Patterns of Variation: The textual sources of the Chinese Buddhist Canon as seen through the CBETA edition

Christian WITTERN

The laureate of this collection of essays has for many years taken a keen and encouraging interest in my work relating to the digitization of the Chinese Buddhist scriptures, seeing it as a way to enhance the way research is currently done and as a new opening of possibilities. In this tribute, I therefore take this as an opportunity to recount some of the history of these endeavors and will also try to consider some research questions that have been hitherto difficult to deal with.

The first promise of digital text as I encountered it in the late 1980s was the immediate gratification of finding usage examples of a certain phrases by simply entering a keyword, pressing a button, and waiting a while for the result to miraculously manifest itself on the screen. We have since built on this promise and have now the potential to access a major portion of the world's cultural heritage, including the heritage of the Chinese cultural hemisphere and especially the texts of Buddhist provenance. In some respect, this is only a potential, since new barriers have been erected, including but not limited to the technical, political, economical and legal hindrances. While I will not proceed to discuss these issues here, it is important to remember that the digital medium is changing not only the way we do research, but also the whole environment within we act, changes that might provide more substantial influence on our work than we might wish.

Buddhist scriptures have been introduced to China and translated into Chinese over a period of more than 1200 years. Translation activity started in the Western Han period (1st century A.D.) and continued with various degrees of intensity until the 13th century.

When this translation process started, the Chinese had no coherent image of Buddhism and did not know that Buddhism had already

developed into a number of competing schools in India, with sometimes contradicting teachings and commandments. Towards the beginning of the translation activity, any text that seemed interesting had been translated, with many translations being redone several times over the centuries as the terminology and the understanding developed.

While this is not the place to go into a detailed history of the Buddhist canon and its translation into Chinese,[1] a few words to frame the narrative might be in place. It was about the end of the 5th century, that the growing corpus of translations of Buddhist texts from Sanskrit and Prakrit into Chinese had reached sizable proportions. Around this time, the need was felt to organize the received scriptures, scrutinize the contents and evaluate whether they were authentic translations or mere fabrications of Chinese origin. Although the actual collection of scriptures into some sort of canon did not occur right away, the fact that they were recorded in bibliographic catalogs contributed to gradually having the scriptures considered as an entity of itself, with heavily guarded entrance gates. At this point the collected translations were called 'All of the sutras' (一切經 *yiqiejing*). One example of usage for this term from Dunhuang has the date of 479. About half a millennium later, during the Song period, first usage of the modern term 大藏經 *dazangjing* can be found, which is today commonly used to refer to the Chinese Buddhist Canon.

To preserve authoritative copies of the scriptures and prevent corruption, a kind of printing technology was adapted toward the end of the 6th century A.D., when monks in a small monastery of Northern China embarked on a project to carve the most important scriptures in stone. The project was carried out over more than 300 years and today

[1] A lot of research is being done in this field, which makes it impossible to even attempt a full listing here. Among the most important recent studies, which include references to the research history, are Chikusa Masaaki 竺沙雅章 *Sō Gen bukkyōshi kenkyū* 宋元佛教文化史研究 (2000), Li Fuhua 李福华 and He Mei 何梅, *Hanwen fojiao dazangjing yanjiu* 汉文佛教大藏经研究 (2003) and Stefano Zacchetti *In Praise of the Light. A Critical Synoptic Edition with an Annotated Translation of Chapters 1-3 of Dharmarakṣa's Guan zan jing* 光贊經, *Being the Earliest Chinese Translation of the Larger* Prajñāpāramitā, (2005) pp. 74-132. There is also the research bibliography by Nozawa Yoshimi 野沢佳美, *Daizōkyō kankei kenkyū bunken mokuroku* 大蔵經關係研究文献目録 (1993) and subsequent additions.

we have access to about 14000 stone slabs, with hundreds of texts that comprise a major portion of the Chinese Buddhist Canon.[2]

This remained a comparatively isolated project however, and there was apparently no plan that detailed the contents and sources of the carvings; it also remained unclear in what way the texts to be cut into stone slabs where selected. Only with the beginning of woodblock printing in the 10th century, complete copies of the Chinese Buddhist Canon became available for the first time. Shortly after the establishment of the Song Dynasty (960), work began in the remote province of Sichuan on imperial orders, which resulted in more than 1000 separate texts or more than 5000 scrolls to be carved, printed and distributed all over the country; the first set was of the so-called 開寶蔵 Kaibaozang, completed in 983. Since then, more than 20 new printing sets have been produced in China, Korea and Japan, each slightly differing in content and arrangement, although new admissions to the canon had been tightly controlled ever since the Song dynasty.

The oldest edition, for which even the printing blocks are completely preserved, is that of the Tripiṭaka Koreana, the Korean edition of the Chinese Buddhist Canon, which was already critically collated from several sources. The woodblocks for this edition have been carved in the middle of the 13th century; it comprises 1521 texts in more than 6500 scrolls.

The edition now most widely used as a standard reference to the Chinese Buddhist Canon is the *Taishō Tripiṭaka* (大正新修大藏經 *Taishō shinshū daizōkyō*), edited by J. Takakusu and K. Watanabe, Tokyo 1924-1932. This has been revised, rearranged and edited according to modern philological and text-critical principles; the total number of works contained is 3053 in 85 volumes. While it does not contain all Buddhist scriptures of importance (there is a substantial amount of commentaries, records and historical texts in the *Supplement to the Chinese Buddhist Canon* (卍續藏經 *Zokuzōkyō*), which have not been included in the Taishō Tripiṭaka), it has served as the textual source for most of the digitization projects that attempted to digitize the Chinese Buddhist Canon.

It might be useful to turn to the background, history and aims of the CBETA project for a moment.

---

[2] Cf. Lothar Ledderose, *Carving Sutras into Stone before the Catastrophe*. Proceedings of the British Academy, vol. 125, pp. 381-454 (2004).

