日本・中国・台湾コンピュータ異体字シソーラスの制作

1N−5

安岡孝一
京都大学大型計算機センター
yasuoka@kudpc.kyoto-u.ac.jp

安岡素子
フリー

1 はじめに

漢字の入ったデータを検索する際に問題となるのが、**異体字**の存在である。例えば「浜」を検索する場合に「濱」も同時に検索したい、ということは漢字検索の基本的要求の1つである。

JIS X 0208の漢字6355字しか用いることのできなかった時代には、異体字の存在はそう大きな問題とはならなかった。というのもJIS X 0208には、このような異体字はたかだか560組ほどしかなく、しかも規格票に異体字関係がとりあえず明記されていたからである。しかしこの数年の間に、コンピュータで用いることのできる漢字の数は飛躍的に増大し、それにともなって異体字関係も複雑となった。例えばWindows-NTではUnicodeの漢字20902字が全て使えるが、その結果「浜」の異体字としては「濱」以外に中国・台湾の漢字コード規格から「濵」と「濱」が追加されているため、漢字検索の手間が増大しているのである。しかも、Unicodeや中国・台湾の漢字コード規格には異体字関係が全く記載されていないため、検索者の方で適当に異体字を含めて検索しているのが現状である。

筆者らはこれまでに、中国のGB 2312と日本のJIS X 0208の相互変換[1]や、日本・中国・台湾の漢字コードの差異に関する研究[2]をおこなってきた。これらの経験を活かし、筆者らは現在、日本・中国・台湾の漢字コード規格に含まれる漢字70000字の異体字シソーラスの開発をおこなっている。本稿ではこの異体字シソーラスと、そのWWWによるヴィジュアライゼーションである「漢字袋」について述べる。

Japan-China-Taiwan Computer Kanji-Hanzi-Hanntzyh Thesaurus
Koichi Yasuoka & Motoko Yasuoka
Kyoto University Data Processing Center
Kyoto 606-01 JAPAN

2 異体字シソーラスの実際

2.1 対象とする漢字コード

筆者らの異体字シソーラスの対象は、以下の規格の漢字全てである。

日本 JIS X 0208-1990 (以下(JIS90))
JIS X 0208-1983 (以下(JIS83))
JIS C 6226-1978 (以下(JIS78))
JIS X 0212-1990 (以下(JIS+))
中国 GB 2312-80 (以下(GB))
GB/T 12345-90 (以下(GB1))
GB 7589-87 (以下(GB2))
GB/T 13131-91 (以下(GB3))
GB 7590-87 (以下(GB4))
GB/T 13132-91 (以下(GB5))
GB 8565.2-88 表A4 (以下(GB+))
台湾 CNS 11643-1992 第一〜第七字面 (以下(CNS1)〜(CNS7))
BIG5 (以下(BIG5))

これらの規格において、国際規格ISO/IEC 10646-1:1993 UCS-2 (以下(UCS))あるいは中国の漢字内碼拡展規範 (以下(GBK))との関係が定義されている漢字については、それらのコードもシソーラス中に含めた。また、日本のJIS X 0208:1997は、漢字コードとしては(JIS90)と同一であるため、シソーラスには含めていない。

2.2 異体字関係の同定

筆者らが用いた異体字同定の基礎資料は、日本の(JIS90)附属書2中の**異体字**と(JIS+)附属書3中の**同義漢字**、中国の**第一批異体字整理表**と**簡化字総表**、台湾の**異体国字字表**である。ただし、異体字関係は日本・中国・台湾でそれぞれ違いがあるため、適宜**大漢和辞典**・**大漢語字典**・**中文大辞典**等を参考に、最終的な決定は筆者らの独断と偏見でおこなった。当然のことながら、1つの漢字が複数の異体字群に属することもありうる。

異体字関係にあると考えられる漢字のうち、以下の基準にあてはまるものに関して

は特に同字であるとみなした。すなわちこれらは、１つの漢字に複数の漢字コードが割り当てられているものと解した。

i) (JIS90) (JIS83) (JIS78) で同一の字形である漢字。
ii) (JIS+) の漢字で (JIS78) との関係が規格の解説表に記載されているもの。
iii) (GB) と (GB1)、(GB2) と (GB3)、(GB4) と (GB5)、およびそれらと (GBK) の間で、同一の字形である漢字。
iv) (GB+) の漢字で (GB2) あるいは (GB4) から集められたもの。
v) (CNS1) の部首用文字で (CNS1) ～ (CNS7) に同一の字形があるもの。
vi) (BIG5) と (CNS1) あるいは (CNS2) の漢字。
vii) (UCS) の漢字で他の漢字コードへの写像が定義されているもの。

