and (3) segmental and prosodic information including segmental, tone, break index, and other miscellaneous labels, which can be edited by using Praat. These annotation files contain all the information required to create the database tables described in Step 3.

Step 2: Additional annotations were conducted: (1) syntactic annotations to dialog and read speeches, which were missing in CSJ–XML, and (2) marking of ‘dislocated’ tonal events, which are ‘physically’ located outside of the accentual phrases to which they ‘logically’ belong.

Step 3: All the annotation files were automatically converted to table files that can be directly imported into the

speaker (e.g., speaker ID, sex, birth generation, birth place), and two containing single- and multiple-rater-based impression ratings about talks (e.g., fluency, politeness, spontaneity).
database. Relation tables were also automatically extracted based on temporal inclusion between segments.

**Step 4:** Finally, CSJ–RDB was constructed from the table files created in Step 3. CSJ–RDB was implemented using SQLite.

We developed general-purpose tools for conducting procedures in Steps 3 and 4 that are customizable according to different format of the annotation files and different configuration of the database (Den and Koiso, in press). We can employ these tools to create other spoken corpora database, e.g., multi-party, multi-modal conversation corpus.

### 5. Efficacy of CSJ–RDB

CSJ–RDB enables us to easily and efficiently conduct complex searches required for corpus-based studies of spoken language. In a typical situation in corpus-linguistic studies, we may want to extract the boundary tones (“falling,” “rising,” etc.) of the final accentual phrases in all the clauses ending with the final particle “ne.” This is very difficult in CSJ–XML, since such a query refers to two different hierarchies, i.e., clause-top and IP-top ones, which cannot be easily combined in CSJ–XML. With CSJ–RDB, simple joining of several segment and relation tables can achieve it. The simple data structure of CSJ–RDB significantly facilitates the management and the searching of the corpus. Since its release in 2013, more than 170 researchers, including those with both computational and humanities backgrounds, have used CSJ–RDB. We believe that CSJ–RDB will greatly promote the development of corpus-based studies of spoken Japanese.

### 6. Acknowledgements

This work is supported by Grant-in-Aid for Collaborative Research Project of NINJAL “Empirical study on the role of prosodic features in conversations” led by Hanae Koiso, JSPS Grant-in-Aid for Scientific Research (B) “Constructing cognitive and communicative models for dialog based on utterance-unit annotations” led by Yasuharu Den, and Grant-in-Aid for Collaborative Research Project of NINJAL “Foundations of Corpus Japanese Linguistics” led by Kikuo Maekawa.

### 7. References

Den, Y. and Koiso, H. (in press). An environment for the usage of spoken discourse corpora that effectively utilizes existing tools (in Japanese). Journal of Natural Language Processing, 21(2).

Kaplan, D., R. Iida, R., Nishina, K., and Tokunaga, T. (2010). Annotation process management revisited. In Proceedings of the 7th International Language Resources and Evaluation Conference (LREC 2010), pages 3654–3661, Valletta, Malta.

Maekawa, K., Kikuchi, H., and Tsukahara, W. (2004). Corpus of Spontaneous Japanese: Design, annotation and XML representation. In Proceedings of the International Symposium on Large-scale Knowledge Resources (LKR2004), pages 19–24, Tokyo.

Maekawa, K. (2003). Corpus of Spontaneous Japanese: Its design and evaluation. In Proceedings of the ISCA and IEEE Workshop on Spontaneous speech processing and recognition (SSPR-2003), pages 7–12, Tokyo.

NINJAL. (2006). Construction of the Corpus of Spontaneous Japanese. National Institute for Japanese Language and Linguistics (NINJAL) Report 124 (in Japanese).

Noguchi, M., Miyoshi, K., Tokunaga, T., Iida, R., Komachi, M., and Inui, K. (2008). Multiple purpose annotation using SLAT—Segment and link-based annotation tool—. In Proceedings of the 2nd Linguistic Annotation Workshop, pages 61–64, Marrakech, Morocco.