The use of Kewen as knowledge structures for comparing textual variants

John A. Lehman

Ching-Chun Hsieh

Abstract

The increasing pace of digitalization of East Asian texts presents users with both opportunities and problems. One of the latter is the comparison of different versions of the same underlying text. This is particularly difficult for coordination of variant texts in the Buddhist canon. The same ur-text is often found in different levels of development, in versions from different traditions, and in different translations. Often (as in the sequence of translations into Chinese), these differences are mixed up in series of versions. Because the earliest versions are often found in rather inexact translations and the versions in the "original" language (e.g. Sanskrit) often represent much later developments, reconstruction of the "original" text in the Western manuscript tradition is often impossible. The varying structures, significant differences in vocabulary and syntax, and the frequent omissions make traditional textual comparisons tedious.

This paper proposes the use of Tiantai kewen as external markup to facilitate textual comparisons. The advantages are that kewen are:

- Readily available
- Structure- rather than text-based
- Directly useful to some target audiences

We describe a prototype computer-based implementation of this proposal, and discuss how it can be extended for practical use.

Overview of translations

In the pre-modern world, two of the largest translation projects ever undertaken were those which resulted in the translation of the Buddhist canon into Chinese and Tibetan. The translation into Chinese was not a single systematic project, but rather a series of projects large and small, organized and casual, which took place between the third century and the twelfth century. Some of these projects resulted from the deliberate selection and collection of Indic texts and were carried out using systematic vocabulary and translation methods (e.g. that of Xuanzang); others resulted from casual translation of texts of unknown provenance. Popular texts were often translated many times; there are, for example, **9** different **Chinese** translations of the *Diamond Sutra* and **11** of the *Heart Sutra*.

These translations differ considerably due to textual evolution of the originals, the evolution of translator sophistication, the development of vocabulary, and changes in stylistic conventions.

The textual evolution of the originals reflects the fact that the Indic-language texts chosen for translation changed over the almost 1,000 years of the translation process. Some of these changes were due to variations in oral and manuscript transmission. Others were due to differences in doctrinal traditions between different schools, the existence of which was not recognized by early Chinese collectors and translators. Still other changes were due to the evolution of Mahayana philosophy, and to the tradition that longer texts were more authoritative than "abridged" shorter versions. For all of these reasons, it is entirely likely that an Indic language text chosen by a third century translator would be substantially different from a text with an identical or similar name chosen by a tenth century translator.

The evolution of translator sophistication refers to the fact that there were no scholars at the end of the Han Dynasty who were well educated in both the Chinese and the Indian traditions. Moreover, the linguistic structures of Indian languages and classical Chinese are very different. Most of the Buddhist missionaries and other bringers of texts spoke little Chinese; their Chinese followers, even if educated, spoke little if any Sanskrit/Prakrit. As a result, the initial rendering of a text was probably into oral Chinese whose nature is described in the verse: “Don’t fear heaven, don’t fear earth, only fear barbarians trying to speak Chinese!” Final written “translation” was then usually by Chinese who could not understand the original, but who tried to put the literal translation into idiomatic Chinese, which often lacked the necessary vocabulary.

One of the problems faced by all of the early translators was the development of vocabulary to express concepts from Indian philosophy. The large incidence of Sanskrit and Pali terms used to discuss Buddhism in English is evidence of the problem, which was much worse for 3rd century Chinese. Not only did Indian philosophy concern itself with issues unknown to Chinese philosophy (which thus lacked terms to describe them), but the nature of the writing system made phonetic borrowing more difficult than is the case for European languages. Early translators tried to adapt terms from Daoism (just as 19th century Western translators of Buddhist materials tried to use terms from Christianity); later translators recognized the fundamentally different world views of the two traditions and adapted an increasingly systematic approach to acquiring Sanskrit terminology through either borrowing meaning or transliteration.

A related source of differences in translations resulted from changes in stylistic conventions for such translations. These changes arose primarily from **stylistic** differences between Sanskrit and Chinese, as well as changes in the audience for whom the translations were intended. Briefly, classical Chinese put much more emphasis on brevity and elegant expression than did Sanskrit style. As a result, early Chinese translators found the literalness and repetition in the Indic originals off-putting, and so resorted to paraphrase instead of literal translations. This was especially true of those translations aimed at winning the hearts and minds of the educated (and powerful) Chinese gentry. Later, as the target audience changed to Chinese Buddhist readers, and as Chinese scholars recognized the importance of understanding what the original texts had said, the translation style became much more literal.

