Language, Information and Computation(PACLIC 11), 1996, 219-228

Markup of Korean Dictionary Entries

Beom-mo Kang
Korea University
bmkang@kuccnx.korea.ac.kr

Abstract

Dictionary markup (encoding) is one of the concerns of TEI (Text Encoding Initiative), an international project for text encoding. In this paper, we investigate ways to use and extend TEI encoding scheme for the markup of Korean dictionary entries. Since TEI suggestions for dictionary markup are mainly for western language dictionaries, we need to cope with problems to be encountered in encoding Korean dictionary entries. We try to extend and modify the TEI encoding scheme in the way suggested by TEI. Also, we restrict the content model so that the encoded dictionary might be viewed more as a database than as a simple computerized, originally printed, dictionary.

1. Introduction

TEI (Text Encoding Initiative) is an international project which aims to provide some guidelines for encoding various kinds of texts in electronic forms. It conforms to the ISO standard of Standard Generalized Markup Language (ISO 1986). The primary result of this project, published in 1994, Guidelines for Electronic Text Encoding and Interchange (TEI P3) edited by Burnard and Sperberg-McQueen, covers many kinds of texts in depth. Dictionary markup is one of them. In this paper, we investigate ways to use and extend TEI encoding scheme for the markup of Korean dictionary entries. To accomplish this objective, it is necessary to consider the logical structure of the Korean dictionary entries.

Although TEI suggestions for dictionary encoding (markup) is very comprehensive to cover various kinds of dictionaries, its original commitment is to consider only western language dictionaries (Ide and Veronis 1995: 168). We need to cope with problems to be encountered in encoding Korean dictionary entries in conformance with TEI. We try to extend and modify the TEI encoding scheme in the way suggested by TEI. In addition, we restrict the content model so that the encoded dictionary might be viewed more as a database than as a simple computerized (originally printed) dictionary.[1][2]

2. Top Level Elements of a Dictionary Entry

Before presenting the model of Korean dictionary entries, let us go over the basic dictionary scheme provided by TEI.

(1) TEI P3: Basic Structure of a Dictionary

```
<text>
<body>
    <entry> ... </entry>
    <entry> ... </entry>
    <superEntry>
            <entry> ... </entry>
            <entry> ... </entry>
            ...
    </superEntry>
    <entry> ... </entry>
    <entry>
            ...
            <hom> ... </hom>
            <hom> ... </hom>
    </entry>
    ...
</body>
</text>
```

The text of a dictionary consists of a number of <entry> elements, each of which can consist of a number of <hom>(i.e. homonym) elements.

Some dictionary elements, called dictionary top level elements can appear at the level of entry, homonym, and sense, disregarding superentry element for the moment. The following DTD definitions show this:

(2) TEI P3: Print Dictionaries DTD Top Level

```
<!ENTITY % superentry 'INCLUDE' >
<![ %superentry; [
<!ELEMENT %n.superentry;  - O  ((%n.form)?, (%n.entry)+)          >
...]]>

<!ENTITY % entry 'INCLUDE' >
<![ %entry; [
<!ELEMENT %n.entry;       - O    (%n.hom; | %n.sense; |
                                 %m.dictionaryTopLevel)+
                                                +(anchor)     >
...]]>

<!ENTITY % hom 'INCLUDE' >
<![ %hom; [
<!ELEMENT %n.hom;         - O    (%n.sense; |
                                 %m.dictionaryTopLevel)*
                                                 -(entry)    >
...]]>

<!ENTITY % sense 'INCLUDE' >
<![ %sense; [
<!ELEMENT %n.sense;       - -    (%n.sense; | %m.dictionaryTopLevel
                                 | %m.phrase | #PCDATA)*          >
...]]>
```

The definitions of <entry>, <hom>, and <sense> include dictionary top level elements, which are defined as follows:

```
(3) TEI P3: Dictionary Top Elements

<!ENTITY % x.dictionaryTopLevel ''                                    >
<!ENTITY % m.dictionaryTopLevel '%x.dictionaryTopLevel def | eg
          | etym | form | gramGrp | note | re | trans | usg | xr'  >
```

As defined by m.dictionaryTopLevel entity reference, these top level elements are <def>, <eg>, <etym>, <form>, <gramGrp>, <note>, <re>, <trans>, <usg>, and <xr>. For Korean dictionary entries, these 10 elements are all are needed. Overall, some elements can be used as they are provided by TEI, and others need to be modified according to specific needs of a Korean dictionary, in the way to be specified from now on.

