構造化4バイトコードによる多言語表現法の提案

斎藤　秀紀
国立国語研究所

1．はじめに

我が国における電子処理を目的にした漢字符号の標準化は、1978年に制定されたJIS C6226 （後、JIS X0208)に始まる。現在、JIS X0208 の利用は多様化し、古典・漢籍などの電子処理にも使用されている。しかし、JIS X0208 は、日常使われる日本語の交換を目的に作られたため、長期間データを安定した状態で保存する機能や古典・漢籍を符号化する充分な文字種が確保されていない。また、これまでの運用経験から幾つかの重大な問題があることが知られている。

主なものは、(1) 第一水準に政令で規定する文字を包含させ二重基準を内包させたこと。(2) 符号間に文字を追加する機能が欠如していること。(3) 変更された情報が他の情報に影響しない局所化に対する機能が欠けていること。(4) 読み・画数などの統一配列基準が規定されていないこと。(5) 文字集合の拡張機能や多言語化に対応できないことが挙げられる。また、JIS X0208 は、利用実態の変化に対応する処置として5 年ごとに見直しが行われるが、1983年の改正で同一符号上での字形の入れ替えや字形の変更を行ったため、旧版で作成されたデータからの移行を困難なものにした [斎藤:1994]。

本稿では、これらの問題点を解決でき、長期のデータ保存や古典・漢籍や大規模の漢字字典を電子化する機能を構造化した4 バイトコードを提案する。4 バイトコードに対する要求は、(1) 古典・漢籍および大規模の漢字字典を符号化できること、(2) 東アジア漢字使用国（中国・台湾・日本・韓国）で使われている漢字を統一的に符号化できること、(3) 長期のデータ保存に対応でき、規範と変化の二つに対応できること、(4) 既存の2 バイト系漢字符号を併用できること、(5) 装置に依存しない理論符号を設定でき、属性情報も字形と同じ枠組みで符号化できること、(6) 規範性と利用者が任意に規定できる文字集合と漢字符号を規定できること、の6 点である。

新しい漢字符号に導入する機能の決定は、諸橋轍次編「大漢和辞典」（以下大漢和）の構成方法とJIS X0208 の問題点を改善するために必要な事柄および国立国語研究所で使用した表外字とこれまで行った各種の日本語処理の経験を基本にした。

2．4バイトコードに要求する基本機能

4 バイトコードは、既存の2 バイト系漢字符号を管理する理論符号としての位置付け、規範と利用者や専門分野別の文字集合を規定する枠組みを実現する。文字の追加や利用上の変化への対応は、漢字符号に枝番号を設け、文字集合の拡張には4 バイトコードを使った。4 バイトコードの符号化領域は、既存の漢字符号との併用をはかるためG3領域を使用した（図1)。大漢和の検字番号から4 バイトコードの各桁を16進数'21'から'7E'に調整する処理は、式1 と式2 を使い、94進数と16進数変換を行った。新しい漢字符号に要求する機能は、図2 に示す3 種類の構造にまとめた。各構造と機能は、以下の通りである。

(1) 各桁の2 の8 ビット目が'01'である2 バイトコードを2 個結合し内部符号に対応させる構造（図2-1)、(2) 既存の2 バイト系漢字符号を1 バイトの識別符号で統轄する構造（図2-2)、(3) 漢字辞書やコードブックに付けた10進数5 桁の検字番号を整数部3 バイトで表す構造である（図2-3)。4 バイトコードの小数部には、異体字と対応する各国語（中国・台湾・日本・韓国）を配当し、異体字に対する見出しを整数部においた（図3)。4 バ

