Current status and future prospects
of the Hanzi Normative Glyphs (HNG) Database

ISHIZUKA Harumichi

(Summary)
Following a presentation at the 2004 Autumn meeting of the Society for Japanese Linguistics, the Internet version of the Hanzi Normative Glyphs (HNG) database (headed by ISHIZUKA Harumichi) was launched in March 2005 (http://www.joao-roiz.jp/HNG/). Since then, every year new texts and relevant data have been added to the database. The objectives and methodology of this work was first published, with Ishizuka as the first author, in Nihongo no kenkyū 日本語の研究 (2005, vol. 1, no. 4), the official journal of the Society for Japanese Linguistics. Following the increasing amount of texts and data (62 texts, 4,554 unique characters, 432,596 character forms), this paper is an introduction to the current status of the project, its findings and future prospects.

In ISHIZUKA (2005), we documented the existence of an early Tang standard of writing and described how this standard shows significant changes in the Kaicheng Stone Classics of the middle Tang, which eventually served as the basis for Song printed works; in Japan the early Tang standard was established as the Japanese standard of writing and, despite the changes that occurred in China, was preserved until the advent of modern printing culture. With the addition of new texts and data to the database, this view basically remains the same. In addition to this, we now have shown the standard of writing in texts of the Nanbeichao, the Sui and early Tang dynasties; the standard in Chinese language material written by the non-Chinese peoples around China; as well as the standard in Chinese language manuscripts written in Japan, starting from earliest times to the beginning of the modern era. For the sake of comparison, the author also includes informal documents in the database and, using the rate of character variants, attempts to show that while a standard is clearly detectable in standard texts, it is much weaker in writings of informal nature. This will also serve to demonstrate that the standard of writing is applicable for ascertaining the formal vs. informal nature of a text.

Furthermore, since the HNG database in itself does not offer any particular conclusions, the author suggests possible research topics, such as the comparison of Dunhuang and Shōsōin manuscripts, as examples for a meaningful application of data.

1. Definitions
Since the terminology related to Chinese characters is defined by researchers in a variety of ways, we are using the terms "form" 書体, "glyph" 字体, and "shape" 字形, as set forth in ISHIZUKA (1984):
Form 書体 (shuti): The conventional style of the shape of characters. It is usually defined in reference to a corpus. (E.g. kaishu 楷書, caoshu 草書)

Glyph 字体 (ziti): Within the scope of one form, the conventional norm of writing each character.

Shape 字形 (zixing): Within the scope one glyph, the physical appearance of how a particular character was written (or printed).

The terms “form,” “glyph,” and “shape” represent three different levels, each one of which is completely independent of the other. On the basis of this understanding, we propose the definition of character “type” 字種 as follows:

Type 字種 (zizhong): The sum total of glyphs recognized by society as one character, which are interchangeable and usually have the same pronunciation and meaning.

Character type is what in everyday life people recognize and understand as an individual character, even though it is rarely defined in specific terms. In this paper, “type” refers to the variety of glyphs collectively.

HNG is a database that provides information on character glyphs. The characters are grouped according to their type (zizhong), with their shapes (zixing) representing the way they actually occur in manuscripts.

2. Overview of HNG
The “Database of the Normative Glyphs in Hanzi Script” (abbreviated as HNG) is a useful tool for observing the standard of writing in each time period and geographical regions. It was initially based on the “Ishizuka Register of Chinese Character Glyphs” 石塚漢字字体資料, the result of the author’s work of twenty some years, which was built by volunteers from the Department of Computational Linguistics at the University of Hokkaido. The Institute for Asian and African Linguistics at the Tokyo University for Foreign Studies provided support for developing the online version of the database (http://www.joao-roiz.jp/HNG/) that was opened to the public in 2004.

The “Ishizuka Register of Chinese Character Glyphs” (hereafter abbreviated as “Ishizuka Register”) was developed with the aim of documenting the standard glyphs in different time periods and geographical regions, including their changes in different times and regions. In the course of analyzing standard texts belonging to the cultural sphere of Chinese characters, character types (zizhong), glyphs (ziti) and their number of occurrences were accumulated on paper cards, amounting to a total number of 500 thousand examples derived from 79 texts of Chinese classical writings, Buddhist scriptures, Japanese manuscripts, etc. In addition, in order to study the standards of writing, not only official but a number of informal documents were included as well.

