The Structure and Function of the Interlinked Electronic CJK-English and Buddhist CJK-English Dictionaries

A. Charles Muller, Professor of Humanities, Toyo Gakuen University

Our current age offers us dramatic new possibilities in terms of the exchange and development of textual research resources, as we can now gather and transmit information with an ease and rapidity which was inconceivable through earlier media. Although most people will no doubt always prefer to have a hard copy of a book or lengthy article to sit down and read, lexicographical and encyclopedic style reference materials, which have relatively brief and compartmentalized data formats, are extremely well-suited for the digital domain, as they can be furnished with search and retrieval capabilities which are impossible in paper form. The twin CJK lexicographical models presented here stand as an early model of the possibilities such digital reference materials.

History

The two digital lexicons are a (1) dictionary of pre-modern East Asian characters and compounds, compiled primarily from the literary, philosophical and historical fields of inquiry (hereafter abbreviated as CJKDict) and (2) a dictionary of East Asian Buddhist terms (hereafter abbreviated as BDict), both of which are compiled with the same structural format. I began the compilation of these two dictionaries in 1986 at the outset of my graduate studies. Since my focus was East Asian Buddhism, I needed to read original Buddhist and non-Buddhist documents in Classical Chinese. Also, since I had a fairly strong interest in Confucianism and Taoism and was studying all of the fundamental classical texts of these traditions, I was necessarily exposed to large amounts of non-Buddhist vocabulary.

I was well aware even as an undergraduate student of the extreme dearth of lexicographical materials available in English for my fields. For Classical Chinese, there were the *Giles* and *Mathews* dictionaries, both of which were venerable compilations, but clearly outdated, and limited in many ways. Similarly, for Buddhism, there was the *Soothill* dictionary, also over fifty years old, in many ways quite uneven, and lacking in pronunciations and scholarly references. There were also a few other more recently published Buddhist dictionaries, but these were generally limited in scope to a particular tradition, or sparse in terms of scholarly explanation. Because of the relatively small scope and inferior quality of these compilations, the English-speaking researcher in East Asian philosophy and religion is to the present day, forced to rely exclusively on East Asian reference materials. While the ability to work with

East Asian secondary materials and reference works should always be a basic requirement in our field, the paucity of English-language reference works seems to be rather incommensurate with the present level of scholarship and general interest in these areas.

With this awareness in mind, I decided at the outset of my graduate studies that it might not be a bad idea to write down and save everything I looked up. Especially since, as a relative beginner, I would be investigating a lot of basic terms that an advanced specialist would not even think about. Working in this way, after just one year of study I had accumulated a considerable glossary, which I entered into digital format for fast access and compact storage. I continued to enter all Buddhist and non-Buddhist CJK terms into a single compilation like this for about four years. In 1990 at SUNY Stony Brook, I was asked to teach an undergraduate course in Classical Chinese. With *Mathews* and the Fenn *Five-Thousand* dictionaries both out of print at the time, I found myself in the awkward position having to teach such a course without the benefit of a comprehensive dictionary. But reflecting on my own compilation, I realized that I already had more than enough Classical Chinese terms on hand to suffice for such an introductory course, so I took the important step of culling out the non-Buddhist terms from the Buddhist terms and creating two separate compilations. I then checked through the draft of the Classical Chinese lexicon and supplemented it with all the characters, which were going to appear in my introductory course texts.

The following summer, I completed my translation of five Chinese Classics,[1] and in the course of this work, added all the KSC-representable[2] CJK logographs in these works. It was at this point that it had really turned for the first time, into a full-fledged dictionary. At the same time, I was continuing in the work of research and translation of a wide range of East Asian Buddhist classical texts, and from these I continued to add to both dictionaries. This work of compilation proceeded in this fashion past the completion of my degree in the summer of 1993 until the summer of 1995, when I first became aware of the WWWeb and HTML publication. At this point I made my first primitive conversion of BDict into HTML format and installed it on my newly established web site, where it has remained since that time.[3] CJKDict was during this time being prepared for commercial CD-ROM publication, a process that helped greatly to improve its overall quality.[4] One year later, sensing a lack of

[1] See *http://www.acmuller.gol.com/mullertext.htm#fiveclassics*

[2] At this point I was working with an early Korean word-processing program developed by IBM-Korea called KWP.

