Building a General Database System of Chinese Character Dictionaries in Early Japan: Tenreibanshōmeigi in the HDIC Project

Shoju IKEDA and Yuan LI

Abstract: This paper introduces the Integrated Database of Hanzi Dictionaries in Early Japan, also known as the HDIC project, which is composed of three main dictionaries of the Heian period. We provide a detailed report of the full-text publication of one dictionary: Tenreibanshōmeigi in the HDIC.

Keywords: Yupian, Unicode, CJK Unified Ideographs, digital humanities, open access

(Received on November 30, 2017)

1 Integrated Database of Hanzi Dictionaries in Early Japan (HDIC)

HDIC is a Unicode-based project that includes three dictionaries: Tenreibanshōmeigi 篆隷万象義, Shinzenjikyō 新撰字鏡, and Ruijumyōgishō 類聚名義抄. Over 70,000 Chinese characters can be processed after the release of Unicode 3.1. By materializing the full-text database of these dictionaries, a worldwide platform for the study of the dictionaries in early Japan can be created. These dictionaries are crucial sources of the history of the Japanese language, especially in the fields of variant characters, phonemes, and lexicons.

With the spread of Unicode, chief Chinese dictionaries like Shuowenjiezi 說文解字 and Guangyun 広韻 have been provided open access. Meanwhile, though many old manuscripts are well preserved in Japan, owing to problems such as variant characters and erratum, the original data of old manuscripts is falling behind. Authors predominantly study only old manuscripts that are handed down.

On 1 September 2016, the full-text data of Tenreibanshōmeigi was released (http://hdic.jp/), which is the first time the full text of a dictionary in early Japan has been made public. One of the authors, Shoju Ikeda, who is the leader of a Japanese language research group, began a database of dictionaries in the 1990s by engaging in the study of the information processing of Chinese characters, like JIS and Unicode. Furthermore, we worked on the database of Tenreibanshōmeigi, which is an abridged version of the Yuanben Yupian 原本玉篇.

As an introduction of HDIC, this paper provides an explanation of the construction of the Tenreibanshōmeigi database, including the unification standards, transliteration principles, full-text,
publication system, and future issues. We hope that the data, which are public, will be of practical use to researchers, in the domains of both premodern Chinese character dictionary studies and information processing.

2 Ancient Chinese character documents and Chinese character dictionaries in HDIC

2.1 The digitization of ancient Chinese character documents

2.1.1 The development of printed-books-based digitization

The study of ancient Chinese character documents is closely relevant to the development of information technology. With the establishment and spread of Unicode, the online index of the full-text of Chinese classics and Buddhist scriptures has been materialized rapidly. Online systems like ‘Zdic (zdic.com)’ in China, ‘Scripta Sinica database (hanji.ihp.sinica.edu.tw)’ in Taiwan, ‘Chinese Text Project (ctext.org)’ in the United States, and ‘The SAT Daizōkyō Text Database (21dzk.l.u-tokyo.ac.jp/SAT)’ in Japan, are noteworthy. However, the systems above rely on the printed books or the revised text. The researches of Sinology and Buddhist studies are based on the printed books.

2.1.2 The current situation of manuscript-books-based digitization

In Japan, from the Nara period, many high-quality old manuscripts have been well preserved. The dictionaries of early Japan, which authors predominantly study, are essential linguistic materials, but they are almost handed down as old manuscripts. As is usual with manuscripts, problems such as variant characters and erratum are inevitable. Deciphering them is difficult, and digitization is delayed. For this reason, for the digitization of manuscripts, imaging is preceding effectively.

Take Tenreiбанshōmeigi as an example. Several kinds of facsimiles of it have been published until now. The earliest facsimile book is the Sō bun Series 崇文叢書 from 1926, and it is registered to the digital collection of the National Diet Library (NDL) and made public. However, the realization of e-texts of manuscript-based materials is still a while away.

The study of manuscript dictionaries in early Japan was mostly done under Japanese scholars, while recently, scholars in the fields of philology, the Chinese language, and the Japanese language in China and Taiwan have also started paying attention to manuscript dictionaries in early Japan. For example, a research book about Shinsenjikyō was published (Zhang, 2012). Studies related to Ruigumyōgishō are also being conducted. Because there are no Japanese readings recorded in Tenreiбанshōmeigi, there are even more research papers about this book. A fully revised text book was published (Lv, 2007). However, as there is no e-text, there are constraints to information processing.

