漢字コードのメタコード化の方法

斎藤　秀紀
国立国語研究所

2Q−2

1.漢字コードの現状と問題点

情報交換用に設定された漢字コードはJIS X0208 、JIS X0212 、GB2312 (中国) 、KSC 5601 (韓国) 、米国におけるUNIX、ユニコード(Unicode) がある。また、国際標準化機構(ISO) においても各国の漢字コードを統合した世界共通の漢字コード(DIS10646)の作成を進めている。

JIS X0208 は、1978年に制定され、1983年に第一回の改訂がおこなわれた。新版は、旧版との間に互換性を崩す問題があった。原因は、JIS X0208 における各水準への正字と異体字の配当規則が常用漢字、人名地名漢字の字体改訂に対し水準間で文字の入れ替えを避けられないものにした。そのほか、JIS X0208 には、多国言語の混在処理、漢字の属性情報 (画数・部・読み) の規定、字体の整理とコード配当の方法など改善すべき点が多い。

また、JIS X0208 とISO 案にも(1) コード間への追加機能とコードの拡張法の欠如(2) 旧規格で作成したデータとの互換性の維持(3) 各種の国家規格とISO 規格の調整方法の確立(4) 古典文献を含む学術情報への適用力の問題がある。本報告は、現行の漢字コードを包含し、文字パターン・コード・漢字の属性情報を規定できる4 バイトコードの提案をおこなう。4 バイトコードを作成するための条件および使用目的は、次の5 項目である。

(1)JIS X0201に従う2 バイトコードを包含すること。

(2) コード間に追加機能があり、文字の追加によっても基本配列を維持できること。

(3) 市販の漢和辞書をコードブックに使用するため検字番号の内部コード化と漢字の属性情報の規定、文字、コードの管理手段を確立できること。

(4) ローカルネットワークおよび異機種間共有データベース用メタコードに使用できること。

(5) 長期のデータ保存用コードに使用できること。

2.コードの構造化の方法

新コードは、既存の2 バイトコードを包含し、旧コードからの移行が容易なことが必要になる。そのために新コードは、複数の異なるコードを構造の要素として含んでいることが重要になる。4 バイトコードは、'01'領域を2 個組み合わせ2 次コード化する (図1)。この組み合わせは、2 バイトコードを2 バイト、4 バイト、2 バイト・4 バイト混在コードの3 種に拡張できる (以下2 バイトコード化領域は、2 の8 ビット目のパターン'00'、'01'、'10'、'11'を使用する)。

4 バイトコードの基本構造の設定は、2 バイトコードを基本に識別コードを付加し3 バイトコードを作る (図2)。次に、辞書の検字番号をJIS X0201 に従う3 バイトコードに圧縮し (式1)、第1 のコードに重ね合わせる。3 バイトコードに文字追加用枝番号1 バイトを付けコードを4 バイト長に調整する (式2)。最後に、4 バイトコードの2 の8 ビット目を'0101'に設定し2 バイトコードとの識別情報とする (図5)。

2 個のコードを重ね合わせることは、4 バイトコードを94区点8,836 字を基本とする面と、検字番号を3 バイト化した連続コードの二種を定義できる。また、構造化した4 バイトコードは、2 バイト、4 バイトコードの混在処理、2 バイトコードの表外字規定、2 バイトコードを4 バイトコードの要素としての位置を明確に規定できる。これらの機能は、4 バイトコードを2 バイトの表外字に使用することによって、2 バイトから4 バイトコードへの移行と併用を容易にする。なお、4 バイトコードの整数部3 バイトで表現できる文字数は83,0584 字、小数部は94字である。

Procedure for Metacode Conversion of the Kanji Character Code
Hidenori SAITO
THE NATIONAL LANGUAGE RESEARCH INSTITUTE

G0 0 0	0 1
1 0	G1 1 1

G0 ：JIS X0208 領域
'10'、'11'：2ﾊﾞｲﾄｺｰﾄﾞ領域
'01'：4ﾊﾞｲﾄｺｰﾄﾞ領域

図1. 2バイト・4 バイトコード化領域

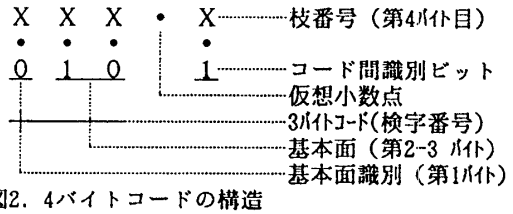

図2. 4バイトコードの構造

```
INPUT "ｹﾝｼﾞﾊﾞﾝｺﾞｳ:",I
I =I-1
FOR J= 1 TO 3
"HEX$(VAL("&H"+HEX$(I MOD 94))+&H21)"     (1)
I=I ¥ 94
NEXT J
GO TO 10
"枝番号圧縮値＝HEX$（枝番号）＋重み21"    (2)
```

