Use of XML and Relational Databases for Consistent Development and Maintenance of Lexicons and Annotated Corpora

Masayuki Asahara*, Ryuichi Yoneda*, Akiko Yamashita*, Yasuharu Den†, Yuji Matsumoto*

* Graduate School of Information Science, Nara Institute of Science and Technology, Japan
8916-5 Takayama, Ikoma, Nara, 630-0101, JAPAN
{masayu-a, ryuich-y, akiko-ya, matsu}@is.aist-nara.ac.jp

† Faculty of Letters, Chiba University
1-33 Yayoicho, Inage-ku, Chiba 263-8522, JAPAN
den@L.chiba-u.ac.jp

Abstract

In this paper, we present a use of XML and relational database for developing and maintaining Japanese linguistic resources. In languages that do not provide word delimitation in texts (e.g. Chinese and Japanese), consistent delimitation definition of words in a lexicon is a critical issue to build POS tagged corpora. When we change the definition of word delimitation in the lexicon, we need to modify the tagged corpora to make them consistent with the lexicon. We propose a use of relational database to perform these modifications in tandem. Hence, in the Japanese language, there are several standards for word delimitation definition. To accommodate more than one definition of word delimitation, we compose a compounding word lexicon in the database. The compounding word lexicon includes dependency structures of compounding words.

1. Introduction

In languages that do not provide word boundaries in texts (e.g., Chinese and Japanese), consistent definition of word formation in the lexicon is crucial to natural language processing. Moreover, there is no consensus on the definition of word delimitation in annotated corpora for these types of languages. The linguistic database, to keep multiple definitions of word delimitation, is much in demand as the requirement for word delimitation tends to differ according to application areas.

To make matters worse, nominal compounding word of Japanese consist of various kinds of nouns with prefixes and suffixes. In many cases, nominal compounding word form different types of nouns from their constituent nouns, such as proper nouns or technical terms. Since complicated compounding words have syntactic structure within themselves, such structure should be specified in the linguistic resources.

In this paper, we present a use of XML and relational database for consistent development and maintenance of linguistic resources of Japanese language, that are lexicons and annotated corpora. The database is also used for building statistical models for Japanese language analyzers. This paper describes two major facilities of our current implementation.

First, the system maintains two kinds of data in the database: lexicons and annotated corpora. The lexicons provide grammatical information as well as canonical form and the construction of the entry words. In order to keep consistency of word delimitation definition in the lexicon and the annotated corpus, we define the relationship between the words in those two linguistic resources through the relational database. It is necessary to ensure changes in one resource to be reflected to another. Such maintenance is semi-automatically achieved by the use of relational database.

Second, we have to cope with multiple definition of word delimitation, as different applications call for different definition of word delimitation. To accommodate more than one definition of word formation in the lexicon, we facilitate a hierarchical definition of word formation of compounding words in the database. The lexicon with the most fine-grained word definition is taken as the base lexicon, and the compounding words are defined as binary tree structure consisting of the base entries. Users may select the grain size of the words so as to meet their requirements. Meanwhile, there are many contracted words in spoken language. Contracted words are derived from two or more words. We treat the original forms of contracted words too in this framework.

In section 2., we represent word delimitation definition of Japanese used in our database. In section 3., we show our database schema for Japanese POS tagged corpora. In section 4., we present our compounding words lexicon. Finally, we give conclusions.

2. Word delimitation definitions of Japanese

In this section, we present word delimitation definitions of Japanese language which we use in our database.

In Japanese language, there are several standards of word delimitation definition. Unfortunately, there is no consensus on the definition of word delimitation. Each word delimitation definition has its own POS tagset. For Japanese natural language processing, two POS tagsets are widely used, which are adopted by Japanese morphological analyzer. One is Masuoka-Takubo POS tagset(Takashi Masuoka and Yukinori Takubo, 1992) adopted by morphological analyzer *JUMAN*(Sadao Kurohashi and Makoto Nagao, 1999). The other is IPA POS tagset(Yuji Matsumoto and Masayuki Asahara, 2001) adopted by morphological

analyzer *ChaSen*(Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka and Masayuki Asahara, 2002). *ChaSen* output is much fine grained than *JUMAN* output. Table 1 shows the difference of word delimitation definitions between *JUMAN* and *ChaSen*.

