Development of a Syntax-directed SGML Editor for Processing Korean Ancient Documents

Young-Sik Hong, Keum Suk Lee, and Yong Kyu Lee
Computer Engineering Dept. and Electronic Buddhist Text Institute (EBTI)
Dongguk University, Korea

Abstract

It is very important to use an efficient text editor for processing Korean ancient documents written in Chinese characters. In Korea, there are several commercialized text editors which can handle only 4,888 Chinese characters based on the KSC (Korean Standard Code Set)-5601. Even though some text editors can handle more than 5,000 Chinese characters, Chinese characters processed by these text editors can not be browsed with the existing Internet navigators, such as Netscape and Microsoft Internet Explorer due to incompatible codes.

We designed a new text editor with SGML markup functions to help build Korean ancient Buddhist full-text databases. Our text editor handles about 20,000 Chinese characters based on the Unicode and has embedded SGML-based markup and error checking functions. It's input window consists of four subwindows, such as editing, SGML-tag list display, DTD display, and the source text without tags.

Our text editor is to be improved to handle XML-tags and used to input parts of the Hankuk Bulgyojonso (Korean ancient Buddhist corpus) which will be put on our server and may be accessed through network connections around this August.

1. Introduction

Most Korean ancient documents have been written in Chinese characters and several institutions, such as the Research Institute for Tripitaka Koreana, the Seoul Systems Inc. and the Electronic Buddhist Text Institute of Dongguk University build their own full-text databases of the Korean ancient documents. It is necessary to use an appropriate text editor to key in ancient source documents and build the full text databases. However, most Korean text editors based on KSC-5601 do not support more than 4,888 Chinese characters. Even though some editors based on 4 byte code can handle more than 30,000 Chinese characters, Internet navigators, such as Netscape and Microsoft Internet Explorer, can not display such 4 byte code characters due to the incompatible code system.

We design a new text editor with SGML-based markup functions to solve these problems and to help build Korean ancient Buddhist full text databases which are under construction by the Dongguk EBTI. This article describes our Chinese character input system based on the Unicode, design principles of the syntax-directed SGML editor and its implementation technique.

2. Chinese character input system based on the Unicode

Most commercialized text editors in Korea use the character code set defined in the Korean Standard Code Set, KSC-5601, where each character is represented with 2 bytes and the size of code space for Chinese characters is 4,888. But the Unicode system allows about 20,000 Chinese characters and it is supported by Windows NT 4.0. Also it is possible to browse Chinese characters using the right software technique.

To key in a Chinese character we need two steps. After a Korean character is keyed in, a subwindow displays Chinese characters with the same pronunciation as the given Korean character, and then the correct Chinese character with the specified meaning is selected and saved on the file. Even though this input method seems inefficient, most Korean text editors employ such user interface to input Chinese characters. For this, we have grouped Chinese characters of the same pronunciation according to the corresponding Korean character. Figure 1 shows a part of the rearranged Chinese character codes with their corresponding Korean character.

/ / AC00
8EFB 52A0 8CC0 9050 9050 53EF 5BB6 5CA2 5E4F 5FA6
93B5 6935 6687 73CE 73C8 9D10 9D1A 70A3 7241 726B
7271 728C 75C2 7615 7638 767F 7822 7A3C 7B33 7B34
801E 8175 8238 827D 82DB 8304 9553 9160 83CF 846D
86B5 8857 8888 8A36 8C6D 6A9F 6B4C 6BE0 6CC7 6E2E
6ED2 8C91 8CC8 8DCF 8DD2 8EFB 9160 5609 560F 9E9A
698E 4E2A 4EEE 4F3D 4F73 4FA1 5047 50F9 53DA 8FE6
5475 547F 5496 54E5 54FF 5777 6118 6119 6228 62C1
659D 67B6 67B7 67EF 5AC1 9AC2 5A7D 99D5 9B7A
/ / AC01
606A 9AC9 9B25 73A8 73CF 71EA 7833 78A6 7910 791C
792D 80F3 8173 8316 578E 57C6 5859 95A3 9601 883C
88BC 89BA 89D2 89F3 6BBC 6BC3 7474 8E7B 8F03 9FA3
69B7 5095 523B 5374 537B 5404 54AF 6128 6164 6354
6409 64F1 658D 65A0
/ / AC04
8271 9592 9593 5E72 6073 5B6F 5978 59E6 59E7 980E
99B0 976C 9B1D 8FC0 6695 6746 73E2 9DAB 7647 7642
76AF 76F0 770B 77AF 784D 78F5 7900 79C7 7A08 7AFF
7B34 7C21 7C33 7CAF 7D06 7E5D 7FA5 809D 81E4 8316
578E 57C6 5859 95A3 9601 8425 8551 85D6 8677 884E
8949 89B8 8A6A 8AEB 8C64 8C7B 6A8A 6D86 8D76 8D95
9F66 69A6 6F97 4F83 501D 5058 520A 56CF 6173 61C7
681E 687F 64C0 653C 65A1 67EC 9845 5A36 5A5C
/ / AC08
5D51 5DB1 9AB1 9821 9A14 9B1D 696C 668D 66F7 9DA1
7332 7366 78A3 790D 79F8 79F8 7A2D 7AED 7FAF 9782
97A8 845B 8910 6BFC 6E34 8F35 874E 4E10 4E6B 5303
5304 54AD 559D 573F 6274 63B2 63ED 64D6