Under these circumstances, the full-text data of Tenreiбанshōmeigi was released (http://hdic.jp/) by a research group whose the authors are affiliated to Hokkaido University, Japan. This is the first time the full text of a dictionary in early Japan has been made public.

2.2 Digitization of ancient printed books and Chinese character manuscript dictionaries in HDIC

To decipher Chinese character dictionaries in early Japan, it is necessary to refer to the previous dictionaries compiled in China, especially the Qieyun 切韻 system rime dictionaries and Yupian system dictionaries. Therefore, we constructed a database of these Chinese dictionaries at first and then started the digitization of Tenreiбанshōmeigi, Shinsenjikyō, and Ruigumyōgishō. In the case of the study of
Japanese pronunciations of Chinese characters, it is essential to refer to *Guangyun*, which records the Middle Chinese system. The methodology we used in HDIC was to make a database of Chinese dictionaries at first and then construct a database of dictionaries in Japan based on it (Figure 1).

HDIC Project is a general database, and it includes both medieval China dictionaries and early Japan dictionaries. Dictionaries information are shown in the order of title, abbreviation and manuscript or print as below.

**Medieval China dictionaries**
- *Yuanben YuPian* (YYP), manuscript
- *Yuanben Yupian* Quoted Fragments in other books (YQF), manuscript
- *Songben YuPian* (SYP), print

**Early Japan dictionaries**
- *Kôsanjibon Tenrei Banshô meigi* (KTB), manuscript
- *Tenjibon Shinsen Jikyô* (TSJ), manuscript
- *Zushôryobon Ruiju Myôgi shô* (ZRM), manuscript, original version and incomplete text
- *Kanchinbon Ruiju Myôgi shô* (KRM), manuscript, revised version and complete set

We are also considering of the connections with other groups about China dictionaries as below.

**China dictionaries**
- *ShuoWen JieZi* (SWJZ), print
- *Songben GuangYun* (SGY), print
- *KangXi ZiDian* (KXZD), print
- *QieYun* fragments (QY), manuscript
3 The brief history of HDIC from *Tenreibanshōmeigi*

3.1 Unicode and digital humanities in Japan

To express characters through a computer, the character encoding corresponding to characters is necessary. From Unicode 1.01 (1992), the number of Chinese characters that can be processed (URO) is 20,902. From Unicode 3.1, by the addition of Extension B to CJK Unified Ideographs, the number went above 70,000 (Lunde, 2008). Recently, by adding the CJK Unified Ideographs, Extension A-F, the number again went above 80,000 from Unicode 8.0.

Against the background of the development of the Unicode mentioned above, in the field of digital humanities in Japan, many research groups across several different fields like the Japanese language, Chinese language, Buddhist studies, and information science that dealt with issues like character encoding, character processing, image database, image processing, and electronic texts promoted collaborative research until now.

Under these circumstances, one of the authors, Shoju Ikeda, the leader of a Japanese language research group, started the construction of a *Tenreibanshōmeigi* database, which is the predecessor of HDIC from 1994. 1994 was the period when only JIS X 0208 could be processed. However, some researchers believed that a large character set can be used in the future. Below, along with Unicode and digital humanities in Japan (Table 1), a brief review of KTB (each version and the attempts at open access) is provided.

**KTB-Temporary version (entries) Ikeda (1994)**

In 1994, the *Tenreibanshōmeigi* database, a temporary version of the database, was released, along with contents of all entries. Information like the location of entry, Daikanwa number, the Yupian volume number, and the JIS graphic character code are included. From the investigation, it is clear that the entries in Book 1-Book 4 can be processed by JIS X 0208 around 30% and by ISO/IEC10646-1 (or Unicode ver. 1.01) around 70%.

**KTB-Mojikyo version (entries) Ikeda (2003)**

From the results of the investigation, it is clear that only 46 characters out of the entries of *Tenreibanshōmeigi* are not registered in Konjaku Mojikyo ver. 3. That is, by using Mojikyo, almost all the entries in *Tenreibanshōmeigi* can be processed. However, the data processing is very limited.

**KTB-UCS version (entries) Ikeda (2011)**

In the information processing environment at the time, around 70,000 characters could be processed by CJK Unified Ideographs in Unicode. According to the results of the investigation, almost all the entries could be processed by using Unicode.

