Construction of XML documents for the study of prosody Using the *Corpus of Spontaneous Japanese*

KIKUCHI Hideaki
Faculty of Human Sciences
2-579-15, Mikajima, Tokorozawa-shi,
Saitama, JAPAN, 359-1192
kikuchi@waseda.jp

MAEKAWA Kikuo
Dept. Language Research,
National Institute for Japanese Language
Tachikawa, Tokyo, JAPAN
kikuo@kokken.go.jp

Abstract

Corpus of Spontaneous Japanese (*CSJ*) is a richly annotated speech and language database of spontaneous speech. It contains segmental and prosodic labels, clause boundary labels, dependency-structure labels, among many other annotations. In order to facilitate verification and information retrieval of the complex data, the annotation data are represented by means of an XML format. We propose a newly devised XML document that reflects the prosodic structure of the original utterance. The first half of the paper is devoted for the discussion of design and implementation issues of the new document, which is followed by the presentation of the result of pilot analyses regarding the linguistic variation of boundary pitch movements in Japanese. The results of the pilot analyses show that the new XML documents can be a useful tool for the study of prosodic structure.

1 Introduction

"*Corpus of Spontaneous Japanese*" (*CSJ*) has rich annotation such as transcription, POS information, clause boundary information, impressionistic rating, speaker information, segmental labels, intonation labels, and more (Maekawa, 2003). The increase in richness, however, made it more and more difficult to keep the consistency of the whole corpus across different annotations. 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 represented explicitly and systematically. We therefore designed XML documents that could represent dependency relations among different types of annotation data (Maekawa, 2004). By using these XML documents, it becomes easier to retrieve information which is concerned with different types of annotations, the position of lexical accents in different syntactic environments, for an example.

Information of intonation labels represented in the XML documents of the *CSJ* are based on the X-JToBI prosodic labeling scheme (Maekawa et al., 2002), the extended version of the J_ToBI labeling scheme (Venditti, 1995).

X-JToBI (and J_ToBI) is a method of prosodic transcription developed for Tokyo Japanese, based upon the design principles of the ToBI system (Silverman et al. 1992). The two main components of the ToBI system are the tones and break indices (BI) labels. The former provides the tonal structure of an utterance, and the latter specifies the hierarchical structure of the utterance by specifying the relative strength of the prosodic boundaries.

Here, it is to be noted that in Japanese, prosodic unit known as accentual phrase plays an important role in the transmission of linguistic as well as paralinguistic information; phonetic analysis based upon accentual phrase is indispensable in the study of Japanese prosody.

It is therefore very important to devise a new XML document that can represent explicitly prosodic structure of an utterance. It is also desirable that the new documents can be derived automatically from the existing XML documents.

This paper describes the specification of the *CSJ* XML documents in section 2 and the design and implementation of the devised new documents based on accentual phrase unit in section 3. The result of pilot analysis using the new XML documents shows that use of them is effective for the study of prosodic structure related to other type of information such as morpheme information.

2 XML documents in *CSJ*

Various types of annotation in the *CSJ* are mutually related. Our policy of XML design is to encode everything in a complex XML format that we call the "Base XML" documents. They are represented in linguistic hierarchical structure to make it easy and convenient to retrieve linguistically dependent some types of annotation information.

Figure 1 shows the principal part of the hierarchical data structure of the Base XML document, and the table 1 shows a partial list of

XML attributes at each node of the hierarchy. Also, figure 2 shows an example of XML documents.

XSLT(Extensible Style Language Transformation) is a language for transforming XML documents into other XML documents. Figure 3 is an example XSLT whose purpose is to extract lemma, conjugate form, phonetic transcription, and perceived accent position of all adjectives.

In the *CSJ*, 'word' is represented as short-unit word. On the XML documents, a short-unit word is encoded as a SUW element. POS information, dictionary form, conjugate form, and phonetic transcription are encoded as attributes named SUWPOS, SUWLemma, SUWConjugateForm, and PhoneticTranscription respectively. Perceived accent position in a word can be got from information of intonation labels and encoded as a PerceivedAccPos attribute in a XJToBILabelWord element. The XSLT script shown in the figure 3 will find SUW elements which its attribute of SUWPOS is equal to adjectives in a document, will output value of SUWLemma, SUWConjugateForm, and PhoneticTranscription attribute of each element in CSV style. At last, the script will output value of PerceivedAccPos attribute of XJToBILabelWord elements belonging to each SUW element.