First of all, the following dictionary top level elements elements are to be used without modifications for the markup of Korean dictionary entries: 1) <def> for definition; 2) <trans> for translation (not used in a monolingual dictionary); 3) <eg> for usage examples; 4) <note> for any kind of notes; 5) <re> for related words. Other dictionary top elements are to be modified as follows.

For the <form> element, besides usual <orth> (orthography) and <pron> (pronunciation) elements, we need a sub-element which contains a form showing long vowels and a major morphological constituent break, which is usually marked for entries in Korean dictionaries. We call this element <lenHyph>. In the following example, we provide the part tagged by <lenHyph> along with parts of <orth> and <pron>.

```
(4) Eg. 용감스럽다 'courageous'
    <form>
            <orth>용감스럽다</orth>
            <lenHyph>용:감-스럽다</lenHyph>
            <pron>-따</pron>
    </form>
```

Another top level element is <gramGrp> for grammatical information of an entry. Besides <pos> and <subc> elements provided by TEI P3, we seem to need some elements which specify the kinds of irregular inflection for some verbs and exemplary inflected forms. These are marked by tags of <irreg> and <irrForm>, as shown in the following example.

```
(5) Eg. 춥다 'cold'
    <gramGrp>
            <pos>형</pos>
            <irreg>ㅂ 불</irreg>
            <irrForm>추우니, 추워</irrForm>
    </gramGrp>
```

The usage note of an entry (<usg>) can contain academic domains (special fields) in which this item is used, other domains (such as marking for 'old Korean'), and dialect areas, which are prominent in Korean dictionaries. These

can be encoded with new tags defined within <usg>. They are <domAca>, <domEtc>, and <dialArea>. They can be used as follows.

(6) Eg. 수선화 'daffodil'
<usg><domAca>식물</domAca></usg>

Eg. 소내기 'shower'
<usg><dialArea>평안</dialArea></usg>

Eg. 가람 'river'
<usg><domEtc>옛말</domEtc></usg>

Since the content and format of etymology in a Korean dictionary is constrained in certain ways, some modifications of the DTD definitions of the <etym> element are needed. For this element, we add a new attribute 'hdType' whose value should be one of: 'hj' (for hanja, i.e. of Chinese origin: content being given in Chinese characters), 'foreign' (of any other foreign origin), and 'kor' (of Korean origin proper).

(7) Eg. 라디오 'radio'
<etym hdType=foreign><lang>영</lang>radio</etym>

Eg. 사고 'thought'
<etym hdType=hj>思考</etym>

Eg. 샹송 'chanson'
<etym hdType=foreign><lang>프</lang>chanson</etym>

For a limited number of kinds of cross reference in a Korean dictionary, we can define various empty elements which mark the kinds of cross reference to be used in the dictionary. Among them are 'synonym', 'antonym', 'long form', 'short form', 'honorific form', etc. One of these elements should be used in the first part of <xr> element. In the first example provided below, a tag specifying antonymy is used.

(8) Eg. 성공 'success'
<xr><xrant><ref>실패</ref></xr>

Eg. 바지직
<xr><xrlarge><ref>부지직</ref></xr>
<xr><xrstr><ref>빠지직</ref></xr>

3. Modifying the TEI DTD

Until now, we have seen what (dictionary top) elements need to be added or modified for Korean dictionary entries. From now on, I will show how, in order to accommodate the above discussions, we can modify the TEI DTD along the way suggested by TEI dtd extension mechanism. In concrete, these modifications amount to editing a TEI.exntenstions.ent file and a TEI.extensions.dtd file.