[3] Presently located at *http://www.acmuller.gol.com/index.html*

[4] Published by EAST KK, Tokyo Japan, in August 1996. At the time of this writing still available for order through the internet at *http://www.est.co.jp*

interest on the part of the publisher in further update and development of the project, I installed it onto the Web site in HTML format for public access, alongside BDict.

Since then, I have been continuing to work towards the further development of both dictionaries in terms of both quantity and quality of content as well technical improvement, especially in regard to indexing and hyperlinking. As time passes, the degree of internal hyperlinking continues to increase. Of equal importance however, is the matter of external hyperlinking, not only between the two sister compilations, but also between them and similar CJK lexicons around the Web. During the past year, the CJK dictionary has been interlinked on an individual character basis with Jim Breen's *WWW JDIC Server* (*http://www.dgs.monash.edu.au/~jwb/wwwjdic.html*), Rick Harbaugh's *Etymological Dictionary* (*http://zhongwen.com*) and Chuck Polisher's *I Ching Lexicon* (*http://www.ns.net/~cpolish/DEFAULT.HTM*). More such linkups are expected in the near future.

Setting aside for the moment their special digital capabilities, both dictionaries already surpass many of their hard-copy counterpart lexicons in terms of basic content. The number of single characters with full information contained in the CJKDict is already (to date) 8,000--more than is contained in *Mathews* (although *Mathews* still has many more compound entries). The content of the definitions in themselves are, for the most part, far more complete than any other current CJK-English dictionary with pre-modern focus, being derived from a wide range of authoritative Chinese, Korean and Japanese lexicons as well as through the direct reading of primary textual sources. They also possess the unique aspect of providing readings in all three prominent northeast Asian languages--Chinese, Korean and Japanese.

If BDict were printed out today with single spacing at 12 pt. on A4 paper it would come out to a little over 550 pages, which far exceeds *Soothill* or any of the other English language dictionaries which treat East Asian Buddhism--and it is continuing to grow rapidly. CJKDict would print out at about 1300 pages in the same format--which means that it is also quite a substantial compilation as compared to presently available Han character-English dictionaries. Furthermore, both are distinguished by being the only presently available dictionaries of their kinds which deliberately strive for a balanced treatment of the Chinese, Korean and Japanese cultural spheres.

More important though, than this basic superiority in volume, is the fact that both dictionaries take full advantage of digital functionality--most important of which is

hyperlinking. Here, there are two general categories: the hyperlinking from the indexes, and the hyperlinks attached to important terms within the definitions themselves. Hyperlink-based indexes are far faster and more flexible than traditional hard-copy indexes, since all one has to do is click--there is no searching for pages and then terms within a page. The hyperlinks attached to the important definition terms also allow one to go immediately to check related concepts--and one does not need to keep one's fingers stuck in three places in the dictionary to save the location of other terms still under investigation. Given this combination of attributes, these dictionaries surpass already existent lexicons in many ways.

Structure of the Dictionaries

Both dictionaries are presently stored in the same text-encoded (close to SGML) format. I experimented for a couple of years with keeping the master file of CJKDict stored in a database program (MS Access), but gave up on this in favor of tagged text. While the database allowed for quick sorting, easy selection of certain sets of information for comparison and ready identification of errors of structure and content, it was extremely limited in the case where major global changes were necessary, where extensive work needed to be done with the definitions, or where definitions were unusually long. The database format is probably more appropriate in the case in the case where the compiler is doing purely lexicographical work. But in a case where the dictionaries are in constant use in the course of research, it is much easier to have ready access to them if they are kept in a text-document format.

Furthermore, in the case of both dictionaries, the writing of any of the compound words is rarely a closed case, as most previously-entered Buddhist and Chinese philosophical terms, place names, school names, text names and personal names can always benefit from some kind of new information and editing. There are always new places to be found to add hyperlinks and index tags, and since these are not added by typing, but through macros, development of the compilation inside a database becomes quite impractical. Therefore they are stored in the text markup format which I describe below. The structure presently used, while not yet SGML-valid, is based on SGML principles such that it can be readily be set for SGML validity as the need arises.[5] That is, structural and semantic elements of the text are tagged with numerous on <element> and off </element> bracket structures, which may be devised without limit by the user, as long as they are implemented consistently. Of

[5] The reason I do not yet keep the work in fully compliant SGML markup at the moment is simply a lack of need, as many SGML markup structures can be represented more simply in the case where an SGML browser is not in use. When XML support becomes standardized in the major Web browsers, that will be enough incentive to make the basic storage formats for both works into fully XML (very close to SGML)-compliant.

course, this is the way that HTML markup is done, and many of the tags used are already HTML tags. With this kind of tagging structure within text format, the material can be easily converted into the desired publication format, whether it be hard copy, database, HTML or (hopefully soon) XML.

