From Text to Image to Analysis: Visualization of Chinese Buddhist Canon

Lancaster, Lewis
buddhst@berkeley.edu
University of California, Berkeley

This presentation is based on software interface development by a team at the University of California, Berkeley. The database which was used for this technology is the digital version of the Korean Buddhist canon written in Chinese characters. The tool shown was built with the help of a two year grant of support (2007-2009) from the National Science Foundation. International collaboration has included the Institute of Tripitaka Koreana in Seoul who provided scanned images of rubbings taken from the original printing blocks at Hae-in Monastery. The software metadata is based on the previous publication *The Korean Buddhist Canon: A Descriptive Catalogue* (Lancaster, 1979).[1] A digital version of this catalogue was made by Charles Muller of Tokyo University who has made it freely available on the internet (Muller, 2004).[2] The project has been a part of the Electronic Cultural Atlas Initiative (ECAI) and received support from that group's *Atlas of Chinese Religions* research funded by the Luce Foundation. This atlas is being constructed in collaboration with the GIS Center at Academia Sinica in Taiwan and will provide references to the place names associated with the production of the translations and compilations included in the canon. Continued research on developing the software is being done in cooperation with the School of Creative Media and the Department of Chinese Translations Linguistics at City University of Hong Kong. It is important to understand that no project of this kind could possibly be undertaken without these multiple and widespread collaborations.

In the example being described in this presentation, we use the software to focus on the digital version of the 13th century Korean printing block edition of the Buddhist canon (Lancaster, 1996).[3] The canon, represented on blocks, contains more than 52 million characters/glyphs carved onto 166,000 surfaces each producing a page of text when printed. The number of lines, containing up to 14 glyphs, on the plates number over three million. The entire set of the canon is divided into 1,514 different texts representing dated translations and compilations made over a period of seven centuries. The size of the data, the temporal span of its composition, and the history of acquisition in Korea of the hundreds of texts from China, provide us with a reasonable challenge for the interface design.

The previous approach to the study of this canon was the traditional analytical one of close reading of specific examples of texts followed by a search through a defined corpus for additional examples. When confronted with 166,000 pages, such activity had to be limited. As a result, analysis was made without having a full picture of the use of target words throughout the entire collection of texts. That is to say, our scholarship was often determined and limited by externalities such as availability, access, and size of written material. In order to overcome these problems, scholars tended to seek for a reduced body of material that was deemed to be important by the weight of academic precedent.

In the current digital age, however, the limits on "what can be considered" in the Korean Buddhist canon have been significantly removed. We can consider all of the texts, all of the words, and all of the metadata in every search. Consequently, the practices of traditional scholarship for the canon have begun to falter. When the entire canon had been digitized in the last decade of the 20th century, the process of search and retrieval of target words and phrases was transformed. Nonetheless, problems remain for Buddhist scholars using this digital version. In many cases, the menu which appears after a search of a term can contain thousands of references. The references presented as a display of each line where the word occurs can still occupy long hours of time to analyze and put into some form of presentation.

We are in need of new ways to display search results that will allow scholars to quickly perceive such things as the patterns of occurrences, examples of clustering, view of target words with adjacent companion words, graphic models of profiles of sequence,

computation of occurrences not only for the whole of the set but also broken down by text and date. The displays give the researcher aids in evaluating and analyzing the patterns for each word.

As a first step, we look for the number of times that each of the characters appears in the canon. As each search is made, a report appears in visual form on a "ribbon" of blue dots, where each of the dots represents one of the 52 million glyphs. The "ribbon" is more than a picture for each "blue dot" has 35 fields of metadata behind it. It is marked for exact placement on the original printing block, date of translation, name of translator for the text in which it appears, UNICODE number, place of translation, name of text containing the example, etc. The dots are arranged by "panes" that correspond to the more than 160,000 pages of the version preserved in Korea. The dot is an abstract image that permits the user to see patterns of occurrence without the barrier of complex display of natural language glyph constructions (Lancaster, 2009).[4]

It is at this first step that we note the distinct shift in methodology. The initial move on the part of the scholar, who uses this interface, is to turn directly to the data itself rather than to reference works. This is accomplished because the software provides a process of searching through the entire set of data at once. There is no step of consulting a reference work such as a concordance before proceeding to the text itself.

This visual becomes the first factor in the scholar's "work flow" planning. It shows whether the glyph(s) are scattered throughout the canon, whether there are heavy concentrations in a few places, whether there are only a few examples that appear in widely separated examples in terms of texts, time, and translators. Securing this much information within a few seconds can be compared to the hours of effort it would take to construct such an analysis of patterning, even with an internet search for each example of the term. In every case, the "blue ribbon" and other graphics are displaying a large amount of data in the visual form. We can "see" the occurrences of our target search within the 52 million glyphs and immediately understand the nature of the patterning.

Words in the canon have a history and we can begin to spot the ways in which they evolve. Since each dot has multiple metadata fields, they need not be seen only in the sequence of the canonic arrangement. They can be rearranged by time of translation, by translator, or place of translation. The visuals will quickly show that some words grow in number of occurrences over time and others fade into a more marginal role.

There have been surprises from the use of the tool on the canonic words. The computation for an important expression can help us identify apocryphal texts, or texts that show the characteristics of translation rather than compilation. "Companion" words that become associated with a target word can be used to identify the primary meaning of an occurrence.

Future plans call for the exploration of making this tool multi-lingual and provision for having it available as an open source and free add-on to datasets.

Funding

- This material is based upon work supported by the National Science Foundation under Grant No. 0840061.
- Support for the Atlas of Chinese Religions given by grants from the Henry Luce Foundation, Inc. New York.
- Support for Atlas of Chinese Religions provided in part by the GIS Center of Academia Sinica, Taiwan.

References

Lancaster, Lewis, Park, Sungbae (1979). *The Korean Buddhist Canon: A Descriptive Catalogue*. Berkeley: Berkeley University Press.

Lancaster, Lewis (1996). 'The Buddhist Canon in the Koryo Period'. *Buddhism in Koryo: A Royal Religion*. Kikun Sub, Chai-shin Yu (eds.). Berkeley: Berkeley University Press.

Lancaster, Lewis (2009). *SGER: Text Analysis and Pattern Detection: 3-D and Virtual Reality Environments*. `http://ecai.org/textpatternanalysis/`.

Muller, Charles (2004). *Digital Version of The Korean Buddhist Canon: A Descriptive*

Catalogue. http://www.acmuller.net/descriptive_catalogue/index.html.

Notes

1. Lewis Lancaster and Sungbae Park. The Korean Buddhist Canon: A Descriptive Catalogue. Berkeley: University Press. 1979.
2. See http://www.acmuller.net/descriptive_catalogue/index.html for the digital version of the The Korean Buddhist Canon: A Descriptive Catalogue on his server in Tokyo. (2004).
3. A description of this canon is found in my article "The Buddhist Canon in the Koryo Period," Buddhism in Koryo: A Royal Religion, edited by Kikun Suh and Chai-shin Yu. Korea Research Monograph #21, Institute of East Asian Studies, University of California: Berkeley, 1996.
4. See examples of the interface and report of progress in my report to NSF: SGER: Text Analysis and Pattern Detection: 3-D and Virtual Reality Environments (2009). http://ecai.org/textpatternanalysis/