First, <lenHyph> (a form element), and <irreg> and <irrForm> (grammatical information elements) are treated by modifying "x.formInfo" and "x.gramInfo" entities so that they can contain needed element names.

```
(9) <lenHyph>, <irreg>, <irrForm>
▶TEI.extensions.ent file
    <!ENTITY % x.formInfo 'lenHyph |'>
    <!ENTITY % x.gramInfo 'irreg | irrForm |'>

▶TEI.extensions.dtd file
    <!ELEMENT lenHyph - -      (%phrase.seq;)>
    <!ATTLIST lenHyph          %a.global;
                               %a.dictionaries; >

    <!ELEMENT irreg - - (%phrase.seq;)>
    <!ATTLIST irreg            %a.global;
                               %a.dictionaries; >

    <!ELEMENT irrForm - - (%phrase.seq;)>
    <!ATTLIST irrForm          %a.global;
                               %a.dictionaries; >
```

<lenHyph> is a subelement of <form>, which is defined in the TEI DTD using m.formInfo entity. So, x.formInfo entity is defined as containing lenHyph. Similarly, <irreg> and <irrForm>, subelements of <gramGrp> is added as part of x.gramInfo entity.

Usage specifics, such as <domAca>, <domEtc>, and <dialArea> are defined as subcomponents of <usg>, a new model of <usg> being defined.

```
(10) <domAca>, <domEtc>, <dialArea>
▶ TEI.extensions.ent file
    <!ENTITY % usg  'IGNORE'>

▶ TEI.extensions.dtd file
    <!ELEMENT %n.usg - O (%paraContent; | domAca | domEtc | dialArea)+ >
    <!ATTLIST %n.usg           %a.global;
                               %a.dictionaries;
         type          CDATA              #IMPLIED
         TEIform       CDATA              'usg'>

    <!ELEMENT domAca - -       (%phrase.seq;)>
    <!ATTLIST domAca           %a.global;
                               %a.dictionaries; >
    <!ELEMENT domEtc  - -      (%phrase.seq;)>
    <!ATTLIST domEtc           %a.global;
                               %a.dictionaries; >
    <!ELEMENT dialArea  - -    (%phrase.seq;)>
    <!ATTLIST dialArea         %a.global;
                               %a.dictionaries; >
```

The case of <etym> is a modification to attributes that this element has. As discussed above, a new attribute hdType is added and its value ranges over kor, hj, and foreign. The content model of <etym> is not changed at all.

(11) <etym>

▶ TEI.extensions.ent file

```
<!ENTITY % etym 'IGNORE'>
```

▶ TEI.extensions.dtd file

```
<!ELEMENT %n.etym;      - O    (%paraContent | %n.usg; | %n.lbl;
                               | %n.def; | %n.trans; | %n.tr; |
                               (%m.morphInfo) | %n.eg; | %n.xr;)* >

<!ATTLIST %n.etym;             %a.global;
                               %a.dictionaries;
         hdType                (kor | hj | foreign)    #REQUIRED
         TEIform               CDATA                   'etym'     >
```

For <xr>, cross reference words, we restrict them to less than 20 kinds shown below. Each kind is marked by an empty element specifying its kind, which appears in the first part of <xr> element.

(12) Kinds of Cross Reference

xrsee	See the word!
xrstd	Standard form
xrxstd	Nonstandard form
xrant	Antonym
xrsame	Same word
xrsyn	Synonym
xrshort	Short form
xrlong	Long form
xrstr	Strong form (onomatopoeia)
xrstr2	Another strong form
xrsoft	Soft form
xrlarge	Large form
xrsmall	Small form
xrhon	Honorific form
xrint	Intimate form
xrchg	Changed form
xrcfwd	Word for reference
xrvar	Variant form

To make these elements available, the element declaration of <xr> is changed as follows. Notice that newly added elements such as <xrsee>, <xrstd>, etc are defined as empty elements.