イトコードで表現できる文字集合は、整数部830,584 字と小数部94字である。また、 G3 領域と他の領域の識別は、各バイトの2 の8 ビットを使った。

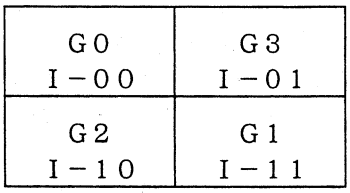
整数部＝ＨＥＸ（検字番号ＭＯＤ９４）＋　'重み16進数21'　　（式1）

小数部＝ＨＥＸ（小数部挿入位置番号）＋　'重み16進数21'　　（式2）

HEX:10 進数値を16進数に変換する関数

MOD:10 進数表現された検字番号の剰余を求める関数

検字番号:10 進数5 桁で初期値が'0' の数値



表外字コード領域

G３

内字コード領域

G０，Ｇ１，Ｇ２

図1　2 バイトコードと4 バイトコードの符号化領域

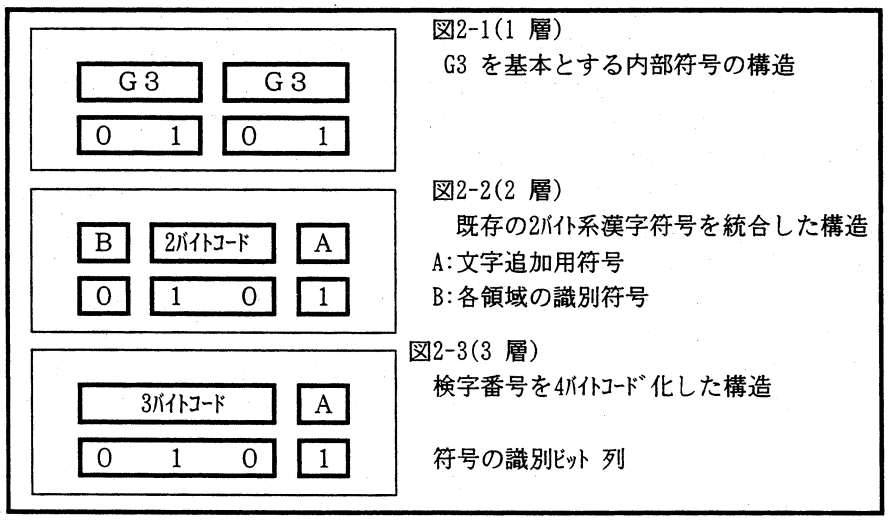


図2-1(1 層)
G3 を基本とする内部符号の構造

図2-2(2 層)
既存の2ﾊﾞｲﾄ系漢字符号を統合した構造
A:文字追加用符号
B:各領域の識別符号

図2-3(3 層)
検字番号を4ﾊﾞｲﾄｺｰﾄﾞ化した構造

符号の識別ﾋﾞｯﾄ 列

図2　4 バイトコードの構造

３．異体字の１字体１符号表現

4 バイトコードの1 字体1 符号による表現は、4 バイトコードを整数部3 バイトと小数部1 バイトに細分し、整数部に見出しに相当する字体（実際は特定の文字で代用する）と、小数部に異体字（異なる形をもつが同じ意味をもつ漢字）や中国・台湾・日本・韓国で使用されている漢字を配当する。この方法で細分化された4 バイトコードは、基準となる漢字字典との接続面の明示と、2 バイトコードを統括管理する理論符号としての性格をもつ。この符合化法は、見出しと4 バイトコード整数部を固定した状態で文字の追加や変更情報の履歴を小数部に累積させることができる。また、整数部に配当した見出しは、共通字形を使った情報交換や、小数部に設けた機能で旧規格の漢字符号で作成したデータの継続使用と文献や資料などを正確に符号化する処理を実現できる。

4 バイトコード小数部への文字配当は、部首順に配列した最大94個の文字集合を'1' から'94'にあてる。図3 は、ISO/IEC 10646-1 で規定した「剣」について符号配当を行なった例である。小数部への文字配当は、文字の追加が行なわれることを想定し、間隔をあけて登録した。