JUMAN	積極的に	融資	して	いる
ChaSen	積極 | 的 | に	融資	し | て	いる

Figure 1: The Difference of Word Delimitation – *JUMAN* v.s. *ChaSen*

The lexicon, which has multiple definitions of word delimitation definition, is in demand. Our database is developed to maintain lexicon and corpora with UniDic(Yasuharu Den, Takehito Utsuro, Atsushi Yamada, Masayuki Asahara and Yuji Matsumoto, 2002) POS tagset and word delimitation definitions. UniDic introduces four layers of word delimitation definitions to cover demands of many domains. Below, we show base idea and word delimitation definitions of UniDic.

2.1. Base Idea of UniDic

In the field of natural language processing, machine readable lexicons are the foundation resources and are developed by many organizations. Many lexicons are developed for written languages. Whereas lexicons for spoken languages be far in the rear of ones for written languages.

In view of the present situation, UniDic Project aims at developing electrical lexicon which is acceptable to phonetician, linguist and computational linguist. The project set the two goals as follows:

- electrical lexicon for linguistic research
  consistent word delimitation definition and sharable POS tagset
- electrical lexicon for speech and acoustic processing
  reading, pronunciation and accent informations

Such an ideal dictionary is unrealistic. Especially, the word delimitation definition cannot be sharable among separate domains. For example, when we think about the matter in terms of morphological analysis, prefix and suffix should be defined as one word. Nevertheless, when we put Japanese reading or accent information for these word, such words should be defined as part of compounding word. To cope with the dilemma, we introduce four layers of word delimitation definitions.

2.2. Word Delimitation Definitions of UniDic

There is no word delimitation definition which can be sharable among separate fields. In UniDic scheme, we defined four word delimitation definitions as follows:

- Layer 0: Morpheme
  Layer 0 defines not “word” but “morpheme”. Then, POS tag cannot be defined for this layer.
- Layer 1: Simple Word
  Layer 1 defines the smallest unit of words without compounding. We defines the layer 1 as basis of compounding words.
- Layer 2: Compounding Word
  The words of layer 2 are defined by layer 1 as following two rules:
  - match following regular expression:
    `(prefix)*(Noun|Adjective|Verb)+(suffix)*`
  - and for each adjacent word pair, left side word has dependency with right side word.
- Layer 3: Nominal Compounding
  Named entity, idiom and collocation.

	若	手	の	会	は	積	極	的	だ
Layer 1	Noun		PostP	Noun	PostP	Noun		Suffix	AuxV
Layer 2	Noun		PostP	Noun	PostP	Adjective			AuxV
Layer 3	Noun				PostP	Adjective			AuxV

Figure 2: The Differences of Word Delimitation – UniDic

Layer 0 defines morphemes. One Chinese character(Kanji/HanZi) is defined as one morpheme. For morphemes, which are composed by Katakana or Hiragana characters, are segmented into minimal units. The unit in layer 0 has no POS tag information, because these are defined not as words but as morphemes.

Layer 1 defines base words to make compounding words. The words in layer 1 are minimal units which can put POS information. UniDic POS tagset is designed for the unit. Note that, POS tag names are defined in Japanese language. But, in this paper, we use English POS name translated from the Japanese original tag name.

A word of Layer 2 is composed by words of layer 1. Second rule restricts the word of layer 2 within a left branching structure. The composition comes from the fact that a right branching structure restricts “rendaku” phenomena[1] and accent moving(Haruo Kubozono, 1991). Figure 3 shows an example of *rendaku*. On one hand, the left branching structure of compounding is defined as one word in layer 2 definition, which has two *rendaku* phenomena. On the other hand, the right branching structure of compounding is defined as two words, then one *rendaku* is restricted because of the right branching. Then, the unit of layer 2 is suit for putting reading and accent informations.