In order to provide information on character types, glyphs and their standardization for users with different backgrounds, the “Ishizuka Register” was digitized, thus creating HNG as an open
access database. Today, in April 2008, searchable data comprises information on 4,554 unique characters (i.e. types) with 432,596 occurrences from a total of 62 texts from different time periods and geographical regions. The database provides access not only to character types and their number of occurrences but also to other bibliographical information, such as geographical region, time period, edition or manuscript version. This is a unique tool with which users can specify their criteria to query data from all cultures and regions where Chinese characters had been used. The Japan Society for the Promotion of Science provided a research grant for opening the database to the public, and the Institute for Asian and African Linguistics at the Tokyo University for Foreign Studies supplied technical support for this task.¹

3. Opening the HNG to the public
HNG annually publishes new data in batches, according to specific topics. In 2004, these consisted of demonstrating the existence of an early Tang standard of writing; the progression from the early Tang standard to that of the Kaicheng stone classics; the adoption of the Kaicheng standard as the Southern Song printed standard; the patterns of implementing the Chinese standard in old Japanese manuscripts. In 2005, the database documented the Nanbeichao standard of writing prior to that of early Tang; and recorded the specific changes in the standard in Japanese manuscripts. In 2006, the main emphasis was laid on documenting the spread of standard glyphs from China to its neighbours, and for this purpose Korean and other non-Chinese material were also included. In 2007, our objective was to grant open access to an even larger body of material.

In the course of making data to available to the public, we are using the “rate of character variants” as a criterion for determining the nature of the texts. The rate of character variants is the rate at which multiple glyphs appear within the same text. It is calculated according to the following formula:

\[
\text{Rate of character variants} = \frac{\text{Total No. of variants}}{\text{Total No. in text} - \text{Total No. of sole occurrences}} \times 100
\]

“Sole occurrences” are character types (zizhong) that appear in a text only once and since this way no information is available regarding their variation, they are excluded from the calculation. “Character variants,” on the other hand, refer to characters that appear in the same text as different glyphs. A low number of variants in a text is considered significant.

For texts open to the public, HNG currently displays the number of character types, the number of glyphs, the total number of occurrences, as well as the number and rate of variants.² The data

¹ The HNG Editing Committee consists of ISHIZUKA Harumichi (Chairman, Professor Emeritus at Hokkaido University), TOYOSHIMA Masayuki (Associate Professor, Institute for Asian and African Linguistics, Tokyo University for Foreign Studies), IKEDA Shouji (Professor, Graduate School of Letters and Faculty of Letters, Hokkaido University), SHIRAI Jun (Lecturer, Faculty of Arts, Shinshu University), ITO Chiyuki (Assistant Professor, Institute for Asian and African Linguistics, Tokyo University for Foreign Studies). Cooperators include TAKADA Tomokazu (National Institute of Japanese Language), YAMAGUCHI Keita (Researcher, Graduate School of Letters and Faculty of Letters, Hokkaido University), OKAGAKI Hirotaka (same as above), TAKAGI Yui (Postdoctoral Program, Graduate School of Letters and Faculty of Letters, Hokkaido University), SAIKI Masanao (same as above).

² Because of the ongoing maintenance and corrections, the current figures may be different from those in ISHIZUKA et al.(2005), TAKATA and OKAGAKI (2006), etc. For detailed information on each text, see OKAGAKI (2008) and
reveals the low rate of character variants in the Kaicheng Stone Classics, as well as the differences in number and rate of variants within two versions of the same volume (juan) of the Nihonshoki. Since the rate of character variants is an indicator of the degree of attention paid to the standardization of glyphs, and the low rate of variants demonstrates a strong awareness, we set the criterion for determining standard (i.e. formal) vs. non-standard (i.e. informal) texts at 1.00%. The Kaicheng Stone Classics exhibit a low rate of character variants, confirming that these were highly standardized texts. In contrast with this, the inconsistency of glyphs would be the indicator of the inconsistency of the standard, leading to a higher rate of variants. We can see such high rate of variant characters in informal documents such as Hanshu Yang Xiong 漢書楊雄.

