Corpus of Spontaneous Japanese: Design, Annotation and XML Representation

Kikuo Maekawa[*], *Hideaki Kikuchi*[†*] *and Wataru Tsukahara*[‡*]

[*]Department of Language Research, National Institute for Japanese Language
[†] School of Human Sciences, Waseda University
‡Graduate School of Information Systems University of Electro-Communications
{kikuo, kikuchi, tsuka}@kokken.go.jp

Abstract: Corpus of Spontaneous Japanese (CSJ) is a richly annotated speech and language database of spontaneous speech. It contains more than 650 hours of spontaneous Standard Japanese. In addition to digitized audio, CSJ contains two-way transcription of about 7.5 million words, two-way POS annotation, speaker information, and impressionistic rating of the way the talks are being spoken. Moreover, there is a true subset of the CSJ, called the Core to which further annotations are provided. The Core annotation contains segmental and prosodic labeling, clause-boundary labeling, dependency-structure analysis, and so forth. In order to facilitate verification and information retrieval of the complex data, the annotation data are represented by means of the XML format. Simple examples of data verification and information retrieval using XML are shown.

1. Introduction

Since 1999, National Institute for Japanese Language, Communications Research Laboratory, and Tokyo Institute of Technology have been collaborating on an national project known as the "*Spontaneous Speech: Corpus and Processing Technology*" (1999-2003) under the general supervision of Professor Sadaoki Furui [1]. Compilation of the *Corpus of Spontaneous Japanese*, or CSJ, is one of the most important outcomes of this project whose mission was to provide seeds for the processing technology of spontaneous speech. In the rest of this paper, we will firstly introduce the general overview of the corpus, and, secondly, introduce the XML representation.

2. CSJ

2.1. Layered structure

Figure 1 shows layered structure of the CSJ. As a whole, CSJ contains digitized speech (16kHz, 16bit), transcription, POS annotation, information about the speakers and talks for more than 7.5 million words (SUW, see below) or 658 hours. In addition, there is a true subset of the corpus known as the Core, to which the cost of annotation is concentrated. The CSJ-Core consists of talks of about 500K words, or 44 hours, and includes segmental and intonation labels (known as the X-JToBI [2]), clause-boundary labels, dependency-structure annotation, and so forth.

The quantity of the whole CSJ was designed to fulfill the minimum requirement of the acoustic- and language-modeling of spontaneous speech, and, the quantity of the Core was designed to be the maximum amount to which manual annotation can be applied during the five years.

Also, manually analyzed POS data of the Core was used for the development of an automatic POS tagging program with which we analyzed the rest of the corpus [3]. As a matter of fact, since we could accomplish manual analysis of the Core much earlier than we expected, about one million words were analyzed manually. The shaded area of figure 1 corresponds to this part of the corpus.

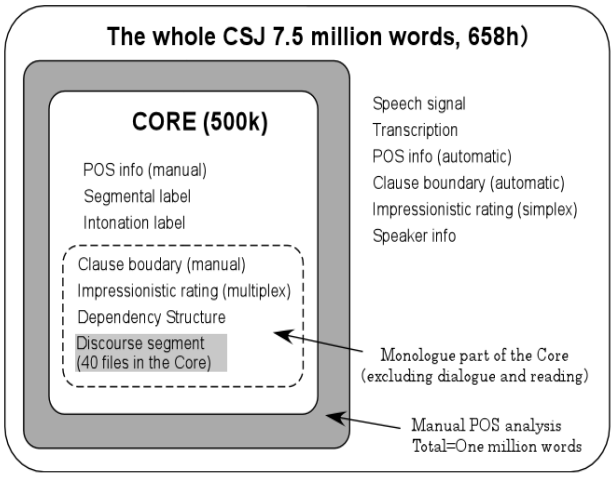

Figure 1: Layered structure of CSJ (See 2.4 for SUW).

2.2. Speech types

As shown in table 1, more than 95% of the speech material is devoted for spontaneous monologues, of which the two main types being APS (Academic Presentation Speech) and SPS (Simulated Public Speaking). Monologue was chosen as the main material because it was monologue that was the main target of speech recognition. Moreover, it was APS, the live recording of academic talks done in various academic societies that was regarded to be the primary objective of speech recognition.

