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

斎藤 秀紀†

JIS X 0208には，符号間への文字の挿入機能の欠如や中国・韓国語への拡張機能の問題がある．本研究では，諸橋轍次編「大漢和辞典」の検字番号に基づく構造化4バイトコードの符号化法と活用法を述べた．最初に，4バイトコードの構造を(1)大漢和辞典の検字番号を16進化94進数変換した整数部3バイトと小数部1バイト符号(2)既存の2バイトコードを統合した符号(3)2の8ビット目が'01'である符号を2個結合した3種にまとめ，これらが画数・読み・部首情報の漢和辞書による基準化，多国語の表現，一字体一符号化を実現できることを示した．次に，符号間への文字の追加機能をつかった一字体一符号化法が，文字集合と漢字符号の固定や文字集合間の互換性を維持した文字集合の分割に利用できることを述べた．最後に，4バイトコードの機能をワークステーション (EWS-4800) とパーソナルコンピュータ (PC-9800) 間のデータ伝送を通して検証した．その結果，4バイトコードは，JIS X 0208-1978の見直しで生じたデータ間の互換性を崩す問題を解決でき，異機種間の共通符号にも使用できる見通しをえた．

A Proposal for a Structural 4-byte Code Based on the *Daikanwa-Dictionary* Index Number

HIDENORI SAITO†

In place of JIS X 0208, I propose a structural 4-byte code containing three types of codes: (1) codes consisting of a 3-byte integer part and a 1-byte decimal part, derived by converting the Daikanwa Dictionary index number into 94-ary code and then into hexadecimal code; (2) codes integrating existing 2-byte code, and (3) codes combining two codes whose 8th bit of 2 is '01'. The 4-byte code makes it possible to standardize the number of strokes, readings, and radicals based on the Daikanwa Dictionary, to include characters used in various countries, and to realize one-character-one-code correspondence. This allows insertion of additional characters between codes, stabilizes relationships between character groups and kanji codes, and separates character groups while maintaining mutual exchangeability of characters. To test the code, deta were transmitted between a workstation and a personal computer, successfully maintaining mutual exchangeability of the data.

1. はじめに

コンピュータによる漢字処理は，古典文献・漢籍や図書情報の処理，特許情報処理，日本語教育と教材開発，外国人のための辞書の作成，多国間で情報交換を行う電子メールなど多様化している．また，中国や韓国など漢字使用国の間で交換する技術文献や商用書簡を一元的に表現し管理できる漢字符号の開発が必要になっている．

しかし，現在使用されている情報交換用漢字符号JIS X 0208は，中国 (GB 2312)，韓国 (KSC 5601) を含む多国語への拡張機能や大規模の漢和辞書を電子化できる符号化領域の確保，表外字に対する符号化法，符号間への文字挿入機能，コードブックを基準にした文字や属性情報の規定，長期のデータ保存を目的にした文字管理機能などが欠けている．そのほか，JIS X 0208には，二つの水準と文字配列基準の設定，第一水準の文字の割り当てに「政令で規定する」とした文字集合の配当に係わる問題がある[1)~3)]．

文字集合の拡張は，漢字処理の多様化を目的に日本工業規格 (JIS) で補助漢字 JIS X 0212-1990 (漢字5,801字，非漢字266字) を定めた[4)]．また，多国語化の提案は，和田[5)]，UNIX System V[6)]，国際標準化機構 (ISO: International Organization For Standardization)[7),8)] が行っている．しかし，いずれの提案も連続した2バイトコードに文字集合を対応させているため JIS X 0208 と同様の問題をもっている．

本稿では，現行の情報交換用漢字符号と文字集合の拡張および属性情報 (画数・部首・読みなど) を漢和辞書で規定することを目的に，整数部3バイトと小数

† 国立国語研究所情報資料研究部電子計算機システム開発研究室
Section for the Development of Software for Language Data, Department of Data Orientation, The National Language Research Institute

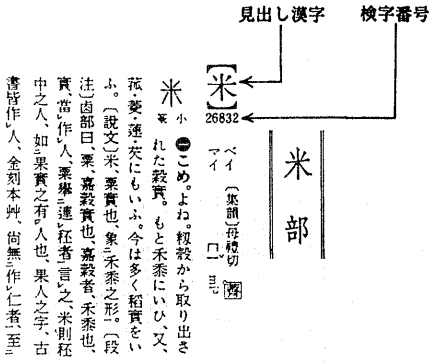

図 1 大漢和辞典の見出しと検字番号(文献10) から引用)
Fig. 1 An example of a character entry and index number in the *Daikanwa Dictionary*.

