# 多国語を統一する情報交換用４バイトコードの研究
## －日本語と中国語処理における

国立国語研究所
斎藤秀紀

## １．まえがき

日本および東アジアで使用されている情報交換用漢字符号は、符号化の対象となる字形が自国のものを中心に構成されている。多言語化の方法には、補助コードセットによる方法を拡張UNIX コードが採用している。また、JIS X0202で規定した拡張機能に従う場合がある。そのほか、ISO/IEC 10646-1 のように一つの符号域で多言語を表現する方法もある。

しかし､拡張符号を使った多言語化の方法は、国名の判別が識別符号で判別できるのに対して、統一符号系では、字形からはどの国の文献・資料を符号化したかを知ることができない。これは、各国語の漢字の読みや意味などを引用するさい字形から該当する辞書を識別できないことになる。

本稿では、構造化４バイトコードを使った中国語と日本語入力実験を通して基本機能の検証と、字形がどの国の文献・資料で使用されたかを識別する属性情報の符号化法について検討する。本年度に行った実験は、以下の５項目である。

(1)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。
(2)サーバにおいた漢字辞書情報をクライアントからオンデマンドで引用する実験。
(3)4 バイトコードを使った中国語と日本語を混在入力のモデルによる多言語表現機能の確認。
(4)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。
(5)二つの利用者規定の文字集合に常用漢字と旧字体を対応させ、文字の切り替えの応用実験。

## ２．実験システム環境の概要

ブログラム開発と実験環境は、イーサネット(10Mbyte/s)上に接続した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)を使用した。日本語と中国語を混在入力するプログラムは、サーバにおくプログラム「4Bserver」とクライアント側プログラム「4Btext」の二つに分けた。第一段階の実験用プログラムでは、イーサネットに接続された他の装置からの影響を避けるため、１台のワークステーションにクライアント・サーバ環境を設けた。

また、サーバには、6,349 字の漢字に 54 個の属性情報を付加した総合漢字辞書をおいた。また、クライアント側漢字辞書は、プログラム起動時にサーバから常用漢字 1945 字と対応する中国語（簡体字）および属性情報を抽出した。クライアイントにおいた辞書に未登録漢字が発見された場合には、オンデマンドでサーバへの要求をだす相補処理を行った。

## ３．キーボードから読みを入力する操作

図１は、変換条件指定部で'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。中国語の読み'chu' は、このリストから資料で使用されている漢字の字形を選択することができる。これらの漢字入力処理は、最初に、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。次に、変換指定が「音読み」または「訓読み」の場合、仮名変換された読みをキーとし、漢字辞書から読みが一致する漢字をすべて候補一覧表示部に表示する。変換指定が「ピンイン」の場合、キーボードから入力したアルファベットの綴りをキーにして辞書検索を行い、一致するものを全て候補一覧部に表示する。キーで候補一覧から必要とする漢字をカーソルで選択する。

決定した漢字について、属性情報付きの４バイトコードを生成する。次いで、変換条件指定部で「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、選択した漢字の読み、＜音読み＞、＜訓読み＞、＜ピンイン＞を符号化するため、該当する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する。

図２に示した画面の左上部は、中国語を４バイトコードで表現した例である。左下部は、漢字変換用に入力した読みと該当する漢字辞書の内容を示したものである。画面右には、２バイトコード、４バイトコードおよび国名、部首、画数、選択した読みを示した。ただし、日本語と中国語の混在処理では、「貴」、「賞」の文字フォントを日本漢字の読みで代用したため、国名が'Ｊ' と表示されている。

## ４．属性情報の符号化の方法

符号化した属性情報は、４バイトコードの作業領域をG3としたため、奇数バイトを'21' から'7E ' に、偶数バイトは、'Al' から'FE' の範囲に調整する。ある。本実験では、４バイトコードの第４バイト目（小数部）を使い、94の符号領域に属性情報を表す領域に 64 個、国名と終端符号に 30 個分をあてた。本実験で符号化の対象とした日本語と中国語の属性情報は、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' 登録を許すものとし、属性情報を記録する位置に辞書に記載されている情報の位置を指定する。

例：楚：[ソ，イバラ,，シモト，スワエ,，ＣＨＵ］の順序で登録されている。[ソ］を引用する場合は、'０' を指定する
（開始番号は'０'）。

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の二つを使った。日本語には、表示用国名を'Ｊ' とし、内部符号は 16 進数'0XFE' をあてる。また、中国語は'C'を国名表示に使い、'0XFD'を内部符号とした。

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コードをあてた。

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。

## ５．属性情報と漢字符号の対応関係

本節では、実験システムのファイル入出力で使用するデータ単位（以下、この単位を「セル」と規定する）を、(l)4 バイトコード、(2)2 バイトコード、(3)属性情報付き４バイトコード、(4)属性情報付き２バイトコードに分けた。

l)4バイトコード：

４バイトコードは、３バイト部分で見出しとなる漢字字形をあて、４バイト目に見出しに対応する各国語の漢字や異体字を配当している。４バイトコードを単独で使用する場合と属性情報が付加される場合の識別は、４バイトコードの最後の桁（小数部）の内容で識別する。