On the other hand, SPS, layman talks about everyday topics like 'my most delightful memory' or 'If I live in a deserted island', was needed for three reasons. Firstly, the distribution of speakers in APS is strongly skewed with respect to their age and sex: most speakers were young male graduate students especially in engineering fields. Secondly,

the vocabulary of APS differed considerably from one field to another. To achieve these objective, age and sex of the SPS speakers were maximally balanced.

Lastly, we wanted to make recording of talks whose speaking style was lower than that of APS, which was spoken in relatively high speaking style. Comparison of different speaking styles is essential in the linguistic analyses of linguistic variations, for example. See literature [4] for the difference of variations observed between APS and SPS.

Table 1: Speech types and amount of speech material

SPEECH TYPE	N Speakers	N Talks	Hour
APS	838	1,007	299.5
SPS	580	1,715	324.1
Interview on APS	* (10)	10	2.1
Interview on SPS	* (16)	16	3.4
Task-oriented dialogue	* (16)	16	3.1
Free dialogue	* (16)	16	3.6
Reproduction	* (16)	16	5.5
Reading	* (244)	491	14.1
Total	1,418	3,287	658.8

*Numbers in parenthesis are counted as speakers of APS or SPS

Table 1 shows that the remaining 5% of CSJ was devoted for dialogues and readings. This part of CSJ was recorded for the sake of the comparison with the monologue part. As for dialogue, four different types were recorded, i.e., interview about the content of APS, interview about SPS, task oriented dialogue, and free conversation. As for reading, two types were recorded, i.e., reading of two short texts excerpted from the books about natural science, and, reproduction of APS and/or SPS that were previously recorded and transcribed. The speakers of the dialogue and reading data were chosen from the speakers of the monologues.

2.3. Transcription

The recorded speech was divided into transcription units at the locations of longer-than-200 ms pauses, and, each unit was transcribed using two different formats. 'Orthographic' transcription was written in Kanji (Chinese logograph) and Kana (Japanese syllabary) characters as in the ordinary Japanese text, but in the CSJ transcription, detailed rules of the usage of Kanji and Kana were applied to minimize the variation of the usage of Kanji and Kana, which is a notorious characteristics of Japanese texts from a information retrieval point of view. In 'phonetic' transcription, on the other hand, transcription was given exclusively in Kana (Katakana) characters. This format is used to record the phonetic details of speech material involving not only linguistically motivated variation, but also reduced, truncated, and/or incorrect pronunciations.

Table 2 shows some of the transcription tags used in the CSJ (Note some tags use Japanese characters). Some of these tags are applied only for phonetic transcription, and some others only for orthographic transcription. Detailed discussion of the CSJ transcription is given in literature [5].

2.4. POS information

Considering the nature of word-formation in Japanese, two-way POS analysis was conducted. Transcription texts were annotated using two POS systems differing in the length of morphological unit, namely, SUW (short-unit word) and LUW (long-unit word). SUW approximates the entry-form of Japanese dictionary, and, LUW represents compounds made up of more than two SUW. For example, the phonemic string of /toHkyoHkoHgyoHdaigaku/ is analyzed into three units — /tokyoH/ (Tokyo), /koHgyoH/ (*technology*), and /daigaku/ (*university*)— as SUW, but constitutes a single compound noun as LUW, i.e., *Tokyo Institute of Technology*. Note, that there are LUW of other POS categories. The string /niyoQte/, which consists of three SUW — particle /ni/, verb /yoru/, and particle /te/ — is analyzed as a single particle LUW whose meaning is '*by mean of*'.

The POS category adopted in the CSJ consists of three layers. In the first layer, words are classified as one of the following POS categories: noun, pronoun, adjectival noun, adnominal, adverb, conjunction, interjection, verb, adjective, auxiliary, particle, prefix, and suffix. In the second layer, verb, adjective, auxiliary, and suffix are classified in terms of their conjugation types. The third layer is concerned with three classifications: the classification of particles, classification with respect to the morphophonological alternation (*onbin*) of verb, adjective, auxiliary, and suffix, and, the specification about word coalescence.

Table 3 shows the total numbers of SUW and LUW contained in each speech type of CSJ. APS and SPS contain nearly the same number of morphological unit.