Layer 2 is designed for not only speech and acoustic processing but also chunking. When we do chunking from layer 1, BIO model fits the branching structure. BIO model, which is widely used in NP or named entity chunking, is a model to put following tags into fine grained word sequence:

[1]The process voices an initial voiceless consonant of the second member in a certain class of compounding words.

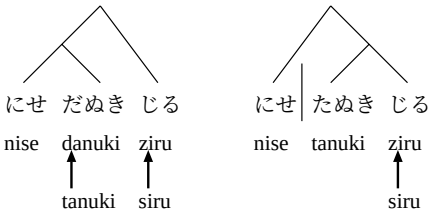

Figure 3: Dependency structures and “rendaku”

- *B* : the beginning of the unit
- *I* : the inside of the unit
- *O* : the outside of the unit

Figure 4 shows relation between branching structures and layer 2 chunks. The word, which do not match the regular expression, are put tag O. When a word match the regular expression, the word, which is left element of a subtree, is put tag B, otherwise the word is put tag I. In this sense, chunking based on BIO tag is suit for layer 2 definition.

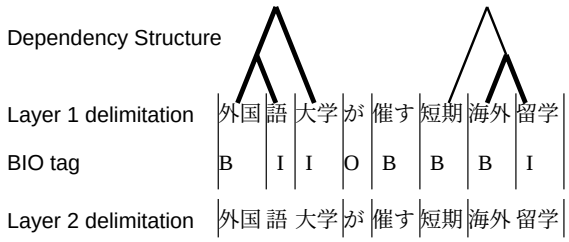

Figure 4: Layer 2 and BIO tag

Layer 3 defines much longer units like named entities, idioms and collocations. The definition has no relation with the dependency structure of word. Then, the relation between BIO tag and dependency structure is weakened. Figure 5 shows layer 3 chunks and BIO tags.

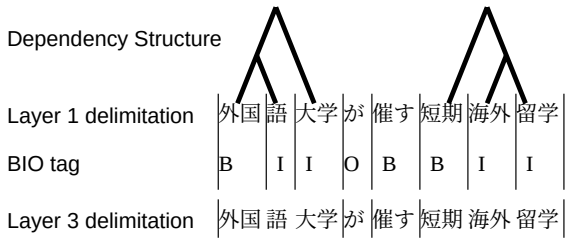

Figure 5: Layer 3 and BIO tag

3. Maintenance schema for Japanese POS tagged corpora

When we develop corpora based on UniDic POS tagset, we must maintain POS tagged corpus for each layer with separate word delimitation definition. POS tag informations of the layers will be overlapping on original texts.

Stand-off annotation(Henry S. Thompson and David McKelvie, 1997) enables us to cope with the overlapping problem. Stand-off annotation is the method separating markup from the material marked up.

First, we present former works of stand-off annotation. Second, we present our database schema for Japanese POS tagged corpora.

3.1. Stand-off Annotation

Stand-off annotation framework is initially formalized in a field of SGML. We illustrate the framework by the example of Henry et al. (Henry S. Thompson and David McKelvie, 1997).

Consider following marking sentence structure:

```
..
<w id='w12'>Now</w><w id='w13'>is</w><w id='w14'>the</w>
..
<w id='w27'>the</w><w id='w28'>party</w><w id='c4'>.</w>
..
```

We can mark sentences in a separate document as follows:

```
..
<s xml-link='simple' href="#ID(w12)..ID(c4)"></s>
<s xml-link='simple' href="#ID(w29)..ID(c7)"></s>
..
```

Their application enables us to see this document collection as a single stream with the words nested inside the sentences:

```
..
<s>
<w id='w12'>Now</w><w id='w13'>is</w><w id='w14'>the</w>
..
<w id='w27'>the</w><w id='w28'>party</w><w id='c4'>.</w>
</s>
<s>
..
</s>
..
```

They showed three reasons why separating markup from the material marked up:

1. the base material cannot copy to introduce markup because of read-only and/or very large
2. the markup may involve multiple overlapping hierarchies
3. distribution of the base document may be controlled, but the markup is intended to be freely available

Our strong reason to introduce stand-off annotation is to cope with overlapping problem of Japanese POS information.