部1バイトを基本に仮想小数点をもつ4バイトコードを提案する．4バイトコードの利用は，諸橋轍次編「大漢和辞典」(以下大漢和) の検字番号 (大漢和収載の文字につけた番号：**図1**) に基づく情報処理用の漢字符号の設定，属性情報の規定，中国・韓国語を含む多国語の取り込み，大規模の漢和辞書の電子媒体化と文字発生装置への応用を想定した．4バイトコードの構造は，漢字符号と利用面の接続点を明確にするため次の3種にまとめた．

(1) 大漢和と接続するための構造．
(2) 既存の2バイトコードを包含するための構造．
(3) コンピュータの内部符号を表現するための構造．

また，4バイトコードの小数部をつかった一字体一符号化の方法は，JIS X 0208 の二つの水準と，二つの文字配列，JIS X 0208-1978 と 1983 年版の間で生じた互換性の問題を解決できることを示し，文字集合に対して共通指標となる文字概念 (字体) を見出しとして行った字形のグループ化，交換性をもった文字集合の全体と部分の設定，部分集合の索引化と属性情報による文字配列に利用できることを述べた．さらに，小数部を使用した文字の追加機能は，符号間へ文字を挿入した場合も基本配列を維持できることからデータ保存用符号に利用できることを示した．

そのほか，10進数で表現された大漢和の検字番号を 94 進数と 16 進数に変換する処理 (以下 16 進化 94 進数変換) は，4バイトコードを JIS X 0201 に従う 16 進数 21 から 7E へ基準化する方法と，図3の構造を数量的にモデル化する方法に使用した[9),10)]．

実験は，2バイトと4バイトコードを混在させた日本語用例 (KWIC) をテストデータとし，16 進化 94 進数変換アルゴリズムの確認，各バイトの2の8ビット目をつかった2バイト・4バイトコードの識別，通信回線を使用する問題点の洗いだしを行った．また，シフト JIS コード，拡張 UNIX コード，JIPS コード (日本電気(株)が規定した漢字符号) などの端末の漢字符号と4バイトコード間の符号変換処理を削減するためには，G1，G2 領域の符号を JIS X 0208 に変換し図 3-II 構造にまとめる処理が有効であることを確認した．

2. 現在使用されている情報交換用漢字符号

「システムソフトウェアの標準化に関する調査研究」によると，現段階における漢字コードの拡張法は，今後の拡張字数が 8,836 字以内では第二セットをつくるか空き領域を使用する．8,836 字を超える場合は，第二・第四セットを設定，符号の長さは3バイトないし4バイトとする案が検討されている．同様に中国では「GB 2312-1980 (基本集)」に対して第二補助集および第四補助集の制定を審議中であり，第二補助集 7,237 字，第四補助集 7,039 字の 14,276 字を印刷字母を含め中国の標準として使用する．選択は「漢語大辞典」の 58,000 字から行う．配列は，中国文字改革委員会制定 (1983) の漢字部首をとる．台湾の CCCII は，7ビット3列をつかい 28,737 字を，米国は，REACC-83 (7ビット3バイト) で漢字を 13,650 字，かな文字 174 字，ハングル文字 2,026 字を符号化している[11)]．

北京図書館では，GB 2312-1980 と独立に北京中文符号を採用し，簡体字・繁体字 32,000 字を符号化している．多国語の表現は，和田が制御符号を含む JIS X 0208 と GB 2312 を2バイトコード化する方法を提案している．また，UNIX System V リリース 3.1 では，同様の目的で拡張 UNIX コードを提案している．日本語処理は，JIS X 0208 の文字集合をセット1に，表外字はセット3にあて 17,672 字を符号化する．ISO では，各国語の漢字符号を併用するため既存の情報処理用の文字集合を包含する4バイトコードを計画中である．また，第2版ではユニコード (Unicode) との共用を考慮し，'00' から 'FF' までの2バイトを使用する案が進められている．

東京外国語大学は，漢和辞書の検字番号を16進4バイト化し，そのなかで中国漢字を表現している[12)]．また，国立国語研究所では，大漢和の検字番号5桁を部首と部首からの相対位置にわけ，数値化した漢字2文