As seen in figure 1, tone labels of X-JToBI are encoded as XJToBILabels elements, which are children of Phone element. It is because the parent of an XJToBILabel element was decided according to the time position of the XJToBILabel element. This design is convenient for the analysis of disagreement between linguistic phrase boundaries and prosodic phrase boundaries. On the other hand, however, it is often difficult to retrieve tone labels

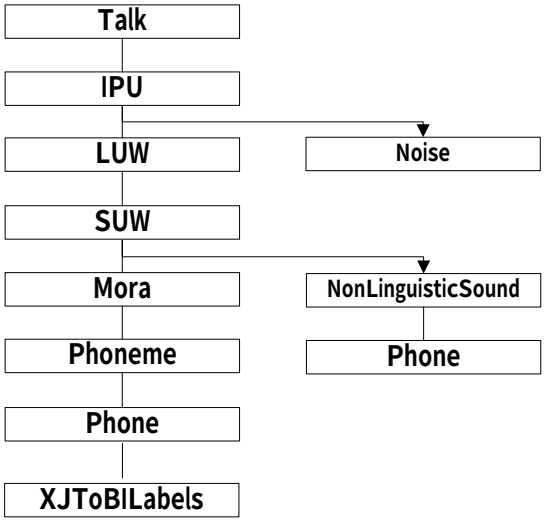

Figure 1: Principal part of the hierarchical data structure of the Base XML.

near the accentual phrase boundaries. Therefore, we designed a new XML document that can represent the prosodic structure of given utterances. The new XML documents, which are derived automatically from the corresponding old XML documents, are organized based upon the accentual phrase structure of the utterances. In the next section, the design and implementation of the new documents will be described..

Table 1: The attributes of the Base XML.

Element	Attribute	Comment
Talk		
	RecordingDate	
	SpeakerID	
	BirthDate	
	WaveFilePath	
IPU		
	Channel	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	
	SUWConjugateType	
	SUWConjugateForm	
	LexicalAccPos	Dictionary position
	TagDisfluency	Element of (D)
	TagFiller	Element of (F)
	TagIncorrect	Element of (W)
	TagIncorrectNorm	Supposed-to-be intended from (W)
	APID	Accentual phrase ID
Mora		
	MoraEntity	
	Uncertain	Uncertainty in label selection
	PerceivedAccPos	X-JToBI label
Phoneme		
	PhonemeEntity	
Phone		
	PhoneEntity	
	Devoiced	Devoiced vowels
	PhoneStartTime	
	PhoneEndTime	
	StartPosUncertain	
	EndPosUncertain	
XJToBI*Labels*		
	Entity	
	Time	
Noise		
NonLinguisticSound		

```
<Talk RecordingDate="2000-01-01" SpeakerID="0001"
BirthDate="1950-01-01" 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" APID="304">
        <Mora MoraEntity="イ" PerceivedAccPos="1">
          <Phoneme PhonemeEntity="i">
            <Phone PhoneEntity="i"
            PhoneStartTime="244.073871"
            PhoneEndTime="244.154540">
              <XJToBILabelTone Time="244.092331"
              F0="210.791">%L </XJToBILabelTone>
              <XJToBILabelTone Time="244.114585"
              F0="216.325"> A </XJToBILabelTone>
            </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" APID="304">
        <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" APID="304">
        <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">
              <XJToBILabelTone Time="244.494613" F0="151.82"
              ToneClass="FBT"> L%" </XJToBILabelTone>
            </Phone>
          </Phoneme>
        </Mora>
      </SUW>
    </LUW>
  </IPU>
</Talk>
```

Figure 2: Annotation data in XML format.

```
<?xml version="1.0" encoding="EUC-JP"?>
<xsl:stylesheet version="1.0"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
xml:lang="ja">

<xsl:output method=" text " indent="yes" encoding=
"EUC_JP"/>

<xsl:template match="SUW">
 <xsl:if test="@SUWPOS='形容詞'">
  <xsl:value-of select="@SUWLemma"/>
  <xsl:text>,</xsl:text>
  <xsl:value-of select="@SUWConjugateForm"/>
  <xsl:text>,</xsl:text>
  <xsl:value-of select="@PhoneticTranscription"/>
  <xsl:text>,</xsl:text>
  <xsl:value-of select="descendant::XJToBILabelWord[1]/
       @PerceivedAccPos"/>
  <xsl:text>,</xsl:text>
  <xsl:text>&#x0a;</xsl:text>
 </xsl:if>
</xsl:template>

</xsl:stylesheet>
```

Figure 3: Example of XSLT script.

3 Construction XML documents based upon the AP unit

In this section, procedure of constructing new XML documents based upon accentual phrase(AP) unit is described. Figure 4 shows relationship between elements in the XML documents.