Let us now take a look at a sample entry at the text encoding level, to identify its components (the circled numbers (e.g.) included are temporarily placed for reference purposes and are not part of the original document).

<entry ID="07501672C-061056027"> <gph>本性</gph> [w] <pron lang="ch-wg">pen-hsing</pron> [p] <pron lang="ch-py">ben3xing4</pron> [k] <pron lang="kr">본성</pron> <pron lang="kmr">pons ng</pron> [j] <pron lang="jp-kk"> </pron> <pron lang="jp-rm">honshô</pron> <sense>" <ind1=trm>original nature</ind1=trm>," or " <ind1=trm>inherent nature</ind1=trm>" (<ind1=skt>prakº ti</ind1=skt>). An originally present fundamental quality of something, often equivalent to the concept of "self-nature" (自性 - Skt. <ind1=skt>svabhâva</ind1=skt>; Pali <ind1=pal>sabhâva</ind1=pal>). Buddhism, and especially Mahâyâna, generally rejects the concept of an inherent nature as being a mistaken perception. But on the other hand, in accordance with the general Chinese philosophical perception of the human nature as being originally good, certain texts will allude to the mind's inherent purity or quiescence. For one discussion of original nature, see the <ind2=txt-chn><wg>Yüan-chüeh ching</wg></ind2=txt> 圓覺經 at <cancol>T</cancol> 842.17.913c.</sense> <ref>iwa750 ZGD1164b naka1263c</ref> <resp>acm(entry)</resp></entry>

The above entry from the master text contains many of the typical elements. Let us go through the most important of these.

In accordance with database and SGML principles, each dictionary entry has a unique ID number, as in the example above, the ID number for the character 本 is (07501672C). Both of my dictionaries use the same ID construction system, a nine-digit string. The first three (075) digits represent the traditional radical (部首) number; the next two (01) indicate the number of strokes ( ) after the radical and the final four are the character's Unicode hex number (672C). The structure of this number allows for sorting by traditional radical and stroke, my preferred method of searching and arranging CJK lexical information. The inclusion of the Unicode hex number allows for linking with external dictionaries, as the

usage of the Unicode hex number has become standardized among our small group of interlinked Internet dictionaries.

The next element, indicated by the tags <gph></gph> demarcates the character graph(s) of the entry. This tag may be converted to a font value or anchor at the time of HTML publication, or to a field for database import. Also, when I use the dictionary for my own research work, I normally use the text-master version (rather than HTML version) by means of an extensive array of macros, so it is also necessary for the head word to be distinctively tagged to enable the proper function of search macros.

The third section is that of the readings, of which there are presently six kinds: Chinese Wade-Giles (still quite popular in the field of English-language Buddhology), Pinyin, Korean Han'gul, Romanized Korean (McCune-Reischauer), Japanese Katakana and Romanized Japanese (Hepburn). I will also make an effort to add Vietnamese readings in the future.

The next is the main explanatory portion of the entry, bracketed by the TEI-recommended tags <sense></sense>. Included in this section are all meanings, synonyms, Indic and Tibetan equivalents and cross-references.

A large number of the elements of the "sense" section are marked with tags for indexing purposes: person names, place names, English renderings of technical terms, Indic and Tibetan terms. These are also cross-classified by cultural/linguistic regions of Indic, Chinese, Korean and Japanese. Since there is no single sample entry, which includes all of these tags, we have selected one, which provides a few: English equivalent, Sanskrit and Pali. Index tags are subdivided into two general categories: "ind1" and "ind2", the difference in number indicating the indexing level. Level one indicates that the indexed term is a direct reference/synonym or translation of the headword, whereas level two just marks the fact that it is a technical term appearing in the definition. The difference between the two can be compared to hard-print lexicographical indexes, which differentiate, by boldface and plain-face text.

One of the most important facets of this kind of dictionary is the possibility of hyperlinking. The method presently used is HTML/HTTP, connected to the filename and an anchor attached to each entry. If it seems worthwhile, this can be changed in the future to SGML linking, which goes directly to the ID number of the target entry.

