

対立する二つの機能を構造化4バイトコードに実装する方法

斎藤秀紀

国立国語研究所・情報資料研究部
〒115-8620 東京都北区西が丘3-9-14
saito@kokken.go.jp

要 旨： JIS X0208 は、一般の国語表記と情報交換を目的に作られたため、大規模な漢字字典や古典・漢籍を符号化するために必要な文字集合を規定していない。また、JIS X0208 は、符号間への文字の登録機能や登録に対して基本配列を維持する機能をもたない。しかし、漢字符号が、多様化する処理に対応するためには、JIS X0208 の基本機能と拡張すべき機能を一つの符号に実装できることが重要である。本論では、対立する関係にある、①文字の拡張に対する適正規模の文字集合の規定、②規範としての符号化基準に対する利用実態の変化に対応できる機能、の二つの条件を、構造化した4バイトコードに併存させる方法を述べた。また、4バイトコードに設けた構造に、漢字と異体字及び漢字と読み・部首・画数を規定する方法についても述べた。

キーワード：4バイトコード、漢字符号、符号の構造化、論理符号

Methodology for Representing Two Opposing Functions

Using a Structured Four-Byte Code

Hidenori Saito

Department of Data Orientation,

The National Language Research Institute

Abstract: JIS X0208 was conceived to deal with Japanese characters in general use and to facilitate exchange of information; thus, this standard does not contain characters sets necessary to code large-scale kanji character dictionaries or corpora of classical Japanese and Chinese texts. Further, JIS X0208 does not have the functional capacity for inserting characters between codes nor can it support registering characters in standard ordering. In order to respond to increasingly diverse kanji processing needs, however, it is critical that both the basic and expanded functions of JIS X0208 be contained within a single code. This paper sets forth a method for combining two opposing functions, namely, 1) determining an appropriate standard for an expanded character set: 2) functional ability to respond to changes in actual usage in terms of code standards, and to meet these two conditions within a structured four-byte code.

Key Words: four-byte code, kanji character code, structured code, logical code

1 はじめに

一般的の国語表記や情報交換用の共通符号として開発された JIS X0208 は、現在、古典や漢籍などの処理にも使用されている。しかし、一般的の国語表記を目的に作られた JIS X0208 は、多様な文字種が使われる古典・漢籍・仏典を符号化できる十分な漢字を規定していない。また、漢字 1 文字に連続した 2 バイトコードを対応させた JIS X0208 や漢字符号は、大幅な文字集合の拡張や符号間へ文字を追加できないなどの問題がある。さらに、連続番号を使った符号化方法は、文献や資料を長期間保存する間に生じる表外字の登録に対して文字と符号の関係を維持することが困難である [1]。

これら漢字符号に関する問題は、現在主流になっている英数字を基本にした符号化方法の限界を示すものであり、漢字を符号とする処理では字形の簡略化や新たな漢字の作成・発見によって大規模化する文字集合や逐次的に発生する文字の追加・登録に対応できることが重要である。また、新たに作成する漢字符号は、既存の漢字符号で有効と認められた基本機能の継承と指摘された問題点を解決できる機能を併せもつことが必要になる。そのほか、漢字処理を効果的に行うためには、支援機能として、符号化の対象となる情報を辞書やコードブックを使って一元的に管理する方法の導入や、漢字と異体字や読み・部首・画数・言語名などの属性情報を統一的に符号表現できることが要求される [2]。

本論では、漢字符号の問題として指摘されている以下に示す二つの条件を 4 バイトコードに併存する方法を述べ、これらの条件は構造の要素と構造及び関係を使って規定できることを示した [3]。

- ①大規模な文字集合と適性規模の文字集合の規定。
- ②規範としての符号と文字の規定と符号への文字の追加機能。

また、漢字符号は、構造化することによって、利用者規定の文字集合間の情報交換における符号変換処理を削減できることを述べ、この符号化方法が、クライアント・サーバ環境で運用するための基本形態となることを示した。そのほか、論理符号に対応させた各種の情報を長期間安定状態で運用するためには、符号化文字集合や属性情報をデータベースの編成方式や装置で使用する内部符号から独立させることが有効であり、実現の方法として、符号化の対象となる情報を転置リストに展開し、構成要素の 2 項対を論理符号とする 3 層表現と、論理符号と内部符号の間に 2 バイトコードを置く 3 層表現の二重化の方法を述べた [4]。さらに、論理符号と 2 バイトコードの併用機能によって、既存の 2 バイトコードで作られたデータの継承が可能になることを示した。