1) Move elements and its attributes of LUW and IPU

   As shown in figure 4, AP element which corresponds to accentual phrase unit is inserted between the Talk element and the SUW element. Instead, LUW element and IPU element are removed and their attributes are encoded as attributes of the SUW element.

2) Construct accentual phrase unit

   As shown in figure 2, sequential ID of AP element is encoded as APID attribute of the SUW element in the corresponding old XML document. Values of APID are incremented if the entity of the XJToBILabelBreak element targeting the SUW element is greater than "2".

3) Generate tone label information corresponding to accentual phrase unit

   All XJToBILabelTone elements are copied at the child position of phonologically corresponding AP element.

Figure 5 shows an example of the newly constructed XML documents. New element "AP" is inserted as the child of the Talk element and it has SUW and XJToBITone elements as its child

elements. X-JToBI annotation is encoded as the attributes of these two nodes, but they are also recorded in 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.

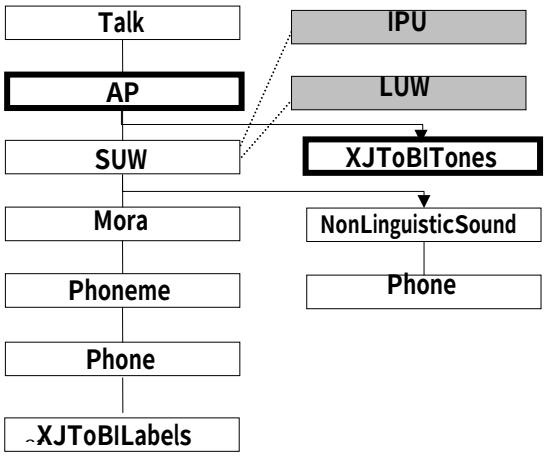

Figure 4: Derived XML for study of prosody.

4 Preliminary analyses using reorganized XML documents

In this section, we describe the result of basic analysis of accentual phrase unit in 174 reorganized XML documents in the *CSJ*.

First, table 2 shows number and duration of accentual phrase unit in each type of speech. APS means Academic Presentation Speech and SPS mieans Simulated Public Speaking. Number of accentual phrase unit is over one hundred thousands. Number of AP units in each utterance unit is 2.43, and Number of words in each AP unit is 3.16 in average. Also, mean duration of AP unit is 700[ms] and there are significant difference between each type of speech. There are 6781 cases that one AP unit is over boundary of utterances and 9741 cases that one long word unit is over boundary of AP units. These facts suggest usefulness of the XML documents in AP units. Finally, frequency of BPM(Boundary Pitch Movement) is investigated from the reorganized XML documents. Table 3 shows the result of investigation. The results showed that L%H% is appeared more frequently in APS than in SPS. It is also noteworthy that L%LH%, a newly introduced BPM in the X-JtoBI system, was found in both speech type of APS and SPS.

```
<Talk RecordingDate="2000-01-01" SpeakerID="0001"
BirthDate="1950-01-01" WaveFilePath="wav/S03f0119.wav">
  <AP APID="304">
    <XJToBITone Time="244.098746" F0="210.791"
    ToneClass="IBT">%L</XJToBITone>
    <XJToBITone Time="244.114585" F0="216.325"> A
    </XJToBITone>
    <XJToBITone Time="244.485417" F0="151.82"
    ToneClass="FBT">L%</XJToBITone>
    <SUW SUWPOS="代名詞" SUWDictionaryForm="イツ"
    SUWLemma="何時" LexicalAccPos="1" APID="271"
    Channel="L" IPUStartTime="244.050"
    IPUEndTime="245.009" LUWPOS="代名詞"
    LUWDictionaryForm="イツ" LUWLemma="何時">
      <Mora MoraEntity="イ" PerceivedAccPos="1">
        <Phoneme PhonemeEntity="i">
          <Phone PhoneEntity="i"
          PhoneStartTime="244.073871"
          PhoneEndTime="244.154540">
            <XJToBILabelTone Time="244.092331"
            F0="210.791">%L </XJToBILabelTone>
            <XJToBILabelTone Time="244.114585"
            F0="216.325"> A </XJToBILabelTone>
          </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>
    <SUW SUWPOS="助詞" SUWDictionaryForm="モ"
    SUWLemma="も" APID="304" Channel="L"
    IPUStartTime="244.050" IPUEndTime="245.009"
    LUWPOS="助詞" LUWDictionaryForm="モ"
    LUWLemma="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>
    <SUW SUWPOS="助詞" SUWDictionaryForm="ノ"
    SUWLemma="の" APID="304" Channel="L"
    IPUStartTime="244.050" IPUEndTime="245.009" LUW
    LUWPOS="助詞" LUWDictionaryForm="ノ" LUWLemma="
    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">
            <XJToBILabelTone Time="244.494613" F0="151.82"
            ToneClass="FBT"> L%" </XJToBILabelTone>
          </Phone>
        </Phoneme>
      </Mora>
    </SUW>
  </AP>
</Talk>
```

Figure 5: Example of reorganized XML document.

