Using the TEI Scheme in Compiling a Korean Dictionary

Beom-mo Kang

Dept. of Linguistics, Korea University, Seoul, 136-701 KOREA

KEYWORDS: TEI, dictionary, Korean

AFFILIATION: Korea University

E-MAIL: bmkang@ling.korea.ac.kr
(or bmkang@nlp.korea.ac.kr)
FAX NUMBER: +82-2-921-4376

1. A Dictionary Project

At Korea University in Seoul, we are currently compiling a Korean monolingual dictionary. We are trying to use computers as much as possible to make the compilation process very efficient and ultimately to make a good dictionary. In this process, we have used and intend to use the TEI scheme in the following two ways.

2. Text Headers

First, we follow the lead of compilers of major dictionaries such as COBUILD (Sinclair 1987) in building and using a corpus as a resource of authentic examples and other valuable information such as sense frequency. The problem is how to encode texts on the computer so that we can extract relevant information efficiently. In the process of building a Korean language corpus called "KOREA-1 Corpus" (Kim and Kang 1996), now of size of 10,000,000 words, we have used tags provided by TEI P3 (Sperberg-McQueen and Burnard 1994), mainly for text header information (<teiHeader>). In the body of a text several tags such as <q> and <l> have been used in some cases but only the tag <p> has been consistently inserted.

Besides many <teiHeader> tags that we have adopted as they are, we have used a modified tag <catRef> to classify Korean texts according to our needs of classification. In short, we use four digits to represent 1) written/spoken distinction, 2) media–newspaper, magazine, book, unpublished material, and others (including originally prepared in electronic form–, 3) fields(topics)–general, literary, humanities, social sciences, natural sciences, etc–, and 4) more detailed field (content) classification within a major field.

For example, the following encoding means that the source of the present text is a book whose topic is in the field of history.

```
<teiHeader>
.....
  <catRef scheme='kcrc' target='k1355'>
    책, 인문, 역사</catRef>
.....
</teiHeader>
```

Notice that while the original <catRef> is defined as an empty element in TEI P3, in our revised scheme we allow it to contain a short explanatory content.

Since it is possible to prefix a KWIC line with this kind of classification code and to sort lines according to it, it is potentially a useful means for lexicographers writing definitions and usage notes of a lexical item. An example of a KWIC concordance follows:

```
1331  베려 갔다. 계곡으로 썩은 나무 등걸을 [주우러] 가기도 했다.</p> <p>성진이 푹에
1326 떨어졌다. 공이 떠내려갔다. 친구가 공을 [주우러] 갔다. 그런데 내가 공을 맞추려고
1331 어진 흰 고무신 한 짝. 이제 신쭈를 같이 [주우러] 다니던 광수형도, 살결이 희디흰
1977 다. 스크랜튼 부인은 할 수 없이 학생을 [주우러] 수구문(水口門) 밖에 나갔다. 콜
1326 들이 나가서 당기고 있었다. 나도 고기를 [주우려고] 그물 가로 들어가니 작은형님이
1357 피 모자를 땅바닥에 떨어뜨렸다. 모자를 [주우려고] 몸을 숙이자 마주 오던 사람이
1326 들은 그물 가로 우르르 모여들어 고기를 [주우려고] 했다. 우리 식구는 세 사람이나
2607 와요. (목에 들렀던 스카프가 떨어진다. [주우려는데)] 진수: (재빨리 집어주고는)
1331 </p> <p>미려는 서포리 바닷가에서 조개 [주우며] 외롭게 혼자 놀던 어린시절의 기나
1119 집을 알게 된 뒤로 시장을 돌고 쓰레기를 [주우며] 죄 짓지 않는 삶을 살고 있다"고
1321 /p><p>"여러분도 어디에서나 남의 물건을 [주우면] 바로 주인을 찾아 주거나 선생님
1370 번가에서 아름다운 조개 껍질과 조약돌을 [주우면서] 놀고 있는 소년에 불과하다. 위
1241 지만 해도 논둑길을 따라 도토리와 밤을 [주우면서] 아버지와 나란히 성묘를 다녀오
```

3. Dictionary Entries

Second, since we want to use computers in writing and publishing the dictionary as well as in building a corpus, the problem that we are now faced with is how to represent a dictionary entry on the computer in its electronic form. Since TEI P3 offers ways to encode dictionary items, we intend to adopt the TEI encoding scheme and use it in some stages of dictionary compilation.