**HDIC Established Ikeda (2014)**

Took the completion of the inputting of *Tenreibanshōmeigi*¹ as the opportunity, summarized the

---

¹ Refer to Li (2015a, 2015b) for the interim reports.
<table>
<thead>
<tr>
<th>Year</th>
<th>Size of character set</th>
<th>Standard</th>
<th>CJK Unified Ideographs</th>
<th>Number of Chinese characters</th>
<th>PC operating system</th>
<th>Digital humanities in Japan</th>
</tr>
</thead>
<tbody>
<tr>
<td>1963</td>
<td>ASCII</td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1978</td>
<td>Around 6,000</td>
<td>JIS C 6226-1978</td>
<td>6,349</td>
<td>MS-DOS 1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1981</td>
<td>JIS X 0208-1983</td>
<td></td>
<td></td>
<td>6,353</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1983</td>
<td>JIS X 0208-1997</td>
<td></td>
<td></td>
<td>6,355</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1984</td>
<td>JIS X 0208-1990</td>
<td></td>
<td></td>
<td>6,355</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1988</td>
<td>JIS X 0212-1990</td>
<td></td>
<td></td>
<td>5,801</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1991</td>
<td>Unicode ver. 1.0</td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1992</td>
<td>Around 20,000</td>
<td>Unicode ver. 1.0</td>
<td>URO</td>
<td>20,902</td>
<td>Windows 95</td>
<td>JALLC³, KTB — Temporary version (entries)</td>
</tr>
<tr>
<td>1994</td>
<td>Unicode ver. 2.0</td>
<td></td>
<td></td>
<td>20,902</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td>JIS X 0208:1997</td>
<td></td>
<td></td>
<td>6,355</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1996</td>
<td>Unicode ver. 3.0</td>
<td></td>
<td></td>
<td>27,484</td>
<td></td>
<td>JINMONKō¹</td>
</tr>
<tr>
<td>2000</td>
<td>Unicode ver. 3.1</td>
<td></td>
<td></td>
<td>70,195</td>
<td></td>
<td>Windows XP; MacOS X</td>
</tr>
<tr>
<td>2003</td>
<td>Unicode ver. 4.1</td>
<td></td>
<td></td>
<td>70,217</td>
<td></td>
<td>Kanji DB²; KTB — Mojikyo version (entries)</td>
</tr>
<tr>
<td>2004</td>
<td>Unicode ver. 5.1</td>
<td></td>
<td></td>
<td>70,225</td>
<td></td>
<td>HNG³</td>
</tr>
<tr>
<td>2005</td>
<td>Unicode ver. 5.2</td>
<td></td>
<td></td>
<td>74,374</td>
<td>Windows 7</td>
<td>Takuhon-moji Database⁴; CHISE ID⁵</td>
</tr>
<tr>
<td>2007</td>
<td>Unicode ver. 6.0</td>
<td></td>
<td></td>
<td>74,596</td>
<td></td>
<td>GlyphWiki⁶; SAT⁷</td>
</tr>
<tr>
<td>2008</td>
<td>Unicode ver. 6.0</td>
<td></td>
<td></td>
<td>74,596</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>JIS X 0208:2012</td>
<td></td>
<td></td>
<td>70,195</td>
<td>Windows 8</td>
<td>KTB — UCS version (entries)</td>
</tr>
<tr>
<td>2011</td>
<td>Unicode ver. 8.0</td>
<td></td>
<td></td>
<td>80,358</td>
<td></td>
<td>CHJ²²; NIJL-NW³³; HDIC Established</td>
</tr>
<tr>
<td>2014</td>
<td>Unicode ver. 9.0</td>
<td></td>
<td></td>
<td>80,358</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>Unicode ver. 10.0</td>
<td></td>
<td></td>
<td>87,861</td>
<td></td>
<td>KTB — UCS version (full-text)</td>
</tr>
</tbody>
</table>

¹Masayuki Toyoshima, Hokkaido University (present affiliation is Sophia University); ²IPSJ SIG Computers and the Humanities; ³Japanese Association for Literary and Linguistic Computing; ⁴Japanese Association for East Asian Text Processing; ⁵Information Processing Society of Japan, Special Interest Groups, Computers and the Humanities; ⁶Taichi Kawabata, NTT; ⁷Hanzi Normative Glyphs, Harumichi Ishizuka, Hokkaido University; ⁸Character Database of Digital Rubbings, Koichi Yasuoka, Kyoto University; ⁹Character Information Service Environment, Tomohiko Moroioka, Kyoto University; ¹⁰Koichi Kamichi, Keio University (present affiliation is Daito Bunka University); ¹¹The SAT Daizōkyō Text Database Committee, University of Tokyo; ¹²Corpus of Historical Japanese, National Institute for Japanese Language and Linguistics (NINJAL); ¹³Project to Build an International Collaborative Research Network for Pre-modern Japanese Texts, National Institute of Japanese Literature (NIJL).

