NGSM and Cluster Analysis: Its Usage in the Digitization of Variant Texts in the SAT (Taisho Daizokyo Text Database)

By ISHII Kosei

[Abstract]

The *Taisho Daizokyo* includes many cases of variant translations of scriptures and treatises. But even in the case of what is ostensibly the same translation, the *Taisho* is filled with variant versions from Song, Yuan, and Ming editions, including printed and hand-copied versions. Although most of the important versions of the texts are available for comparison within the *Taisho* (which is originally based on the Korean Canon), when, in the *Taisho* we find references to other versions of the texts, a digital search of the *Taisho* does not necessarily turn up the full set of information that we need. Therefore, when we are studying important scriptures and treatises, we really need to have digitized versions of all the variants available to us, just as a starting point. Once these variant texts are fully digitized, we can then properly analyze them using NGSM （N-Gram based System for Multiple document comparison and analysis) and cluster analysis. Using this technology, the systematic relationship of a family of texts can be quickly displayed in a chart format. It furthermore becomes possible to uncover the system underlying the arrangement for the works contained in each individual canonical collection. In this presentation, some of these cluster analysis techniques will be introduced.

SAT is the abbreviation for *Samganikikritam Taisho Tripitakam*, which is Sanskrit for digitized *Taisho Tripitaka*. The organization developing SAT is called ACUBAT, that is the Association for Computerization of Buddhist Texts. However, since the name of ACUBAT has not become widely know (such that even the members themselves sometimes forget it), the organization developing SAT is in general called SAT as well. For this reason, we members of ACUBAT refer to the organization as SAT when we make a presentation like this. This is because it is convenient, and also because we ourselves are very attached to the name SAT.

Although the input work began at SAT prior to that done at CBETA, it has been delayed due to a shortage of funding, wherein there were times when work had to stop almost completely. Even though CBETA, with which SAT has cooperative ties, has already published its volumes 1 to 55 and volume 85, SAT has for the past few years published only a few volumes a year. However, thanks to the supporting organization that finally began its activity last year, prospects for obtaining funds look much brighter. We can probably expect to complete the input of the remaining volumes of Japanese Buddhist scriptures within three years.

Thanks to the great progress made in the digitization of Buddhist scriptures, it has become extremely easy to search for references. However, many problems still remain. For instance, when multiple numbers of Chinese versions are present in one scripture, *Taisho* records all of them, but when characters used differ according to the version, such as Song Tripitaka, Yuan Tripitaka, and Ming Tripitaka, with the same translation, *Taisho* merely indicates the differences of the characters by version at the bottom of the page. Therefore, even if *Taisho* is fully digitized so as to allow for searching, if a certain Buddhist scripture quotes a passage from a scripture of one version with different characters than the digitized version, that reference cannot be retrieved. Those lists of differences are extremely useful, but they also have limitations.

Let us take a well known scripture, the *Heart Sutra*, as an example. It says that when Avalokitesvara examined the five aggregates, he found them to be totally empty. But a monk named Zhizhou, of the Chinese Yogacara school in the Tang dynasty, wrote in his work: "The *Prajnaparamita Sutra* says, ‘he examined the five aggregates and so forth and found them to be totally empty." When we search various *Prajnaparamita Sutra*s contained in the *Taisho*, this phrase cannot be found. But, in fact, this phrase can be found in some variant versions of Xuanzang’s translation of the *Heart Sutra*. For example, the text traditionally used in Horyu-ji Temple states: "He examined that five constituents, and so forth are totally empty," with the addition of "and so forth." This is based on the interpretation of the Chinese Yogacara school, so monks of the school during the Tang dynasty used the text with the added words "and so forth."