The last years of the 20th century have seen various efforts towards a complete digitization of the Chinese Buddhist Canon. Professor Lewis R. Lancaster, then of the University of California, was among the first to realize the potential of digital texts and the enormous need for exchange, cooperation and standardization in this field. In 1993, he assembled delegates from various Buddhist electronic projects in different languages and scripts, and founded the Electronic Buddhist Text Initiative (EBTI) as a forum for exchange of information and sharing of technology among these projects. Subsequent meetings of the EBTI have been held at Haein-sa 海印寺, Korea in 1994, Fokuang Shan 佛光山 Taipei in 1996 and Otani University 大谷大学, Kyoto, Japan in 1997 and, together with PNC, ECAI and SEER at Academia Sinica, Taipei in January of 1999; a similar joint conference was held the following year at the University of California Berkeley in January 2000, while the conference in 2001 was a EBTI only meeting hosted by Dongguk University in Seoul, Korea. The next meeting is scheduled to take place in 2008.

Table 1. Timeline of the early years of the digitization of Buddhist Scriptures

Beginnings in Japan and Taiwan:	
Kyoto:	Zenbunka kenkyūjo 禪文化研究所
Tokyo:	Indogaku Bukkyōgakugakkai 印度學佛教學學會
Kaohsiung:	Fokuang Shan 佛光山

Beginnings in USA and Korea:	
Berkeley, U.S.A:	UC Berkeley, Lewis Lancaster
Seoul und Haien-sa 海印寺	Research Institute of the Tripitaka Koreana
1993:	Founding of the Electronic Buddhist Text Initiative (EBTI)

Workshops and meetings:	
1993	Berkeley,
1994	Haiensa und Seoul,
1996	Taipei,
1997	Kyoto,
1999	Taipei
2000	Berkeley
2001	Seoul

Publications and activities:	
1995:	International Research Institute for Zen-Buddhism, Hanazono University, Kyōto: ZenBase CD1 (ca. 80 Chan/Zen Texte, Research Tools, Bibliographien, Indices)
1996:	Daejanggyong Research Institute, Seoul: Tripitaka Koreana[3] (Complete edition of the Tripitaka Koreana based on the photomechanichal reprint Seoul 1965).
1997:	Foundation of the Daizōkyō tekisuto detabesu kenkyūkai 大藏經テキストデータベース研究會 (Saṃganikīkrtaṃ Taiśotripiṭakaṃ: SAT[4]) in Tōkyō as an umbrella organization for a number of individual projects, subsequent publication of texts on the web site.
1998:	Formation of the Chinese Buddhist Electronic Text Association (CBETA) in Taipei.
1999:	First publication of an almost complete electronic edtion of the Taishō Tripitaka: Fomei dazangjing 佛梅大藏經, Hongkong.
2000:	CBETA publishes 56 volumes of the Taishō Tripitaka.
2000:	Second edition of the Tripitaka Koreana as a set of 15 CD-ROMs.

In February 1998, Venerable Shi Heng-ching 釋恆清, Taiwan University and Venerable Shi Huimin 釋惠敏, National Institute of the Arts, founded the Chinese Buddhist Electronic Texts Association (CBETA), to coordinate efforts in Taiwan and promote the creation of a new scholarly digital edition of the Chinese Buddhist scriptures. The present author was attending the founding meeting, joined CBETA in April 1998 and serves to date as an adviser to this project. CBETA was not planning to start from scratch with the input of Buddhist texts, but rather aimed at collecting and proofreading materials that had been put into electronic form elsewhere, thus ensuring a high reliability throughout the database.

CBETA had received a grant from the Yin-Shun Foundation of North-America and the initial plan was to consecutively release the complete canon of Chinese Buddhist scriptures (again according to the Taishō collection) within three to five years. A first release of six volumes of the Taishō Tripiṭaka, both on CD-ROM and on the Internet was made in December 1998, subsequent years have seen a steady flow of more extensive releases. At this point, 56 volumes of the Taishō collection and

[3] http://www.sutra.re.kr/english/default.asp
[4] http://www.l.u-tokyo.ac.jp/~sat/index.html

all the texts from the Zokuzōkyō that had not been already included in the Taishō are available without charge through the CBETA homepage.[5] This means that the superset of all the text contained in both collection is now available free of charge in a highly reliable format to everybody who has access to the Internet. And for those without convenient access, there is still a widely distributed CD-ROM, which is also free of charge.

As of early 2007, a complete breakdown of the texts contained in the CBETA electronic edition is as follows:

Table 2. Contents of CBETA Electronic Tripitaka as of 2007

Number of texts	3597
Number of juan	14034
Number of characters	147 721 972[6]
Individual characters used	35755

CBETA is making every effort to encode and markup the text using internationally accepted and widely used open standards like XML and TEI. CBETA is also closely cooperating with SAT on such important issues like the representation of rare characters in the texts.

The CBETA project grew largely out of a volunteer effort of people with interest in the digitization of Buddhist texts. On a BBS forum run out of Ven. Heng-ching's office at Taiwan University, discussions had been going on and even some preliminary tests had been completed which culminated in the release of the texts of Vol. 9 of the Taishō Tripiṭaka.

Inspired by the availability of the *ZenBase CD1*, which provided an example of how to achieve high quality digital texts even with limited ressources, a version of the Korean Tripiṭaka CD-ROM converted into the Taiwanese encoding Big5 and the ongoing effort by a Taiwanese businessman, Mr. Hsiao Chen-Kuo 蕭鎮國, who was sponsoring the keyboarding of the Taishō Tripiṭaka in mainland China, it was realized that although data were becoming rapidly available, there was the need to gather these data and pipe them trough a quality assurance process to make them reliable and suitable for use as a resource for academic research.

[5] http://www.cbeta.org/

[6] This is the count of Chinese characters only, discarding punctuation, space characters, numbers, etc.

In order to minimize the manpower and effort, it was decided from the start to rely heavily on supporting programs developed in-house and customized to the needs of team members.

