多言語を統一する情報交換用４バイトコードの研究
－漢字の属性情報に対する符合化法の提案－

国立国語研究所
斎藤秀紀

１．はじめに

JIS C6226-1978 は、1978 年に日本語処理のための情報交喚用漢字符号として制定された。(1987 年 JIS X0208 に改称)。しかし、JIS X0208 は、漢字符号に連続番号を使用しているため、文字の追加・更新に対する初期状態や履歴を管理することが困難な状態にある。また、1993 年に国際標準化機構が発表した国際符号化文字集合(ISO/IEC 10646-1, JIS X022l-1995)は、各国の漢字識別方法を規定していないため、実装される文字フォントが包摂基準による「字形統合」の範囲か「ソースコード分離規則」を併用するのが曖昧である。さらに、データに各国語が混在しているデータは、符号からはどの国の漢字であるのが判別できない。

本論文では、調査・研究や漢字字典に記載されている漢字の属性塁清報を漢字データベースで規定し、これらを形式的に符号化する方法を提案する。符号化実験は、サーバに置いた JIS C6226-1978 の 6349 字に 40 種の属性情報を附加した漢字データベースを基本に、以下に示す 16 種の属性情報を対象に行った。

(1)JIS X0208-1983 で改定された情報、(2)総画、(3)部首部分を省いた画数、(4)教育漢字識別符号、(5-6)学習漢字の識別符号２種、(7-9)当用漢字の識別符号と各補正符号２種、(10)常用漢字識別符号、(11-14)人名漢字識別符号４種、(15)国名、(16)読み。

なお、中国語は、GB 2312 を基準に読みはワークステーション(日本電気株式会社製 EWS4800/32OEX)付属の漢字入力システム用電子辞書から引用した。

２．漢字データベースと転置ファイル

文献・資料の電子化と再現処理には、明確に規定された符号と文字の関係をデータ作成者と利用者の双方が共通に使用できる環境をもつことが重要である。また、符号と属性情報を安定状態で運用するためには、情報の追加や更新に対して符号と文字の初期状態が維持されていることが必要になる。そのためには、符号が、装置で使用する内部符号や漢字データベースの編成方式から独立していることが要求される。

本論文では、漢字データベースの索引として設けた転置ファイルから形式的に属性情報を符号化するため、漢字データベースで規定した属性情報の符号化（以下、属性符号）に対して３層モデルを導入する。符号長は、４バイトコードとし、構造要素、要素を結合した構造、構造間の関係を対応させ、構造の要素には符号に要求する機能の単位を当てた（図 1)。

第１層は、論理符号を漢字データベースの編成方法から独立させるため、表１に示した転置リストの構成要素である属性情報と漢字データベースへの接続情報の２項関係を論理符号とした。属性情報の位置は、漢字データベースに記載されている見出しからの位置を項目番号とし、同一属性をもつ場合はま項目番号からの相対位置（最大９４種）の二つを使った。相対位置は、データへの属性情報を埋め込む処理用に使い、１バイトで表した。また、漢字符号は、３バイト(830,584 字）で表現した。

第２層は、属性情報１バイト論理符号の３バイトを結合したものと拡張 UNIX コードの２バイトを結合した符号系である。拡張 UNIX コードは、クライアントて使用する漢字符号であり、論理符号はサーバで使用する符号である。また、データに埋め込む属性情報は、漢字符号部分の拡張 UNIX コードと論理符号をデータ符号と共用させ、属性を表す１バイトを使った。