(13) <xr>

▶ TEI.extensions.ent file

```
  <!ENTITY % xr   'IGNORE'>
```

▶ TEI.extensions.dtd file

```
  <!ELEMENT %n.xr;   - -  ( (xrsee | xrstd | xrxstd | xrant | xrsame |
                            xrsyn | xrshort | xrlong |
                            xrstr2 | xrstr | xrsoft | xrlarge | xrsmall |
                            xrhon | xrint | xrchg | xrcfwd | xrvar | xrof),
                            (%paraContent | %n.usg | %n.lbl)* )   >
```

```
<!ATTLIST %n.xr;          %a.global;
                          %a.dictionaries;
        type               CDATA              #IMPLIED
        TEIform            CDATA              'xr'          >

<!ELEMENT xrsee      - O  EMPTY                             >
<!ATTLIST xrsee           %a.global;
                          %a.dictionaries; >
.....
```

4. SGML Parsing and Processing

Since TEI is an instance of SGML, Standard Generalized Markup Language, any TEI document can be processed by an SGML-aware software. Since the encoding of Korean dictionary entries we have pursued here is TEI-conformant, any document encoded in this way should be processed by such softwares. NSGMLS (Clark 1995), an SGML parser, and SoftQuad Panorama Pro (SoftQuad 1995), an SGML viewer, have been tested for such a purpose successfully.

Some sample Korean dictionary entries marked up (encoded) in this way follows.

(14) Sample Entries

```
<entry>
    <form><orth>심뇌</orth><lenHyph>심뇌</lenHyph>
          <pron>-뇌/-눼</pron></form>
    <etym hdtype=hj>心惱</etym>
    <gramGrp><pos>명</pos></gramGrp>
    <def>마음으로 근심하는 것. 또는, 그 근심.</def>
    <re><form><orth>심뇌하다</orth><lenHyph>심뇌-하다</lenHyph></form>
        <gramGrp><pos>동</pos><subc>자</subc><subc>타</subc>
                 <irreg>여불</irreg><irrForm>심뇌하여</irrForm></gramGrp>
    </re>
</entry>
<entry>
    <form><orth>파다</orth><lenHyph>파다</lenHyph></form>
    <gramGrp><pos>동</pos><subc>자</subc></gramGrp>
    <sense n='1'><def>(구덩이나 구멍을) 만들다.</def>
        <eg><q>땅을 괭이로 <oRef>.</q></eg></sense>
    <sense n='2'><xr><xrsyn><ref>새기다</ref>, <ref>만들다</ref></xr>
        <eg><q>도장을 <oRef>.</q></eg></sense>
</entry>
```

And what you see on the monitor when you view the file by Panorama Pro is as follows:

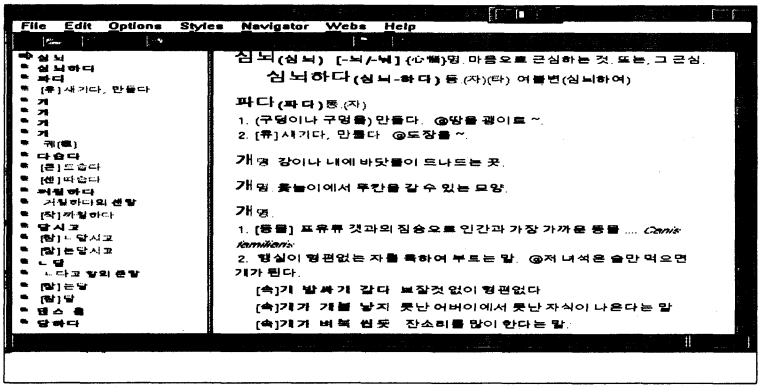

Figure 1 Dictionary Entries Viewed by Panorama Pro

As can be seen, many components of a dictionary entry can be highlighted by color and font variations and by some symbol prefixed.

Regarding the possibility of using the TEI method in dictionary compilation, that is, in the case of a lexicographer's writing of an entry, the encoding would be very difficult if we work on a text editor or a non-SGML editor and try to type the tags directly. Instead, we might adopt the following work procedure.