字形	剣		劍	劍	劍	劍	劍	劍	劍

（3ﾊﾞｲﾄ 見出し部分）　（小数部1ﾊﾞｲﾄ 異体字と各国語登録部分）

図3 小数部への文字配当例定

4．　文字集合に対する規範性と利用者規定の方法

図2-1 と図2-2 で示す二つの構造は、4 バイトコードに重ね合わせることができるため、同一構造に8,836 字単位の符号領域と漢字1 文字を表現する符号を二重に規定できる。この二重規定の方法は、図2-2 と図2-3 の構造を文字集合に対する全体と部分に対応させることによって、規範となる文字集合と既存の2 バイト系漢字符号や後述する利用者規定の文字集合を4 バイトコードの枠組みのなかで明確に規定することができる。そのほか、二重規定は、4 バイトコードによる2 バイトコードの管理と2 バイトコードと4 バイトコードの混在使用、4 バイトコードを2 バイト系漢字符号に対する理論符号としての利用を可能にする。

5．日本語と中国語処理と属性情報の符号化実験

多言語化には、補助コードセットやISO/IEC 10646-1 が採用した同一符号域で表現する方法がある。本稿で提案する1 字体1 符号化法も、一つの符号で多言語を表現することを基本としている。しかし、拡張符号を使った多言語化は、識別符号で国名を判別できるのに対して、統一符号系では、字形から、どの国の文献・資料を符号化したかを知ることができない。本実験は、中国語と日本語入力実験を通して構造化4 バイトコードに対する基本機能の確認と、属性情報の符号化と再現処理から漢字符号に属性情報を付加することの有効性を確認する。プログラムの実験環境は、1 台のワークステーション（日本電気製:UP4800/610,128MB,SPECrate-int92:4,165,SPECrate-fp92:5,035）に日本語と中国語を混在入力するプログラム（サーバ・プログラム「4Bserver」とクライアント・プログラム「4Btext」）を実装した。行った実験は、以下の3 項目である。

(1)　4バイトコードを使った中国語と日本語入力モデルによる多言語表現機能の確認。
(2) 中国語と日本語表現した2 バイトコードと4 バイトコードから、辞書に記録されている属性情報を引用し、データを再現する処理の実験。
(3) 属性情報付き4 バイトコードを使ったクライアント・サーバ間のデータ伝送実験。

5.1　キーボードから読みを入力する操作

図4 の変換条件指定部は、'chu' を共通情報とする漢字の字形一覧を示したものである。これらの漢字入力処理は、変換条件指定部で、入力する文字の種類「音読み」、「訓読み」、「ピンイン」を指定し、キーボードから読みを入力する。入力したスペースまでを一変換単位とし、ローマ字を仮名文字に変換する。「音読み」または「訓読み」が指定された場合には、国名情報'J' を、「ピンイン」の場合に'C' を与える。次に、変換条件指定部で選択した漢字の読み、を符号化するため、漢字辞書に記録されている読みの位置情報を属性情報として記録し、属性情報付きの4 バイトコードを生成する。

なお、本実験では、4 バイトコードの第4 バイト目（小数部）の符号領域94に属性情報用として64個、国名と終端符号用に30個分をあてた。また、属性情報は、辞書項目の位置を示すリンク情報として使用した。符号化の対象とした属性情報は、次の4 種である。

(1) 読み情報：4 バイトコードの整数部を大漢和の検字番号に再変換し、辞書情報を引用するための情報として使用する。辞書に登録されている読みが複数個ある場合には、'0' から' 255'個の登録を許すものとする。例：楚:[ｿ,ｲﾊﾞﾗ,,ｼﾓﾄ,ｽﾜｴ,CHU] の順序で

登録されている場合［ソ］は'0'を指定する（開始番号は'0'）。また、符号化する属性情報は、辞書に記載されている情報の位置を記録する。

(2) 国名情報：その漢字がどの国の言語として使用されたかを符号化する。本実験では、符号化の対象になる国を、中国・台湾・日本・韓国のなかから日本と中国を対象とした。日本語と中国語の国名表示は、'J'と'C'とし、内部符号を16進数'0XFE'と'0XFD'にあてた。