Moreover, the 432,596 total occurrences of the 4,554 unique characters in 62 texts, once again, not only confirm the existence of a standard in different time periods and geographical regions, but also show that the standard changed according to different times and regions.

Data opened to public in 2004

<table>
<thead>
<tr>
<th>Category</th>
<th>Name of text (date)</th>
<th>Abbreviation</th>
<th>Number (type, glyph, total number)</th>
<th>Variants (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>初唐写本</td>
<td>今西本妙法蓮華経卷五(671) 今西本妙法蓮華経三(675)</td>
<td>〈宮廷今西〉</td>
<td>(585字, 592字, 5685字)</td>
<td>114字 (2.03%)</td>
</tr>
<tr>
<td>寺院文書楊雄篇(初唐)</td>
<td>〈宮廷今西〉</td>
<td>1573字, 1701字, 4510字</td>
<td>206字 (4.57%)</td>
<td></td>
</tr>
<tr>
<td>開成石經</td>
<td>論語(837)</td>
<td>〈開成論語〉</td>
<td>(1322字, 1328字, 14325字)</td>
<td>5字 (0.03%)</td>
</tr>
<tr>
<td>周易(837)</td>
<td>〈開成周易〉</td>
<td>1404字, 1420字, 23248字</td>
<td>43字 (0.18%)</td>
<td></td>
</tr>
<tr>
<td>北宋版</td>
<td>東簡寺版阿毘達摩大毘婆沙論卷百七(1100)</td>
<td>〈東簡毘婆〉</td>
<td>(357字, 368字, 6979字)</td>
<td>42字 (0.60%)</td>
</tr>
<tr>
<td>齊民要術卷五(12C初)</td>
<td>〈齊民要術〉</td>
<td>1404字, 1420字, 23248字</td>
<td>43字 (0.18%)</td>
<td></td>
</tr>
<tr>
<td>開元寺版道神足無極變化経卷四(1126)</td>
<td>〈開元神足〉</td>
<td>674字, 692字, 5528字</td>
<td>57字 (1.03%)</td>
<td></td>
</tr>
<tr>
<td>南宋版</td>
<td>華厳經內章門等雜孔目卷一(1146)</td>
<td>〈華嚴孔目〉</td>
<td>(779字, 814字, 16967字)</td>
<td>107字 (0.63%)</td>
</tr>
<tr>
<td>日本書紀(写本)</td>
<td>岩崎本巻二十四(10C)</td>
<td>〈岩崎紀24〉</td>
<td>(1099字, 1173字, 5401字)</td>
<td>116字 (2.15%)</td>
</tr>
<tr>
<td>兼方本巻三(1286)</td>
<td>〈兼方紀3〉</td>
<td>1143字, 1166字, 10006字</td>
<td>55字 (0.55%)</td>
<td></td>
</tr>
<tr>
<td>日本書紀(版本)</td>
<td>慶長勅版巻(1599)</td>
<td>〈勅版紀〉</td>
<td>(1141字, 1163字, 9920字)</td>
<td>65字 (0.66%)</td>
</tr>
<tr>
<td>日本書写本</td>
<td>和銅経大般若経巻二百五十(712)</td>
<td>〈和銅250〉</td>
<td>(161字, 166字, 7476字)</td>
<td>10字 (0.13%)</td>
</tr>
<tr>
<td>高山寺本大教王経巻一(815)</td>
<td>〈金剛大教〉</td>
<td>495字, 508字, 6645字</td>
<td>52字 (0.78%)</td>
<td></td>
</tr>
<tr>
<td>東簡寺版写大教王経巻一(12C)</td>
<td>〈佛説大教〉</td>
<td>(794字, 845字, 4291字)</td>
<td>118字 (2.75%)</td>
<td></td>
</tr>
</tbody>
</table>