¥: 整数わり算計算式
HEX$:10 進数、16進数変換式　MOD:剰余計算式

図3.検字番号の94進16進数変換処理

検字番号	94進数	16進数	21重み
1	000000	000000	212121
2	000001	000001	212122
93	000092	00005C	21217D
94	000093	00005D	21217E
8836	009393	005D5D	217E7E
830584	939393	5D5D5D	7E7E7E

図4.検字番号の94進数16進数変換値

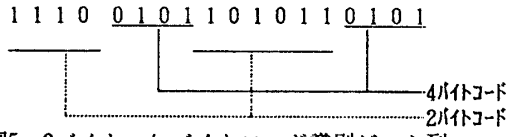

図5. 2バイト・4 バイトコード識別ビット列

3.レコードの書き込み・読みだし処理

文字発生装置は、ROM 化されているため、利用者が大量の文字登録やコードを変更する場合、外部記憶装置を使用する必要がある。外部記憶装置へのデータの記録には、レコードのブロッキング、デブロッキング、セクタ指定の効率的な処理が必要になる。以下レコード処理には、検字番号の進数変換による内部コード化の方法が使用できることを示す（図6)。

データを記録する外部記憶装置のセクタ長（変数S)、レコード長（変数R)、検字番号（変数I)を初期値として与え検字番号の最小値を'1' から '0'に変更する（式3)。外部記憶装置上のセクタ指定は、検字番号をレコード番号として使用する。外部記憶装置のセクタ番号(Sn)は、"レコード長(R)"に処理対象となる "検字番号" を掛け "必要な全レコード領域" を求め "セクタ長" で割る（式4)。次に "セクタ長" を「法」とし剰余（Pn）を求める（式5)。セクタへのレコードの転送桁数は、セクタ番号'S0'と'S1'を比較し、同値の場合 'S0' 番のセクタへ、異値では'S0'と'S1'に分割する（式6)。'S0'番目のセクタへの転送桁数は、'P1-P0' である。'S0'および'S1'番目のセクタへのレコード転送桁数は、'S-P0'と'R-(S-P0)'である。

```
  "ｾｸﾀﾁｮｳ:S, ﾚｺｰﾄﾞﾁｮｳ:R, ｹﾝｼﾞﾊﾞﾝｺﾞｳ:I"
  I=I-1                                    (3)
  S0=R*I ¥ S:  S1=R*(I+1) ¥ S              (4)
  P0=R*I MOD S:  P1=R*(I+1) MOD S          (5)
IF S0=S1
  THEN PRINT"ｾｸﾀ:";S0;"ｲﾁ:";P0;"ｹﾀ:";P1-P0;
  ELSE PRINT"ｾｸﾀ:";S0;"ｲﾁ:";P0;"ｹﾀ:";S-P0;
                          "+";R-(S-P0);    (6)
GO TO 30
```

図6.レコードの書き込み・読みだし処理

4.おわりに

4 バイトコードは、検字番号の内部コード化の方法によって表外字、外部記憶装置へのレコードの書き込み、読みだしへの進数変換の導入、多国語の包含、大規模の漢字辞典の電子媒体化を可能にした。また、辞書とコンピュータの結合は、漢字の属性情報の規定と辞書による情報交換の道を開いた。さらに、異機種間情報交換用のメタコード利用の可能性を高めた。今後は、正字、異体字の4 バイトコードへの配当方法の検討、異機種間ネットワーク上のメタコードとしての効果と問題点を探る予定である。

参考文献
1)斎藤秀紀: 東アジア漢字圏に対する統一コードの提案、情報処理学会論文誌投稿中(1991).
2)斎藤秀紀: 漢字コードの拡張法、情報処理学会論文誌投稿中(1992).
3)斎藤秀紀: 異機種間共有データベース構築における4 バイトコードの利用、情報処理学会論文誌投稿中(1992).