Table 2: Tag set of the CSJ transcription

TAG	USAGE
(D)	Word fragment
(W)	Reduced, truncated, or incorrect pronunciation
(?)	Uncertainty of perception
(F)	Filled pauses
(M)	Meta-linguistic expression
(O)	Foreign language or archaic Japanese
(A)	Use of alphabets in the orthographic transcription
(K)	Exceptional use of Kana in the orthographic transcript.
(笑) (泣) (咳) (あくび)	Speaking while laughing Speaking while crying Speaking while coughing Speaking while yawning
(L)	Whispery voice
<H>	Non-lexical lengthening of vowel
<Q>	Non-lexical lengthening of consonant
<FV>	Vowel whose phonemic status is not identifiable
<息> <笑> <泣> <咳>	Breathing noise Laughter (not speaking) Cry (not speaking) Cough (not speaking)

Table 3: Numbers of SUW and LUW by speech type

SPEECH TYPE	SUW	LUW
APS	3,279,364	2,654,823
SPS	3,605,729	3,115,302
Interview on APS	27,907	24,287
Interview on SPS	43,817	38,782
Task-oriented dialogue	30,326	25,981
Free dialogue	47,776	42,494
Reproduction	49,487	40,326
Reading	157,991	131,890
Miscellaneous	282,728	239,989
Total	7,525,125	6,313,874

2.5. Impressionistic rating

By impressionistic rating is meant listener's subjective evaluation of the way a talk is being spoken. All spontaneous talks of the CSJ were evaluated, at the time of recording, by one of the recording staffs using an evaluation sheet. The evaluation included five-scale ratings with respect to the speaking style, spontaneity, speaking speed, clearness of articulation, amount of dialectal lexical items, and amount of field-specific jargons. In addition to these scales, there was also a list of 32 evaluation words, from which the rater was asked to choose (as many) words that seemed to fit the impression of the talk. The enlisted words include '*halting, fluent, monotonous, expressive, confident, assured, not assured, nervous, relaxed*' and so forth.

Although analyses of linguistic variations in the CSJ revealed that some of these ratings could be used as excellent dependent variables in the statistical analyses of variations [4], the impressionistic rating data is not free of controversy: different talks were rated by different raters (it is especially true of APS recorded in parallel sessions), and a talk was rated only by one rater. To avoid these problems, more reliable rating data is provided for 181 monologue talks of the Core. In this rating task, eight raters rated three parts (each about one minute long) excerpted from a talk using 20 pairs of evaluation words and seven-scale rating. See literature [6] for the detail of the multiple rating experiments.

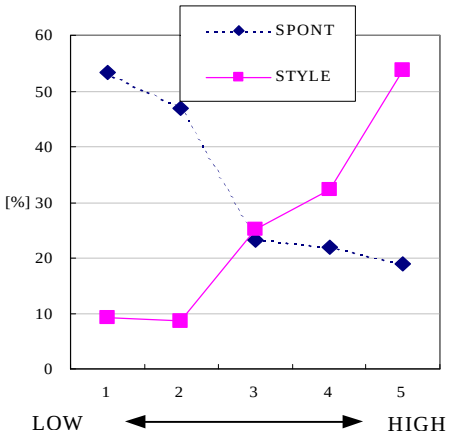

Figure 2: Occurrence rate of the 'PNLP' version of HL% boundary tones as a function of the impressionistic ratings of speaking style and spontaneity (abscissa).

2.6. Segment and intonation labels

Segmental and intonational labeling of the Core is the most cost-consuming annotation of CSJ. Most of the segment labels are phonemic, but some labels are concerned with phonetic events like release of stop closure, distinction between voiced affricates and fricatives (which are non-contrastive in Japanese), and, voicing of vowels. These phonetic labels were introduced for the sake of the study of phonological variations.

The intonation-labeling scheme, called X-JToBI, is an extended version of the J_ToBI scheme [2]. Important new features of the X-JToBI include the followings. A) Time decomposition of the bi- and tri-tonal tones (E.g. L%HL% boundary tone has three time stamps corresponding to L%, H, and L%). B) Introduction of a new symbol, '>', called *extender*, that shows the prolongation of boundary tones. C) Ramification of break indices. D) Introduction of new break indices for the treatment of various disfluency phenomena like word fragment, word-internal pauses, and filled pauses. E) Extension of the inventory of boundary pitch movement. F) Classification of the variation of boundary pitch movement.