Stand-off annotation is widely introduced in the field of corpus maintenance(Nancy Ide and Patrice Bonhomme and Laurentj Romary, 2000). Bird et al. (Steven Bird and Mark Liberman, 2001) proposed *Annotation Graphs* which allows to encode in a same structure various information. In XML framework, standoff annotation is formalized as Xpointer (Steve DeRose and Eve Maler and Ron Daniel Jr., 2001).

3.2. Stand-off Annotation Framework for Japanese POS Tagged Corpora

We introduce stand-off annotation framework for Japanese POS tagged corpora in order to solve three problems. First is to maintain POS tagged corpora based on several word delimitation definitions. Second is to permit coexistence POS information and other phonetic informations in corpora. Third is to keep consist word delimitation definition between lexicons and corpora.

Figure 6 shows stand-off annotation for Japanese POS tagged corpora. We defined character ID for each characters in the original text. POS tag information is defined by "Begin ID", "End ID" and "POS Tag " information. Original text and POS tag information are separated in different tables in our relational database.

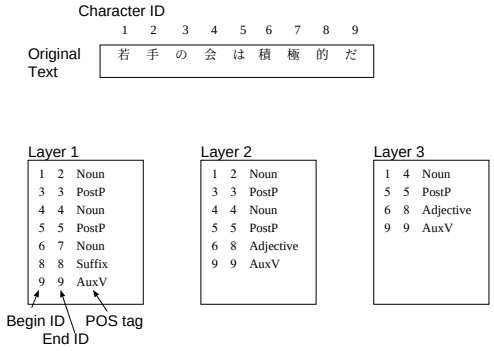

Figure 6: Stand-Off Annotation for several word delimitation definitions

Stand-off framework enables us to keep multiple word delimitation definition on one raw text. Moreover, we can add other information to the corpora without overlapping restriction.

When we maintain Japanese POS tagged corpora, it is difficult to keep consistency of word delimitation definition. Against this problem, we introduce links between corpora and lexicons. Figure 7 shows the links between corpora and lexicons. The links make the word delimitation in tagged text consistent with the lexicon.

Nota that, practical word lexicon keeps more information – form (i.e. Japanese reading), pronunciation, conjugation informations. We classify conjugation informations into *CTYPE* and *CFORM*. *CTYPE* stands for conjugation type how the word conjugate. *CFORM* stands for conjugation form. On the relational database, we define word ID and CFORM ID as primary key. Then, tag information is represented by pointers to the primary key. We used character ID as the anchor of pointers. When we maintain corpora for spoken language, the anchors will be defined as time and track on audio data.

4. Compounding Word Lexicon – relationships among multiple word delimitation definitions

In the preceding section, we present how to deal with multiple word delimitation definition for one raw text. In this section, we present how to deal with the relationships

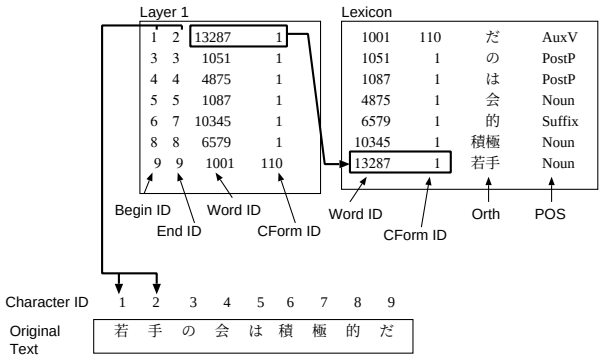

Figure 7: Linking between corpora and lexicons

among these definitions. We maintain the relationships as compounding words lexicon with dependency structure. First, we show the classification of compounding words based on dependency structure of the words. Second, we present the database schema to deal with the dependency structure. Third, we present our XSLT usage to extract lexicons.