Table 1 Unicode and digital humanities in Japan

Outline and issues of HDIC, and gave the description of the perspective of Shinsenjikyō and Ruijumyōgishō.

KTB-UCS version (full text) Li and Ikeda (2016)

Gave the description of the full-text and publication system of the database based on Unicode. After this long period of preparation, the time for full-text publication arrived.
3.2 Publication plan of HDIC

Database will be made public in stages. First, all the entries and full text of a small part of selected radicals will be made available as samples to the public (Trial Version Part 1). Then, more full texts of a larger part of the selected radicals will be made available as samples to the public (Trial Version Part 2). Finally, the full text will be made public (Trial Version All), and the necessary corrections will be made (Revised Version). The revising work will depend on Github, and the general information access will be realized through the following website: http://hdic.jp/.

The schedule of the research results publication is shown below (Figure 2). By the end of March 2018, the Shinsenjikyō’s full-text will be released (Trial Version All). By the end of March 2019, the Zushoryōbon Ruijumyōgishō’s full text will be released (Trial Version All). By the end of March 2022, the Kanchinbon Ruijumyōgishō will be released (Revised Version).

![Figure 2 HDIC publication plan](image)

4 Textual study of Tenreibanshōmeigi

4.1 Item structure

The item structure of Tenreibanshōmeigi is relatively simple (Figures 3 and 4).

Headword, fanqie, and meaning play the part of explaining the glyph, pronunciation, and equivalent word of the headword, respectively. The variant form provides the supplement of the glyph.

![gj](image)

Figure 2 HDIC publication plan

e.g. 1 \(^{(A)}\) 畢 \(^{(B)}\) 力絹反。\(^{(C)}\) 慕也。\(^{(a)}\) 績字也。（Tenreibanshōmeigi, Book 1, f. 75r）

Headword ‘\(^{(A)}\) 畢’ indicates the glyph of the character and fanqie 反切 ‘\(^{(B)}\) 力絹反’ indicates the pronunciation. ‘\(^{(C)}\) 慕也’ indicates meaning, and annotation ‘\(^{(a)}\) 績字也’ provides the variant form. By using the Tenreibanshōmeigi Database, the textual study of glyphs, pronunciations, and meanings is carried out. Below, we will discuss the textual study of headword (4.2), fanqie, and meaning (4.3) with examples.

---

2 Related to Tenreibanshōmeigi and Yuanben Yupian, the main preceding studies are listed below:


4.2 Headword identification

The *Tenreibanshōmeigi* was compiled by Kūkai 空海 in Japan around 830. As is well known, the *Tenreibanshōmeigi* is an abridged version of the *Yuanben Yupian*, which was compiled in 543 by Gu Yewang 顧野王 in China. Because the *Yupian* was lost in China and only about seven volumes have been preserved in Japan, this is a very valuable resource for obtaining a glimpse of the *Yupian*’s original form. At the same time, it is pointed out there are many errata in *Kōsanjibon*. Previous researches about the decipherment of the headwords of *Tenreibanshōmeigi* (Shirafuji, 1977; Miyazawa, 1977; Ikeda, 2014; Lv, 2007) show that there are different instances in the identification of the headwords. To improve the precision of the identification of headwords, it is necessary to piece together the identifications. Regarding the preceding researches mentioned above, the following is a comparison list of the ways of form, data format, decipherment of headwords, image of headwords, and decipherment of annotations (Table 2).

### Table 2 Comparison list of headword identifications in preceding researches

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Form</td>
<td>index</td>
<td>list</td>
<td>database</td>
<td>text</td>
</tr>
<tr>
<td>Data format</td>
<td>handwritten</td>
<td>handwritten</td>
<td>electronic text</td>
<td>printed book</td>
</tr>
<tr>
<td>Headwords decipherment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Headwords image</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Annotation decipherment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