An example of F) is the classificatory label 'PNLP' (penult non-lexical prominence) placed in the miscellaneous tier. This label is used exclusively with HL% bitonal boundary tone and shows that the time alignment of the two tones is different from the ordinary case: while ordinary HL% is realized within the time domain of the last syllable, the peak H of the PNLP version is aligned to the penult position. This kind of variation is worth being labeled, because the occurrence rates of the two variants shows contrastive difference with respect to the speaking style as shown in figure 2. In this figure, occurrence rate is defined as the ratio of occurrence of the 'PNLP' version to the total number of HL%.

2.7. Other annotations

Although CSJ-Core contains many more annotations, it is impossible to touch them because of limited space. See [7] and [8] respectively for the annotation of the clause-boundary and discourse segment boundary.

3. XML representation

As shown in the previous sections, CSJ has rich annotation. The increase in richness, however, made it more and more difficult to keep the consistency of the whole corpus across different annotations. This problem arises mainly because we had to conduct various annotation works simultaneously to finish everything within the finite time of research project.
Another problem of huge complex corpus is the information retrieval. It becomes more and more difficult to make retrieval in an effective way as the corpus size becomes larger. It is, thus highly desirable as well as necessary that the relationships among multiple annotations are expressed explicitly and systematically.

It is widely acknowledged that XML is suitable both for information retrieval (by use of XPath and XQuery, for example), and, verification and transformation of complex data (by use of XSLT). But using XML can not be a perfect answer by itself, because simultaneous achievement of these two goals is nearly impossible. The requirements of the two goals are mutually inclusive, so to speak. For verification purpose, the XML representation needs to include all annotation information hence results in a huge structure, which is not suitable for information retrieval purpose.

Our policy of XML design is to represent everything in a complex XML format that we call "Base XML" and use it for the purpose of verification. For effective information retrieval, various research-oriented XML documents are to be derived from the Base XML by means of XSLT.

3.1. The Base XML

Figure 3 shows the principal part of the hierarchical data structure of the Base XML and table 4 shows the list of XML attributes at each node of the hierarchy.

'Talk' is the root element and has attributes about the talk being recorded. A 'Talk' consists of more than one 'IPU', or inter-pausal unit', which is the unit of transcription (IPU is separated by longer than 200ms paus) and has its start- and end-time as attribute. An 'IPU' consists of more than one 'LUW', which in turn contains more than one 'SUW.' These two elements contain various POS information.

A 'SUW' is made up of more than one 'Mora'. 'Mora' has attributes like 'Entity' (e.g. /a/, /ku/, /se/, /N/, /to/) and 'Perceived Accent Position' among others. Information about 'Perceived Accent Position' is a part of the X-JToBI intonation labels and is available only for the Core.

A 'Mora' consists of more than one 'Phoneme' which contains as attribute its 'Entity' like /a/, /k/, /u/, /s/, /e/, /N/, etc. A 'Phoneme' contains more than one 'Phone,' which is the label of segmental labeling. The attributes of 'Phone' element involve the start- and end-times in addition to its 'Entity'.

The 'Phone' element in figure 3 is the linking point of the X-JToBI intonation labels: all intonation labels, with the exception of the 'Perceived Accent Position' mentioned above, are located below the 'Phone' elements. This is because most of the X-JToBI labels have time information about their occurrence position, and, 'Phone' is the only element that has fine-grained time information in figure 3.

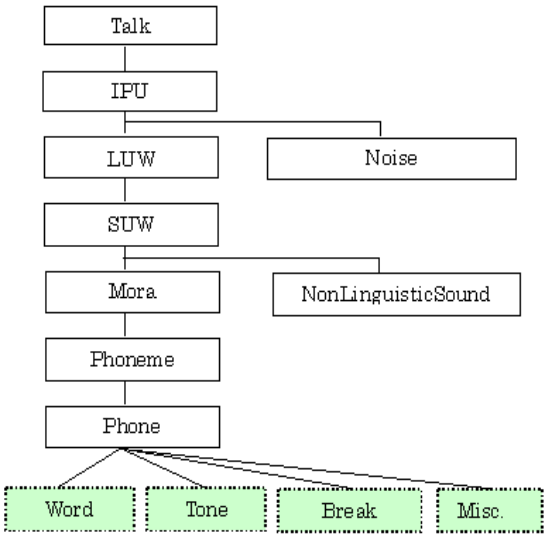

Figure 3: Principal elements of the Base XML.