There are three areas in which the methods applied by the CBETA team might serve as an example to other similar projects. These are:

- computer assisted proofreading
- consequent application of structural markup
- systematic handling of non-system characters

Each of these items will be discussed in more detail below.

As indicated above, CBETA tries to use information technology to minimize effort, manpower and cost while at the same time maintaining highest quality standards. In the proofreading process, this was achieved through a workflow that would try to optimize the time to find and correct errors in the sources. The proofreading process assumes that at least two, preferably three electronic versions of a text are available. If this is not the case, the 'Input group' is asked to prepare such copies as necessary. The electronic source files are then compared with a highly configurable program written in-house for that purpose and differences in these files are marked, separately for all input files. The proofreader than opens a program, that allows display of a scanned image of the original and the electronic text side by side. Using this program, she can jump to the locations that have been marked and have the original page displayed accordingly, in a way similar to OCR proofreading system. This allows minimizing the time used in the most time consuming process during proofreading, that is locating errors and finding out what version is correct.

The screenshot in Figure 1 shows a part of the interface for the proofreader. In the left part is the scanned image, with the red hair cross pointing to the character in question. The right part shows the result of the comparison of three files, with the one set of differences highlighted in black. Since some characters are very difficult to discriminate on the screen, identical occurrences in two of the source texts are highlighted in the action dialog. It is thus just one keystroke for the operator to make the necessary change in the file and move on.

There are of course some problems with this approach, for example it does require three independently created copies to be available and the

Figure 1: Proofreading

copies should also be created with different input techniques—it would not be useful to simply scan the same text multiple times and then use a OCR program to recognize the text. Identical errors in all three copies are of course not identified. Preliminary tests with subsequent manual proofreading of the texts have shown that the results are as reliable as expected, with an error quote approximately at 1 in 10000.

CBETA realized that using standardized markup for the creation of digital resources for Buddhist studies was a necessary condition for any further development of methodologies. Researchers will need to be able to incrementally add comments, definitions, pointers to related material as well as other meta-information about a text. Markup, in combination with other knowledge representation strategies can express the inherent information and retrieve it in ways that enable surprising new discoveries.

When CBETA was established, I was the only one in the team with experience in the application of markup to electronic text. I took it upon me to convince the team members, to use the TEI Guidelines as a base for this project. At the beginning, all markup was applied within the Research & Development group, but this turned out not to be practical. Since the proofreading group worked so closely with the texts, it was decided that in terms of order to reach an efficient workflow, it was the best place to apply markup to the texts. The tools in use by the group, however, would not allow to use SGML/XML and this would also place a very heavy entry-

barrier in terms of required skills (not to mention the language barrier, since at that time, virtually no relevant documentation was available in English).

At that time, the standard format for the texts as used by the proof-reading team was to have a location reference to volume, text number, page and line at the beginning of each line of the electronic text. It was then decided to extend this identifier by some columns and put shortcuts for structural markup there. Headings, footers, bylines and so on could easily be identified. The paragraphs in the source texts were clearly marked, so this information should simply be transformed into the 'simple' markup.[7] Figure 2 shows the beginning of text number 2067 from volume 51. The last three columns before the Chinese text starts are used for the 'simple' markup. Without going into too much detail here, 'P' will turn into markup for a paragraph, 'Q' starts a new division, 'A' is the author or translator and so fort on. If no markup is applicable, the '#' mark is used, the underscore is for lines that continue to belong to the previously mentioned markup entity. This ad-hoc markup is then transformed to XML based on a customized version of the TEI DTD with a `perl` program; all further editing then is done on the XML files. The file generated from the above text is shown in Figure 3, with some additional editing applied.

To understand the analysis that is to follow, it will be necessary to introduce some of the details of the encoding of the text critical apparatus as employed by CBETA. The encoding takes as its point of departure the text critical apparatus found in the Taishō source text. An example of the encoding that is found in the Taishō is given in Figure 4.

In this case, the Taishō text has the following apparatus: 「殖＝植【三】【宮】」. In plain English, this means that the character 殖 used in the base text[8] was actually written as 植 in all the versions used for comparison, only the Korean edition has 殖, but the editors of the Taishō decided to keep this.

---

[7] During the process of introducing this workflow, it was found that there was a need for fine-tuning the markup beyond the unit of the line, since in some cases new paragraphs would not start a new line, so later some markers were introduced that appeared within the lines of the text.

