The Statistical Analysis of Large-Scale Groups of Buddhist Texts

Moro Shigeki
(Hanazono University, Kyoto)

A philological study needs a hypothesis. When a scholar chooses a text from innumerable possible options and begins to read it, behind this choice there are usually hypotheses based on knowledge of that text or previous research, personal interest and so on. In the same way, the retrieval of an electronic text often does not produce significant results without any prior knowledge of its contents and linguistic expressions.

In recent years, the digitization of Chinese classical texts, including Buddhist texts, has progressed rapidly, and it seems that we have now a much more favorable environment for research. In this new environment it is possible to deal with large groups of texts, far beyond the knowledge of individual researchers, and the traditional method of research, based on an initial hypothesis, has become obsolete. Thus there is the need to develop a new methodology to handle such large amounts of materials.

In the field of computer science the technology called "data mining" is worthy of attention. Thanks to statistical analysis and data-matching, this technology helps the formation of hypotheses and concepts by discovering patterns in large databases. Although a number of studies have been carried out about the statistical analysis of classical texts, little is known about its application to hypothesis-formation in the field of philological studies.

In this presentation, I would like to examine a possible way of handling large-scale text databases, and in order to do this I will try a method for producing hypotheses through classification by statistical cluster analysis applicable to philological research. As an example, I will use all of the translations attributed to Xuanzang's and make a comparison between computer-based classification and traditional methods of research.