Figures 4 and 5 show example raw annotation data and their Base XML representation respectively. The top panel of figure 4 is the two-way transcription of CSJ corresponding to 3 IPU (#91-#93), of which the first one consists of two small syntactic unit (*bunsetsu*) /icumono/ and /basyode/. The second panel is the SUW information corresponding to the first unit /icumono/. The LUW information is omitted because in this case SUW and LUW coincide perfectly. The third panel shows the segment labels of /icumono/, of which symble '<cl>' stands for the release of closure. Lastly, the fourth panel shows the X-JToBI labels included in the time domain corresponding to /icumono/, where symbol 'A' stands for the timing of the F0 peak due to lexical accent.

All these annotation information are represented as XML elements and properties in figure 5. Tone labels are represented as property "XJToBIEntity" of "XJToBITone" element, which is a child of "Phone" element.

Table 4:Examples of the attributes of the Base XML

ELEMENT	ATTRIBUTE	Comment
Talk	RecordingDate	
	SpeakerID	
	SpeakerSex	
	BirthYear	
	BirthPlace	
IPU	Channnel	Dialogues have two channels
	IPUStartTime	
	IPUEndTime	
LUW	LUWPOS	
	LUWConjugateType	
	LUWConjugateForm	
	LUWDictionaryForm	in Kana
	LUWLemma	in Kanji & Kana
SUW	SUWDictionaryForm	in Kana
	SUWLemma	in Kanji & Kana
	SUWPhoneTrans	Phonetic transcription
	SUWPOS	
	SUWConjugationType	
	SUWConjugationForm	
	LexicalAccPos	Dictionary position
	TagDisfluency	Element of (D)
	TagFiller	Element of (F)
	TagIncorrect	Element of (W)
	TagIncorrectNorm	Supposed-to-be intended form (W)
	TagForeign	Element of (O)
Mora	MoraEntity	
	Uncertain	Uncertainty in label selection
	Whisper	Whispered voice
	PerceivedAccPos	X-JToBI label
Phoneme	PhonemeEntity	
Phone	PhoneEntity	
	Devoiced	Devoiced vowels
	PhoneStartTime	
	PhoneEndTime	
	StartTimeUncertain	
	EndTimeUncertain	

3.2. Data verification by XML Schema and XSLT

XML Schema is a language whose purpose is to define a class of XML documents [9]. It can be a powerful tool of data verification. Among the capabilities of XML Schema are type definition of XML elements and constraints about the order and number of occurrence of child elements.

Figure 6 is an example XML Schema whose purpose is to validate 'Talk' elements by checking if they have IPU and

attributes of specified data types, ‘RecordingDate’ as the ‘date’ type data, for example.

XSLT is a language for transforming XML documents into other XML documents [10]. Figure 7 is an example XSLT whose purpose is to check if a ‘Mora’ /no/ has two ‘Phone’ grandchildren elements whose attribute are “n” and “o” in this order of occurrence. Note that this XSLT prints out a warning message when if finds a problem.

3.3. Information retrieval by means of XPath

XPath and XQuery are the two main query languages for XML document. Of these, we refer only to XPath: a language for addressing parts of an XML document, which is usually used within a XSL stylesheet that specify how to display an XML document. XPath enables us to retrieve those elements that fulfill designated conditions in a XML document. Figure 8 is an example of XPath that retrieves all SUW containing devoiced vowel.

Lastly, figure 9 shows the structure of a sample research-oriented XML document designed specifically for intonation study. New element ‘AccentualPhrase’ is inserted as the child of “Talk” and has ‘SUW’ and ‘Tone’ as its child elements. X-JToBI annotation is expressed as the attributes of these two nodes in addition to the lowest elements of the tree, which are time-linked to ‘phone’ elements. This redundancy enables quick retrieval of both the label and time information of the X-JToBI annotation.