字で表外字を表している．数値化は，出現頻度の高い漢字を1から450の数字にあて，部首「一：001：計」から「龠：271：合」と，相対位置「001：計」から「450：力」を表現している[13]．

これらの文字集合は，JIS X 0201 やこれに準ずる符号を基本に，切り替え符号で拡張や多国語化をはかっている．しかし，いずれも文字・属性情報・漢字符号を総合的に管理する方法やこれらの情報を漢和辞書などで独立に規定する考え方が欠けている．

3. 情報交換用漢字符号の問題点

3.1 JIS X 0208 の問題点

JIS X 0208 は，日本における情報交換用漢字符号として利用されているが実用面でいくつかの問題がある．以下は事例である．

（1）文字集合は，使用頻度の高い一群を第一水準に，そのほかを第二水準に配当した．また，第一水準を読み順，第二水準を部首順とする二つの配列基準をとり，第一水準への文字配当に「政令で規定するもの」とする規則をもうけた．

JIS X 0208 の文字集合を二つの水準にわけ読みと部首に配列したことは，文字の使用効率と検字の容易性を向上させることにあった．しかし，第一水準の文字配当に「使用頻度の高い漢字」と「政令で規定する」とする規則の併用は，常用漢字表の改訂で水準間の文字の入れ替えを避けられないものにした．

見直し作業による JIS X 0208-1983年版は，294 字の字形変更と 84 区点以降へ2文字の追加があった．野村の調査では（1）常用漢字に関するもの14字，（2）人名用漢字に関するもの 16 字，（3）通用字体の準用に関するもの151字，（4）異体字の位置に関するもの44字，（5）部分字形の統一に関するもの47字，（6）独自の理由に関するもの 22 字である（第一水準206字，第二水準88字）[14]．294字の字形の変更は，JIS X 0208-1978 と1983年版で出力字形と文字配列が異なる問題が生じた．

（2）文字の追加機能と属性情報の規定に関する問題

JIS X 0208 および JIS X 0212 に対する文字の追加方法には，利用者にゆるされている登録領域をつかう方法（JIS X 0208，JIS X 0212 の空き領域 2,600 字程度）と，JIS X 0202(35,344 字）を使用する方法がある．しかし，JIS X 0208 では，略字や俗字などの字形を整理し（李の調査によると JIS X 0208 の異体字の種類は 555 種 1,159 字である[15]），JIS X 0202 と併せた空き領域を使用した場合も大漢和収載の50,311字や漢籍・古典文献を符号化する機能はない．

また，符号間への文字挿入機能の欠如は，利用者とシステムによる文字の二重登録や重複する文字の移動を余儀なくさせた．1983年の4文字の登録を含め84区点への6文字の移動は，互換性を崩す要因となった．配列順序は，堯：5272，槇：15310，遙：39035，瑤：21149，凜：1717，熙：19192 である（数字は大漢和の検字番号）．JIS X 0208-1990年版における情報交換用の文字集合は，第一・第二水準・84 区点の6文字・利用者登録・補助漢字の5ブロックにわけられた．

一方，漢字処理における文字配列は，部首・画数・読みを使用することが多い．しかし，JIS X 0208 は，属性情報の規定がなく，部首も「｜，疒，内，走」が欠落している．読み順を使用した場合は，利用者と作成者の間で読みの選択が異なることがある．漢字符号では，配列が読みと部首順になり，検字，データ配列，印刷物に対する人手作業が困難になる．これらは，JIS 0208 の規格の改訂や文字の追加に対して配列順序の維持を考慮しなかったことによる．

3.2 DIS 10646 の問題点

DIS 10646（Draft International Standard：ISO で決定される規格案）の第2版は，ユニコードとの結合をはかり符号化領域を2進数の '00' から 'FF' とする2バイトコードを提案している．しかし，DIS 10646 は，現在各国で使用している情報交換用文字集合を統合したため，既存の漢字符号の問題もそのまま持ち込むことが予想される．特に，連続した 16 ビットコードに文字集合を割り当てたことは，字形の変更や追加文字への対応がよわく，規格の改訂にともなうデータ保守と新旧データ間の互換性を維持するうえで問題が残る．また，DIS 10646 は，16 ビットコードを使用しているため JIS X0201 規格に従う漢字符号からの移行費用が膨大になる可能性がつよい．