Following the <sense> division of the entry, we come to the references section, in which all

occurrences of the entry term in the major Buddhist reference works are included. I was a little bit lax about including these during the earlier stages of my work, and so am now going back and filling many of these in. In the above sample, we have references to three works: The Iwanami *Bukkyô Jiten* **岩波**: **教　典** (p. 750) The Taishûkan *Zengaku Dai Jiten* **大修館**: **大　典** (p. 1164b) and Hajime Nakamura's *Bukkyogo Dai Jiten* **中村元**: **佛教語大　典** (p. 1263c).[6]

The final part of the entry indicates the person(s) responsible for its contents, using initials which are identified in the front matter of the work. In the case of this term, it was I who created the entry, and no other persons edited or added to any aspect of it, so no other names are included. There are other entries however, which have as many as three names attached.[7]

HTML Publication

The dictionary is stored, edited and added to in the above format, sorted by the ID numbers. Conversion of the entire dictionary is done by running a macro, which changes all tags as necessary to support and optimize HTML display and create the indexes. At its present stage, the entire production takes about an hour for the BDict and about three hours for CJKDict, using VBA macros on a 300MHz Windows system. The underlying code in the above-presented sample, after the HTML publication macro, looks like this:

本性 [w] pen-hsing [p] ben3xing4 [k] **본성** pongseong [j] honshô ||| "<ind1=trm>original nature</ind1=trm>," or "<ind1=trm>inherent nature</ind1=trm>" (prakº ti). An originally present fundamental quality of something, often equivalent to the concept of "self-nature" (**自性** - Skt. svabhâva; Pali sabhâva). Buddhism, and especially Mahâyâna, generally rejects the concept of an inherent nature as being a mistaken perception. But on the other hand, in accordance with the general Chinese philosophical perception of the human nature as being originally good, certain texts will allude to the mind's inherent purity or quiescence. For one discussion of original nature, see the <wg>Yüan-chüeh ching</wg> **圓覺經** at T 842.17.913c. [References] iwa750 ZGD1164b naka1263c [Responsible] acm(entry) <p>

---

[6] The references for the Iwanami and Taishûkan dictionaries are taken from the extensive index of Buddhist dictionaries developed by the IRIZ project at *http://www.iijnet.or.jp/iriz/irizhtml/irizhome.htm.*

[7] For the initial establishment of much of the above structure I am indebted to Dr. Christian Wittern of Chung-hwa Institute, who found my dictionary on the Web when it was in its earliest and most primitive state. Christian converted the entire work into SGML format and also provided many Pinyin readings.

Which in HTML display, appears as:

本性 [w] pen-hsing [p] ben3xing4 [k] **본성** pongs [j] honshô ||| "original nature," or "inherent nature" (*prakº ti*). An originally present fundamental quality of something, often equivalent to the concept of "self-nature" (**自性** - Skt. *svabhâva*; Pali *sabhâva*). Buddhism, and especially Mahâyâna, generally rejects the concept of an inherent nature as being a mistaken perception. But on the other hand, in accordance with the general Chinese philosophical perception of the human nature as being originally good, certain texts will allude to the mind's inherent purity or quiescence. For one discussion of original nature, see the *Yüan-chüeh ching* **圓覺經** at *T* 842.17.913c. [References] iwa750 ZGD1164b naka1263c [Responsible] acm(entry)

While the Unicode character set has made significant improvements in terms of its basic Latin diacritical set (as least as far as East Asian languages go) unfortunately it offers little support for the diacritics needed for Indic and Tibetan, which means that representation of these characters in HTML is still going to present problems. However, we should be able to handle this problem when we begin to set up these documents in XML (see below).

The generation of the Classical Chinese dictionary works in an almost identical process, differing mainly in the number and type of indexes created in the process. CJKDict differs in nature from BDict in that it is aimed to deliver much more specific information on characters themselves, and therefore has many more detailed fields containing stroke, radical, variant, coding, readings information, etc.