[8] According to *Showa hōbō mokuroku* 昭和法寶目録, Vol. 1, p. 220, this is the text of the Korean Tripitaka, which has been compared text critically to the Song, Yuan and Ming editions, as well as the Song edition found in the Imperial Household Library【宮】.

```
T51n2067_p0012b15N##No. 2067
T51n2067_p0012b16J##[16] 弘贊法華傳卷第一
T51n2067_p0012b17_##
T51n2067_p0012b18A## 藍谷沙門惠詳撰
T51n2067_p0012b19P#1 圖像第一 第一卷 翻譯第二 第二卷
T51n2067_p0012b20P#1 講解第三 第三卷 修觀第四 第四卷
T51n2067_p0012b21P#1 遺身第五 第五卷 誦持第六 ( 第六卷第七卷第八卷 )
T51n2067_p0012b22P#1 轉讀第七 第九卷 書寫第八 第十卷
T51n2067_p0012b23_##
T51n2067_p0012b24Q## 圖像第一
T51n2067_p0012b25P#1 西域祇洹寺寶珠寶塔內說此經像
T51n2067_p0012b26P#1 西域擬前說法金像
T51n2067_p0012b27P#1 西域鷲 [ 山 / 爭 ] 山說此經像
T51n2067_p0012b28P#1 宋釋惠豪造靈鷲山圖
T51n2067_p0012b29P#1 後魏太祖造耆闍崛山圖
T51n2067_p0012c01P#1 晉殷夫人造法華臺 宋謝婕妤造法華寺
T51n2067_p0012c02P#1 後魏太常卿鄭瓊造法華堂
T51n2067_p0012c03P#1 晉釋惠力造多寶塔
T51n2067_p0012c04P#1 宋劉佛愛造多寶寺多寶塔
T51n2067_p0012c05P#1 齊舍人徐儼造石多寶塔
T51n2067_p0012c06P#1 唐悟真寺釋法誠造多寶塔法華塔 ( 并 ) 法華
T51n2067_p0012c07P#1 臺唐國子祭酒蕭瑝造多寶塔
T51n2067_p0012c08P#1 宋路昭太后造普賢像 宋釋道悶作普賢
T51n2067_p0012c09_## 齋
T51n2067_p0012c10P#1 宋釋僧苞作普賢齋
T51n2067_p0012c11P## 案祇洹圖云。前佛殿東樓上層。有白銀像。像
T51n2067_p0012c12_## 內有七寶樓觀。樓觀內有寶池寶花。花上有
T51n2067_p0012c13_## 白玉像。池中蓮花內。有白銀塔。於塔心中。有
T51n2067_p0012c14_## 真珠塔。塔內有釋迦多寶二像。說法花經第
T51n2067_p0012c15_## 七會者。又云。妙法華經。事同花嚴。波若多會
T51n2067_p0012c16_## 說之。今之所翻。當第三會。又云。複殿四臺五
T51n2067_p0012c17_## 重。上層有吠摩尼珠。此珠。過去諸佛。曾於
T51n2067_p0012c18_## 中說法花。三變淨土。隨經所有。於中具現。
T51n2067_p0012c19P## 案西域書傳。中天竺摩揭陀國恒河南有故
T51n2067_p0012c20_## 城。周七十餘里。荒蕪歲久。基趾尚存。昔人壽
T51n2067_p0012c21_## 無量歲時。號拘蘇摩補修羅城。唐言香花宮
T51n2067_p0012c22_## 城。逮人壽數千歲時。更名波吒釐子城。是巴
T51n2067_p0012c23_## 連弗邑也。去此城西南四百餘里。渡尼連禪
T51n2067_p0012c24_## 河。至伽耶城。城西南二十餘里。至菩提樹。金
T51n2067_p0012c25_## 剛座等。菩提樹東。渡大河入大林野。行百餘
T51n2067_p0012c26_## 里。至 [ 奚 * 鳥 ] 足山。[ 奚 * 鳥 ] 足山東北百餘里。至大山。入
```

Figure 2: A text file with simple markup

In the CBETA electronic edition, this fact is established in a variety of different ways. For one thing, where the Taishō editors used the reference 【三】as shorthand for the three editions of the Song, Yuan and Ming,[9] these editions are treated separately in the CBETA edition. Also, the notation is translated into a form that is machine-readable and can be used to re-

[9] To be more precise, these are the socalled Zifu Zang 資福藏 (Song 宋) printed ca. 1241-1252, Puning Zang 普寧藏 (Yuan 元), printed ca. 1277-1290 and Jingshan Zang 徑山藏 (Ming 明), printed from 1589.

```
<lb n="0012b24" ed="T"/><div1 type="other"><mulu level="1" label="1 圖像 " type=" 其他 "/><head> 圖像第一 </head>
<lb n="0012b25" ed="T"/><list>
<item id="itemT51p0012b2501"> 西域祇洹寺寶珠寶塔內說此經像 </item>
<lb n="0012b26" ed="T"/><item> 西域擬前說法金像 </item>
<lb n="0012b27" ed="T"/><item> 西域鷲 &CB00123; 山說此經像 </item>
<lb n="0012b28" ed="T"/><item > 宋釋惠豪造靈鷲山圖 </item>
<lb n="0012b29" ed="T"/><item> 後魏太祖造耆闍崛山圖 </item>
<pb n="0012c" id="T51.2067.0012c" ed="T"/>
<lb n="0012c01" ed="T"/><item> 晉殷夫人造法華臺 </item><item> 宋謝婕妤造法華寺 </item>
<lb n="0012c02" ed="T"/><item> 後魏太常卿鄭瓊造法華堂 </item>
<lb n="0012c03" ed="T"/><item> 晉釋惠力造多寶塔 </item>
<lb n="0012c04" ed="T"/><item> 宋劉佛愛造多寶寺多寶塔 </item>
<lb n="0012c05" ed="T"/><item> 齊舍人徐儼造石多寶塔 </item>
<lb n="0012c06" ed="T"/><item> 唐悟真寺釋法誠造多寶塔法華塔 <note place="inline"> 并 </note> 法華 </item>
<lb n="0012c07" ed="T"/><item> 臺唐國子祭酒蕭璟造多寶塔 </item>
<lb n="0012c08" ed="T"/><item> 宋路昭太后造普賢像 </item><item> 宋釋道悶作普賢
<lb n="0012c09" ed="T"/> 齋 </item>
<lb n="0012c10" ed="T"/><item id="itemT51p0012c1001"> 宋釋僧苞作普賢齋 </item></list>
<lb n="0012c11" ed="T"/><div2 type="other"><p id="pT51p0012c1101"> 案祇洹圖云。前佛殿東樓上層。有白銀像。像
<lb n="0012c12" ed="T"/> 內有七寶樓觀。樓觀內有寶池寶花。花上有
<lb n="0012c13" ed="T"/> 白玉像。池中蓮花內。有白銀塔。於塔心中。有
<lb n="0012c14" ed="T"/> 真珠塔。塔內有釋迦多寶二像。說法花經第
<lb n="0012c15" ed="T"/> 七會者。又云。妙法華經。事同花嚴。波若多會
<lb n="0012c16" ed="T"/> 說之。今之所翻。當第三會。又云。複殿四臺五
<lb n="0012c17" ed="T"/> 重。上層有吠摩尼珠。此珠。過去諸佛。曾於
<lb n="0012c18" ed="T"/> 中說法花。三變淨土。隨經所有。於中具現。</p></div2>
<lb n="0012c19" ed="T"/><div2 type="other"><p id="pT51p0012c1901"> 案西域書傳。中天竺摩揭陀國恒河南有故
<lb n="0012c20" ed="T"/> 城。周七十餘里。荒蕪歲久。基趾尚存。昔人壽 </p>
```