そのほか，DIS 10646 の問題には，表外字管理と漢字の属性情報の規定，コードブックによる情報管理法に関するものがある．以下は，予想される問題である．

（1）2進数符号を基本とするため JIS X 0201 に従う複数バイトを使用した情報交換用漢字符号との間に符号交換が必要になる．

（2）符号間に追加機能がないため文字の追加が文字配列の基本性を崩すおそれがある．

(3) 国家規格やISO規格の変更に対する互換性の維持の方法が明確になっていない.
(4) データ保存・運用管理に必要な漢字の属性情報が規定されていない.
(5) 表外字の登録領域が少ない.
(6) 各国で使用している標準字形との相違が多国語の混在処理に心理的な負担をかける可能性がある.
(7) DIS 10646 が示す2バイトコードによる多国語表現は，4バイトコードに配当される文字集合との関係が明確になっていない.
(8) JIS X 0202 で規定された制御文字が DIS 10646 の符号化領域と重複する.
(9) 辞書の電子媒体による出版は，辞書をもとにした情報交換用漢字符号への発展が予想される.

4. 4バイトコードの作成方法

4.1 漢字符号の構造化の考え方と大漢和の結合方法

JIS X 0208 と DIS 10646 に共通する問題は，符号間への文字の挿入機能や文字集合に対する属性情報の規定，コードブックの使用に対する考え方がないことである．この問題を解決するためには，漢字符号・文字・属性情報をコードブック上に固定しデータの作成者と利用者の双方に明示することが必要になる．コードブックは，属性情報・漢字符号・文字集合の規定とともにデータを長期間保存・管理する方法に使用できる．また，JIS X 0208 の文字集合を二つの水準にわけたことは，文字選定の段階で利用目的に応じた階層化の考え方があったことが伺える．この条件は，4バイトコードの特性を構造として示し，利用者にインタフェースの形で明示する手段として重要な機能となる.

このような，要求事項は，漢和辞書のなかに見ることができる．辞書の索引配列は，部首・読み・画数を使用し，本体は部首順に配列している．JIS X 0208が読みと部首による配列を採用したことは，漢和辞書の本体が部首配列を，索引が複数配列を使用したのと同じといえる．見出し，属性情報，検字番号，索引，本体の文字配列を総合的に規定するためには，情報交換用漢字符号と文字集合の配列を漢和辞書と一致させることが必要になる．以上のことを実現するためには，JIS X 0208で確定した機能と新しく導入すべき機能を構造として明示し，4バイトコードに反映させることが重要である.

4バイトコードが JIS X 0208 から引き継ぐ機能は，二つの文字配列と二つの水準の設定，第一水準の文字配当にある制限文字集合による情報交換の考え方である．JIS X 0208 で採用した同一規格内での「読み」と「部首」による文字配列法は，同一字形で異なる読みをもつ中国語や韓国語の混在処理，特殊な読みに対する複数の索引設定に拡張できる．また，第一水準の制限文字集合の考え方は，専門領域別に文字集合を規定する場合や，利用者による任意の文字集合の抽出，文字集合を制限した常用漢字などと共通する.

新しく導入すべき機能は，コードブックによる字形や属性情報の規定，検字番号の情報交換用漢字符号への利用，符号間への文字の挿入機能と一字体一符号化の採用，文字集合の分割利用と分割した文字集合に対する利用者配列の導入，多国語の混在および2バイト系情報交換用漢字符号からの移行性が課題になる．以下は，4バイトコードの設計条件である.

(1) 4バイトコードは，単独の使用と既存の2バイトコードと併用できること.
(2) 2バイト・4バイトコード間の識別に各バイトの2の8ビット目のビットを使用できること.
(3) 4バイトコードは，JIS X 0201 に従うこと.
(4) 大漢和と結合するための符号と内部符号は，検字番号を基本とすること.
(5) 4バイトコードは，符号間に追加機能をもつこと.
(6) 中国・韓国語を統一表現できること.

そのほか，漢和辞書のコードブック化には，長期のデータ保存に対する支援手段としての利用がある．データとなる符号化された文字は，入力と再現時点で一致している．この前提のもとでデータは，電子媒体に記録される．しかし，規格改訂による字形の変更は，情報交換用に規定された字形と資料の関係を崩し，データを再現するための基準が変化する．過去のデータ作成環境を再現するためには，資料上の字形と漢字符号に対応する字形を固定し，コードブックによる情報管理と電子媒体化が必要になる．電子媒体化されたコードブックは，コンピュータの文字発生装置との併用によってデータを作成した環境を再現できる.