HTML vs. CGI

A prominent difference between these dictionaries and others of similar type available on the Internet is that it relies completely on a text-generated hyperlinked format, as opposed to a CGI-forms input format. The main reason I have not pursued CGI searching is that the design of the dictionaries was originally created for the kind of research where one can investigate a broad range of knowledge concerning a topic. Or, when perusing the dictionaries lightly, with no special aim in mind, one can just continue to point and click--to points of new curiosity without end. So with HTTP "travel" when a person comes across the term shown in our example "**本性**" he or she can also naturally see on that page, with just little bit of scrolling, all the other compounds available that start with the character **本**. When information is delivered by a CGI script, it is fragmentary and disconnected from its surrounding text, and does not usually include a selection of hyperlinks within the displayed entry to directly browse to a related topic. The CGI method also tends to move away from

the graphical nature of the Han character. It is quite often the case when we are looking up a Han character, which we have not seen or written for some time, that we may not have a precise recollection of its graphical form or pronunciation, but can quickly recognize it once seen. In this sense, I have found various kinds of graphics and pronunciation-based indexes to be preferable to the rigid form-entry system. Another related virtue of the HTML approach is precisely its technical simplicity: the researcher with a minimum of computer skills may easily download the entire dictionary to his or her local system and run it as is. Except in the case of a person with considerable computing expertise, this is not possible with CGI-forms. Nonetheless, the most ideal situation would probably be that of having both options available, since a forms-entry system with a powerful search engine can often be much faster than simple HTML.

Present Search Options

At the moment, the interface to both dictionaries is an index page, which allows a browser-based search through a wide range of pronunciation indexes, radical-stroke indexes, four-corner, etc. BDict also offers, as mentioned above, indexes for texts, persons, places, schools, and technical terms for many of the various cultural manifestations of Buddhism.

In my own everyday research, I rely much more on the master-text files, rather than the HTML-converted files for a few reasons: (1) being in text format, they can be searched much more quickly by means of macros which select on-screen characters; (2) in accessing directly into the master files, editing can be done to improve the quality of the content of the entries, and (3) in the case where the entry is not yet included in the lexicon, it can be added automatically from the on-screen text. It happens that I presently work in MS-Word, but since I am working in text format, the macro functions could be easily replicated in any other word processing application.

One further, readily implementable option is to prepare the research texts in advance by running a macro through them, which tags each character so that it is linked to its appropriate location in the dictionary. This would be especially useful for teaching beginner students how to read classical Chinese texts. As long as the text being studied and the dictionary are located on the same computer system, the searching of the characters is instantaneous. For more sophisticated treatment of such texts, they would need to be marked up manually, to include distinct phrases and compound words. In any of the search formats, this type of dictionary represents a huge leap over its paperbound predecessor.

The Next Stage: XML

HTML has plenty of limitations, which is why the master version of the dictionary is not stored in HTML. HTML presently just does not have enough complexity and flexibility to adequately support a compilation of this sort. SGML, along with the guidelines of the Text Encoding Initiative (TEI), on the other hand, has everything we need for such a lexicon and more. The problem with SGML though, is that it is presently only used by a small group of specialists, and is only supported by a few, little-known forms of browsers which the average humanities researcher cannot possibly have the time, energy, know-how and money to implement on his/her own. It is for this reason that up till now I have not made an effort maintain the dictionaries in well-formed SGML.

It appears that this situation will change in the near future as the WWW is preparing to open up the functionality of the Internet considerably through the introduction of XML (Extensible Markup Language). Like HTML, XML is a kind of subset of SGML, but it is a subset that makes much greater use of SGML's power. It does this mainly by allowing the freedom to publishers of making the parameters of their documents self-defining, and by providing functions, which permit much more database-like operation. Thus, in using XML, we will have much more leeway in designing or our fonts, types of links, embedded programs, indexing functions and in working with large-size documents. Along with this will be included all of HTML's present functions (most importantly, hyperlinking), so nothing will be lost in the process. The increased functionality of XML will also allow for more complete interoperability with popular word-processing software. In fact, it is conceivable that a fully XML-based word-processor could be developed (if those companies should choose to take advantage of the fact). The implementation of XML is well in progress, and it is expected that version 5 of the two major browsers will fully support XML.

If XML implementation becomes widespread as expected, it is quite likely that I will convert the storage format into "well-formed" XML. This means that users will still be able to download the dictionaries to their own systems as fully functional digital lexicons, but with much greater functionality. The adoption of a practical, interoperable and widely used format will also be of great importance for the further development and usage of such dictionaries, since it will encourage the adoption of similar formats by those who are creating related compilations, which will in turn enhance the possibilities for mutual cooperation in large projects. For example, Classical Chinese and Buddhist Classical Chinese dictionaries which are composed in Japanese, Korean, Chinese, or any other language could be easily integrated just by the mere fact of being structured by XML and using a compatible ID system.