---

3 Li (2017).
By using the Tenreibanshōmeigi database, the differences of identification of headwords in the preceding researches are listed above (corresponding to the Yupian fragments). The result is that the different items number reaches 254 and almost takes up over 10% of all the headwords. Then, these 254 items are classified into four groups. Group 1: items can be unified, Group 2: items can be recognized as variant characters, Group 3: items can conflict with other characters, Group 4: items are difficult to identify. Illustrations are show below (Table 3).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Unification</td>
<td>虫</td>
<td>鼻</td>
<td>鼻</td>
<td>鼻</td>
<td>鼻</td>
</tr>
<tr>
<td>2 Variant</td>
<td>虫</td>
<td>御</td>
<td>御</td>
<td>御</td>
<td>御</td>
</tr>
<tr>
<td>3 Conflict</td>
<td>鼻</td>
<td>割</td>
<td>割</td>
<td>割</td>
<td>割</td>
</tr>
<tr>
<td>4 Difficultness</td>
<td>割</td>
<td>割</td>
<td>割</td>
<td>割</td>
<td>割</td>
</tr>
</tbody>
</table>

Below are the measures that we take in HDIC at present to deal with these differences in the preceding researches.

1 Decide the representative character and take the different identifications into the same group.
2 Decide the representative character and take the different identifications into the same group.
3 Represent the different identifications respectively. This group contains the items that need to be paid special attention.
4 Represent the different identifications respectively. This group contains the items that are open to question.

4.3 Revision of annotations

(1) Collation with phonetic materials

To revise the mistranscription of fanqie by collating materials such as Guangyun data, Qieyun data, Ueda (1986) from HDIC and preparing and improving the phonetic data of the Yupian system dictionaries.

e.g. 2 鼻 明弟反。羿字。 (Tenreibanshōmeigi, Book 5, f. 36v)

In Tenreibanshōmeigi, fanqie is ‘明弟反,’ and in Shinsenjikyō, it is ‘胡弟牛弟二反’.

Revision notes are entered below: ‘明弟反：《新撰字鏡》作「胡弟牛弟二反」。上田作「胡弔」，又云「原明弟誤」。

(2) Collation with Chinese classics

By collating with the Chinese classics, the textual content is gradually revised.

e.g. 3 踏 子石反。如也。長置行也。 (Tenreibanshōmeigi, Book 2, f. 53r)

It is ‘君在。蹴踏如也。與與如也’ in Lunyu Xiangdang 論語·鄉黨。 It is clear that ‘如也’ is the meaning of ‘... manner’, while not the explanation of ‘踏’. Therefore, this is an example of mistranscription.

Yupian system dictionaries have been studied in both Japan and China actively. However,
importance is placed on the compiled histories of the two countries in the dictionaries respectively. The research materials and results do not often refer to each other. As was mentioned above, by piecing together the research results in both Japan and China, it is in order to improve this condition.

4.4 Entry type

As written in Tenreibanshōmeigi’s ‘Tenrei,’ the headwords in the book should be written in the seal script and clerical script. However, in Kösanjibon, which existed until now, only about 6% of the total contains headwords in the seal script, while the headwords in the clerical script form the main part. When transcribing Kösanjibon or even before, problems such as omission, mistranscription, and overlapping of headwords and inclusion of headwords into annotations occurred. Therefore, while organizing the headwords in Tenreibanshōmeigi, it is necessary to organize not only the headwords written in the large clerical script but also the ones that are omitted, as well as the ones embedded into the annotations (Li, 2015b).

The headwords can be grouped into three basic classes: regular headwords (about 88%), embedded headwords (about 4%), and omitted headwords (about 3%). We may then add the information of the seal or clerical script to these groups in a subordinate position (Group a, b, c, mentioned later).

The remaining group is the supplementary classifications of variant scripts (about 5%) supplied in the Songben Yupian (Group d). This is a measure to take into consideration the Yuanben Yupian and Songben Yupian as well, which also belong to the Yupian system dictionaries.

Full details of the four groups are given below:

**Entry type**

Group a — Regular headword
1 Regular (= Regular_clerical)
2 Regular_seal

Group b — Embedded headword
3 Embedded_clerical
4 Embedded_omitted
5 Embedded_seal

Group c — Omitted headword
6 Omitted
7 Omitted_regular

Group d — Variant of the Songben Yupian
8 Songben-Yupian
5 Tenreibanshōmeigi’s full-text and publication

5.1 Unification rules and decipherment principle

Unification, was a concept that was originally used in information processing, but now it has also been adapted to character identification in decipherment. Regarding the abstracted character form, even though the concrete forms (handwritten or printed) are slightly different, the range that should be recognized exists. For example, the Kangxi radical 162 车 can be written as ‘@’, or ‘_’ (differences in the dots are one or two), but in either case, it should be recognized as the same radical. The ranges described above are called ‘unification rules’.