なお，コードブックに利用できる最大の漢和辞書に，諸橋轍次編「大漢和辞典」がある．大漢和のコードブック化には，和製漢字の補填や大漢和にない漢字を定義する必要があるが，情報処理用に使用できる文字集合の大枠を確定させ，属性情報の規定や JIS に

おける字形の逐次追加，自由領域への利用者登録を最少限におさえることができる．大漢和には，補遺版および各親字間に追加する文字の位置を示す追加用記号「′」「″」欠番がある．これらは，編集上生じたと思われるが，総字数の 50,311 字の内容は本体 48,902 字，補遺版 1,062 字，追加 504 字，欠番の総数は 157 字である．JIS C 6226 との比較では，大漢和・新字源・大字典のどれにも収録されていない漢字が 94 字ある．

4.2 4バイトコードの一字体一符号化の方法*

文字集合の一字体一符号化の方法は，変動が少ないと考えられる字体を4バイトコードの見出しにあて，字形を小数部に配当する．整数部に見出しをあてることは，漢字符号と文字集合の大枠を規定できる．また，小数部は，配当した種々の字形を符号化の基準として使用でき，字形変更や追加処理を局所化できる．字形の変更の局所化は，漢字符号と字体の関係を一対一に固定でき，分割した文字集合の部分と全体，部分間で漢字符号の交換性を維持できる．部分となる文字集合は，本体と独立に配列した場合も符号と文字の関係は不変である．これによって，漢字符号は文字表現としての機能を明確にし，文字配列は JIS X 0208 の暗黙の方法から属性情報で明確に規定できる．

一字体一符号化した文字集合に対する文字の追加は，小数部の空き番号への登録を基本とし，字形の変更は符号の切り替えで対応する．改訂前の字形を使用する場合は，変更の必要はない．この方法は，小数部に文字を登録するさい重複がおきにくく長期のデータ保存用符号として重要な機能となる．

図 2 は，漢字「剣」の常用漢字表でかかげる通用字体(常用漢字表および人名用漢字別表でかかげる字体)を指標とし，小数部「1, 3, 5, 7, 9, 11」に異なる字形を総画順に配当した例である．整数部の6桁の数字は，大漢和の検字番号 '2076' を基本字形とし 16 進化 94 進数変換した値である．見出し漢字につけた括弧内の数字は，JIS の区点番号，大漢和の検字番号，4バイトコードである．検字番号の '0000' は，大漢和にない字形である．小数部に中国・韓国語を登録する場合は，各国で許容できる見出しを選ぶ必要がある．なお，大漢和をコードブックに使用する場合は，辞書も一字体一検字番号化する必要がある．

剣(23-85,2076)213728.21 劒(78-63,0000)213728.23
剣(49-91,0000)213728.25 剣(49-88,2228)213728.27
劍(49-90,2243)213728.29 劍(49-89,2245)213728.2B

図 2 4バイトコードへの文字配当例
Fig. 2 Examples of characters and their 4-byte codes.

4.3 検字番号を4バイトコードへ変換する方法

4バイトコードは，利用面のインタフェースと漢字符号の特性を規定するため，**図 3** で示した3種の構造にまとめた．Ⅰは，大漢和と結合のための符号を構造として表現したものであり，Ⅱは既存の2バイトコードとの併用をはかるための構造である．Ⅲは，ⅠとⅡの構造を重ね合わせ内部符号表現したものである．

Ⅰは，検字番号を 16 進化 94 進数変換し，整数部を3バイトに変換した構造である．4バイトコードの整数部で表現できる字数は 830,584 字である．10 進表現した5桁の検字番号を 16 進化 94 進数に変換する手続きは，検字番号の初期値を '1' から '0' に調整する（例えば検字番号 '830584' は '830583' となる）．次に，検字番号の下位2桁を JIS X 0208 の符号化範囲である 7E と 21 の差 94 を「法」とする計算を行い剰余を 16 進数に変換する．結果には，JIS X 208 で規定した符号の最小値に調整するため重み 16 進数 '21' を加える（式(1)）．この計算を各2桁単位 ('05', '83') と繰り返し3バイトコードをつくる．

小数部は，追加個数の取りうる範囲を最大 94 に制限したため，1から94の値を16進に変換値し重み 21 を加えた値となる（式(2)）．図3で示した記号「A」は，4バイトコードの小数部である．

整数部＝HEX(検字番号 MOD 94)＋重み 21　　(1)

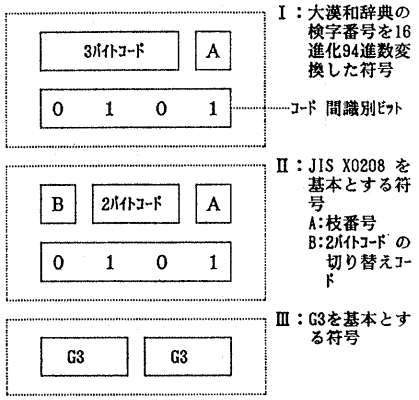

図 3 3種の4バイトコード
Fig. 3 The three types of 4-byte code.