Transcription

```
0091 00244.050-00245.009 L:
いつもの                    & イツモノ
場所で                      & バショデ
0092 00245.270-00245.581 L:
(D ねろ)                    & (D ネロ)
0093 00245.800-00247.076 L:
寝転がっていますと          & ネ<Q>コロガッテイマスト
```

SUW POS information

```
S03f0119 0091 00244.050-00245.009 L:-001-001    いつ
        いつ      イツ      何時      イツ      代名詞
                                    2001-05-14 12:01:18+09
S03f0119 0091 00244.050-00245.009 L:-001-005    も
        も        モ        も        モ        助詞
                  副助詞    2001-04-09 10:59:11+09
S03f0119 0091 00244.050-00245.009 L:-001-007    の
        の        ノ        の        ノ        助詞
                  格助詞    2001-04-09 10:59:11+09
```

Segmental label

```
244.073871  #
244.154540  i
244.187874  <cl>
244.240331  c
244.268501  u
244.328683  m
244.372218  o
244.418315  n
244.493862  o
```

X-JToBI label

```
244.092331  %L
244.114585  A
244.494613  L%
```

Figure 4: Raw annotation data

```
<Talk RecordingDate="2000-01-01" SpeakerID="0001"
BirthYear="1950F" WaveFilePath="wav/S03f0119.wav">
  <IPU Channel="L" IPUStartTime="244.050"
  IPUEndTime="245.009">
    <LUW LUWPOS="代名詞" LUWDictionaryForm="イツ"
  LUWLemma="何時">
    <SUW SUWPOS="代名詞"SUWDictionaryForm="イツ"
    SUWLemma="何時" LexicalAccPos="1" >
      <Mora MoraEntity="イ" PerceivedAccPos="1">
        <Phoneme PhonemeEntity="i">
          <Phone PhoneEntity="i"
          PhoneStartTime="244.073871"
            PhoneEndTime="244.154540">
              <XJToBITone XJToBIToneEntity="%L"
              LabelTimePos="244.092331"/>
              <XJToBITone XJToBIToneEntity="A"
              LabelTimePos="244.114585"/>
          </Phone>
          </Phoneme>
      </Mora>
      <Mora MoraEntity="ツ">
        <Phoneme PhonemeEntity="c">
          <Phone PhoneEntity="cl"
          PhoneStartTime="244. 154540"
          PhoneEndTime="244.187874"/>
            <Phone PhoneEntity="c"
            PhoneStartTime="244.187874"
            PhoneEndTime="244.240331"/>
        </Phoneme>
        <Phoneme PhonemeEntity="u">
          <Phone PhoneEntity="u"
          PhoneStartTime="244. 244.240331"
          PhoneEndTime="244.268501"/>
        </Phoneme>
      </Mora>
    </SUW>
  </LUW>
  <LUW LUWPOS="助詞" LUWDictionaryForm="モ"
  LUWLemma="も">
    <SUW SUWPOS="助詞" SUWDictionaryForm="モ"
    SUWLemma="も">
      <Mora MoraEntity="モ">
        <Phoneme PhonemeEntity="m">
          <Phone PhoneEntity="m"
          PhoneStartTime=" 244.268501"
           PhoneEndTime="244.328683"/>
        </Phoneme>
        <Phoneme PhonemeEntity="o">
          <Phone PhoneEntity="o"
          PhoneStartTime="244. 328683"
          PhoneEndTime="244.372218"/>
        </Phoneme>
      </Mora>
    </SUW>
  </LUW>
  <LUW LUWPOS="助詞" LUWDictionaryForm="ノ"
  LUWLemma="の">
    <SUW SUWPOS="助詞" SUWDictionaryForm="ノ"
    SUWLemma="の">
      <Mora MoraEntity="ノ">
        <Phoneme PhonemeEntity="n">
          <Phone PhoneEntity="n"
          PhoneStartTime=" 244.372218"
          PhoneEndTime="244.418315"/>
          </Phoneme>
        <Phoneme PhonemeEntity="o">
          <Phone PhoneEntity="o"
          PhoneStartTime="244. 418315"
          PhoneEndTime="244.943113">
            <XJToBITone XJToBIToneEntity="L%"
            LabelTimePos="244.494613"/>
          </Phone>
        </Phoneme>
      </Mora>
    </SUW>
  </LUW>
</IPU>
</Talk>
```

Figure 5: Annotations in XML format

```
<xs:element name="Talk">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="IPU" maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute name="RecordingDate" type="xs:date"/>
    <xs:attribute name="SpeakerID" type="xs:string"/>
    <xs:attribute name="BirthDate" type="xs:date"/>
    <xs:attribute name="WaveFilePath" type="xs:string"/>
  </xs:complexType>
</xs:element>
```