'Al－CE' のとき、４バイトコードの漢字符号として使用する。

'D1－FE' のとき、属性情報が漢字符号に付加されていることを示す。

2)2バイトコード：

実験システムでは、拡張 UNIX コードをクライアントで使用する利用者規定の２バイトコードとして使用する。

3)属性情報付き４バイトコード：(l)読み・国名情報が入る場合

セル長は、６バイトを使用する。５バイト目は、読み情報を表し、漢字辞書上の最大 64 種と結合が可能である。６バイト目は、国名情報を表す。最大 30 カ国の識別が可能である。７－８バイト目は、部首と画数を表す。(2)部首・画数・読み・国名情報」が入る場合

セル長は、10バイトである。

１ －４バイトで漢字符号を表す。

５ －８バイトで部首・画数を表す。

９ －10 バイトで読み・国情報を表す。

4)属性情報付き２バイトコード
　２バイトコードに属性情報が付いた」セルでは、属性情報を結合するための識別符号として３バイト目を未使用のバイトとし、16 進数'21'で埋めた。４バイト目は、'D0'をおき属性情報の有無を識別するために使用する。４バイトコードの構成内容は、先頭２バイトに漢字を符号化した２バイトコードを格納し、５バイト目以降の属性情報は、属性情報付き４バイトコードと同様の内容をとる。

６．属性情報付きの漢字符号の認識・切断方法
　1)入カファイルからデータを読む。
　2)各バイトの２の８ビット目が
　(1)'１－１'の場合－＞２バイトコードである。処理 1)へもどる。
　(2)'０－１'の場合－＞２バイトをメモリに格納し、次の２バイトを読み込み４バイトを作る。
　読み込んだ４バイトの、各バイトの２の８ビットは、'0,1,0,1'である。
　3)4 バイト目の値が
　(1)'A1－CE' の場合－＞４バイトコード（属性情報無し）である。セルの終端'E1FE' を検出した場合、処理 1)へもどる。
　(2)'D1－FE'の場合－＞属性情報が後続する。※印の処理を実行する。
　(3)'D0'の場合－＞属性情報付き２バイトコードである。処理※）を実行する。※）さらに２バイト読み込み、２バイト目が
　(1)'E1－FE' の場合－＞セルの終端に達した。処理 1)へもどる。
　(2)'A1－E0' の場合－＞後続する属性情報を読む。処理は※印の位置にもどる。

７．まとめ
　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サーバ・クライアント環境での実行に対応できることを述べた。
　そのほか、多言語処理の出力形態の一貫として漢字符号に漢字の属性情報や、その漢字が使われた資料の国名を付加する実験を行い、同一符号系で多言語処理を行う場合に各国語の辞書を選択する上で不可欠の処理となることを述べた。
　属性情報を漢字符号の利用環境により特定の辞書から情報を引用生成する方法は、本論文で初めて提案されたものであり、これにより、伝統的な漢字の特性を説明する手段として認められている、「音」、「義」との関係を符号化する方法に道を開くものとなる。また、今後、異体字の分類には、意味を考慮することが必要になることが予想される。この方法は、その基本として重要な役割をはたすものである。
　そのほか、属性情報の符号化法は、自動処理や簡易印刷で必要となる各種の情報を文章に埋め込むことを形式化する予定である。

参考文献

1)多言語間の情報交換を統一的に行うための構造化４バイトコードの研究－中間報告－、文部省科学研究費（創成的基礎研究費)「国際社会における日本語についての総合的研究」(研究代表者：水谷修，課題番号 08NP0701)第４班：分担（印刷中)．
2)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)．
3)斎藤秀紀:1 字体に 1 符号を対応させる漢字符号化の方法，計量国語学会誌，第 19 巻５号,pp.223－233(1994)．