Figure 3: The text from Figure 2 converted to XML

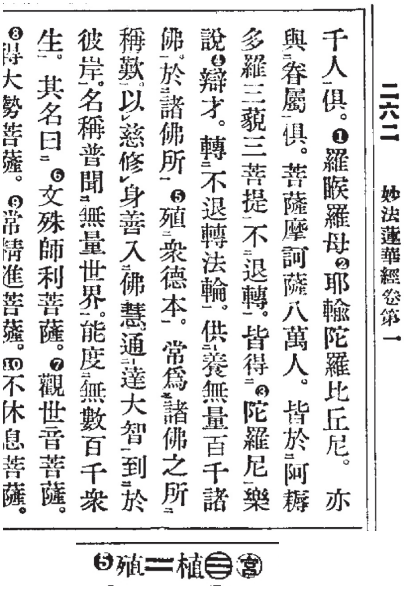

Figure 4: An excerpt from the beginning of the *Lotus Sūtra* in the *Taishō Tripitaka*, Vol. 9, p. 2.

construct all of the texts that are reported through the Taishō edition.[10] In the form used in the CBETA source files, this would then look as follows:

```
<app>
 <lem>殖</lem>
 <rdg wit="【宋】【元】【明】【宮】">植</rdg>
</app>
```

Figure 5: Text critical markup according to the TEI

The text critical apparatus is written here according to the *Guidelines for Electronic Text Encoding and Interchange*[11] After using this method for a while, the CBETA editors found that what they were doing was not simply transcribing the existing, printed Taishō edition to a different representation, but in fact the creation of a new edition: Editorial judgment was used to establish readings that differed from those giving in the Taishō.[12] It was thus necessary to refine the representation as follows:

```
<app from="beg0002005" to="end0002005">
  <lem wit="【大】">殖</lem>
  <rdg resp="Taisho" wit="【宋】【元】【明】【宮】">植</rdg>
</app>
```

Figure 6: Text critical markup as initially employed by CBETA

[10] This works of course only so far as the Taishō is faithfully reporting the editions consulted. For a truly new text critical edition, not only the Taishō source texts, but also the wealth of material that has become available since the Taishō edition was established would have to be taken in account. While this might be feasible for individual texts, it would be difficult to do it properly for the whole CBETA collection. But for the argument here, this question is irrelevant, since the Taishō edition as reported through CBETA is taken as a given and as the object of analysis.

[11] See Burnard and Sperberg McQueen (2002), Vol. 1, p. 481ff, esp. 484.

[12] While this seems obvious in some sense, it is by no means the usual mind-set for those involved in the creation of digital editions. The editors of the SAT database, for example, quite explicitly adopted the editorial policy of "reproducing the text of the Taishō Canon as it is, with all its mistakes" (a statement by the late Ejima Yasunori 江島惠教, personal communication, June 1998). Just as any other reprint, even a facsimile reprint, creates a new edition, all the more is a digital edition, with its need to interpret and asses every single character by necessity a new edition. On this point, see also Shillingsburg, *From Gutenberg to Google*, p. 12.

As can be seen,【大】is used as a sigil for the Taishō edition, and in addition the fact that the other readings are reported through the Taishō and not directly, is recorded through the "resp" attribute. This now allows unmistakably the recording of editorial judgments, as for example in the following case:

```
不從佛聞法，常行不善<app>
              <lem resp="CBETA.maha" wit="【CBETA】【麗】【磧砂】"
       cf1="K09_p0753a08 事" cf2="Q09_p0151a12 事">事</lem>
  <rdg wit="【大】">時</rdg>
</app>，色力及智慧，斯等皆減少。[13]
```

Figure 7: Emendation of the Taishō text as employed by CBETA

Here, one of the CBETA editors corrected what appears to be a misprint in Taishō by comparing it to the *Tripitaka Koreana* and the *Qisha edition* of the Lotus Sutra; the reference to the location of this stanza in these texts is given as well, to facilitate verification.

There is no explicit statement on the editorial aims and principles followed by CBETA, but from the existing evidence it can be clearly deducted that the guiding principle of the textual editing was to achieve an ideal text that is free of errors and internally consistent, rather than the reconstruction of some earlier stage in the textual transmission of the scriptures.[14]

The *CBETA Chinese Electronic Tripitaka Series* with its detailed critical apparatus provides ample material not just for reading, locating of quotations and other traditional research activities, but also for new ways of "reading" through program-driven analysis. There are many possibilities, but I will here limit myself to only a few examples. Given the countability

[13] This is on T. 09, p. 24c10. CBETA also silently changed the interpunction. The XML files containing the description to this level of detail have been released to the public and can be consulted at http://www.cbeta.org/xml. In these files, the text critical apparatus has been moved to the end, to facilitate the reading and processing of the text. In addition, the footnotes as they appear in the Taishō are preserved as well.

[14] In text-critical editorial theory there is some debate on what the aim of a text-critical edition is. A useful summary of Anglo-American practice (authorial oriented) and German practice (text oriented) is given in Hans-Walter Gabler et.al. (ed.), *Contemporary German Editorial Theory*, Ann Arbor 1995, pp. 2-12 and *passim*.

of textual variants, one of the first questions that come to mind is: Which texts have the most textual variants (and thus probably the highest degree of corruption/interest/debates)? Which texts do have little recorded variants? And, focusing on the textual variants themselves, which are frequent textual variants? Are there patterns that can be observed? I will try to answer some of these questions in this section, but will first give a breakdown of all textual witnesses recorded in the CBETA edition and their count (Appendix, Table 1).[15]

In the 2471[16] texts analyzed here,[17] there are a grand total of 666376 variant locations recorded.[18] The table gives the witnesses for these variants, in descending order. As can be seen, Taishō【大】is by far the most frequent, which simply means that in most cases, the CBETA editors did not need to interfere with the text. It is nevertheless remarkable that the【CBETA】witness ranks surprisingly high with more than 10000 conjectures in more than 1200 text made to the source edition of the Taishō, which are in most cases misprints.[19] This means that roughly

[15] To avoid interruption of the text flow here, the larger tables have been placed as an appendix at the end. Since considerations of space do not allow printing the whole tables here, they are available as electronic text files at http://www.kanji.zinbun.kyoto-u.ac.jp/~wittern/papers/patterns/.