Regarding the handwriting of Chinese characters, there are individual differences. Manuscripts dictionaries are the same, even though differences can occur. However, when deciphering these characters, and in the course of digitization, it is necessary to remove this kind of influence caused by handwriting. Therefore, when deciphering ancient manuscripts, shifts between character codes and graphic shapes occur. To avoid the hindrance caused by this kind of shift, it is necessary to use the ‘unifications rules’ mentioned above.

Moreover, in the process of the digitization of manuscript books, how to process the variants is a problem that cannot be avoided. For example, like ‘於’ and ‘扌’ and ‘因’ and ‘扌’ can be encoded as ordinarily used character forms and earlier forms. According to the different decipherment principles, it can be followed by either form. If the principle to preserve earlier forms is adopted, it should be described as ‘扌’ and ‘扌’ while if the principle to interpret ancient documents is adopted, then it should be described as ‘於’ and ‘因’.

E.g. 4 车 略逆反。龺也。 (Tenreibanshōmeigi, Book 6, f. 137r)

絵絵絵 三同。去逆反。入：龺葛布。 (Shinsenjikyō, Volume 4, f. 4v)

絵 <l>去逆反。絵。<l> (Kanchihinbon Ruijumyōgishō, Höchū, f. 66r)

絵 俗通。 (Kanchihinbon Ruijumyōgishō, Höchū, f. 66r)

絵 悖二歴。 (Kanchihinbon Ruijumyōgishō, Höchū, f. 66r)

In addition, against Chinese dictionaries, dictionaries in the Heian period are endowed with diversity and complexity. One example ‘絵’ is shown in Tenreibanshōmeigi above is like ‘絵 略逆反。龺也.’ and is shown as the solo-variant headword, while Shinsenjikyō in the middle, three variants of headwords ‘絵絵絵’ is lined up in the same item, followed by the Ruijumyōgishō, with the variants ‘絵’ ‘絵’ ‘絵’ separated into three individual items to form variant item groups. The delay in manuscript dictionaries in early Japan is related to this feature. The unification relation between the graphic shape and decipherment forms in manuscripts, and the problem of coordination of different dictionaries, are issues that should be solved in the process of digitization of manuscript dictionaries.

About the decipherment principle of the current Tenreibanshōmeigi database, it basically follows the Kangxizidian 康熙字典 scripts. However, in some parts, when the differences in the variant shapes were consciously described when the dictionary was compiled, the character form familiar to the original shapes was adopted. Plural character forms adopted is one of the strengths of HDIC. Furthermore, to do the decipherment more flexibly, abstraction should be classified into several degrees, and classifying the manuscript dictionaries hierarchically is necessary.
5.2 Processing conditions based on Unicode

Over 70,000 Chinese characters could be processed after the release of Unicode 3.1 in 2001. The Kōsanjibon Tenreibanshōmeigi database was constructed within the Unihan range offered by Unicode. In transliteration, it follows the standards of Kangxizidian scripts, and some of the characters are processed following the original scripts. Additionally, some characters are indicated with the ideographic description sequence (IDS), which cannot be processed by CJK Unified Ideographs. IDS is a method that uses IDC (Ideographic Description Character, 井|水|火|土|田|日|月|木, from U+2FF0 to U+2FFB) and parts of characters to describe the whole character.

In fact, the percentage of characters in the dictionaries that can be processed with Unihan is given below: Tenreibanshōmeigi, 99.2%; Songben Yupian, 99.8%; and Shinzenjikyō, 89.2%. The numbers of headwords that cannot be processed is 128 in Tenreibanshōmeigi, 46 in Songben Yupian, and 2,624 in Shinzenjikyō. The details are shown in Tables 4 and 5.

Table 4 Unicode processing conditions in HDIC

<table>
<thead>
<tr>
<th>DB</th>
<th>CJK</th>
<th>IDS</th>
<th>Others</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>KTB</td>
<td>15,872 (99.2%)</td>
<td>80 (0.5%)</td>
<td>48 (0.3%)</td>
<td>16,000</td>
</tr>
<tr>
<td>SYP</td>
<td>22,954 (99.8%)</td>
<td>23 (0.1%)</td>
<td>23 (0.1%)</td>
<td>23,000</td>
</tr>
<tr>
<td>TSJ</td>
<td>21,654 (89.2%)</td>
<td>1,468 (6.0%)</td>
<td>1,156 (4.8%)</td>
<td>24,000</td>
</tr>
</tbody>
</table>