Lexicographers use a popular wordprocessor, which is non-SGML, but they write in a strict format and use a few special symbols which specify various dictionary elements. A program converts this file to a text file which conforms to the modified TEI DTD discussed above. Lexicographers view the content of the writing in a formated way by an SGML browser. If we find inconsistencies and faults in the file, we correct the wordprocessor file and convert it and view the content again, and so on. This procedure can be adopted at a later stage of dictionary compilation and is necessary if one considers an SGML publishing of the dictionary in the long run (Alschuler 1995). Also, it is a step toward making a machine-readable dictionary to be used for theoretical and computational linguistic research purposes (Boguraev 1994, Atkins and Zampoli 1994, Fillmore and Atkins 1994, Sinclair 1987).

5. Conclusion

In this paper, I have presented some ways to accommodate TEI encoding scheme for the markup of Korean dictionary entries. It is an SGML application, so it is in general an approach to structured information. Since dictionary entry is a typical instance of structured information, the SGML method can be applied most easily and successfully. However, since the structure of a dictionary entry is relatively flexible too, we need to be careful not to be too strict in modeling the structure. TEI's definition of dictionary top level elements seems to give enough flexibility to cope with such a problem. Other SGML approaches relating to the markup of Korean dictionaries, such as Choe (1996) and Choe & Choe (1996), which assume a more strict structure for a Korean dictionary entry, might encounter a problem in actual markup of Korean dictionary entries.

[Notes]

[1] Ide and Veronis (1995) and Chapter 12 (Print Dictionaries) of TEI P3 (Sperberg-McQueen and Burnard, eds., 1994) discuss three views of dictionaries: (a) the typographic view; (b) the editorial view; (c) the lexical view. The first view is concerned with the two-dimensional printed page while the last view is concerned with underlying information represented in a dictionary, without concern for its exact form. The editorial view is in between.

[2] The markup scheme presented here has been investigated in conjunction with an on-going project of Korean dictionary compilation at Korea University. Also, the TEI scheme has been used for the encoding of texts in "Korea-1 Corpus" compiled in this project (Kim and Kang 1996).

[References]

Alschuler, L. 1995. *ABCD..SGML: A User's Guide to Structured Information*, London: ITP.

Atkins, B.T.S. and A. Zampoli. eds. 1994. *Computational Approaches to the Lexicon*, Oxford: Oxford University Press.

Boguraev, B. 1994. Machine-Readable Dictionaries and Computational Linguistics Research, in A. Zampoli, N. Calzolari, and M. Palmer. eds., 119 - 54.

Choe, B. 1996. Markup of Machine Readable Dictionaries and Texts, presented at 1996 Symposium on the Standards for Information Processing in Koran, July 1996, Seoul.

Choe, B. and K. Choe (1996) The Logical Structure for the Construction of a Machine Readable Dictionary, Ms. KAIST, Taejon, Korea [written in Korean].

Clark, J. 1995. NSGMLS - a Validating SGML Parser, software available at ftp://ftp.jclark.com/pub/sp/.

Fillmore, C. J. and B.T.S. Atkins. 1994. Starting where the Dictionaries Stop: The Challenge of Corpus Lexicography, in B.T.S. Atkins and A. Zampoli. eds., 349 - 393.

Ide, N. and J. Veronis. 1995. Encoding Dictionaries, in *Computers and the Humanities* 29-2, 167-179.

ISO. 1986. ISO 8879: *Information Processing - Text and Office Systems - Standard Generalized Markup Language (SGML)*, ISO.

Kim, H. and B. Kang. 1996. KOREA-1 Corpus: Design and Composition, in *Korean Linguistics* 3, 233-258 [written in Korean].

Sinclair, J. ed. 1987. *Looking Up*, London: Collins.

SoftQuad (1995) Panorama Pro, SGML software.

Sperberg-McQueen, C.M. and L. Burnard. eds. 1994. *Guidelines for Electronic Text Encoding and Interchange* (TEI P3), Chicago and Oxford: TEI.

Zampoli, Antonio, N. Calzolari, and M. Palmer. eds. 1994. *Current Issues in Computational Linguistics: In Honour of Don Walker*, Pisa: Giardini and Dordrecht: Kluwer.