Difficulties with variorum text comparisons

The Western approach to textual comparison (as supported in the TEI approach to markup for variants) is based on the tradition variations which dates from the Renaissance of reconstructing original texts from manuscript. These efforts arose with the development of printing; given the ability to reproduce multiple copies of identical texts, scholars attempted to undo the “corruption” of a millenium of monastic copying and issue authoritative texts, both for the Latin and Greek traditions, and more importantly (especially with the text-based intellectual warfare of the Reformation) for the Greek and Latin Bibles. This approach assumes a lost base text which can be reconstructed through analysis of variations in subsequent editions.

While appropriate for the elimination of printing errors in editions of the Tripitaka, this approach is not effective in comparing the vastly different versions of many of the texts that the Tripitaka contains. The major problem is the lack of a base text. Existing manuscripts of the Sanskrit "originals" are always much later than the Chinese translations. These translations were often based on "incomplete" lost originals, on earlier smaller versions of the surviving Indic or Tibetan texts, and suffered from the limits of memorized versions, lost sections of originals, as well as problems arising from the fact that the Chinese Tripitaka is a mishmash of texts from varying traditions. Moreover, the texts continued to develop in India for almost a thousand years after they started being translated into Chinese.

For all of these reasons, as well as the problems in translation style discussed above, reconstruction of a base text from the Chinese translations is difficult if not impossible. Moreover, it is often irrelevant to the needs of scholars studying the texts, who are concerned not with an original infallible Word of Buddha, but with the development and adaptation of Buddhist ideas throughout the temporal and cultural sphere through which they were extended. The TEI approach to variant analysis is therefore not very helpful.

An additional practical issue is that of external versus internal markup. Comparison of differences between different versions of the same text is a form of analysis which is not a definitive part of the text, and which should not be included in the text body. This calls for a form of markup where external knowledge about texts is linked to the various versions without affecting the structure of those versions.

Kewen approach to variant analysis

Kewen (科文) refers to a style of study aids for sutras developed within the Tiantai school of Chinese Buddhism. Kewen are in the form of an outline, which today is usually printed as a tree-structured diagram (Figure 1【1】). They range in length from a single page for the *Heart Sutra* to 81 pages for the *Flower Adornment Sutra*. A thousand years of Tiantai scholarship has produced kewen for almost all of the sutras in the Chinese Tripitika.

A Kewen is divided into two parts. One part represents the structure of the text【2】; the second the doctrinal structure which forms the basis for Tiantai commentary. Where different commentators have produced different kewen for a single text, the structural portion should be identical, while the doctrinal portion may vary depending on the understanding of the commentator. However, because the methodology for developing the doctrinal portion of the analysis was laid out in considerable specificity by **Zhiyi**, the doctrinal structures do not vary enormously.

A Kewen thus provides hierarchical meta-data on the structure of a sutra. It has five advantages as a source of such meta-data: it is available from published literature, it provides a common structure for different versions of a text, it does not require choice of base text, and it is directly usable by two groups of potential users: students of the Tiantai tradition, and students within the Tiantai and Chinese integrative traditions.

The fact that kewen are available from the published literature (i.e. from Chinese editions of the Sutras) means that they can be entered as published, without requiring a project to analyze the structure of each sutra. Given that there are **21** volumes of around 1,000 pages each in the sutra section of the Taisho edition of the Chinese Tripitika, this represents **substantial potential savings** of time which can better be spent on analysis of the texts.

The provision of (ideally) a common structure for texts, especially versions of the same text, **along** with the fact that this does not require choice of base text, means that the problems discussed above with regard to textual evolution are bypassed. Adapting the structure represented by the Kewen, and linking it to each version of the target text, allows the analysis to be based on the structure, bypassing questions of how that structure is instantiated in any one translation or edition.

The fact that the results of so using kewen are directly usable by two groups of potential users: students of the Tiantai tradition, and students within the Tiantai and Chinese integrative traditions, means that the potential audience (and source of funds) for such a project are increased significantly beyond the small **group** of scholars who study the development of the textual tradition.