[16] This is according to the division of textual units employed in the CBETA electronic edition. Depending on how to consider texts that share the same entry number in the Taishō edition and other minor differences, the total number does vary lightly. The *WWW Database of Chinese Buddhist texts* has 2418 items in this group.

[17] The data on which this analysis based is as of October 10th, 2006. This might be the appropriate place to note that the figures should be read more as indications of magnitude, rather than as exact values.

[18] This number gives the number of locations in the text for which variants are recorded, in the TEI notation used by CBETA, this is the number of `<lem>` occurrences in the texts. Since a given `<lem>` can have multiple correspondences in different textual witnesses, the total number of recorded differences, that is the number of `<rdg>` elements is much larger, 1.511 439 in this case.

[19] It should be noted here, that although CBETA made the editorial decision to prefer these emendations to the erroneous Taishō text, the fact that they are recorded in a precise and machine-processable form makes it a trivial exercise to re-constitute the exact Taishō text where needed and the text browser *CBReader* which is bundled with the CBETA CD-ROM does in fact offer this as a user-configurable option.

1 in 2 texts of the Taishō collection has seen at least one correction, with the average at 5 corrections per emended text. Another fact that might be surprising about this table is that there are quite a number of variant locations—more than 12000 in more than 300 texts—for which the Taishō does not record the textual witness. It would require a lot of research to try to recover these lost witnesses.

One question that might be asked at this point is, what texts are more likely to contain textual variations, or more generally, what kind of patterns are visible with regard to the textual variation. For this purpose, I have calculated the "density" of variation, which gives the number of textual variants divided by total length of the text. Table 2 in the appendix gives a small excerpt of this table, listing the texts with the highest density of variation.

Some interesting observations can be made here. A naïve view might hold that the oldest texts are most likely to have a high degree of variation due to their long exposure to a scribal tradition that might introduce errors to the text, but this is not the case. The oldest texts in the Chinese Buddhist tradition, the 四十二章經 Sishier zhang jing [*Sutra in 42 sections*] translated in A.D. 67 by Kāśyapa Mātaṇga and Dharmarakṣa for example, does not appear in this selection of frequently emended texts at all; a view of the full list reveals that it comes at position 816 of 2471, which means that roughly a third of all texts have a higher density of textual variation. There are only very few texts earlier than Tang at all (5, including doubtful texts), but with 70% by far the larges part of texts entered the record under the Tang. While it is true that quite a large number of texts have been written during this period, they account for only about 36% of all the texts in the Chinese Buddhist canon[20] so there seems to be a significant bias towards texts from the Tang in this list. A closer look at Table 1 reveals however, that not only (or maybe not mainly) the period of

[20] See the C. Wittern, *WWW Database of Chinese Buddhist texts* http://www.kanji.zinbun.kyoto-u.ac.jp/~wittern/can/can2/ind/canwww.htm and C. Wittern, *Entrance Through the Scriptures: Catalogues and Electronic Text as a New Gate to the Buddhist Tradition*, forthcoming in: Chung-hwa Buddhist Journal, No. 21 (2007), Figure 2. This figure only includes the 2471 texts under consideration, of which 893 originated under the Tang. For all canonical texts, the value would be roughly 26%.

origin seems to contribute to the high number of textual variants, but also the type of scripture, or to use the descriptor employed here, the division of the canon a scripture is placed in. By this measure, again about 2/3 of all texts fall under one value, which is the section of esoteric scriptures, which indeed was most productive under the Tang. Another characteristic of the esoteric scriptures is their high proportion of unusual characters used to transcribe dharāni and other esoteric formula, or even Siddham characters. All these are factors that contribute to a comparatively less stable scriptural tradition, which seems to be the reason for this comparatively high textual variation.

There are a lot of more questions to be asked that could be answered by analyzing this material. For the moment, it should suffice to pursue just one more path. Here, an answer is sought to the question of what characters are most frequently mistaken for others, or more to the point, what characters are frequently corrected by CBETA in the Taishō? With this knowledge, we could then proceed to look at the patterns of variation for specific characters and their distribution among text witnesses. Table 3 in the appendix gives the most frequent of those characters that have been frequently used in exchange for each other. The characters are listed separately according to whether they appeared as a lemma or as a variant reading, the table is given in descending order of the total number of occurrences. In addition to that, the question has been asked, whether the given variant is of visual, graphical nature or whether the characters in question do have a phonetic relationship. No clear pattern of preference for either type of variation could be established; both seem to occur with approximately the same frequency.

An attempt was made here to recount some of the background of the development of the *CBETA Electronic Tripitaka* and at the same time to explore some of the new looks at the textual tradition that have become possible due to the specific ways of how this digitization was undertaken. The surface of analytic possibilities have been scarcely scratched, but the resulting raw data and immediate results are made available, so as to allow further experiments and explorations.

APPENDIX

This appendix gives the tables discussed in the text. Due to limitations of space, only excerpts of the data can be given. A complete set, including additional data is available for download at http://www.kanji.zinbun.kyoto-u.ac.jp/~wittern/papers/patterns/.