Data opened to public in 2005

| 敦煌南北朝写本 | P2179極楽論巻八(514) | 〈P2179〉 | (556字, 565字, 6138字) | 40字 (0.65%) |
| S2067華厳経巻十六(514) | 〈S2067〉 | 629字, 643字, 7528字 | 37字 (0.49%) |
| S81大般涅槃経巻十一(506) | 〈S81〉 | 928字, 959字, 6661字 | 58字 (0.87%) |
| P2160摩诃摩耶経巻上(586) | 〈P2160〉 | 1046字, 1088字, 6008字 | 54字 (0.90%) |
| 佛写本 | P2413大楼炭経巻三(589) | 〈P2413〉 | (547字, 574字, 4626字) | 49字 (1.06%) |
| 仏経賢劫経巻二(610) | 〈賢劫経2〉 | 884字, 927字, 7762字 | 86字 (1.11%) |
| P2334妙法蓮華経巻五(617) | 〈P2334〉 | 632字, 647字, 5672字 | 23字 (0.41%) |
| 高昌写本 | 大品経巻二十八(高昌期) | 〈京博大品〉 | (271字, 273字, 1547字) | 2字 (0.13%) |

the top page of HNG’s online version.
In order to be able to use HNG for research, it is important to understand the nature and objective of the database. Firstly, in terms of its basic character set and arrangement, HNG is primarily...
based on UEDA Kazutoshi’s 上田万年 Daijiten 大字典. Characters listed in the Daijiten as identical characters 同字, popular characters 俗字, etc. are treated in the database under the same character type (zizhong) and are allocated the same data position. There are many such cases; the characters 虫→蟲 and 喜→豊 are typical examples. At the same time, there are cases when experience tells us that it is appropriate to deviate from our initial criteria. For example, although the Daijiten lists the character 笑 as the ancient form of 笑, in Japan they are often considered different characters, and thus they are also differentiated in HNG. Similarly, the characters 增 and 垂, 州 and 洲 are also treated as separate character types (zizhong). At the same time, in order to examine glyph standardization more efficiently, the characters 无 and 無 are merged into 無, the characters 修 and 修 into 修. (This rule is observed with only a few exception in specific texts.) In these points HNG differs from the Daijiten and applies its own criteria, and thus it is worth keeping in mind that in certain cases character variants may be interpreted in a different way from how they are ordinarily understood.

Next, let us look at the definition of variation at the level of glyphs. HNG encompasses material from the entire cultural sphere of Chinese characters, with a spread of over a thousand years, beginning with 6th century Nanbeichao manuscripts found in Dunhuang. This material is arranged and presented in a variety of different ways. Distinguishing variants is vital for showing the standardization of glyphs, but there are also exceptional cases when concrete occurrences of characters differ from each other visually and yet they are not accepted as variants.

In contrast with printed editions, in manuscripts the concrete character shapes (zixing) of the same glyph may show a considerable degree of discrepancy according to different handwritings or calligraphic forms (shuti). To illustrate this point, consider the case of 灶 and 灚 which are two concrete examples and which differ from each other in that their last portion is written as 灌 or 灌, respectively. However, since 灌 is the abbreviation of 灌, we can ascertain that this is a difference of calligraphic forms (shuti), and thus consider the two as a single glyph. Beside this, differences in length and intersection of strokes, location of components, and other cases where objective distinction is problematic are not considered as multiple glyphs.

In addition, as shown on Figure 1, when characters can be confirmed as mistakes by comparing the text with parallel sections on other manuscripts, or on the basis of context, they are treated as “erroneous characters” 誤字 without counting them as variants.

![Figure 1: Erroneous characters (left: non-existing glyph in the Zushoryō version of the Nihonshoki 圖書紀24; right: typo in the Japanese manuscript of the Gaosengzhuan from the Moriya Collection 五一續高)](image)

The above sums up the way information is processed in HNG. Within the search results, each glyph is then displayed using a single representative example -- the rest of the examples recorded
on the original paper cards at this point are not included in the database.

4. Possible research topics
Although HNG in itself does not offer any particular conclusions, it provides the possibility for exploring a variety of research topics. Below are a couple of examples.

4.1 HNG and traditional character dictionaries
In the Kangxi zidian and other character dictionaries, the glyph 高 is listed as the standard character (zhengzi) and 高 as the popular (suzi). At the time HNG was first opened to the public, we have already pointed out (Ishizuka 2005, etc) that this distinction could not be applied uniformly to all time periods and geographical regions. Now, looking over the 62 texts included in HNG, this observation appears just as valid as before (Figure 2).

The dictionary descriptions are generally based on the Shuowen jiezi. For example, the Xin jia jiujing ziyang 新加九经字样 says:

Gao means to esteem highly; the character visually resembles the shape of a raised platform used for observation; the second one is an abbreviated glyph from the clerical script; the characters 亭, 毫, etc all derive from 高 by means of abbreviation.