Although TEI suggestions for dictionary encoding are very comprehensive to cover various kinds of dictionaries, its current commitment is to consider only dictionaries of western languages (Ide and Veronis 1995: 168). We are struggling with problems 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 content models to a certain degree so that the encoded dictionary might be viewed more as a database than as a simple computerized (originally printed) dictionary.[1]

Among other things, we revise the <entry> model so that it can have a number of proverbs <prov> and idioms <idiom>, which consistently appear on the entry level in Korean dictionaries.

```
<!ELEMENT %n.entry; - O ( (%n.hom; | %n.sense; |
                          %m.dictionaryTopLevel)+,
                         (prov | idiom)* )
                                  +(anchor) >
```

<prov> and <idiom>, in turn, can be defined so that they can contain any dictionary parts such as <form>, <def>, and <eg>:

```
<!ELEMENT prov - -    (%paraContent |
                      %m.dictionaryParts)* >
<!ATTLIST prov          %a.global;
                        %a.dictionaries; >
<!ELEMENT idiom - -   (%paraContent |
                      %m.dictionaryParts)* >
<!ATTLIST idiom         %a.global;
                        %a.dictionaries; >
```

One major revision which affects the hierarchical structure of dictionary entries would be allowing recursion for <hom>, as is allowed to be self-embedded. In Korean dictionaries, some entries have two levels of homography; namely 1) different parts of speech, and 2) different subcategorizations. For example, some form (one entry) is both a verb and an adjective (and a suffix with a related meaning, too). Sometimes, a verb form can be an intransitive verb, a transitive verb, or an auxiliary verb. Of course, some theoretical considerations might allow us to disregard this kind of complex homography levels and have different entries for different parts of speech, so that we can stick to the TEI scheme. However, respecting the tradition of Korean lexicography, we want to maintain at least the two levels of homography mentioned above.

```
<!ELEMENT %n.hom; - O (%n.sense; | %n.hom |
                      %m.dictionaryTopLevel)*
                              -(entry) >
<!ATTLIST %n.hom;     %a.global;
                      %a.dictionaries;
    type              (homPos | homSubc) homSubc
    TEIform           CDATA       'hom'       >
```

Here is an example, where '....' represents some Korean characters. Grammar codes in <pos> and <subc>, such as 'verb', 'adj', 'trans', 'intrans', are transliterations from the Korean counterparts. (For the elements <lenHyph>, <irreg> and <irrForm>, see below.)

```
<entry>
<form><orth>.....</orth><lenHyph>.....</lenHyph></form>
  <gramGrp><irreg>.....</irreg>
    <irrForm>.....</irrForm></gramGrp>
<etym>.....</etym>
<hom type=homPos n='I'>
  <gramGrp><pos>verb</pos></gramGrp>
    <hom type=homSubc n='1'>
      <gramGrp><subc>intrans</subc></gramGrp>
      .....
    </hom>
    <hom type=homSubc n='2'>
      <gramGrp><subc>trans</subc></gramGrp>
      .....
    </hom>
  </hom>
  <hom type=homPos n='II'>
    <gramGrp><pos>adj</pos></gramGrp>
    .....
  </hom>
</entry>
```

Also, we add a dictionary top level element <sciName> for scientific names, which appear prominently in Korean dictionaries, by defining an x-dot parameter entity in the TEI.extensions.ent file:

```
<!ENTITY % x.dictionaryTopLevel 'sciName |'>
```

For dictionary entry forms, which conventionally show the major morphological immediate constituent break (by a hyphen) and long syllables (by a colon) at the same time, we add <lenHyph> as a member of the class "formInfo" in the same way. In addition, to indicate the irregular inflectional classes and to show typical inflected forms, which usually appear along with grammatical category information, we add <irreg> and <irrForm> as members of the class "gramInfo". These are slight revisions to the TEI suggestions.

```
<!ENTITY % x.formInfo 'lenHyph |'>
<!ENTITY % x.gramInfo 'irreg | irrForm |'>
```

Academic domains (special fields), other domains (such as 'old Korean'), and dialect areas, which are also prominent in Korean dictionaries, are encoded with new tags defined within <usg>. They are <domAca>, <domEtc>, and <dialArea>. Also, since the content and format of etymology (<etym>) and cross reference (<xr>) in a Korean dictionary is constrained in certain ways, some modifications of the DTD definitions of these elements are needed. For <etym>, we add a new attribute 'hdType' whose value should be one of the following: 'hj' (hanja, i.e. of Chinese origin: content given in Chinese characters), 'foreign' (of any other foreign origin), and 'kor' (of Korean origin proper). Incidentally, more than half of the entries in a large Korean dictionary are of Chinese origin and can be written in Chinese characters as well as in Hangul, the Korean alphabet.

