IRG N1868 @ Gyeongju: Presentation of the *CJKV-E Dictionary*

By A. Charles Muller (University of Tokyo)

At IRG 38, Charles Muller presented the *CJKV-E Dictionary* (*http://www.buddhism-dict.net/dealt*), an online academic project aimed at developing, recording, and displaying information on East Asian ideographs and compound words. The CJKV-E constitutes one part of a two-pronged project, the other being the *Digital Dictionary of Buddhism* (DDB; *http://www.buddhism-dic.net/ddb*), which has the same back-end structure, but is focused on the meanings of ideographs and compound words in a Buddhist context.

1. History

The *Chinese-Japanese-Korean-Vietnamese/English Dictionary* [CJKV-E] is a compilation of Chinese ideographs, as well as ideograph-comprised compound words, text names, person names, etc., found in East Asian Confucian, Daoist, Neo-Confucian texts, as well as other historical sources. Its information on individual ideographs is intended to be comprehensive, including pronunciations and meanings from ancient and modern sources from the Sinitic cultural sphere including China, Korea, Japan, and Vietnam. Modern-day compound words are included incidentally, but the coverage of modern vocabulary is not intended to be comprehensive.

The compilation of these two lexicons—now separate entities, originally started out as a single work, initiated by Muller in 1986. In 1995, shortly after the inception of the Web, he converted the material to HTML and placed it on the web with the aim of being able to present the material to scholars for their use, and to gain their feedback. For the first few years the data was maintained in a simple hard-linked HTML format. With the advent of XML in the late nineties, the data was converted to XML, patterned roughly on the guidelines of the Text Encoding Initiative (TEI *http://www.tei-c.org/index.xml*). At this time the data was converted to Unicode utf-8 encoding, and an ID structure was created based on the ISO 10646 hexadecimal system (e.g. 一: ID="c4e00"). Entry-node-level attributes also included radical and stroke number information like the following:

```
<entry ID="c4e00" added_by="cmuller" add_date="1995-07-15" rad="一" radval="01" radno="001"
strokes="00" totstrokes="01">
```

In 2001 we were extremely fortunate to have Dr. Michael Beddow of the University of Leeds to join the project. Michael, an expert in the area of Humanities computing, took the XML data and created a search system using XPath/XLinking, along with Perl, which was, to the best of our knowledge, the first at the time that would search mixed Latin and double-byte East Asian text in XML/Utf-8 encoding. Dr.

Beddow has continued to support the project up to the present, adding numerous enhancements, periodically updating the system, as well as providing web site security.

2. Structure and Content

The CJKV-E Dictionary is distinguished from other Chinese ideographic dictionaries presently found on the web in the fact that is (1) not simply a computerized aggregation, and (2) not simply a reproduction of an older print dictionary. It is being actively developed by scholars in conjunction with the reading of classical texts. Besides its inherent digital advantages, the CJKV-E dictionary already surpasses many of its hard-copy counterpart dictionaries in a number of ways. The total number of entries in December 2011 was 27,622, with 10,900 of these being single ideograph entries. Each of the entries in this CJKV-E dictionary is human-edited, and usually offers far more detailed information than any other comparable lexicon, being developed while consulting a wide range of authoritative Chinese, Korean, and Japanese sources, and usually through the direct reading of primary classical texts. The primary dictionaries consulted in the production of each entry include:

(1) *Hanyu dacidian* (Chinese)
(2) *Daejaweon* (Korean)
(3) *Dae han-han sajeon* (Korean)
(3) *Dai kango rin* (Japanese)
(4) *Dai kanwa jiten* (Japanese)
(5) *Gakken kanwa jiten* (Japanese)
(6) *Kadokawa kanwa jiten* (Japanese)
(7) *Lin Yutang Chinese-English Dictionary* (English)
(8) *Mathews Chinese-English Dictionary* (English)

While a number of the Japanese-oriented modern *kanji* dictionaries that have appeared during recent decades have been of acceptable quality in terms of precision within their respective purviews, they are limited in their scope and orientation to modern vocabulary, and thus are not that useful to those who are doing scholarly research/translation of pre-modern *han-wen* texts, who need to know all of the pronunciations and ancient semantic implementations and readings of a particular ideograph.