Kewen as a markup system for texts:

Kewen can be viewed as a markup system to Buddhist Sutras similar to a SGML markup description, such as a DTD and its associated tag set, to specific classes of text. The relations between the tree structure of kewen to a Buddhist Sutra functions like that of the DTD of SGML to text. A kewen is usually a hermeneutic explanation of Sutra by an honored Master of Buddhism. So, from different Masters, different kewen may by produced for a specific Sutra. Therefore, a kewen can be viewed as a specific markup to explain the content of a Sutra. In this respect, it is not a general markup like that of SGML. However, for different versions of the same Sutra, a kewen functions quite similarly to SGML, because it governs a set of different versions of the same Sutra just like a SGML markup description applies to specific class of texts. Therefore, our system enables users to compare different explanations from the views of different Masters on the same Sutra.

While implementing the markup (functions) kewen provides, each node of the tree of a kewen will be treated as a tag to explain the content of the Sutra. So, the kewen itself provide a tag set for markup. In our system, this tagging procedure is much easier than that of SGML tagging, because user doesn't have to memorize tag names; they all already appear on the tree kewen. In fact, those tags are transparent to users, i.e. they don't have to know the nodes of the tree of kewen are treated as tags. Our system can keep tracking of all the tagging automatically and provide on-line update functions for the convenience of users.

Kewen as a hyperlink system

The linkage between the nodes of kewen to their associated text streams in a Sutra can also viewed as a hyperlink system. However it is not like the usual hyperlink in many ways. First, the relation of this linkage is well defined, while as the relation of a hyperlink is usually unknown for general hyperlink systems. A node of the tree of a kewen can be considered as a knowledge chunk and the whole tree of a kewen can thus be considered as a tree-structured knowledge representation of what a Master known about the content of a Sutra.

Secondly, the tree structure of a kewen functions exactly like that of a thesaurus. Therefore, the hyperlink system provided by kewen is a hyperlink system with a specific thesaurus about the content of a Sutra. So, it is more than an ordinary hyperlink system. Browsing the tree of kewen can provide more knowledgeable operations than ordinary hyperlink systems.

Finally, since our system allows many to many mappings between a set of different versions of a Sutra and a set of different kewens from different authors, such as Masters and Buddhist scholars, a hyperlink in our system is not a one-to-one mapping between two text strings. In other words, a hyperlink in our system is much more structured and complex than an ordinary hyperlink. For example, a text string in a Sutra may have several hyperlinks to different nodes of various kewens, and a node of a kewen may link to different text strings of different versions of a Sutra. This flexibility and the richness of relations are not reachable by ordinary hyperlink systems.

Kewen as a knowledge/content representation system

For example, in the kewen of Heart Sutra by 周芷庵, there are nodes for: the person who is practicing (¯能修之人), what has been practiced (所修之法) and the degree or extent of his/her practice (修行境界), etc. It can be easily seen that these queries can not be realized/understood by any existing morphological retrieval systems. (Usually a morphological retrieval system mainly deals with syntactic structures of a text, only few with superficial semantically information. So, in general, they can not understand the content of a text.) Kewen is a tree-structured knowledge of what a Master understood about a Sutra. If a parser of the phases of kewen can be written, then, kewen can be used as a fundamental part to create an intelligent interface for retrieving the content of a Sutra, or be considered as a knowledge representation of understanding of a Sutra.

Description of computer-based use of Kewen for comparisons of translations

In order to demonstrate the feasibility of this approach to external markup of Buddhist texts, the Document Processing Laboratory at the Institute of Information Sciences, Academia Sinica on Taiwan has developed a prototype system, programmed by Mr. Derming Chuang. This system takes a kewen of the *Heart Sutra* and uses it as a structure to compare 11 Chinese and one Tibetan version of this popular text.