* 本節の論旨は，本論文の採録時までに別稿「1字体に1符号を対応させる漢字符号化の方法」として『計量国語学』に受理されている（1994年6月刊行予定）．

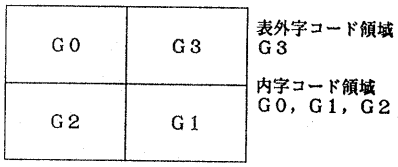

図 4 2バイトコードと4バイトコード領域
Fig. 4 The 2- and 4-byte code regions.

小数部＝HEX(小数部番号)＋重み 21 (2)

HEX：16進数変換用関数，MOD：剰余関数

Ⅱは，既存の2バイトコードを4バイトコードに組み込む構造で基本面2バイトコード，切り替え用コード1バイト，小数部1バイトで構成する．Ⅱの整数部は，8,836 字を 94 枚集めた3次元表現となる．G1，G2 領域の文字集合を構造Ⅱへ統合する操作は，各領域の2の8ビット目を '00' に調整し JIS X 0208，GB 2312，KSC 5601 を，先頭1バイトで識別する．

Ⅲは，2の8ビットが '01' である G3 領域の2バイトコードを2個結合し，ⅠとⅡをコンピュータの内部符号にまとめる構造である．G3 領域を4バイトコードの要素とする方法は，4バイトコード構造の基本をつくり，2バイト・4バイトコードの識別を各バイトの2の8ビット目で処理する方法を確立できる(各バイトの2の8ビット目は，G0，G1，G2 領域が '00'，'10'，'11'，4バイトコードは '0101' である)．この方法は，既存の2バイトコードとの併用と，4バイトコードへの移行を緩やかなものにする(**図4**)．

ここで，2バイトコードから4バイトコードを引用する場合は，出現頻度の低い文字が対象になる可能性が高くデータ量の増加は少ないと考えてよい（国立国語研究所における雑誌・新聞の漢字調査では「現代雑誌九十種の用語用字―漢字表」昭和31年1年分：3,328字「現代新聞の漢字」昭和41年1年分：3,213字である[16),17)]．ただし，既存の2バイトコードと4バイトコードは，文字集合を独立に符号化しているため，4バイトコードを表外字として位置づけ運用上で重複をさけることが必要である．

5. 実験システムの概要

実験では，2バイト・4バイトコードで表現した日本語用例 (KWIC) を回線を通すことによって(1)大漢和の検字番号を16進化94進数変換するアルゴリズムの検証，(2)各バイトの2の8ビット目をつかった2バイトと4バイトコードの識別，(3)各端末の漢字符号（シフト JIS コード，拡張 UNIX コード，JIPSコード）に対し中間符号 JIS X 0208 をつかった符号変換処理の削減，(4)既存の2バイトコードを構造Ⅱに集約する効果を確認した．

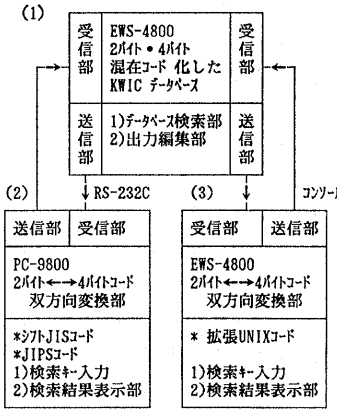

図 5 実験プログラムの概要
Fig. 5 Outline of the test program.