But it is impossible to record all variant texts of a certain scripture. It would become massive in volume if we did, and such a *Tripitaka* would likely become very expensive. In contrast, it would not take up much space, no matter how large the volume, if it were in digitized form. And as PCs nowadays work at extremely high speed, it is possible to search a massive amount of data in a short time. Digitizing all variants would, of course, require substantial labor, but if the basic texts in *Taisho* have been digitized, then it is not that difficult to create variant texts by making use of the notes on variances at the bottom of the page. Regarding variant texts not contained in *Taisho*, we can make comparisons with the digitized texts of *Taisho* and revise the parts that are different. If the basic texts have been digitized, it would be far easier to create the digitized texts of the variants than to input them from the start. Excluding some cases of a small number of variants with minimal differences, it is impossible to provide metadata related to the variants in the texts using XML. At SAT, we are for the time being focusing on digitizing all variants of several important scriptures and getting them up on the Internet, so as to stimulate interest in new uses for digitized scriptures.

If different variants are digitized, it would not only be possible to carry out a more accurate search, but also various analyses would become possible. For instance, if NGSM (N-gram Based System for Multiple Document Comparison and Analysis) is used, comparisons of multiple numbers of scriptures, including variants, can be made easily. NGSM is a system that was developed by several members of the Japan Association for Asian Texts Processing including Mr. Shigeki Moro, my colleague at SAT, and myself. The journal of the association featured on this technique in volume II, and members have been introducing their methods of processing and posting the results in their own websites. Most programs are Perl scripts, and almost of them are posted on the Internet.

As I considered the processing of variants as my primary objective, I called this system the N-gram Based System for Multiple Document Comparison and Analysis (NGSV), but since it became clear that it can be used for various purposes, the name was changed to NGSM. We had advice from Mr. Charles Muller on these English names.

I gave a simple report on the effectiveness of NGSM at the EBTI Conference at Seoul last year. And Mr. Moro began an experiment on graphically representing the processed results by NGSM using cluster analysis. His experiment was so interesting that I was stimulated to attempt a few experiments on my own. The study related to the assessment of authenticity of Nichiren’s work by Professors Masakatsu Murakami and Zuiei Ito was very useful in carrying out our experiments.

First, to display the effectiveness of this method, let’s compare the variants of the *Heart Sutra* as translated by Xuanzang. This chart shows the results of these variants processed by NGSM, which was loaded into MS-Excel. It shows how many times various combinations consisting of two Chinese characters appear in respective variant texts. We can get much information and hints for our research just by looking at this chart. But when it is processed further, many more unexpected facts become apparent.

For instance, when we graphically display this data by cluster analysis, it looks like this. As you can see, it is processed extremely rapidly. The researchers who made great efforts in their studies over much time might feel like shouting, "Give me back all the time I spent!" when they see this demonstration.

It is clear what this chart indicates. Three texts in the Ming, Yuan, and Song Tripitaka are lined up in the left-hand column, and we can see that they are very similar. Three texts next to them belong to the Chinese Yogacara school. The text to its right is the one used at Horyu-ji Temple in Nara, Japan, which belongs to the doctrinal lineage of the Chinese Yogacara school. The one lined up to its right is the text in the *Taisho*. *Taisho*, as is generally known, is based on the *Tripitaka Koreana*. From this chart, we can see that the texts of the *Heart Sutra* in *Tripitaka Koreana* are something unique and different from any other lineage.

Coming next is the text that has been used widely in Japan since the Edo period. To its right is the text found in the annotation by Fazang, of the Huayan school of the Tang dynasty. The text on the far right is one used by the annotations of the *Heart Sutra* discovered in the Dunhuang cave temples, recorded in volume 85 of the *Taisho* Tripitaka, as No. 2747. In *Taisho*, the name of the author is not noted with this text. However, since Professor Fumimasa Fukui had discovered that a fragment of an annotation of Wenzhao (文沼), discovered in the Dunhuang manuscript, matched No. 2747, it was established that it was the annotation of Wenzhao. Professor Fukui states that the background of Wenzhao is a obscure. But looking at this chart, and judging that Wenzhao used the text close to the one used by the scriptures of the Huayan school, it was possibly written by a monk of the Huayan school. In fact, there was a monk named Wenchao (文超) among the disciples of Fazang, and the characters for Wenchao and Wenzhao are similar. As is well known, the Dunhuang manuscripts contain many typographical errors, and many characters with similar shapes or sounds are written mistakenly, so it is not unlikely that Wenzhao was Wenchao, a disciple of Fazang.