2. 大規模な文字集合の規定と適正規模の文字集合の規定

対立関係にある二つの条件の中で、文字集合に対する大規模化の要請には、漢字が利用対象や目的によって制限されることへの懸念がある。また、適性規模の文字集合の必要性を主張する背景には、文字集合と漢字符号に対する取り扱いの容易性や処理の効率化がある。しかし、「大規模な文字集合」の主張には、文字選択の対象となる文字集合の規模の大きさを重要視する中で、利用者が目的や対象に応じて文字集合を選択・使用していることへの省略があり、「適性規模の文字集合」の主張には大規模な文字集合から出現頻度の高い

文字を抽出している操作の中で選択の対象が省略されている。これらのことから、二つの主張は、不可欠の関係にあるものとして、選択の対象となる文字集合から目的の文字集合を抽出する手続きとして形式化できる(図1)。

また、4バイトコードは、3バイトコードとしての利用と、94種の2バイトコードを二重に規定できることから、一つの符号で約83万字の文字集合と、8,836字を単位とする2バイトコード対応の文字集合を重ねて符号化した。83万字の文字集合は、大規模な漢字辞書や各国で使用されている漢字を符号化するために必要な領域と、長期にわたって文字を逐次的に追加するための十分な空間を与える。このような4バイトコードに2バイトコード対応の領域を埋め込む形態で再定義する形式は、4バイトコードと2バイトコードの独立した利用と併用機能を通して既存の2バイトコードで作られたデータを4バイトコードへ継承する機能と、4バイトコード対応の文字集合から利用者が必要とする文字集合を選択し、二次符号を与えるための基本的な枠組みを作ることができる。

4バイトコードで規定する機能や情報を構造の要素と構造を使い規定する方法は、電子辞書やデータベースによる漢字と属性情報の一括管理、クライアント・サーバ環境における電子辞書の統括管理と分散を容易にする。この機能は、クライアントに分散した電子辞書の利用者運用を通して追加する文字の収集・登録・管理や、クライアントで発生した文字の不足をサーバから補填する処理に応用できる。

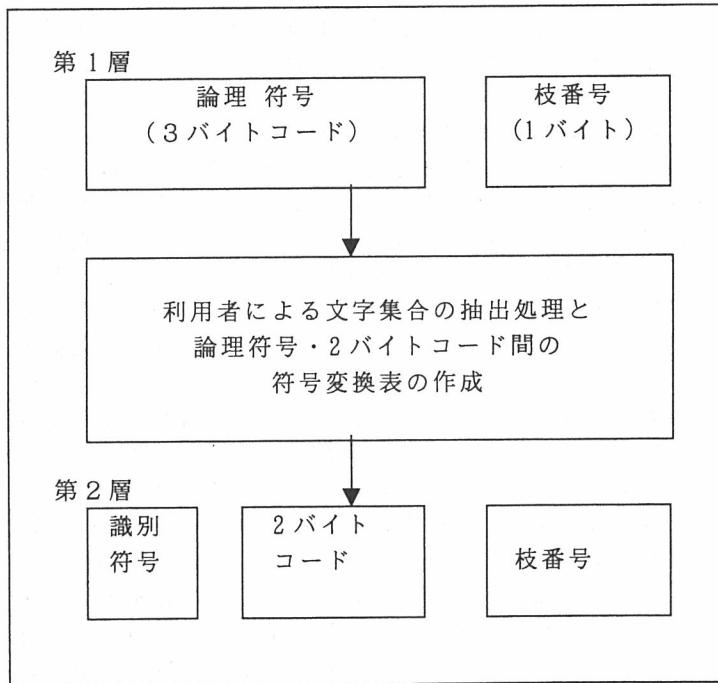


図1 4バイトコードを構成する構造と要素の関係

3. 規範としての符号と文字の規定と符号への文字の追加機能の規定

長期間保存されたデータを正確に再現するためには、漢字符号に対して行われた文字の追加の過程が履歴として記録できることが重要である。また、文字の追加に対しては、規定されている符号と文字の関係が固定されていることが必要になる。さらに、新しく作る漢字符号は、指摘された問題点を解決するための拡張機能と、既存の漢字符号で有効と認められた機能を併せもつことが要求される。