Table 1: Witnesses used in the CBETA Electronic Tripitaka

This table shows the witnesses, by number of texts that use them. The first column gives the number of texts (occurrences), the second the sigil used by CBETA, followed by the number of text locations that refer to this witness.[21]

occurrence	witness	No of instances	Explanation[22]	Notes and examples
2232	【大】	665188	Taishō as witness	
1460	【明】	351486	The 'Ming Edition' A.D. 1601	Ming edition
1405	【元】	309115	The 'Yuan Edition' A.D. 1290	Yuan edition
1403	【宋】	300647	The 'Sung Edition' A.D. 1239	Song edition
1205	【CBETA】	10321	CBETA correction	
844	【宮】	215614	The Old Sung Edition [A.D. 1104-1148] belonging to the Library of the Imperial Household	Old Song edition
517	【甲】	130826		
380	【麗】	2973	The 'Kao-Li Edition' A.D. 1151	Koryŏ edition
343	【?】	12069		

[21] There is a similar, though less exhaustive table in Shi Huimin, et.al. "Techniques for Producing Critical Editions of Digital Versions of Ancient Texts: The Case of the CBETA Electronic Text of the Taishō Canon," *Journal of the study on Kanji Culture* No 1, p. 130. Out of the total of 58 witnesses, the printed table shows only the 29 witnesses that occur more than 5 times. The limits of the current description is visible here: In some cases, the same sigil describes unrelated editions, like 【A】【B】 or 【流布本】. This has to be taken in account for more detailed analysis.

[22] Where applicable, these explanations are taken from the corresponding tables at the end of each Taishō volume.

occurrence	witness	No of instances	Explanation	Notes and examples
213	【聖】	80406	The Tempyō Mss. [A. D. 729-] and the Chinese Mss. of the Sui [A. D. 581-617] and Tang [A. D. 618-822] dynasties, belonging to the Imperial Treasure House Shōsō-in at Nara, specially called Shōgo-zō	Shōgozō collection
178	【乙】	50279		
122	【磧砂】	1612		
74	【丙】	8501		
70	【unknown】	99		
57	【原】	235		
48	【南藏】	246	Appears in vols. T03, T09, T11, T12, T15, T16, T17, T25, T26, T27, T28, T32, T50, T51, T52, T53, T54, T55	Example:T3, p. 110, note 9
38	【聖乙】	12390	Another copy of the Shōgozō collection	Shōgozō collection (2)
26	【知】	4981	The Tempyō Mss. of the monastery 'Chion-in'	Chion-in edition
19	【丁】	2173		
19	【北藏】	112	Appears in vols. T09, T10, T11, T16, T17, T25, T28, T30, T32, T52, T53	Example:T9, p. 500, note 1
13	【明異】	34	Appears in vols. T02, T05, T06, T13, T14, T15	Ex:T2, p. 353, note 6
7	【久】	347	The Tempyō Mss. belonging to the Kuhara Library	Kuhara edition
7	【和】	1335	Ninnaji Mss. by Kūkai and others. C. 800. A. D.	Ninna-ji edition
7	【嘉興】	9	Jiaxing edition	T46, T47, T48, T49, T51
7	【敦】	430	Stein Mss. from Tun-huang	Dunhuang editions
6	【A】	1342		T21, T40, T44
6	【B】	446		T21, T40, T44
6	【流布本】	368	Appears in vols. T12, T29, T31	Ex:T12, p. 265, note 5
6	【石】	8493	The Tempyō Mss. of the monastery 'Ishiyama-dera'	Ishiyama-dera edition

Table 2: Density of Variation

The following table lists the texts with the highest "density" of variation. This is calculated as the number of <app> entries that record a variant of characters, in relation to the number of characters of a text. Listed are only 34 texts with more than 10000 characters and a density of more than 0,015.

count	density	title	date[23]	textkey	witness list	division
785	0,0248	梵語雜名	唐	T54N2135	【CBETA】【元】【大】【宋】【明】【甲】	事彙部
220	0,0198	火吽軌別錄	唐	T18N0914	【?】【乙】【大】【甲】	密教部
277	0,0191	注進法相宗章疏	日本	T55N2181	【?】【大】【甲】	目錄部
685	0,0182	金剛峰樓閣一切瑜伽瑜祇經	唐	T18N0867	【CBETA】【乙】【大】【甲】	密教部
550	0,0177	阿吒婆拘鬼神大將上佛陀羅尼經	唐	T21N1238	【?】【CBETA】【乙】【原】【大】【甲】	密教部
1169	0,0174	大毘盧遮那成佛神變加持經蓮華胎藏悲生曼荼羅廣大成就儀軌供養方便會	唐	T18N0852A	【?】【CBETA】【丙】【乙】【大】【甲】	密教部
236	0,0174	金剛頂勝初瑜伽經中略出大樂金剛薩埵念誦儀	唐	T20N1120A	【大】【甲】	密教部
271	0,0172	大方廣佛華嚴經金師子章	唐	T45N1881	【大】【甲】	諸宗部
1019	0,0169	十一面觀自在菩薩心密言念誦儀軌經	唐	T20N1069	【?】【CBETA】【乙】【元】【大】【宋】【明】【甲】【麗】	密教部

[23] As in traditional Chinese catalogs, this indicates the dynasty for texts originating from China, for other areas, only the area is indicated.

count	density	title	date	textkey	witness list	division
291	0,0167	大使咒法經	唐	T21N1268	【大】【甲】	密教部
1299	0,0166	攝大毘盧遮那成佛神變加持經入蓮華胎藏海會悲生曼荼羅廣大念誦儀軌供養方便會	唐	T18N0850	【?】【CBETA】【乙】【元】【大】【宋】【明】【甲】	密教部
216	0,0166	大黑天神法	唐	T21N1287	【?】【乙】【大】【甲】	密教部
802	0,0165	成就妙法蓮華經王瑜伽觀智儀軌	唐	T19N1000	【CBETA】【乙】【大】【明】【甲】	密教部
356	0,0164	大聖妙吉祥菩薩祕密八字陀羅尼修行曼荼羅次第儀軌法	唐	T20N1184	【?】【CBETA】【丙】【乙】【大】【明】【甲】	密教部
386	0,0163	藥師如來觀行儀軌法	唐	T19N0923	【?】【CBETA】【大】【甲】	密教部
2035	0,0161	洛陽伽藍記	元魏	T51N2092	【CBETA】【丁】【丙】【乙】【內】【大】【己】【戊】【甲】	史傳部
385	0,0161	聖閻曼德迦威怒王立成大神驗念誦法	唐	T21N1214	【?】【丙】【乙】【大】【明】【甲】	密教部
517	0,0160	大威怒烏芻澀摩儀軌經	唐	T21N1225	【?】【乙】【大】【明】【甲】	密教部
4330	0,0159	金剛般若論會釋	唐	T40N1816	【A】【B】【CBETA】【乙】【原】【大】【甲】	論疏部
1039	0,0159	法華玄贊義決	唐	T34N1724	【?】【乙】【原】【大】【甲】	經疏部
201	0,0159	金剛頂經觀自在王如來修行法	唐	T19N0931	【?】【丙】【乙】【大】【明】【甲】	密教部
191	0,0158	天地八陽神咒經		T85N2897	【?】【CBETA】【大】【甲】	古逸部全．疑似部