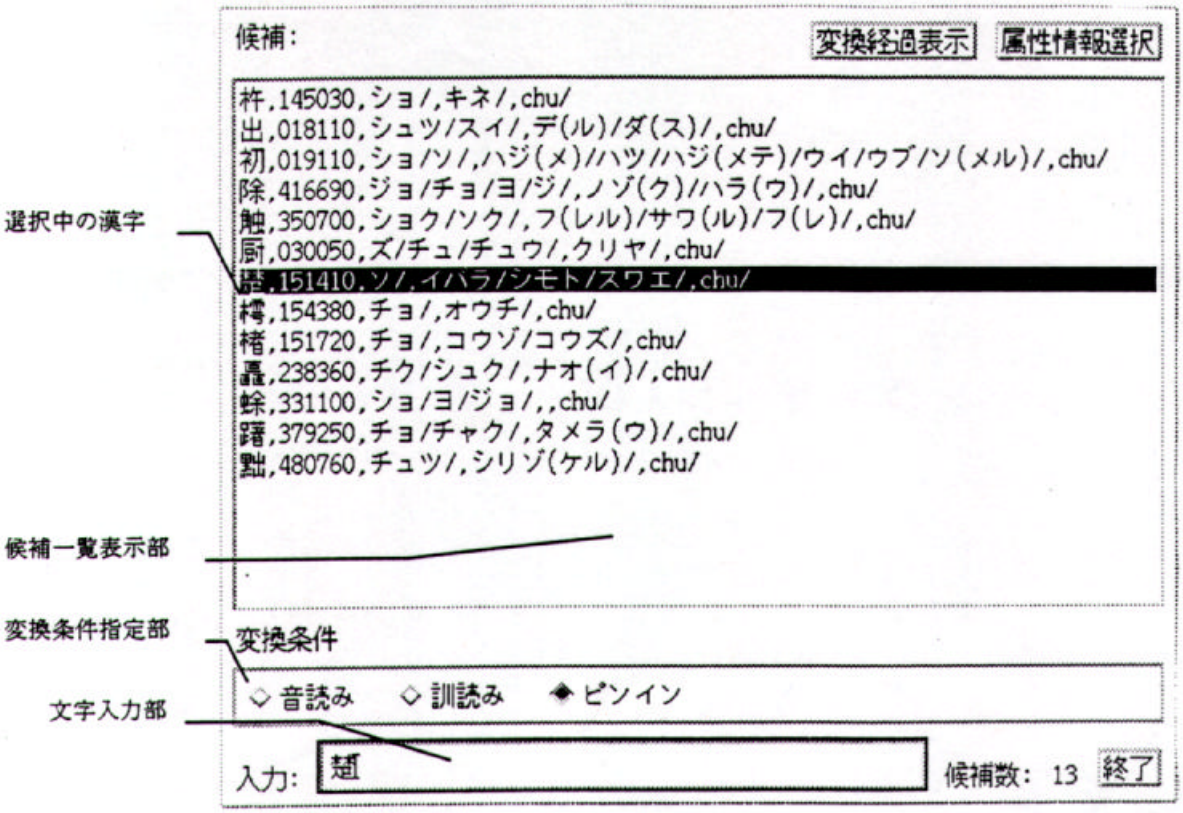


**図１　漢字辞書検索と読みの決定例**

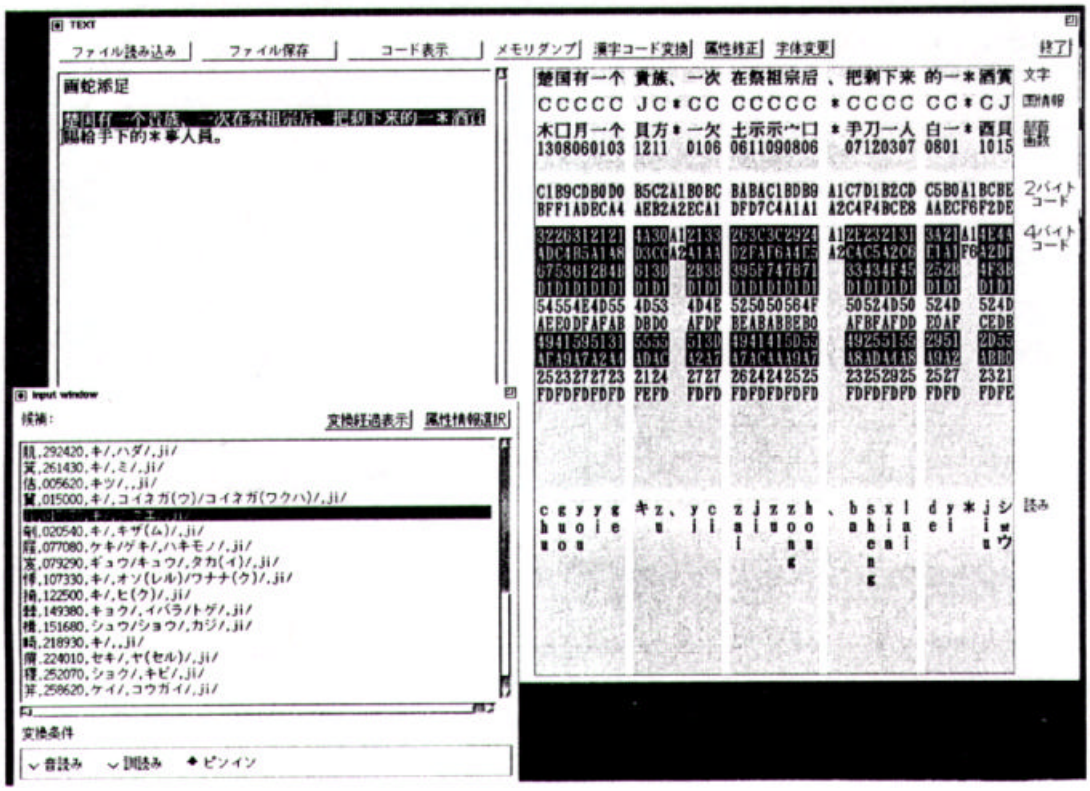


**図２　中国語と属性情報の処理例**