No doubt, in order to get a result that carries statistical weight, it is necessary to process the data after adding grammatical information and various information. Indeed, the studies on Nichiren's writings by Professors Murakami and Ito, as well as the study on variants of the *Tale of Genji* by Professor Testuya Ito, using computers and taking much time and effort to carry out detailed analysis, have achieved wonderful results. But to put it another way, is it not amazing that we can produce such highly interesting results through a very simple process without the time-consuming work of adding grammatical information? Students of Buddhism tend to be constrained by preconceived ideas, but to be free from such preconceptions and get inspiration for their studies, such mechanical analysis serves well. At SAT, we plan to simplify such processing methods by NGSM and release them to the public, so that researchers who are unaccustomed to computers can also use them.

Further, NGSM not only processes Chinese texts, it can also process any other languages. Mr. Sadanori Ishitobi has recently applied our method to his study of philosophical literature of India written in Sanskrit, and has published his method and results on his website.

http://homepage3.nifty.com/ajunamar/inform/n-gram.html

Now let's take a look at the illustration of the variant translations of the *Heart Sutra* processed by Mr. Moro. The variants are divided largely into two, as shown in the illustration, and the two on the left-hand side are simplified scriptures, generally called the Small Edition, and the translations on the right are well formed scriptures, called the Long Edition. Among these, the one on the far left is the 5th century Kumarajiva, and the second one is the translation of the 7th century Xuanzang. Next are the translations of Dharmacandra, from the beginning of the 8th century, followed by Prajnaparamita, from the late 8th century, and then the 9th century Jnanacakra. the two translations on the far right are the 9th century Facheng (法成) and the late-10th century Shihu (施護). In short, they are lined up in almost perfect temporal order from left to right. If the order in which they are lined up were significantly different from the order they were translated, or the nature of the translated version, it must be assumed that there is some reason for this. In other words, it is conceivable that they may have actually been translated by other scholars.

Now, in the illustration of the variants of the *Heart Sutra* we have looked at earlier, the *Taisho Tripitaka*, or in other words, the *Tripitaka Koreana*, was indicated as being of an independent lineage. This is to say, roughly, it retains the oldest form and is positioned between Chinese variants and Japanese popular editions. This fact actually also applies to other Buddhist scriptures. For example, let's take a look at the variants of the philosophical poem Cantong-qi (參同契), by Shitou Xiqian（石頭希遷）, who was a prominent figure of Zen Buddhism in the Tang dynasty.

The four variants on the left in this chart are all contained in Chinese literature, and those literatures are lined up from the left in order of age. Zutang-ji (祖堂集), in the middle, is the one edited in China in the 10th century and printed in Korea in the 12th century, and its woodblocks are still stored in Haein-sa Temple. The texts contained in this Zutang-ji are unique, and there are significant differences from other texts. The two variants on the far right are texts found in the annotations written in Japan in the Edo period. Basically, even here, the Chinese variants are lined up on the left, followed by the texts printed in medieval Korea in the middle, with new Japanese texts to the right. Professor Koyu Shiina, one of the authorities on Zen literature studies surmises that the text in the Zutang-ji has retained the oldest form. His speculation closely matches the results of the cluster analysis.

When many variants are input in such a way, studies that were unthinkable until now become possible. Analysis by NGSM, and analysis that combines NGSM and cluster analysis, are just examples. It will probably become possible, in the near future, to add grammatical information up to a certain level by automatically dividing the words of the texts written in Chinese.

To study variants in such a way would also be useful in re-examining the method of studying Buddhism. Researchers tend to demand the oldest and best-quality texts. But if a text has been well read and has influenced many people for hundreds of years, regardless of its having been rewritten extensively and containing many typographical errors and omissions, the text should not be ignored. In such a case, we need to study the modified text in order to clarify the ideas of that era. In short, the digitized Tripitaka of the future must become something that can serve as an infrastructure that supports such research.