In the compounding word lexicon, we do not mention about morpheme layer 0. Because maintaining the relationships between layer 0 and layer 1 is cumbersome but no use for many users. We use layer 1 as basis and annotate relationships with layer 2 or layer 3.

4.1. Classification of Compounding Words

To keep the relationship between longer and shorter word delimitation definition, we take notice of dependency structure of compounding word. We suppose that compounding words have binary dependency structures. Then, the relationships between longer and shorter word are defined as parental relations on the dependency tree. We defined the classification of compounding words from the dependency tree. Table 1 shows the categories for the classification.

Table 1: Category for Compounding Words

Category	Specification
B	Basis
N	Compounding
P	Parallel Compounding
C	Contracted Word
X	Fragment of Compounding Word
S	Collocation Pair without Dependency

Below, we exemplify the definitions of categories.

4.1.1. Binary Structure of Compounding Words

We represent a compounding word with the constituent words as a binary dependency structure. When left side constituent word has dependency with right side one, we put category "N" for compounding word. Then, we annotate pointers to two constituent words. When the constituent words can be also divided into constituent words,

we recursively define the parental relationship for the constituent words.

To examine the descendant on the dependency structure via database, we can discriminate whether the compounding word has left and/or right branching structure. The discrimination is crucial information to identify layer 2 words of UniDic schema. Figure 8 exemplifies left and right branching structure of compounding words.

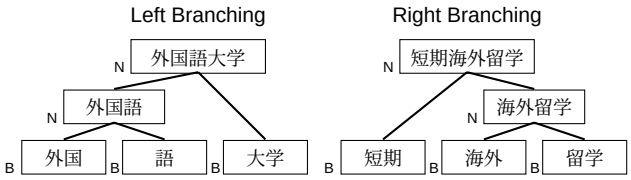

Figure 8: Left and Right Branching of Compounding Words

4.1.2. Ternary Tree of Compounding Words

Because of parallel structure, ternary structures can occur in the dependency structure of compounding words. Nevertheless, we restrict within binary structures to represent compounding words. Then, we extract ternary structure into left branching binary structure. The extracted nodes are put category "P". Figure 9 shows binarization of ternary tree structure.

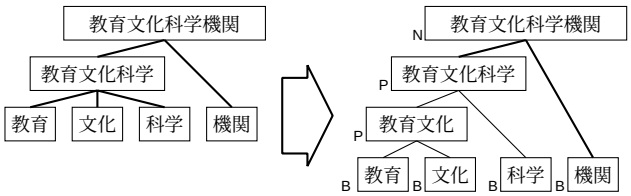

Figure 9: Binarization of Ternary Structure

4.1.3. Contracted Words

Contracted words are peculiar to spoken language in Japanese. Contracted expression 「ちゃう」"Chau" is composed by 「て」"Te" and 「しまう」"Shimau". Our database treat these contracted expressions with category "C" with original words. Figure 10 exemplify the expression "Chau".

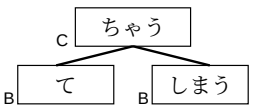

Figure 10: Contracted Word 「ちゃう」 – "Chau"

4.1.4. Collocation, Idiom and Named Entity

When we annotate for collocations, idioms and named entities, some constituent words become fragments of compounding word, which we do not want to register into any layer. In that case, we put category "X" for the fragments. Figure 11 is an example of case particle collocation. The fragment「ついて」"tsuite" is not registered into any layer.

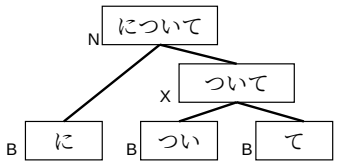

Figure 11: Case Particle Collocation 「について」 – "nitsuite"

4.1.5. Collocation without Dependency

There are some collocations without dependency in the lexicon. These words are registered because of accuracy of Japanese morphological analysis.

Figure 12 shows an example of numeral suffix 「号室」"Goushitsu". The constituent words of this collocation have neither dependency nor parallel relationship. Nevertheless, this collocation has a trigger role of word formation for numeral expression, then morphological analyzer contains such a word. We put category "S" for these words.

