Ideograph Variant Forms and Usage Control in NACSIS-CAT

Akira Miyazawa

Follow this and additional works at: https://scholarsarchive.byu.edu/jeal

BYU ScholarsArchive Citation
Available at: https://scholarsarchive.byu.edu/jeal/vol2006/iss138/4
IDEOGRAPH VARIANT FORMS AND USAGE CONTROL IN NACSIS-CAT

Akira Miyazawa
National Institute of Informatics, Japan

Preface

In the 2005 CEAL meetings in Chicago, I heard from several people that UNICODE, EACC and IME problems were a current issue in the OCLC and RLIN new cataloging systems. Sometimes I was asked a question about how our NACSIS-CAT cataloging system handles these problems. Of course, NACSIS-CAT has nothing to do with EACC, though it has been using UNICODE more than five years. Still, I thought it might be helpful for American CJK catalogers to know about character set handling in NACSIS-CAT. I would be glad if you find something useful in this article.

1. Introduction: NACSIS-CAT

Japanese catalogers in CEAL may not need introduction to NACSIS-CAT. However, for other area librarians, a brief introduction to NACSIS-CAT may be helpful. NACSIS-CAT is a shared cataloging system, similar to the OCLC and RLIN cataloging systems, operated by National Institute of Informatics (NII) Japan. It is part of a government supported program to provide a national information infrastructure for research activities. The system was developed by the Center for Bibliographic Information at the University of Tokyo and started its operation in 1984. National Center for Science Information Systems (NACSIS) was established in 1986 to operate this system, along with other scholarly database services. NACSIS was reorganized to NII, enhancing research function of broad fields of informatics in 2000.

In April 2005, over 1300 libraries (mostly Japanese university libraries including some European and Asian libraries) are contributing to the union catalog, which comprises more than seven million monograph titles, two hundred and eighty thousand serial titles, about eighty million holdings records, 1.3 million author name authority records and twenty-four thousand uniform title authority records. Weekly increase is about ten thousand titles and a hundred thousand holdings records. In addition to the union catalog, more than thirty three million MARC records of various countries, including Japan, USA, UK, Germany, Korea, and China, are loaded to be referenced for cataloging. Recent developments have enabled RLIN and OCLC databases to be referred online for cataloging.

First generation NACSIS-CAT system used JIS (Japanese Industrial Standard) character code with private extension for European languages. Japanese, most Latin script written languages, and Russian could be handled with this character code. In 1996, a new NACSIS-CAT server client based system was introduced, and this new system was enhanced to use Unicode in year 2000. Now, it can handle almost all the languages.

NACSIS-Webcat Plus (http://webcatplus-international.nii.ac.jp/en/) is an open access web interface to the whole union catalog database produced by NACSIS-CAT. NACSIS-Webcat Plus and NACSIS-CAT both utilize CJK unified index, which enables cross variant form search (i.e., to retrieve or with )

2. Unification, Source Code Separation and Z-variant

For a discussion of Unicode usage, some basic concepts and terms should be introduced. The first one is “unification”. This is a basic concept of Unicode [1]. By the glossary of Unicode, it is “the process of identifying characters that are in common among writing systems”. For example Ő in Turkish is unified with Ơ in German and named LATIN CAPITAL LETTER O WITH DIAERESIS. Its code point is U+00D6. Hanzi, kanji or hanja in Unicode is called CJK unified ideograph. That means the repertory was made from Chinese, Japanese and Korean standard character sets with “unification”.

Unification of CJK unified ideographs was not a very simple job when the repertory was made in 1991 [2],
explained in the Unicode Standard [5]. In short, it is to unify very similar shape variants like U+8fB6 with one dot and two dots.

In spite of the unification rules, characters U+6236, U+6237 and U+6238 are not unified in the repertory. This is due to the rule called source separation rule, or round trip rule. It is to ensure that round trip conversion between UNICODE and a source national standard code (such as JIS) does not cause loss of information. In this case, Taiwan national code distinguished these three “characters,” while Japanese, Korean and Chinese in that time made no distinction.

Variant like this case is sometimes called “Z-variant.” By the Unicode website, Z-variant is “Two CJK unified ideographs with identical semantics and unifiable shapes, for example, U+8AAA and U+8AAC.” [6].

Z-variant is caused by not only source separation rule but also by the need to keep compatibility with other sources. Korean pronunciation variants of a hanja necessitate 268 compatibility ideographs which are completely same form. Other reasons of Z-variant are described in the Unicode Standard [5].

There is another type of variant in the glossary called “Y-variant.” It is “Two CJK unified ideographs with identical semantics and non-unifiable shapes, for example, U+732B and U+8C93.” [6]. Characters of , and are also Y-variant examples.

Unihan database is a kind of dictionary attached in the Unicode Standard. It is also available through web page [7]. This database lists some other variant relations such as simplified variant or traditional variant. These variants can be classified as Y-variants.

