The Current Development Situation of Digital Technology in Kunten Materials

Hongyan Liu\textsuperscript{a,} and Qingping Ren\textsuperscript{b}

\textsuperscript{a}School of Foreign Languages, Yanshan University, China
\textsuperscript{b}liuhongyan@ysu.edu.cn, reap123@163.com

Abstract. This paper is about the current development situation of Kunten materials’ digital technology. In East Asia, there is a large number of Kunten Materials which has important research value. The digital processing of Kunten Materials’ practical problems are basic problems in natural language processing and identified problems in semeiology. This technology is based on punctuation text technology from a third party. By this, pure text information can be tagged with Kunten information and can make text structuring. Besides, it is possible to make structuring text information vernacular. Based on this, this paper analyzed the way of undigested ancient manuscripts and solve direction problem.

1. Introduction

Kanbun Kundoku is a common literature reading way in ancient East Asia. To be concise, it means that intellectuals of Chinese characters cultural ring in East Asian countries make the special marks with white or red ink in written word to understand Chinese phonology, semantics and grammar when they learn Chinese characters. It is called as Text Glossing in vernacular readings. Literature with these manuscripts punctuation in historical literature is often called as Kunten Manuscripts. They are important for linguistics, history, education and philology with important research value. In East Asia, there are large number of Kunten Materials left in history. However, compared with the same type of historic text materials, digital technology in Kunten Materials obviously falls behind. For example, in Japan, the number of Kunten materials in Heian Period has reached to more than 5000 as identifiable number of Kunten materials. However, \textit{Vajraśekhara Sūtra} has been the only one that is arranged by digital technology by now.

Thus, it can be seen that current situation of digital technology is in backward although there is a large number of Kunten Materials which has important research value. This is caused by the fact that digital processing of Kunten Materials’ practical problems are more complicated, not only basic problems in language processing but also identified problems in semeiology. On the other hand, these complicated problems which have been solved can boost developments of digital technology in Kunten materials. From current research results, the digital technology in Kunten Materials has been in great progress. This paper will focus on this year’s current development situation of the digital technology in Kunten materials. Based on this, this paper will analyze the way of undigested ancient books’ manuscripts and solve direction problem.

2. Punctuation Textual Standardizing Technology based on LaTeX Instructions

The most fundamental technology skill of digital technology in Kunten materials is punctuation textual technology which can transfer the punctuation in written word to the text which can be identified by machines. The most popular way is LaTeX user-defined way which is realized by third tool of kunten2.sty. It has steady characteristic from current using.

Furigana textual instruction in Kunten materials is mostly in view of basic instructions of LaTeX input command and form which is ameliorated by ruby especially for vertical writing Kunten
materials, called as \textit{MigiNakaTn} constructions. It can not only realize basic functions of punctuation textualization but also can set left or right position of glossing, the size of glossing words and proportion relationship in textual words.

The above picture was examples about Kaeriten mark textualization instructions in Kinsui (1999). Corresponding to different Kaeriten mark, the solving strategies listed above can be divided into the forms of \textit{reten} and \textit{kaeriten}. Kunten form can be transformed into textual form which can be identified by machines, which is the basic of digital technology in Kunten materials.

3. The Conversion in Structuring Text

The digital technology of Kunten materials must realize the digitalization of original Chinese text and also realize the digitalization of matching punctuation digitalization. Therefore this is not the text conversion problem but working process with tag text construction. The workflow of this can be described as five steps. First, it is conversion process of Kunten materials in plain text. Second, it is coding process in Kunten information. Third, it is generated process of .xml format in Kunten information structured text. Fourth, Kunten marks are read by Wokototen reader skill. Fifth, vernacular glossing is in output.

The chart below was the coding process of Kunten information through adding up the text of Kunten materials to realize the digital output of final format of XML (Tajima ed. 2012). Based on this, Takada (2013) set up the following instruction format in structuring text with Kunten information.
Those assured instructions and procedures basically guarantee the conversion from text plus marks into structuring text. This solved problem accelerate the pace of Kunten materials corpus and can make it public by internet. The trial implementation data base in NINJAL is established based by this thought now.

4. Automatic Identifying and Unscrambling Kunten Materials

According to the text above, the Kunten marks need to be identified and make it as the connection in structuring texts during the process of structuring text conversion to make Kunten materials as vernacular glossing language. However, a large cost in time and economy will be taken if it is done by human beings. Therefore, automatic identifying technology is a critical link to make Kunten materials as a public data base. Kunten marks’ automatic identifying technology is different from OCR East Asia word recognition because interval proportion between words in East Asia is fixed but general optics recognition is really hard to identify because the Kunten marks are always around the single Chinese word.

The concept strategy in this technology is in the chart above (Takada 2014). The strategy is that there is certain space around each Chinese word so every Wokototen mark is above or below the Chinese word can be assured in that certain space. Therefore, every different Wokototen mark is in one certain coordinate point. The coordinate point and Kunten mark need to be corresponded, then glossing text can be in Kunten output. Concrete coordinate instructions can be set as follows.
However as most of Kunten materials are in written form, Kunten marks are influence by Chinese character pattern so they are not in certain positions. Here is the problem when it comes to the identification when the certain coordinate standard is used. Therefore, it is the this technology’s future development direction of variable coordinate based on Chinese character.

5. Conclusion
The main part of this literature review is about the current development situation of Kunten materials’ digital technology. This technology has become more and more mature and its essence is based on punctuation text technology from a third party. By this, pure text information can be tagged with Kunten information and can make text structuring. Besides, it is possible to make structuring text information vernacular.

It is worth for us to make further study and to remold it to reorganize and save other Chinese books. it is especially important for us to use the above Kunten materials digital technology for reference. Chinese books’ manuscripts can be formed in a large data base and be used in all kind of research in the near future by establishing systematic way of text and set structuring text of processing flow.

6. Acknowledgement
This research was financially supported by the Provincial Social Science Fund of Hebei (Grant No. HB17YY044).

References