count	density	title	date	textkey	witness list	division
287	0,0157	葉衣觀自在菩薩經	唐	T20N1100	【?】【CBETA】【乙】【元】【大】【宋】【明】【甲】【麗】	密教部
173	0,0157	大聖天歡喜雙身毘那夜迦法	唐	T21N1266	【?】【丁】【丙】【乙】【大】【明】【甲】【聖】	密教部
496	0,0155	仁王護國般若波羅蜜多經陀羅尼念誦儀軌	唐	T19N0994	【?】【CBETA】【乙】【大】【明】【甲】【麗】	密教部
999	0,0154	尊勝佛頂脩瑜伽法儀軌	唐	T19N0973	【?】【CBETA】【丙】【乙】【原】【大】【甲】	密教部
470	0,0154	金剛手光明灌頂經最勝立印聖無動尊大威怒王念誦儀軌法品	唐	T21N1199	【CBETA】【丁】【丙】【乙】【大】【明】【甲】【聖】【麗】	密教部
937	0,0154	金剛頂瑜伽千手千眼觀自在菩薩修行儀軌經	唐	T20N1056	【?】【丁】【丙】【乙】【元】【大】【宋】【明】【甲】	密教部
428	0,0153	千眼千臂觀世音菩薩陀羅尼神咒經	唐	T20N1057B	【元】【大】【宋】	密教部
707	0,0152	甘露軍荼利菩薩供養念誦成就儀軌	唐	T21N1211	【?】【丙】【乙】【元】【大】【宋】【明】【甲】	密教部
253	0,0152	兩部大法相承師資付法記	唐	T51N2081	【?】【CBETA】【丙】【乙】【大】【甲】	史傳部
694	0,0151	蘇悉地羯羅供養法	唐	T18N0894A	【?】【CBETA】【乙】【大】【明】【甲】【麗】	密教部
404	0,0151	金剛頂瑜伽護摩儀軌	唐	T18N0908	【?】【丙】【乙】【元】【大】【宋】【明】【甲】	密教部
730	0,0150	佛説孛經抄	吳	T17N0790	【元】【大】【宋】【宮】【明】【聖】【聖乙】	經集部

Table 3: Characters with high variation

The most frequent characters that frequently appear as either lemma or reading are listed here, given is the character, the number of occurrences as lemma or reading, the total of these two and the type. The list is ordered by total number in descending order; giving only those with a value of 25 or more.

character	lemma	readings	total	type[24]
諡	484		484	both
謚		484	484	both
己	199	61	260	visual
羨	213		213	both
羨		213	213	both
刺	172	35	207	visual
刺	35	172	207	visual
日	169	11	180	visual
曰	11	169	180	visual
陝	179		179	both
陝		179	179	both
辦	154	9	163	both
采	157		157	visual
釆		157	157	visual
祛	142		142	phonetic
袪		142	142	phonetic
已	119	11	130	both
斂	129		129	both
斂		129	129	both
僭	108		108	both
偸	108		108	visual
僭		108	108	both
瑜		108	108	visual

[24] There are three types: 'visual' for variants that can be seen as deriving from the visual appearance of a character, 'phonetic' for those variants, where the reading seems to provide the clue for the variants and 'both' where both possibilities are plausible.

character	lemma	readings	total	type
汩	96		96	visual
汨		96	96	visual
己	23	61	84	visual
已	61	11	72	both
圮	60		60	visual
密	45	15	60	both
蜜	15	45	60	both
圯		60	60	visual
士	41	17	58	visual
土	17	41	58	visual
弈	34	22	56	both
奕	22	34	56	both
摶	53		53	visual
搏		53	53	visual
韄	52		52	
二	18	34	52	none
韄		52	52	
昧	34	17	51	both
忘	28	23	51	both
妄	23	28	51	both
味	17	34	51	both
婇	47		47	phonetic
三	34	13	47	none
二	13	34	47	none
綵		47	47	phonetic
秖	46		46	both
入	32	14	46	visual
人	14	32	46	visual
秖		46	46	both
杞	45		45	visual
若	23	20	43	visual
二	9	34	43	none
未	30	12	42	visual
一	24	18	42	none
末	12	30	42	visual
殊	40		40	none
珠		40	40	none

character	lemma	readings	total	type
緣	27	11	38	visual
戌	23	15	38	visual
如	20	18	38	visual
知	18	20	38	visual
戍	15	23	38	visual
祿	11	27	38	visual
瞻	10	27	37	visual
嗚	27	9	36	visual
鳴	9	27	36	visual
裟	26	9	35	visual
空	21	14	35	–
如	17	18	35	visual
巳	11	23	34	both
辨	9	25	34	both
薹	30		30	
幢	30		30	none
自	22	8	30	visual
憧		30	30	none
苦	20	9	29	visual
歡	17	12	29	visual
歎	12	17	29	visual
惑	28		28	phonetic
空	14	14	28	–
瞻	27		27	visual
成	19	8	27	visual
或	8	19	27	visual
曝	26		26	
券	26		26	both
間	13	13	26	visual
問	13	13	26	visual