Figure 1: Chinese character codes with their corresponding Korean character

Figure 2 displays a sample subwindow for Chinese characters corresponding to a given Korean character to input a Chinese character.

Figure 2: Chinese characters corresponding to a given Korean character

3. Design and implementation of a syntax-directed SGML editor

Since SGML is widely used to define the logical structures of electronic texts, we have decided to use it to markup Korean ancient documents to build full text databases. Our text editor is designed to work as a conventional text editor as well as an SGML markup tool. It consists of four subwindows such as, editing, SGML tag list display, DTD display, and the source text without tags. Figure 3 shows subwindows of our editor.

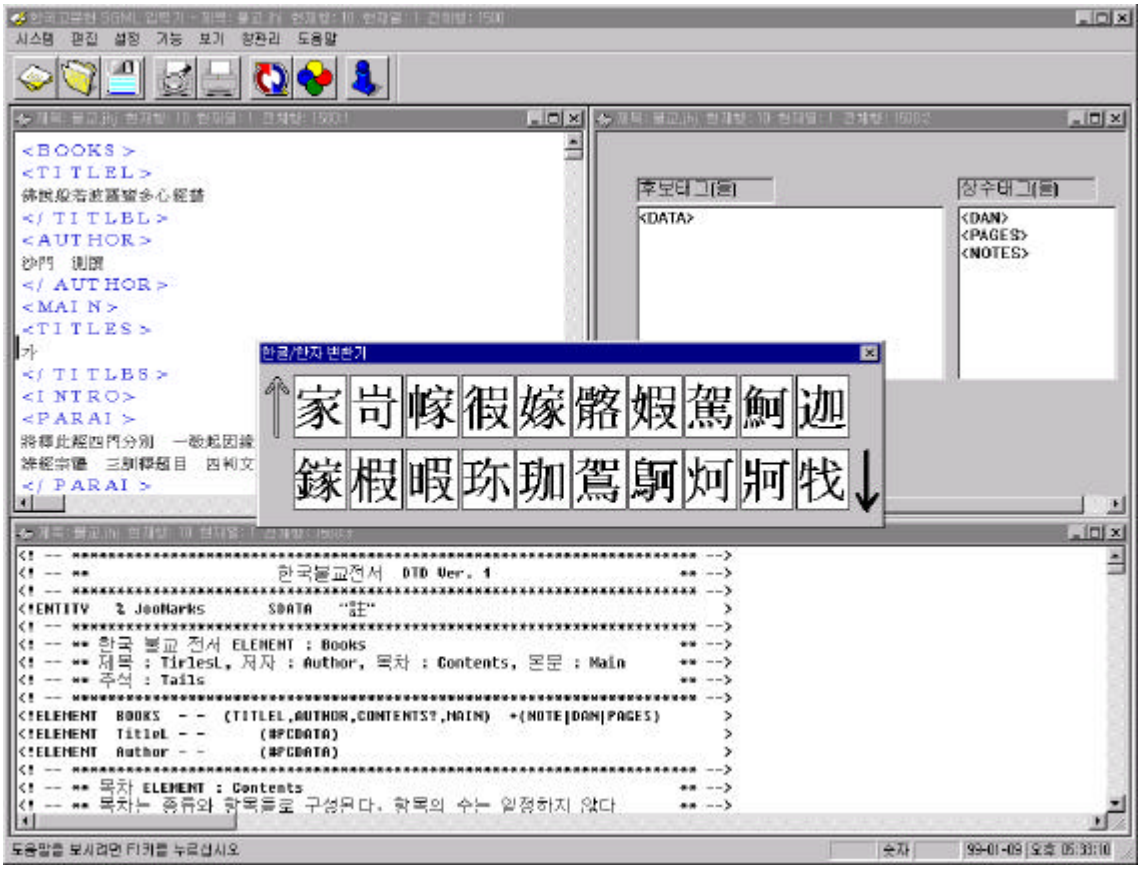

Figure 3: Subwindows of our text editor with SGML-based markup functions

In Figure 3, the editing subwindow is used to key in a source document written in Chinese characters and the subwindow for SGML tag list display is used to enter SGML tags by selecting appropriate tags while editing source texts. Tags listed in this subwindow are obtained from the predefined document type definition (DTD). For example, Figure 5 shows candidate tags and constant tags obtained from the DTD in Figure 4.

```
<!ELEMENT BOOKS - -
          (TitleL?,  Author?,  Main,  Close)
            + (Note|Dan|Pages) >
```

Figure 4: A part of an example DTD

Candidate tags	Constant tags
TitleL Author Main	Note Dan Pages

Figure 5: Tags obtained from the DTD in Figure 4

By selecting tags from the tag list display subwindow, we can prevent tagging errors caused by mismatching tags and typing errors of tag names. Our editor does not allow users to edit tag strings directly to prevent errors, instead it assigns attribute values to tags automatically by parsing DTD statements.