これらの機能を一つの符号で実現するためには、異なる複数の条件や機能を漢字符号に併存でき、一方の情報の追加や変更が、ほかの情報に影響を与えない局所化機能をもつことが必要である。以下は、対立する関係にある二つの機能を4バイトコードに設定した二つの要素と要素間の関係を使い規定する方法である。

4 バイトコードに対応させる二つの基本機能は、階層関係において二つの要素に対応させ、拡張機能を要素間の関係を使い規定した。4 バイトコードの上位要素 3 バイト部分には、安定が要求される見出し漢字約 83 万字を符号化する機能を与え、下位の要素 1 バイトには変更される可能性の高い異体字や各国の漢字を登録・管理する機能をもたらせた。要素間の関係では、符号と文字の固定と文字に対する文字の追加機能を規定し、局所化による上位構造に配当した見出し漢字の固定化を図った。この二つの要素に配当した情報を 4 バイトコードに投影したものが論理符号である。

なお、3 バイト部分で規定する見出し漢字は、各国の合意を得ることが必要であるが、下位の要素に配当した異体字や各国漢字に対する代表形として、東アジア各国で使用する情報交換用の共通字形として利用することも可能である。

4. データベースによる論理符号の規定

文献・資料に出現した漢字や文字を忠実に符号化するためには、符号化の対象となる漢字を可能な限り細分化して登録できることが重要になる。また、符号化の対象になる漢字と属性情報は、符号に追加された情報の履歴や変更された情報の過程が視覚的に把握できることが要求される。さらに、これらの漢字符号をクライアント・サーバ環境で使用するためには、サーバによる情報の一元的な管理とクライアントへの分散機能のほか、クライアントで利用者が付加した情報をサーバで統合できる双方向の情報補填機能をもつことが必要になる。

これらの条件をもとに 4 バイトコードは、規定する情報を異体字や各国漢字と属性情報と順編成ファイル形式で管理し、視覚化するための手段としてコードブックによる一元的な管理を行った。順編成ファイル形式で規定した漢字情報は、見出し 6,349 字に、政令で規定した情報、漢字辞書で規定した検字番号、漢字の読み・部首・画数など 40 項目で構成し、電子漢字辞書としての役割を与えた。論理符号は、索引順編成のデータベースの索引や検索用キーとして利用するため、ファイルで規定した情報を転置リストに展開し、転置リストの見出しと見出しに付属する異体字や属性情報の 2 項対を論理符号として規定した。また、論理符号は、2 バイトコードを配当した上位構造に置き、全ての符号と符号化の対象になる情報を統括管理する機能をもたらせた。

そのほか、JIS X0208 の問題点として指摘された文字集合の拡張機能と符号への文字の追加機能は、2 バイトコードの上位構造に置いた論理符号の要素に配当し、不足機能を論

理符号との併用によって補った[5]。論理符号の下位の要素1バイトには、異体字と各国漢字を可能な限り、詳細な分類基準として利用するため、異体字や各国漢字の出現形の登録を基本とした。論理符号の1バイト部分へ登録する属性情報や異体字が94字を越える場合は、情報を管理するための制御領域を設け、94字単位で管理した（この方法は、将来別 の方法で再実験する予定である）。コードブックへの文字登録は、未登録部分の後ろに追加し、 文字の配列順序が崩れた場合は属性情報による再分類を行うことを想定した。表1の転置リスト形式で表した異体字は、JIS X0208で規定したものを使い、属性情報は主に新字源から引用した[6, 7]。

表1 転置リスト形式で表した漢字の異体字と属性情報

代表形	個數	異体字	
劍	6	劍	劍
弁	5	弁	辨
鉛	4	鉛	礦
崎	4	崎	琦
疊	4	疊	疊
体	4	体	體

— — 以下略 — —

音読み	個数	同音の漢字
ア	13	亞哩娃阿雅蛙窪亞壘猗西鉛闊
アイ	20	娃哀愛挨喝乃哇噫埃曖欸嬾穢藺陌隘靄靉鞋
アク	14	哩阿惡握渥厄壘幄惡扼輶陌鷺齧
		以下略