実験の方法は，2バイトと4バイトコードで表現した日本語用例 1.2 万件をワークステーション (EWS-4800) におき，漢字1文字をキーに用例を検索する(**図5**)．4バイトコードは，変換処理を少なくするため各端末の内部符号を JIS X 0208 に変換したのち4バイトコード化した．該当する用例は，2バイトと4バイトコード混在のまま端末に返送し画面表示する．実験用のデータは，漢字パターンの作成をさけるため，JIS X 0208 で規定した漢字を4バイトコードにあて，英数字・かな文字を JIS X 0208 で表現した．各用例の長さは，前部分を 60 バイト，後部分を 116

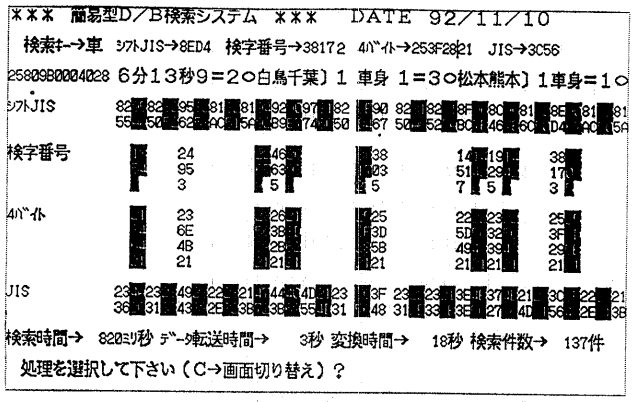

図 6 符号の表示画面
Fig. 6 Display showing the codes.

バイトとし，満たない場合は 16 進数 '00' で埋めた．用例の見出しは，漢字2文字をキーにあて4バイトコードで表現した．実験装置は，EWS-4800 をサーバと端末に使用し，パーソナルコンピュータ (PC-9800) を端末と日本語処理端末 (N-6500) のシミュレート機能に併用させた．

図6は，用例とキー '車' の漢字符号であるシフト JIS コード (8ED4)，大漢和の検字番号 (38172)，4バイトコード (253F2821)，JIS X 0208(3C56)，用例を検索するのに要した時間，2バイトと4バイトコード間の変換時間，検索したデータ件数を表示したものである．4バイトコードの空白は，かな文字部分である．4バイトコードの16進数値は，'253F28' が3バイト整数部，'21' が小数部である．用例1.2万件から137件を検索するのに要した時間は，PC-9800 で 820 ミリ秒，16 用例をサーバから端末に転送する時間3秒，2バイトと4バイトコード間の符号変換時間は 18 秒であった．EWS-4800 のコード変換時間は，PC-9800 と同一条件で 10 ミリ秒である．

6. おわりに

本稿では，漢字符号の以下の点を検討した．

(1) JIS X 0208 および漢字使用国における漢字符号の利用状態を紹介し，JIS X 0208 における配列規準の二重性，利用者登録の定義域のつかい方，通用字体と異体字に対する符号化の方法などが規格改訂で水準間の入れ替えをおこし，互換性を崩す原因となったことを指摘した．また，属性情報や字形を規定する基準書がないことが実務面で問題になることを述べた．

DIS 10646 においても JIS X 0208 と同様の問題があり，実用化には既存の2バイト系情報交換用漢字符号からの移行が問題になることを述べた．

(2) 4バイトコードを提案するにあたって，JIS X 0208 で確定した機能と付け加えるべき機能を構造化する方法が，漢字符号の特性と利用面とのインタフェースを明確にできることを示し，大漢和の検字番号を 16 進化 94 進数に変換する方法が4バイトコードの構造化と大漢和のコードブックとしての利用，図3の構造を数量的に処理できることを述べた．また，4バイトコードの小数部をつかった一字体一符号化の方法は，文字集合全体の配列規定と目的に応じた文字集合の抽出，全体と部分間で独立した配列をとる場合の符号の一貫性を維持できることを説明した．

そのほか，4バイトコードの小数部は，異なる字形リストや符号化領域が 21 から 7E の範囲にある中国，韓国および JIS X 0208 に準拠した漢字符号を統合できることから，各国語の混在が予想される多国間言語辞書や日本語教育のための教材の作成，異体字の研究をともなう総索引の作成，古典文献の研究に有効な手段となることを述べた．

(3) (1)と(2)を検証するために実験を行った．実験では，大漢和の検字番号を 16 進化 94 進数変換する手続きと構造化の方法，2バイトと4バイトコードの混在と識別処理，符号化領域の G0 から G3 領域を JIS X 0208 に統合させ，既存の2バイトコードを領域単位で切り替える方法，2の8ビット目が '01' である領域を4バイトコード化する方法を機能面から確認した．