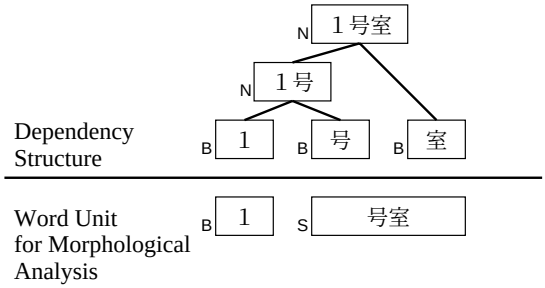

Figure 12: Suffix 「号室」"Goushitsu"and Dependency Structure

4.2. Compounding Lexicon on Relational Database

We presented the methods to represent word relationships by pointers for constituents. We keep these pointers on relational database as they stand. Table 2 shows compounding word table on our relational database. To simplify, we represent "Word ID" for the anchor of pointers. On our practical database, the anchor is a pair of "Word ID" and "CFORM ID" to transact Japanese conjugation.

Annotators are using GUI like figure 13 to annotate compounding relationships. In left below or right below fields, the GUI list up possible prefix or suffix matching words. Annotators select most suitable segmentation by top down strategy.

4.3. Extraction of Constituent Words

We use XML framework for distribution. Original lexicon is exported from relational database into XML file as figure 14[2]. In the XML file, pointers to composed words

[2]Our practical lexicon contains much information: reading, pronunciation, conjugation type and conjugation form.

Table 2: Compouding Lexicion on Relational Databases

Word ID	Category	Orthography	Left Const.	Right Const.
1	N	教育文化科学機関	2	7
2	P	教育文化科学	3	6
3	P	教育文化	4	5
4	B	教育	-	-
5	B	文化	-	-
6	B	科学	-	-
7	B	機関	-	-

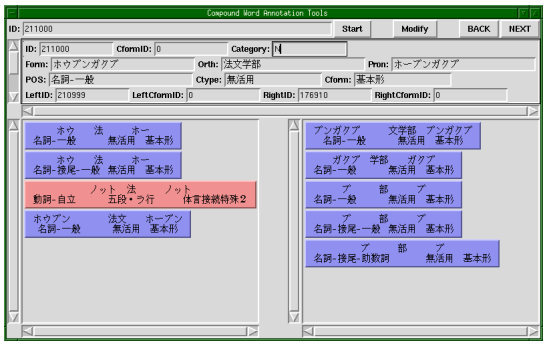

Figure 13: GUI for compounding word annotation

are left as these stand. We use XSLT to extract constituent words from the pointers. An example of xsl file is figure 15. Using XSLT engines, lexicon is resolved as figure 16.

```
<?xml version="1.0" encoding="EUC-JP"?>
<ipacomp>
<word lexid="1" cformid="0" category="N">
  <orth>短期海外留学</orth>
  <compleft lexid="3" cformid="0" />
  <compright lexid="2" cformid="0" />
</word>

<word lexid="2" cformid="0" category="N">
  <orth>海外留学</orth>
  <compleft lexid="4" cformid="0" />
  <compright lexid="5" cformid="0" />
</word>

<word lexid="3" cformid="0" category="B">
  <orth>短期</orth>
</word>

<word lexid="4" cformid="0" category="B">
  <orth>海外</orth>
</word>

<word lexid="5" cformid="0" category="B">
  <orth>留学</orth>
</word>
</ipacomp>
```

Figure 14: Example of compounding word lexicon

```
<?xml version="1.0" encoding="EUC-JP"?>
<ipacomp>
<word category="N">
 <orth>短期海外留学</orth>
 <word category="B">
  <orth>短期</orth>
 </word>
 <word category="N">
  <orth>海外留学</orth>
  <word category="B">
   <orth>海外</orth>
  </word>
  <word category="B">
   <orth>留学</orth>
  </word>
 </word>
</word>

<word category="N">
 <orth>海外留学</orth>
 <word category="B">
  <orth>海外</orth>
 </word>
 <word category="B">
  <orth>留学</orth>
 </word>
</word>

<word category="B">
 <orth>短期</orth>
</word>

<word category="B">
 <orth>海外</orth>
</word>

<word category="B">
 <orth>留学</orth>
</word>
</ipacomp>
```