One problem here is that there is no normative source of Z-variants or Y-variant. Unihan database is not a normative part of the standard. Y-variant is unquestionably language dependent. U+53F6 is simplified character of U+8449 in modern Chinese, but not in Japanese. U+82B8 is simplified character of U+85DD in modern Japanese, but not in Chinese.

Z-variants are less language dependent. Appendix S of ISO 10646 lists examples of Z-variants by source separation rules, but it is not an exhaustive list. In fact, the plan of providing complete source separation list was abandoned by the standard developer due to difficulties foreseen. Judgment about Z-variant or not is subjective to personal view. There are certain number of people who claim U+9AD9 is not Z-variant of U+9AD8, but Y-variant. Z-variants interpretation of Unihan database is too broad, in my opinion. For example, U+85DD is listed as a Z-variant of U+82B8, though it might be just by a mistake. U+4E0A and U+4E04 is also a Z-variant.

The concept of Z-variant and Y-variant is acceptable for most of people. But actual assignment of variant relation to individual character is much more debatable.

3. Character set usage control

NACSIS-CAT system has a concept and a mechanism of character set usage control from the first generation system. It is to limit usage of certain characters in the character set and unify them with other characters in the set. This is introduced mainly to avoid cumbersome distinction between full and half width Latin alphabets.

Japanese encoding systems have two set of Latin alphabets due to parallel usage of double byte and single byte character sets. Double byte Latin alphabets (often called zenkaku or full-width) are distinguished from single byte alphabets. But, such distinction has meaning only in computers. There is no way to tell if an alphabet written on a title page is full width or not.

In NACSIS-CAT system, full-width and normal alphabets are regarded as same characters. To ensure this, the system normalizes input data. Full width alphabets input by users are converted to normal alphabets before stored in the database. This conversion is applied for all the input to the system, including search terms. You will not see full-width alphabets in the output from the system.

Character set usage control is intended to limit the character set, so that the complexity of cataloging work is reduced. Script level normalization by the system ensures consistent usage of the character set.
4. CJK unified ideograph usage of NACSIS-CAT

When NACSIS-CAT started to support UNICODE, its character set usage control was expanded. In brief, Z-variants are unified. That is, for example, U+8AAA and U+8AAC are regarded as the same character. If a user inputs U+8AAA in a data entry, it is automatically converted to U+8AAC by the script level normalization process.

But, our Z-variant list is different from Unihan database. As mentioned before, Z-variants of Unihan database seems to be too broad. We selected our Z-variant list based on the Appendix S of ISO/IEC 10646.

You may ask if this causes loss of information. Yes, it may cause slight loss of information. But, catalog description can not copy everything. In Western languages, upper and lower cases are not preserved. Italic, bold or handwritten styles are not preserved. Information loss by Z-variant normalization is at the same level as those cases.

In fact, there are few complaints about this, and it definitely reduce burden of catalogers.

5. CJK unified index

Script level normalization should not be confused with a unified index to provide , , and type access. Script level normalization is performed for all input data to the system. A unified index also normalizes certain characters, but it is only for the hidden index fields, which are invisible to normal users.

When a catalog record is written to the database, the system automatically extracts index words from the record. Then the words go through the unified index normalization process. This is to unify Y-variants. But our Y-variant list is, again, different from Unihan database. It is broader than Unihan database. As mentioned before, Y-variant is language dependent. We merge Chinese, Japanese and Korean Y-variants to provide one large list.

Here is an example of CJK unified index Y-variant group.

U+79C7  U+827A  U+82B8  U+841F
U+84B7  U+84FA  U+8553  U+85DD

By the unified index normalization process, any occurrence of these characters is converted to the representative character ( U+82B8 in this particular case). The system also applies this process to all the search terms input by the user.

With this unified index, you may get an inappropriate result. The search term retrieves , too. This is not desirable, theoretically. But, in real cataloging application, this flaw seldom causes actual problems. The advantage of not needing to worry about variant forms is far more valuable.

6. Other character usage control in NACSIS-CAT

Character set usage control in NACSIS-CAT is not only for ideographs. Usage of symbols is also controlled. Hundreds of new symbols in addition to scripts became available when NACSIS-CAT introduced Unicode. Symbols include, for example, , , and , which could appear in book titles.

None of these symbols are used in the NACSIS-CAT database. Even if you find a title “I L VE NEW YORK”, the description should be “I love New York” and the heart mark should be described in the note. Similarly, you don’t use Roman numeral in the pagination.

This limitation is to keep compatibility with existing records and also with the records from external source like US MARC.

7. Conclusion
Z-variants and newly added symbols are excluded from the Unicode to provide catalog description character set. This usage control contributes to reducing from a cataloger’s workload details of judging variant forms.

Our usage control is based on Japanese linguistic sense and may not be applied directly to a different linguistic environment. But some kind of character set usage control is necessary for a very large character set like Unicode.

References