Figure 6: Schema Definition by means of XML Schema

```
<xsl:template match="/">
  <xsl:apply-templates select="//Mora[@MoraEntity='ノ']"/>
</xsl:template>
<xsl:template match="Mora">
  <xsl:apply-templates select="./Phoneme/Phone" />
</xsl:template>
<xsl:template match="Phone">
  <xsl:choose>
    <xsl:when test="@PhoneEntity='n'">
      <xsl:variable name="npos" select="position() + 1" />
      <xsl:if test="../Phone [position() = $npos]/@PhoneEntity != 'o'">
        <xsl:message>'o' not found</xsl:message>
      </xsl:if>
    </xsl:when>
    <xsl:otherwise>
      <xsl:message>'n' not found</xsl:message>
    </xsl:otherwise>
</xsl:choose>
</xsl:template>
```

Figure 7: Data verification by means of XSLT

```
/Talk/descendant::SUW[Mora/Phoneme/Phone/@Devoiced="1"]
```

Figure 8: An Xpath expression that retrieves SUWs containing devoiced vowel

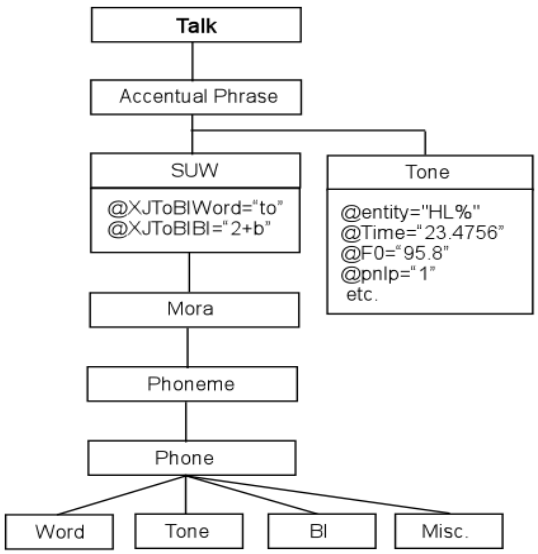

Figure 9: Derived XML for intonation study

4. Concluding remarks

Use of XML in CSJ had two aims. As for data verification, we feel that XML is an excellent solution, but we are not completely sure if XML provides optimal solution for the purpose of information retrieval. The time required for complex search of CSJ can be excessive even if we use the latest XML-native database management system. In view of the rapid development of computing power in recent years, we may be able to be optimistic about the future, but for the present, there are at least two solutions to be tried. One of them is the derivation of research-oriented XML documents mentioned briefly above. The other is the conversion of XML documents into traditional RDB format. We will pursue the last possibility in the years to come.

5. Acknowledgements

Compilation of the CSJ is supported by the Organized Research Combination System Grant from the Ministry of Education, Culture, Sports, Science and Technology.

6. References

[1] Furui, S. et al. "A Japanese national project on spontaneous speech corpus and processing technology", *ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the new Millennium*, pp. 244-248, 2000.

[2] Maekawa, K., et al., "X-JToBI: An extended J_ToBI for spontaneous speech", *Proc. 7th Int. Conf. Spoken Language Processing*, pp. 1545-1548, Denver, 2002.

[3] Uchimoto, K., et al., "Morphological analysis of the Corpus of Spontaneous Japanese", *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 159-162, Tokyo, 2003.

[4] Maekawa, K. et al., "Use of a large-scale spontaneous speech corpus in the study of linguistic variation", Proc. 15th Int. Cong. Phonetic Sciences, pp.643-636, Barcelona, 2003.

[5] Koiso, H. et al., "Transcription criteria for the Corpus of Spontaneous Japanese", *Japanese Linguistics*, 9, pp.43-58, 2001. [In Japanese]

[6] Kagomiya, T., et al., "Kooen onsei ni taisuru insyoohyootei shakudo no sakusei", *Proc. 17th General Meeging of the Phonetic Society of Japan*, pp. 135-140, 2003.

[7] Takanashi, K. et al., "Identification of 'Sentence' in spontaneous Japanese: detection and modification of clause boundaries", *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 183-186, Tokyo, 2003.

[8] Takeuchi, K., et al., "Committee-based discourse purpose assignment: Discourse annotations of spontaneous Japanese monologue", *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 199-202, Tokyo, 2003.

[9] http://www.w3.org/TR/xmlschema-0/

[10] http://www.w3.org/TR/xslt