(3) 部首情報：部首に該当する漢字を拡張UNIXコードで表現する。利用者規定の漢字符号には、疑似的に拡張UNIXコードで表現した。例：部首「母」を表す拡張UNIXコード、'DDD6'で表す。

(4) 総画情報：数値と結合情報ともに画数に対応する数値で表す。例：10画の場合は、16進数'10'を符号化した'0X0A'で表現する。

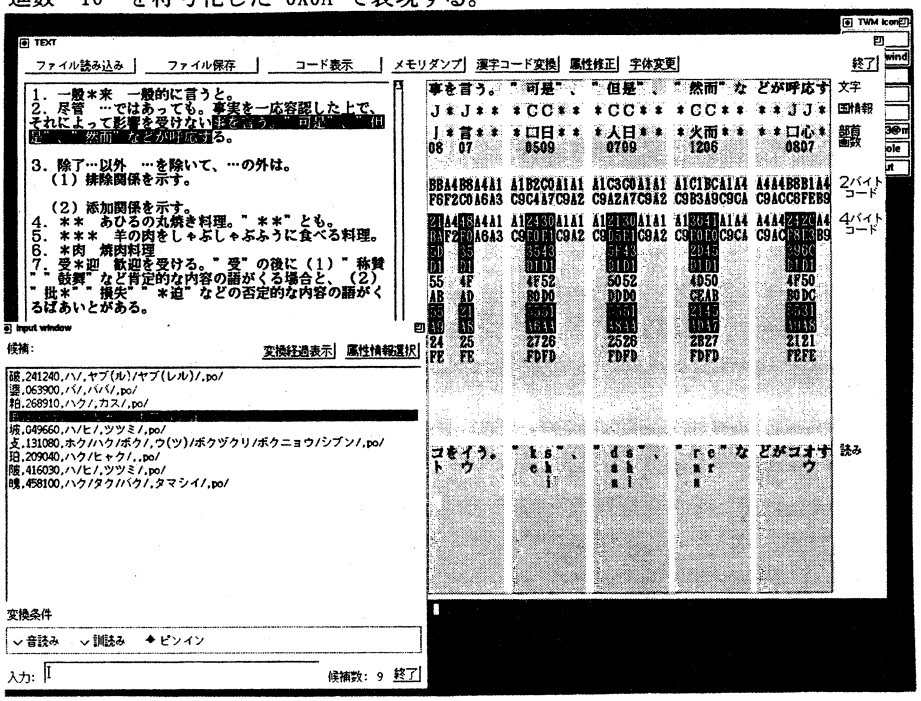


図4　4バイトコードで表現した属性情報
と日本語・中国語の表示例

6．おわりに

本稿では、日本語と中国語を混在入力し、構造化4 バイトコードで多言語を表現する機能の確認と、4 バイトコード対応の文字集合を全体と部分に分ける方法がサーバ・クライアント環境での実行に適合できることを確認した。そのほか、多言語処理の出力形態の一貫としてその漢字が使われた資料の国名や辞書で規定した、読み、画数、部首などの属性情報を付加する実験を行い、同一符号系で多言語処理を行うための支援情報として有効な方法であることを示した。属性情報を特定の辞書から引用するためのリンク情報として規定する方法は、符号の圧縮効果と辞書の種類によって使用対象を選択できる効果が期待できる。属性情報の符号化は、本論文で始めて提案されたものであり、伝統的な漢字の特性を説明方法として認められている「音」、「義」を符号する方法に道を開くものとなる。

参考文献

［斎藤:1994］大漢和辞典の検字番号に基づく構造化４バイトコードの提案，情報処理論文誌，Vol. 35, No. 6, pp. 1119-1126(1994).