ただしわずかでも疑義がある場合には、同字とはみなしていない。

2.3　シソーラスファイルの構造

筆者らの異体字シソーラスファイルは、幅広い利用が可能であるようにプレーンテキストとし、１行が１グループをなすように構成されている。同字関係にあると考えられる漢字コードはコンマで区切られ、異体字関係にあると考えられる漢字コードは空白で区切られている。例えば

```
JIS-4145,UCS-6D5C JIS-6332,UCS-6FF1 JI
S+-4106,UCS-6FF5 GB-1785,GBK-B1F5,UCS-
6EE8 GB1-1785,GBK-9E49,UCS-6FF1 CNS1-7
423,BIG5-C0D8,UCS-6FF1 CNS3-5159,UCS-6
FF5 CNS4-2D6F CNS7-4B43
```

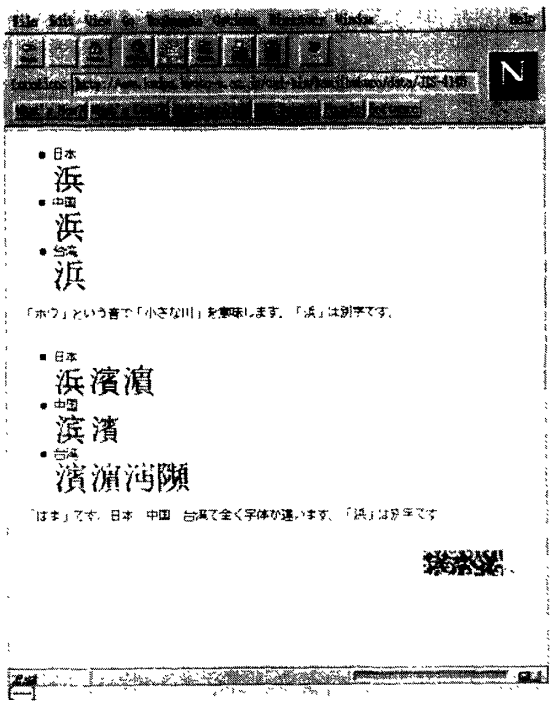
という行 (紙面の都合で複数行に分割されているが、ファイルでは１行である) は「浜・濱・濵・浜・濱・濱・濵・洟・隕」の９字が異体字関係にあることを示している (「浜」は (JIS90) (JIS83) (JIS78) ではいずれも 41 区 45 点で (UCS) では 6D5C、以下同様)。なお、JIS-4145 を含む行はファイル中にもう１つ

```
JIS-4145,UCS-6D5C GB-6826,GB1-6826,GBK
-E4BA,UCS-6D5C CNS3-3226,UCS-6D5C
```

があり、これは「浜・浜・浜」がまた別の異体字関係にあることを示している。

3　漢字袋

前章で述べた異体字シソーラスの WWW によるヴィジュアライゼーションとして、筆者らは**漢字袋**を開発した。漢字袋の各ページはシソーラスの各行に対応しており、それを日本語音、拼音、画数などから検索可能である。「浜」を検索した結果を以下に示す。



ここでは先の例で述べた２つの異体字群が、同時に表示されているのがわかる。なお、ここで表示されている漢字をそれぞれクリックすると、その漢字の漢字コードが表示されるようになっている。

4　おわりに

異体字シソーラスファイルは、ftp://ginkaku.kudpc.kyoto-u.ac.jp/CJKtable/Variants.Z においてフリーで配布している。漢字袋の URL は http://www.kudpc.kyoto-u.ac.jp/~yasuoka/kanjibukuro/ である。どちらも自由に利用されたい。

参考文献

[1] 安岡孝一、一谷素子: GB汉字 ⟷ JIS 漢字相互変換ツールの開発, 全国共同利用大型計算機センター研究開発論文集, No.16 (1994 年 11 月), pp.67-74.

[2] 安岡孝一、安岡素子: 漢字と汉字と漢字, 全国文献・情報センター人文社会科学学術情報セミナーシリーズ, No.3 (1995 年 12 月), pp.37-75.