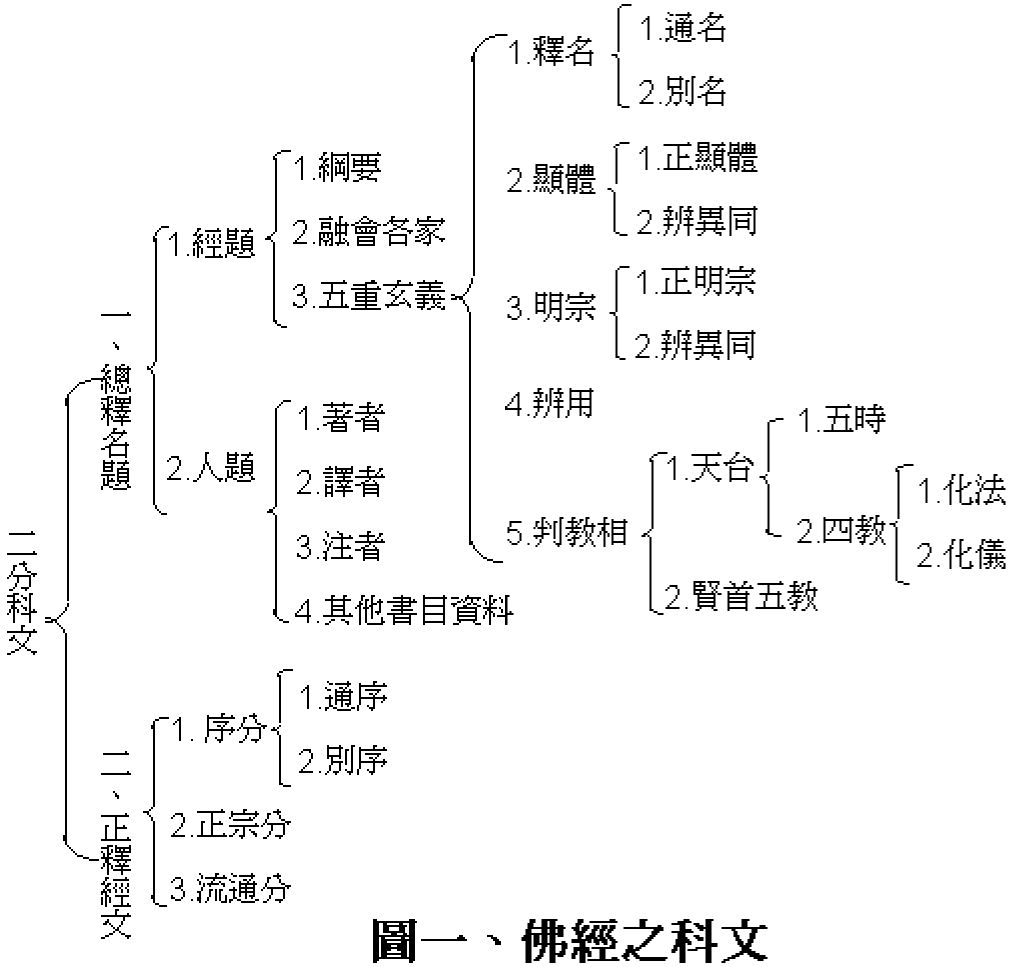
Conclusion and suggestions for further research

This prototype using kewen as a source for hierarchical meta-data on the structure of sutras demonstrates that it does indeed work as a base structure for comparing variations of the same text. In order to make the process practical, the following improvements need to be made:

1. A generalized data structure for kewen needs to be developed.
2. A method for non-programmers to input a printed kewen into this structure needs to be developed
3. A graphical syntax for linking the kewen graph to versions of a target text in the form of external markup to a published edition of digitized texts needs to be developed

4. The prototype needs to be rewritten as a general text comparison engine working off the stored kewen, the external markup, and the published digitized texts.

【1】 Kewen for Heart Sutra



圖一、佛經之科文

【2】The kepan(科判) of Tiantai usually has the following general form and content:

天台宗 智者大師說《法華經》時，以『**名、體、宗、用、相**』五重要項以明經中要旨。此所以用「重」字，是表示**名、體、宗、用、相** 五者有「次第相生」的性質的緣故。後人稱天台之「判教」或「科判」為「五重玄義」。

五重玄義：

一、 **名：**釋名也。即解釋此經名稱和命名的緣由（因實得名）。

二、 **體：**顯體也。即（依經名）顯示此經之本體（因名核實）。

三、 **宗：**明宗也。即說明（依本體）修行之旨趣。

四、 **用：**論用也。即論述（修行）此經之功效。

五、 **相：**判教相也。即（依上述四者）說明此經在佛教中之「相」，以隨眾生根機教化。

如，華嚴宗有：『**小、始、終、頓、圓**』五種教相之判。
天台宗則判為：『**藏、通、別、圓**』四種化法，與
『**頓、漸、秘密、不定**』四種化儀。