——以下略——

訓読み	個数	同訓の漢字
アフ	1	乎
アフ	11	惡於于呼咨嗚嗟噫惡猗羌
アイ	7	始逢間際相藍胥
アイタ	1	間

——以下略——

画数	個数	同一画数をもつ漢字
01	5	ノ丶一乙丶
02	29	匚丁儿十口冂宀力了口乃フヒ宀入刀冂厂ム七ト人 又二九乂八几
03	57	宀女土也匚子下巾才寸叉夕上与干巳己凡子士万山 久乞刃大及三弓小亡丸巠口升川尤么广爻丈夕口兀 于工弋冉互彑彳个夕尸千巳刃

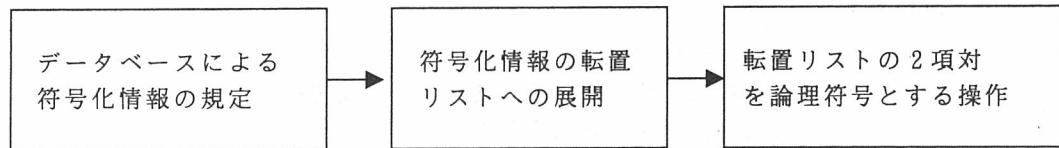


図2 データベースで規定した情報から論理符号を作る操作

5. 漢字符号に設定する構造と論理符号の枠組み

2バイトコードの各バイトは、MSD(Most Significant Digit)の0と1の混在を許し、指示できる領域をG0,G1のほか、G2,G3へ拡張した。また、領域と符号は、文字集合に対する指示と呼び出しを一つにまとめるため各領域と符号の対応を、G0:2121から7E7E、G1:A1A1からFEFE、G2:A121からFE7E、G3:21A1から7EFEに固定した。2バイトコードと4バイトコードの識別は、各バイトのMSDを使った。

ISO 2022 (JIS X0202) に規定する 7 単位系と 8 単位系符号は、2 バイトコードの MSD を省く各バイトが 16 進数 21 から 7E の範囲をとるものと対象とした。4 バイトコードと 2 バイトコードは、それぞれ独立した符号としての利用と、論理符号と既存の 2 バイトコードの関係を使い、対立する二つの機能を拡張機能として規定した。4 バイトコードによる文字の符号化は、3 バイト部分で論理符号に対応させた約 83 万字の文字集合と、相対表現された JIS X0208 や拡張 UNIX コードなどの 2 バイトコード対応の文字集合、94 種を埋め込む形式で規定した。(なお、4 バイトコードに 2 バイトコードを埋め込む形式で規定した符号は、領域を明確に分離する必要がある)。

5.1 複数の 2 バイトコードの規定

図 1 に示した第 2 層は、現在使用されている 2 バイトコード対応の文字集合を構造化 4 バイトコードに埋め込む形式で規定した構造である。第 2 層で規定する漢字符号は、各バイトの MSD を省く 7 ビットが 16 進数 21 から 7E の範囲にあるものを対象とした。94 種の相対表現された 2 バイトコード対応の文字集合は、G0, G1 に呼び出した場合に既存の 2 バイトコード対応の文字集合との互換を図り、3 種の文字集合を G0, G1, G2 に呼び出すことによって、同時に使用できる 2 バイトコード対応の文字集合を 26,508 字に拡張した。

5.2 論理符号と 2 バイトコードの関係

図 1 に示した二つの層は、第 1 層を上位とする階層関係の中で、第 1 層で情報の全体と第 2 層に配当する 2 バイトコード対応の文字集合の統括管理機能を持たせた。また、第 1 層は、論理符号に対応させ、データベースで規定した符号化の対象となる情報の全体を規定した。また、これらの符号化の対象となる各種の情報は、定形形式をとる転置リストに展開し、構成要素である見出しとデータベースとの接続情報の 2 項対を論理符号として規定した。第 2 層は、第 1 層に対する部分として位置付け、既存の 2 バイトコード対応の文字集合や論理符号に対応させた文字集合から利用者が必要とする文字を選択し、2 バイトコードを与える枠組みを作った。

また、4 バイトコードは、2 バイトコードに対するメタコードとして使用し、1 対 1 の符号変換処理に対して変換処理を削減した（1 体 1 の符号変換数：符号の重複を許した N 個の符号から R 個とる組み合わせになり、論理符号経由では、符号の 2 倍の回数になる。変換数は、符号数 4, 5, 6 種では 12, 20, 30 回に対して 8, 10, 12 回である）。