For <xr>, we 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>. The relevant part of the DTD extension is given below:

```
<!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'     >

<!-- "See" the word for definition -->
<!ELEMENT xrsee           - O EMPTY >
<!ATTLIST xrsee         %a.global;
                        %a.dictionaries; >

<!-- synonym -->
<!ELEMENT xrsyn           - O EMPTY >
<!ATTLIST xrsyn         %a.global;
                        %a.dictionaries; >

....., etc.
```

Here is an example with an etymology <etym> and a cross reference of type synonym <xrsyn>. (Again, '.....' are parts in Korean.)

```
<entry>
  <form><orth>.....</orth><lenHyph>..... </lenHyph></form>
  <etym hdType=hj> ..... </etym>
  <sense n='1'><def> ..... </def>
      <eg><q> ..... <oRef>.</q></eg></sense>
  <sense n='2'><xr><xrsyn><ref> ..... </ref></xr>
      <eg><q> ..... <oRef>.</q></eg></sense>
</entry>
```

We might have constrained the "type" of <xr>, e.g. <xr type='syn'>, in the DTD instead of introducing empty elements such as <xrsyn>.

4. Character Representation Problem

Finally, the character representation problem for the 11,172 modern Hangul (Korean Alphabet) characters and tens of thousands of Chinese characters used in Korean texts and dictionaries should be addressed. Unlike Roman alphabets which require only one byte to encode a character, at least two bytes are required for Hangul and Chinese characters. Once UNICODE/UCS (ISO 10646-1) has been adopted by program developers, the character problem would no longer be a serious one, but for the time being, we should be satisfied with the current Korean standards. The standard we are adopting now uses the control code area above ASCII 127 (C2 area).

Therefore, if a default SGML declaration for an SGML parser like nsgmls (Clark 1995) prohibits the use of both control code areas of C1 and C2, we should revise the SGML declaration. The relevant part of the sgml declaration used for parsing by NSGMLS follows:[2]

```
<!SGML "ISO 8879:1986"
CHARSET
BASESET "ISO 646-1983//CHARSET
        International Reference Version
        (IRV)//ESC 2/5 4/0"
DESCSET 0       9       UNUSED
        9       1       9
        10      1       10
        11      2       UNUSED
        13      1       13
        14      12      UNUSED
        26      1       UNUSED -- eof --
        27      5       UNUSED
        32      95      32
        127     1       UNUSED
        128     127     128 -- used by Hangul code:
                                KSSM --
        255     1       UNUSED
```

5. SGML Parsing and Processing

Sample encodings of a Korean dictionary, together with dtd extensions and modifications, TEI.extensions.ent and TEI.extensions.dtd, have been validated by nsgmls, an SGML parser (Clark 1995). In addition, SoftQuad PANORAMA Pro has been able to process the sample encodings successfully. The beginning part of a sample Korean dictionary which is to be parsed by nsgmls is as follows:

```
<!DOCTYPE tei.2 SYSTEM "c:\sgml\dtd\tei2.dtd" [
  <!ENTITY % TEI.dictionaries "INCLUDE">
  <!ENTITY % TEI.corpus "INCLUDE">
  <!ENTITY % TEI.extensions.ent SYSTEM "hdic.ent">
  <!ENTITY % TEI.extensions.dtd SYSTEM "hdic.dtd">
]>
```

References

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

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

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.

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

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.) Since we are not encoding an existing dictionary on the computer but preparing a lexical database which is to be used in printing later, the last view should be adopted in our project.

[2] Michael Sperberg-McQueen helped me to revise the SGML declaration while I was participating in the 1995 Summer Seminar organized by Center for Electronic Texts in the Humanities, Princeton University.