Table 2 Analysis of AP unit in each speech type.

Speech Type	N. Speech	N. AP units	Duration of AP unit([ms])
APS	62	51109	675.3
SPS	106	74246	687.7
Reading	6	6358	745.3
Total	174	131713	685.7

Table 3 Frequency of BPMs in each speech type. (Numbers in () denotes ratio in all BPMs [%])

Speech Type	L%+H%	L%+HL%	L%+LH%
APS	12926(85.1)	2152(14.1)	95(0.8)
SPS	9706(58.7)	6572(39.7)	237(1.6)
Reading	1453(97.9)	30(2.0)	1(0.0)
Total	24085(72.6)	8754(26.3)	333(1.1)

Table 4 Occurence of the FR version of the L%H% BPM. Asterisk at the beginning of words shows that the expression has‘adversative’ meaning.

Last 2 words of AP	FR	L%H%	%FR
*desu ga	132	553	23.9
*keredo mo	121	436	27.8
desu ne	116	1213	9.6
*desu kedo	61	167	36.5
*kedo mo	40	150	26.7
de wa	33	217	15.2
yo ne	30	147	20.4
ni wa	20	165	12.1
no ga	19	146	13.0
te wa	18	168	10.7

Table 5 Occurence of the FR version of the L%HL% BPM. Asterisk at the beginning of words shows that the expression has‘adversative’ meaning.

Last 2 words of AP	FR	L%HL%	%FR
*keredo mo	95	524	18.1
*desu ga	88	431	20.4
*desu kedo	59	256	23.0
*kedo mo	37	206	18.0
*desu ne	26	500	5.2
to ka	6	158	3.8
no wa	6	138	4.3
ni wa	5	64	7.8
*da kedo	5	30	16.7
desu kara	5	63	7.9

Lastly, tables 4 and 5 show the relationship between the so-called ‘floating rise’ (FR hereafter) BPM and the morphological properties at the end of APs. FR is a variant of the L%H% and L%HL% BPMs; in the FR versions, the beginnings of the pitch rise locate in the penult morae rather than the last morae that were the location of the beginning of the pitch rise in the normal variants.

In these tables the numbers of occurrence of the FR variants, the total occurrence of the BPM in question, and, the rate of FR variants are shown as a function of the last 2 words of the APs. Each table shows the top 10 word sequences where the FR variant was the most frequent. It is interesting to note that half of the word sequences in these tables share the semantic property of being ‘adversative’ expression (like ‘but’ or ‘yet’), and the same word sequences showed the highest rates of FR.

5 Conclusion

This paper describes design and implementation issues of the XML documents which are effective for the study of prosodic structure in the “*Corpus of Spontaneous Japanese*”. The described XML documents will be released in the future.

References：

K. Maekawa, H. Kikuchi, Y. Igarashi, and J. Venditti. 2002. *X-JToBI: An extended J_ToBI for spontaneous speech*. In *proceedings of “the 7th International Congress on Spoken Language Processing (ICSLP2002)”*, vol.3, pages 1545-1548, Denver.

K. Maekawa. *Corpus of Spontaneous Japanese: Its Design and Evaluation*. In *proceedings of “ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)”, pages* 7-12. Tokyo.

K. Maekawa, H. Kikuchi, and W. Tsukahara. 2004. *Corpus of Spontaneous Japanese: Design, Annotation and XML Representation.* In *proceedings of* “International Symposium on Large-scale Knowledge Resources (LKR2004)”, pages 19-24, Tokyo.

H. Kikuchi and K. Maekawa. 2003. *Performance of Segmental and Prosodic Labeling of Spontaneous Speech*. In proceedings of “*ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*”, pages 191-194. Tokyo.

J. Venditti. 1995. *Japanese ToBI Labelling Guidelines.* Manuscript. Ohio State University, USA.

K. E. A. Silverman, et al. 1992. *TOBI: A standard for Labeling English Prosody*. In Proceedings of “*the 1992 International Conference on Spoken Language Processing*”, Vol. 2. pages 867-870. Banff, Canada.