また、第 1 層で規定した情報から情報を抽出する処理では、4 バイトコードと 2 バイトコードの対応表を作り、これにクライアントで使う各種の情報を附加したもの電子辞書として使用した。論理符号と 2 バイトコードの関係は、規定する文字集合の全体と部分の関係をクライアント・サーバ環境に対応させることによって、論理符号に対応させた文字集合の全体をサーバで管理し、クライアントの運用に必要な情報の最適化を図った。この処理によって電子辞書は、サーバによる情報の一元的な管理・運用と、クライアント側に置いた電子辞書に未登録情報が生じた場合にオンデマンドでサーバからの情報補填を行うことを可能にした。

そのほか、電子辞書は、データを検索する処理において、(1) 代表形と同一グループに属する異体字をキーとする検索処理、(2) 異体字と同一グループのほかの異体字をキーとする検索処理、(3) 異体字とその代表形をキーに該当するデータを検索する処理、に対しても利用できるよう図った。

6. まとめ

本論では、4 バイトコードを構造の要素と構造及び構造間の関係を使い、対立する関係にある二つの条件、①大規模な文字集合の規定と適正規模の文字集合の規定と、②規範として文字と符号への文字の追加機能、を実装する方法について述べた。①の条件は、規定した文字集合の全体から必要な文字集合を抽出する操作を変換処理に置き換えることによって実現したものである。また、②の条件は、異体字と各国漢字を集合の要素とし、

見出しを集合名とする関係を使い、要素への文字の追加に対して集合名が変化しない特性を利用したものである。

対立する関係にある二つの条件は、文字集合を全体から部分を抽出し、再規定する開構造としての漢字符号の開発に道を開いた。さらに、一つの符号で文字を統一する符号化方法は、各規格で規定されている文字集合の重複排除と文字の逐次的な追加に対して一貫した管理を可能にした。さらに、この手続きは、4 バイトコードで規定した文字集合の全体に既存の 2 バイトコード対応の文字集合を二重に規定することによって、部分と全体との間に文字集合の分散利用と表外字として追加された文字を全体に統合する双方向処理を可能にした。そのほか、4 バイトコードは、構造化することによって以下に示す機能を実現した。

- (1) 東アジア漢字圏で使用されている漢字の統一表現。
- (2) 規範としての文字集合と利用実態の変化に対応できる機能の実現。
- (3) 保存データの再現を保証するため辞書による文字の追加過程の一元管理。
- (4) 論理符号に対する内部符号やデータベースの編成方式からの独立。
- (5) 属性情報と漢字の異体字に対する統一符号化方法の実現と、異体字や各国に対する代表形の設定と各国で共用できる共通字形の規定。
- (6) 利用者による文字と符号に対する二次規定。
- (7) 複数の独立した文字集合の統合と属性情報による文字の統一配列処理。

本論では、対立関係にある二つの条件を実装する方法を述べたが、実用化を図るために異体字と各国漢字が 94 を越える場合の拡張方法と、属性情報の符号化方法を統一することの妥当性を確認する必要がある。特に「寿」の異体字を 100 字、1 千字、1 万字集めた例があり、これらの特別な事例に対応できる機能についても検討の必要がある。これらは、構造化 4 バイトコードの今後の課題として研究を進める予定である。

7. 参考文献

- [1] 斎藤秀紀：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、情報処理学会論文誌, Vol. 35, No. 6, pp. 1119-1126 (1994).
- [2] 斎藤秀紀：漢字の属性情報に対する符号化法の提案、計量国語学, 21 卷, 4 号, pp. 131-144 (1998).
- [3] ジャン・ピアジェ (滝沢武久・佐々木明共訳)：構造主義、白水社 (1970).
- [4] JAMES. MARTIN (國友義久・久保美沙共訳)：データベース [改訂第 2 版], 日本コンピュータ協会 (1983).
- [5] R. L. ワイルダー (吉田洋一訳)：数学基礎論序説、培風館 (1969).
- [6] 小川環樹・西田太一郎・赤塚忠編：新字源、230 版、角川書店 (昭和 60 年 1 月発行).
- [7] 情報交換用漢字符号系 JIS C6226-1978, 日本工業標準調査会審議、日本規格協会発行、(昭和 53 年 1 月 1 日).