Figure 16: Extracted compounding word lexicon

5. Conclusions

In this paper, we presented database schema for Japanese POS corpora and lexicons. Since texts in Japanese language are written without word boundaries, consistent definition of word delimitation is crucial to build POS tagged corpora. To keep the consistency, we used an lexicon that maintains the definition of word delimitation and

```
<?xml version="1.0" encoding="EUC-JP"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:output method="xml" encoding="EUC-JP" />
<xsl:template match="ipacomp">
 <ipacomp>
  <xsl:apply-templates/>
 </ipacomp>
</xsl:template>

<xsl:template match="word">
  <word>
    <xsl:attribute name="category">
      <xsl:value-of select="@category"/>
    </xsl:attribute>
  <xsl:apply-templates/>
  </word>
</xsl:template>

<xsl:template match="orth">
    <orth><xsl:value-of select="."/></orth>
</xsl:template>

<xsl:template match="compleft">
  <xsl:variable name="leftlexid" select="@lexid"/>
  <xsl:apply-templates select="//word[@lexid=$leftlexid]"/>
</xsl:template>

<xsl:template match="compright">
  <xsl:variable name="rightlexid" select="@lexid"/>
  <xsl:apply-templates select="//word[@lexid=$rightlexid]"/>
</xsl:template>
</xsl:stylesheet>
```

Figure 15: Example of xsl file

defined the relationship between the words in the lexicon and those appearing in the tagged text. Because different application domains use different definition of word delimitation, we designed multi-stratal word delimitation definitions to cover demands of many domains. Then, word delimitation definitions will be overlapped on raw texts. We introduced stand-off annotation framework to cope with overlapping problem. To maintain multiple definitions of word delimitation, we composed tables for compound words. The tables also store syntactic structures of compound words, which is also useful to know phonological realization of the words.

We used mainly relational database framework to maintain linguistic resources and we used XML framework only to distribute and transform resources, because relational database cannot do them. For example, whereas the query language SQL in relational database framework cannot retrieve pointers recursively, XSLT can do the recursion.

6. References

Haruo Kubozono. 1991. *The Organization of Japanese Prosody*. Studies in Japanese Linguistics. kurosio Publishers, Inc.

Henry S. Thompson and David McKelvie. 1997. Hyperlink semantics for standoff markup of read-only documents. In *SGML Europe '97*.

Nancy Ide and Patrice Bonhomme and Laurentj Romary. 2000. XCES: An XML-based Encoding Standard for Linguistic Corpora.

Sadao Kurohashi and Makoto Nagao. 1999. JUMAN Ver. 3.61. http://www-nagao.kuee.kyoto-u.acjp/.

Steve DeRose and Eve Maler and Ron Daniel Jr. 2001. XML Pointer Language (XPointer) Version 1.0. http://www.w3.org/TR/xptr/.

Steven Bird and Mark Liberman. 2001. A formal frame work for linguistic annotation. *Speech Communication*, 33:23–60.

Takashi Masuoka and Yukinori Takubo. 1992. *Kiso Nihongo Bunpou – kaitei-ban –*. kurosio Publishers, Inc.

Yasuharu Den, Takehito Utsuro, Atsushi Yamada, Masayuki Asahara and Yuji Matsumoto. 2002. Design of an Electric Dictionary Suitable for Spoken Language Research. In *Proceedings of the Second Spontaneous Speech Science and Technology Workshop*, pages p.p. 39–46.

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka and Masayuki Asahara. 2002. Morphological Analyzer version 2.2.9. http://chasen.aist-nara.ac.jp/.

Yuji Matsumoto and Masayuki Asahara. 2001. IPADIC Users Manual.