これによって，JIS X 0208 および4バイトコードの混在が現行の2バイトコードを小規模用漢字符号に，4バイトコードを大規模用の文字集合として使用できる見通しをえた．さらに，4バイトコードは，メタコード化することによって，JIS X 0208 や ISO の規格改訂による変更情報の管理，大規模の漢和辞書の電子媒体化に必要な符号化の方法，電子媒体化した辞書によるコンピュータ文字発生装置の作成，JIS X 0208 やシフト JIS コードの各漢字符号を包含したネットワーク用符号へと発展が期待できることを指摘した．なお，今後本研究は，実用面での問題を把握するため以下の実験を予定している．

(1) 漢和辞書の電子媒体化に対する情報交換用漢字符号の在り方の検討．
(2) 4バイトコードへ DIS 10646 を取り込む実験．
(3) SQL を使用した4バイトコードの評価．
(4) 英数字やアルファベットを4バイトコードの小数部で符号化する方法の検討．

謝辞 本論文をまとめるにあたり，東京電機大学・守屋慎次教授には論文の問題点を指摘していただいた．また，実験システムの開発には，日本電気株式会社・C&C第二官庁システム事業部第4営業部・一杉宏一部長，中村千恵子さん，C&C事業部第3システム部・加藤隆士主任に大変おせわになりました．ここに記して謝意を表します．

参考文献

1) 中華人民共和国標準信息編碼字符集 GB 2312-1980：技術標準出版社，中国北京 (1981)．
2) Lee, B. W., Kan, l. b. and Lam, K. T.: Information Exchange of Bibliographic Informa-

tion—From an Asian Perspective, *First International Conference on Scholarly Information Network East Asian Applications & International Cooperation* (1987).
3) 情報交換用漢字符号系 JIS X 0208-1990：日本工業標準調査会審議，日本規格協会 (1990).
4) 情報交換用漢字符号系—補助漢字 JIS X 0212-1990：日本工業標準調査会審議，日本規格協会 (1990).
5) Wada, E.: Three Byte Code Considered Harmful and Standardization of the Two Octet Character Sets, *First International Conference on Scholarly Information Network Ease Asian Applications & International Cooperation* (1987).
6) 日本語処理機能を備えた Unix の標準化：日経エレクトロニクス，No. 387, pp. 243-251 (1986).
7) Universal Multiple-Octet Coded Character Set (UCS) Part 1: Architecture and Basic Multilingual Plane Unified Ideographic CJK Character (Version 1. 0), ISO/IEC Draft International Standard 10646-1 (December 1991).
8) 長谷川裕美：ユニコードと DIS 10646 統合のユニバーサル文字セット UCS の全貌，SuperASC-II, Vol. 3, No. 5, pp. 162-170 (1992).
9) 斎藤秀紀：漢字コードのメタコード化の方法，第 46 回情報処理学会全国大会論文集(1)，pp. 23-24 (1993).
10) 諸橋轍次編：大漢和辞典，第 3 刷，大修館書店 (1971).
11) システムソフトウェアの標準化に関する調査研究（拡張漢字コード）報告書：日本規格協会 (1987).
12) コンピュータマニュアルシリーズ別冊 2 文字フォントリスト 2V. 02. 00：東京外国語大学，アジア・アフリカ言語文化研究所・電子計算機室 (1988).
13) 松本 昭：国研用漢字テレタイプと同機利用の言語情報処理，電子計算機による国語研究，国語研究所報告 31, pp. 57-89 (1968).
14) 野村雅昭：JIS C 6226 情報交換用漢字符号の改正，標準化ジャーナル，Vol. 14, No. 3, pp. 4-9 (1984).
15) 李 春澤：韓国標準規格と日本工業規格の漢字について，学術情報センター紀要，Vol. 3, pp. 21-24 (1993).
16) 現代雑誌九十種の用語用字（第二分冊漢字表），国立国語研究所報告 22 (1963).
17) 現代新聞の漢字，国立国語研究所報告 56 (1976).

(平成 5 年 8 月 25 日受付)
(平成 6 年 2 月 17 日採録)

斎藤 秀紀（正会員）

1940 年生．昭和 41 年東京電機大学工学部（二部）電気通信工学科卒業．昭和 40 年国立国語研究所文部教官．現在，同研究所情報資料研究部電子計算機システム開発研究室に勤務．用語・用字調査および電子計算機による言語処理，漢字符号，データベースに関する研究に従事．IEEE，計量国語学会各会員．