Table 5 Details in CJK Unified Ideographs of Table 4

<table>
<thead>
<tr>
<th>DB</th>
<th>URO</th>
<th>Extension A</th>
<th>Extension B</th>
<th>Extension C</th>
</tr>
</thead>
<tbody>
<tr>
<td>KTB</td>
<td>10,160 (63.5%)</td>
<td>2,000 (12.5%)</td>
<td>3,712 (23.2%)</td>
<td>0 (0.0%)</td>
</tr>
<tr>
<td>SYP</td>
<td>13,386 (58.2%)</td>
<td>3,243 (14.1%)</td>
<td>6,325 (27.5%)</td>
<td>0 (0.0%)</td>
</tr>
<tr>
<td>TSJ</td>
<td>15,286 (63.0%)</td>
<td>2,444 (10.1%)</td>
<td>4,010 (16.5%)</td>
<td>6 (0.0%)</td>
</tr>
</tbody>
</table>

5.3 Publication system

5.3.1 Information access

Tenreibanshōmeigi (KTB) is a part of HDIC. The main access of the HDIC Project is http://hdic.jp/. The information shown below can be obtained:

- Progress, events, latest information, etc.
- Introduction of the entire project
- Construction of editorial committee
- References to related papers and presentations
- Links of related sources and tools of Chinese classics, also Buddhist scriptures

The latest version of the full text of Tenreibanshōmeigi can be accessed at http://github.com/shikeda/HDIC in the tab-separated value (TSV) format. The formation is being gradually improved.
The previous version was prepared to follow the principle of preserving the original forms. The later versions were provided while the revision was in progress.

5.3.2 License

Github is a web service for sharing a software project. Therefore, a license is required for an open source project. The open data in HDIC is meant to be improved, and so the data can be copied and distributed freely, and suggestions and bug reports can be received. In the future, we are thinking about writing a license by clearly following GPLv2, as shown below:

These data are distributed under GPLv2. Send bug reports and feature requests via email.

5.3.3 TSV data

We adopted the TSV data format. A TSV file is a simple text format for storing data in a tabular structure. Each record in the table is one line of the text file, and each field value of the record is separated from the next by a tab character. The TSV format is thus one type of the more general delimiter-separated-values format. TSV is a simple file format widely supported in both lexicography and information processing and is often used in data exchange to move tabular data among different computer programs that support the format.

5.3.4 TSV information

For the reference of researchers in other countries and areas, we put the TSV format information at the head of the open data. The numbers 01-10 are for the convenience of the following explanation. The information encoded in 01-10 can be divided into three groups by their contents.

I. Basic information: Location, system, structure information [01-03], transliteration of headwords and annotations in the Tenreibanshōmeigi [04, 07].

II. Relevant information: Information on the location of the corresponding entries in the Yupian system dictionaries (Yuanben Yupian and Songben Yupian) [08, 09].

III. Revision Information: Classification of headwords by the authors, differences in identification in previous researches, revision comments [05, 06, 10].

**TSV format information:**

01 TBID (v_www_xyz): Book (v), leaf (www), recto-verso (x), line (y), and number (z)
02 TB_vol_radical (xx#yyy): Volume (xx) and radical number (yyy)
03 TB_radical: Radical of Chinese character
04 Entry: Headword
05 Entry_type: For details, refer to the previous section
06 Entry_diff: Differences of transliteration with other scholars
07 TB_def: Definition of pronunciation, meaning, and variant (s)
08 SYID (vwwwxyyyzz): Book (v), leaf (www), recto-verso (x), line (yy), and number (z)
09 YYID (Ywwxxxyyy-z): Volume (ww), leaf (xxx), line (yyy), and number (z)
10 TB_remarks: Editor’s notes
Data sample:

Table 6 Details of TSV data of ‘哿’

<table>
<thead>
<tr>
<th>No.</th>
<th>Column</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>TBID</td>
<td>3_024_A22</td>
</tr>
<tr>
<td>02</td>
<td>TB_vol_radical</td>
<td>v9#96</td>
</tr>
<tr>
<td>03</td>
<td>TB_radical</td>
<td>可</td>
</tr>
<tr>
<td>04</td>
<td>Entry</td>
<td>叓</td>
</tr>
<tr>
<td>05</td>
<td>Entry_type</td>
<td>Regular</td>
</tr>
<tr>
<td>06</td>
<td>Entry_diff</td>
<td></td>
</tr>
<tr>
<td>07</td>
<td>TB_def</td>
<td>公可反。喜也、可也。</td>
</tr>
<tr>
<td>08</td>
<td>SYID</td>
<td>a088a023</td>
</tr>
<tr>
<td>09</td>
<td>YYID</td>
<td>Y09103299-1</td>
</tr>
<tr>
<td>10</td>
<td>TBRemarks</td>
<td>喜：當作嘉。原本《玉篇》・宋本《玉篇》作嘉。 (revision comments)</td>
</tr>
</tbody>
</table>