Therefore we provide the readings of each ideograph in each of the East Asian languages, distinguishing semantic regions according to pronunciations as appropriate according to each of those languages. For Chinese, we provide Pinyin and Wade-Giles information. For Korean, we provide pronunciation information in Hangeul, Revised Romanization and McCune-Reischauer. For Japanese, we provide Katakana information for *on-yomi*, distinguishing between *kan-on* and *go-on*; *kun* readings are given in Hiragana. Both are romanized in Hepburn. We also provide Vietnamese romanized readings for most ideographs. After the semantic area of the entry, we provide the names and the page numbers of the lexicons consulted. Since we use a TEI-like XML structure for the entries, responsibility for nodes within each entry is clearly distinguished as are the sources for the information for each node. For example:

```
<entry ID="c91ac" added_by="Charles Muller" add_date="2012-04-18" update="" rad=" 酉 "
radval="07" radno="164" strokes="11" totstrokes="18">
<hdwd>醬</hdwd>
<pron_list>
<pron lang="zh" system="py" pos="1" resp="Charles Muller" source="Gakken,Hanyu">jiàng</pron>
<pron lang="zh" system="wg" pos="1" resp="Charles Muller">chiang</pron>
<pron lang="ko" system="hg" pos="1" resp="Charles Muller" source="Daejawon">장</pron>
<pron lang="ko" system="mc" pos="1" resp="Charles Muller" source="Daejawon">jang</pron>
<pron lang="ko" system="mr" pos="1" resp="Charles Muller" source="Daejawon">chang</pron>
<pron lang="ja" system="kk" read="on" type="kan" pos="1" resp="Charles Muller"
source="Kangorin">ショウ</pron>
<pron lang="ja" system="hb" read="on" type="kan" pos="1" resp="Charles Muller"
source="Kangorin">shō</pron>
<pron lang="ja" system="hi" read="kun" resp="Charles Muller" source="Kangorin">ひしお</pron>
<pron lang="ja" system="hb" read="kun" resp="Charles Muller" source="Kangorin">hishio</pron>
<pron lang="vi" system="qn" resp="việnhánnôm">tương</pron>
</pron_list>
<sense_area>
<sense_group resp="Charles Muller">
<sense resp="Charles Muller" source="Gakken">To marinate, pickle, or soak meat in wine. Salted meat.
Salted preparations.[<xref idref="c91a2">醢</xref>]</sense>
<sense resp="Charles Muller" source="Gakken">Rice, barley, beans, etc., which have been salted and
soaked with rice wine or miso. </sense>
<sense resp="Charles Muller" source="Gakken">Miso, or soy sauce. [<xref idref="c6f3f">漿</xref>]
<quote rend="brackets">不得其醬</quote> <bibl type="canonref">論語，鄕黨</bibl> </sense>
<sense resp="Charles Muller" source="Gakken">A soup with starch. </sense>
<sense resp="Charles Muller">A kind of bean paste.</sense>
<sense resp="Charles Muller" source="Mathews">Any jam-like or paste-like food;
thick sauce. </sense>
<sense resp="Charles Muller">Modern Japanese form is <xref idref="c91a4">醤</xref>.</sense>
<sense resp="Charles Muller">Modern Chinese simplified form is <xref idref="c9171"> 酱
</xref>.</sense></sense_group>
</sense_area><dictref>
<dict><title>Hanyu dacidian</title><page>63596.170</page></dict>
<dict><title>Dai kanwa jiten</title><page>40011</page></dict>
<dict><title>Daejawon</title><page>1787.180</page></dict>
```

```
<dict><title>Mathews</title><page>0661</page></dict><extdict>
<a
href="http://www.csse.monash.edu.au/cgi-bin/cgiwrap/jwb/wwwjdic?1MKU91AC">WWWJDIC</a>
</extdict></dictref>
</entry>
```

3. Access

The dictionary can be searched through *hanzi* headwords, by dedicated indexes for each of the East Asian pronunciation groups, as well as by full text search. Since the ID numbers are constructed from Unicode code points, operators of external resources can also call entries by constructing a link based on the Unicode number. Thus, for example, this entry

```
<entry ID="c9152-7121-91cf" added_by="Yao Zhang" add_date="2012-04-18" update="" rad="酉"
radval="07" radno="164" strokes="03" totstrokes="10">
<hdwd>酒無量</hdwd>
```

…can be called by this string:

```
http://www.buddhism-dict.net/cgi-bin/xpr-dealt.pl?91.xml+id('c9152-7121-91cf')
```

This technique is presently being used to access our system by the SAT Taishō Database, the Smarthanzi application, and Jim Breen's WWWJDIC Server. Monthly-generated indexes of terms contained in the CJKV-E, along with these links, are available to interested operators of other resources (including, of course the UniHan database!).

We have established a password security system in order to block access by abusers of the dictionaries who send in search robots to download the data as well as to encourage regular users to feel a sense of responsibility to make their own contributions to this shared resource. This system operates at two levels:

1. **Limited Use (no user contribution)**: Any user may access the dictionary by entering *guest* as the username with no password. This will allow a total of 10 searches in each of the CJKV-E dictionary in a 24-hour period.

2. **Unlimited Use**:

i. **User Data Contributions** - While our most basic aim in putting these dictionaries on the web is to make this material readily available to everyone, the larger purpose of this project is to bring about a collaborative effort that will lead to the eventual development of a comprehensive body of data. In order to accomplish this, we need contributions toward content development from users. Thus, users may obtain an unlimited-use password by becoming a contributor to the DDB or CJKV-E. For details, see *http://www.buddhism-dict.net/contribute.html*.

ii. **Paid Subscriptions** - Those who are unable to make a contribution, but need unlimited access may pay for a two-year subscription to the CJKV-E and DDB dictionaries, at the rate of U.S. $60 for individuals and U.S. $600 for institutions. A template form for institutions is available here. For other questions, please write to *acmuller@jj.em-net.ne.jp*.

4. Technical Publications

For more detailed background material on the history and development of the DDB and CJKV-E, Muller has published a few papers and has made numerous conference presentations on the topic over the years, which are available through his personal publications page (*http://www.acmuller.net/publications-etc.html*). A monthly newsletter reporting new entries in the dictionary is also available to users, which provides a link to a list of the newest entries. The latest update list is here: *http://www.buddhism-dict.net/ddb/monthlies/ddbcjkveMonthly2012-05.html*. Updates from prior months can be accessed by changing the date values in this URL.