Our prototype of text editor is designed to help build the Hankuk Bulgyojonso. Figure 6 shows a part of the DTD for the Hankuk Bulgyojonso. A simple SGML document according to the DTD is shown in Figure 7.

```
<!ENTITY %JooMarks SDATA “註”>
<!ELEMENT Books – (TitleL?, Author?,
                Contents?, Main) + (Note| Dan| Pages) >
<!ELEMENT TitleL – (#PCDATA) >
<!ELEMENT Author – (#PCDATA) >
<!ELEMENT Contents – (AKind, Items+) >
<!ELEMENT AKind – (#PCDATA) >
<!ELEMENT Items – (#PCDATA) >
…
```

Figure 6: A part of the DTD for the Hankuk Bulgyojonso

```
<BOOKS BOOKNUM = 1>
<PAGES PAGENUM = 1>
<DAN DANNUM = 1>
<TITLEL>
佛說般若波羅蜜多心經
</TITLEL>
<AUTHOR>
沙門
</AUTHOR>
<PARAI>
將此經四門分? 起因緣 二辨經宗體 三訓題目 四文解
</PARAI>
<GENUSI>
言? 起者
</GENUSI>
```

後以無所得故下 顯所得果 所以無 …

Figure 7: A part of the SGML text for the Hankuk Bulgyojonso

Figure 8 shows the Unicode file corresponding to a given part of SGML text

SGML text	<TITLE> **佛說般若波羅蜜多心經**
Unicode file	FFFE 3C00 5400 4900 5400 4C00 4500 4C00 3E00 0D00 0A00 5B4F 2D8A 2C82 E582 E26C 857F 1C87 1A59 C35F 937D 0000 0000 0000

Figure 8: A part of the SGML text and the corresponding Unicode file

Our text editor with SGML markup functions has been implemented using Microsoft Visual C++ 5.0 on the Windows NT 4.0 supporting the Unicode.

4. Conclusions

A new text editor with SGML markup facility has been designed and implemented on the Windows NT 4.0 supporting the Unicode. With this text editor, we can handle about 20,000 Chinese characters with which most Korean ancient documents could be written except some special Chinese characters. According to the result of studies on the statistical aspect of Chinese characters used in Korean ancient documents, most of them use less than 15,000 Chinese characters. To resolve the missing Chinese characters in the Unicode we decided to use a page pointer to a Chinese character dictionary, which is similar to the approach used in the electronic text of the Taisho Tripitaka.

Our text editor with SGML markup functions is the first editor in Korea which minimizes the code problem of Chinese characters and enables electronic texts to be browsed using the existing Internet navigators, such as Netscape and Microsoft Internet Explorer. Our text editor is to be improved to handle XML tags and used to key in the Hankuk Bulgyojonso to build a full text database, which will be put on our server and a part of it may be accessed through network connections around this August.

5. References

[1] Dongguk University Press, *The Hankuk Bulgyojonso (Korean ancient Buddhist corpus), Vol. 1*, 1979.
[2] EditTime, *http://www.timelux.lu*, TimeLUX, Dec., 1997.

[3] C. F. Goldfarb, *The SGML Handbook*, Oxford University Press, 1990.

[4] Y. S. Hong, et al., "Develpoment of the technologies for Korean Ancient Document Management and Retrieval on the Web," Project Final Report, Ministry of Information and Communications, 1998. (In Korean)

[5] Chu-Ren Huang, Keh-Jiann Chen and Shin Lin, "Corpus on Web: Introducing The First Tagged and Balanced Chinese Corpus," PNC Special Meeting in Taipei, Feb. 17-19, 1997.

[6] Michael Murry, "Unicode Issues and the Input of "Han" Character Texts", The 4th EBTI Meeting Kyoto, Oct.23-26, 1997.

[7] Rajeev Nagar, *Windows NT File System Internals: A Developer's Guide*, O' Reilly, 1997.

[8] Panorama, *http://www.softquad.co*, Softquad, Dec., 1997.

[9] The Unicode Consortium, *The Unicode Standard, Version 2.0*, Addision-Wesley, 1996.

[10] C. Wittern, "SMART Project and the Database of Chinese Buddhist Texts." The 4th EBTI Meeting Kyoto, Oct.23-26, 1997.