Accepting the same point of view, the Japanese manuscript of the Ruiju myoigisho 類聚名義抄 from the Kanchiin temple’s collection sees 高 as the standard character. HNG, on the other hand, lists 高 as the standard in the Kaicheng Stone Classics, the Song prints, and the Nihonshoki printed on imperial commission during the Keichō reign. However, the database also shows that in some cases 高 was the standard glyph, as in the Japanese manuscript of the Mile shangsheng jing 弥勒上生经 (738). In the course of the transition from the early Tang standard to that of the Kaicheng Stone Classics, there were also smaller divergences that disappeared in later periods, as
it is demonstrated by the Japanese manuscript from the Tempyō era.

4.2 Comparison of the Dunhuang and Shōsōin manuscripts

Before the standard of the early Tang, a series of standards existed during the Nanbeichao period, as it can be learned from HNG by comparing manuscripts S81 and P2160, S2067 and P2179. At the same time, one cannot document this phenomenon in full confidence on the basis of Dunhuang manuscripts alone. Since manuscripts from the Shōsōin Shōgozō collection have been recently published on CD and DVD, this is a newly opened direction for research. Figure 4 shows the glyphs of character 最 in manuscripts S81 (Southern dynasty), S2067 (Northern dynasty), S2423 (Dunhuang), Mile shangsheng (Tempyō), Zhengsifen 20 (Tang), Zhengsifen 16 (Tempyō), and Huayan Xinluo (Simla). Manuscript Zhengsifen 20, brought back by the monk Ganjin, not only reveals the transitional changes of the high Tang period, which are likewise seen in the Mile shangsheng and Huayan Xinluo manuscripts, but also raises the possibility that its text originally derived from a manuscript which had been -- similar to S81 but unlike Zhengsifen 16 that was based on Genbō’s manuscript (i.e. Chang’an jing 長安經) -- written in the Nanbeichao standard. (Ganjin was a native of Jiangyang county in Yangzhou and stayed at the Dayun monastery in Yangzhou.) The character 突 illustrates this point. On the other hand, the character 正 is an indication that this is not true for all characters.

This kind of comparison once again confirms the uniqueness of the Mile shangsheng manuscript. This manuscript is representative of the transition from the early Tang standard to that of the Kaicheng Stone Classics, whereas such small distinctions cannot be detected in the Huayan Xinluo. The Dunhuang manuscript S2423 is also very interesting. Generally speaking, the Dunhuang manuscripts are most representative of the culture of Central China for the 70 year period between late 7th and early-mid 8th centuries. Manuscripts dating earlier or later this period do not always reflect the culture of China proper. But even during this period, when compared with the numerous copies of Nara-period Pratīyāparamita sutras that survived in Japan, the Dunhuang copies of the same sutra exhibit a certain degree of regional peculiarities.

In either case, looking at the Dunhuang and Shōsōin manuscripts together is also meaningful from the point of textual theory.

<table>
<thead>
<tr>
<th>Char.</th>
<th>S81 (total number)</th>
<th>S2067</th>
<th>S2423</th>
<th>正四分20</th>
<th>正四分16</th>
<th>花嚴新羅</th>
</tr>
</thead>
<tbody>
<tr>
<td>因</td>
<td>01611</td>
<td>(15)</td>
<td>(2)</td>
<td>(1)</td>
<td>(47)</td>
<td>(1)</td>
</tr>
</tbody>
</table>
5. Summary

Above the author described the current status (as of April 2008) of the HNG data that has been opened to the public, and demonstrated the meaning of the rate of character variants within this material. Using this rate as a criterion, it is possible to determine the nature of a text. The standard texts of different time periods and geographical regions reveal the presence of a standard of writing, and the database can demonstrate the changes of this standard in time and space.

Finally, although HNG in itself does not offer any particular conclusions, it provides the possibility for exploring a variety of research topics, such as the relationship of the data with traditional character dictionaries, or the implications of the comparison of Dunhuang and Shōsōin manuscripts for textual theory.

Bibliography

ISHIZUKA Harumichi 石塚晴通 (1999). 「漢字字体の日本的標準」、『国語と国文学』第76巻第5号