01 Book 3, leaf 24, recto, line 2, number 2 (location)
02 Volume 9 and radical number 96 (volume and radical number)
03 可 (radical)
04 叓 (headword)
05 Regular headword (type of headword/entry type)
06 None (differences of transliteration among other scholars)
07 公可反。喜也、可也。 (annotation)
08 Book a, leaf 6, verso, line 2, number 3 (location in SYP)
09 Volume 9, leaf 1-3, line 299, number 1 (location in YYP)
10 喜：當作嘉。原本《玉篇》・宋本《玉篇》作嘉。 (revision comments)

6 Conclusions and future directions

To improve the study of ancient manuscript dictionaries in early Japan, a comparison among Yupian system dictionaries of Yuanben Yupian fragments, Tenreibanshōmeigi, and Songben Yupian is necessary. Until now, attempts of digitization of manuscript dictionaries and databasing of Yupian system dictionaries was insufficient. In HDIC, we make the data of Yupian system dictionaries (Tenreibanshōmeigi, Yuanben Yupian and Songben Yupian) as the foundation of the project and work on Shinsenjikyō and Ruijumyōgishō (Figure 5).

First, the databasing of Songben Yupian (an ancient printed book) was completed; then, based on this, the inputting of Tenreibanshōmeigi was completed. In 2016, Songben Yupian was released in April, and Tenreibanshōmeigi was released in September sequentially. Additionally, considering international compatibility, we gave presentations on international conferences (Li, 2015c; Ikeda, 2015; Ikeda, 2017; Ikeda and Li, 2017). We hope that the data, which has been made public, will be of practical use to researchers, both in the domains of pre-modern Chinese character dictionary studies and information processing.

In future, based on the data opened, issues like the search for IDS parts, structured method by XML, image data, linked data, and the collaboration with the digital collection operated by NDL, should be paid attention to. We are looking forward to provide a worldwide platform for the study of the ancient manuscript dictionaries through HDIC, especially for the dictionaries complied in early Japan.
The numbers on the right side of the abbreviation indicate the entries number and total characters number in text.

References


Ikeda, Shoji. 2015b. Fojingyinyi yu riben guizhushu [The Glossaries of Buddhist Sutra and Hanzi dictionaries in early Japan].


in Humanities, Kyoto University.
Li, Yuan. 2015a. Tenreibanshōmeigi no homon dōtabēsu no kōchiku: Gyokuken ran kan taitō bibliun o chūshin ni. [Construction of Tenreibanshōmeigi text database: Corresponding to Yupian fragments]. Kokugo Kokubun Kenkyū 146: 65-80.
Li, Yuan. 2015c. The creation of a Tenreibanshōmeigi database and its textual study. 9th Conference of the European Association of Chinese Linguistics (EAACL-9), University of Stuttgart (Germany), 24-25 September.

Primary Sources
Yuanben Yupian 原本玉篇. Tōhō bunka sōsho dairoku 東方文化叢書第六, Tokyo: Tōhō bunka gakushū, 1932
Shinsenjikō 新撰字鏡. Tenjibon Shinsen jikō 天治本新撰字鏡増訂版, Kyoto: Rinsen shoten, 1973
Acknowledgements

A previous version of this paper has been presented at 15th International Conference of the European Association for Japanese Studies (EAJS2017), Universidade NOVA, Lisbon, 30 Aug-2 Sep 2017. We are grateful to participants for their useful comments and suggestions. We are also grateful to the owners of the dictionaries mentioned below, for permission to publish the decipherment text.

- Kōsanbibon Tenreibanshōmeigi — Kōsanji Temple authorities Ogawa Chie (the previous chief priest) and Ishizuka Harumichi (the representative director)
- Tenjibon Shinshenjikō — The Imperial Household Archives
- Zushōryobon Ruijumyōgishō — The Imperial Household Archives
- Kanchiibon Ruijumyōgishō — The Tenri Central Library

This work was supported by JSPS KAKENHI Grant Numbers 25370506, 16H03422, and 17F17301.