第３層は、装置で使用する内部符号である。既存の漢字符号との併用を図るため G3 領域を使った。

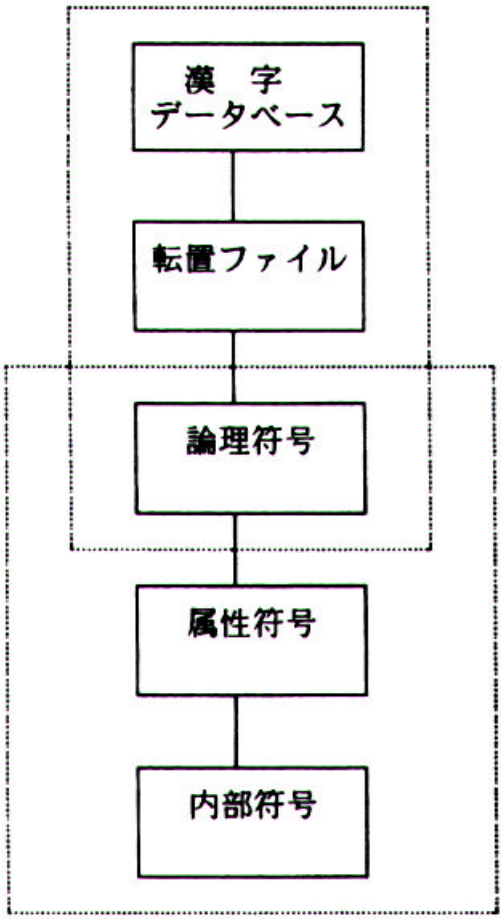

図1 論理符号と属性符号に対する３層構造

表１　転置リストの一例

音読み　個数　同音の漢字

ｱ　１３　亜唖娃阿雅蛙窪亞堊猗而鋺閼

ｱｲ　２０　娃哀愛挨喝乃哇噫埃曖欸瞹矮穢藹阨隘靄靆鞋

ｱｸ　１４　唖阿悪握渥厄堊幄惡扼軛阨鴛齷

ｱﾁ　１　藹

ｱﾂ　１０　亜圧斡亞壓扎藹軋遏閼

――以下略――

訓読み　個数　同訓の漢字

ｱｧ　１　乎

ｱｱ　１１　悪於于吁咨嗚嗟噫惡猗羌

ｱｲ　７　始逢間際相藍胥

ｱｲﾀﾞ　１　間

ｱｲﾃ　２　対對

ｱｳ　１３　逢会遇合遭値會翕靆覿逅遘邂

――以下略――

画数　個数　同一画数をもつ漢字

01　５　ノ丿一乙丶

02　２９　勹丁儿十冂冖匚力口乃冫匕匸刀冂厂厶七卜人又二九メ八几

03　５９　宀女土也勺孑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巛口井川尢幺广廴丈夂口兀于工弋屮互彡彳个夕尸千巳歹

04　１０９　刈文分比父少日夭手匹夫方氏夫无支攵升乏片尺仏什不仄厄殳毋气孔亢欠从仍弍月歹公仆仂犬午互元五幻戸勾孕支仇毛木牛勿卞尤匁丐卅止友予爿斤爻今区天兮六凶之双中火弓丑弔屺井冗允尹巴円廿引日爪扎切心云化戈斗水反牙壬太王丹屯介匂内仁

――以下略――

３．属性情報の符号化の方法

選択された属性情報の管哩は、属性情報の見出し漢字からの位置を示す項目番号と項目番号からの相対位置の二つを使った。項目番号は、属性情報に設けた分離記号',' (カンマ)を数え検出した。複数の属性値をもつ場合は、項目番号からの相対位置を内容による検索で得た。また、漢字データベースから属性情報を読み出す処理は、制御レコードの項目番号とデータの相対位置を使った。項目番号は、ヘッダーの一部として使用する制御レコードに記録し、該当するデータの管理に使用した。

４バイトコードの漢字符号は、３バイトで漢字を表し、４バイト目を 47 個の異体字登録と、以下に続く情報が「属性情報なし」か「属性情報付き」かを判別する機能に使った。漢字符号と属性情報の接続状態は、以下の組み合わせとなる。

(1)属性情報なし２バイト漢字符号

全てのデータを拡張 UNIX コードで表現する。

(2)属性情報なし４バイト漢字符号

１バイト目から３バイト目に論理符号の３バイト部分を埋め、４バイト目に「属性情報なし４バイト漢字符号を表す'A1'から'EC'を入れる。

(3)属性情報付き２バイト漢字符号
　１バイト目と２バイト目に拡張 UNIX コードを入れ、３バイト目と４バイト目に「属性情報に対する接続情報」をす'21D0 'を入れる。
(4)属性情報付き４バイト漢字符号
　１バイト目から３バイト目に論理符号の３バイト部分を埋め、４バイト目に「属性情報付き４バイト漢字符号を示す'D1'から'FE'を入れる。

４．プログラム機能と実験の概要

実験装置は、ワークステーション１台にクライアント・サーバ・モデルを実装した。符号化実験は、漢字データベースで規定した属性の符号化と位置情報から属性情報を再現する二つの処理を行った。本実験では、クライアント側漢字データベースに登録されていない漢字がある場合は、サーバの漢字データベースから補填した。日本語と中国語の識別は入力文字がカナの場合に日本語とし、ピンインの場合は中国語とした。属性情報のデータへの埋め込みは、プログラム起動時に画面表示された 16 種のリストから選択した最大８種の属性情報を表す位置情報を指定順に結合した。クライアント側の漢字符号は、拡張 UNIX コードを使い、サーバでは４バイトコードを使用した。

実験用プログラムは、サーバ(4BServer)用とクライアント(4BText)用の二つを作成した。4BServer では、漢字データベースの検索と管理を行った。4BText は、(1)ファイル入力、(2)ファイル出力、(3)日本語や中国語に対する読み漢字変換、(4)データと属性情報の表示、(5)読み情報の修正処理、(6)新旧漢字の選択表示機処理、の各機能をメニュー選択方式で実行する方法をとった。

図２の左画面は、テストデータの内容である。右画面に、JISC6226 で規定した区点番号、総画情報、部首を省いた画数情報（内画）、テストデータで使われた漢字の読みと国名を表示した。国名は、日本語を、'Ｊ'で表示し、中国語は'Ｃ'を使った。画面右側に示した項目名「4 バイト漢字符号」と「2 バイト漢字符号」はサーバの論理符号とクライアントで使った符号である。画面中央の副画面は、漢字データベースに日本語と中国語を入力する場合に同じ読みをもつ漢字の候補を示したものである。読みの入力は、副画面の変換方法指定ボックスから「音」、「訓」または「ピンイン」を指定し、次に、読みと漢字を選択する方法をとった。

５．おわりに

本研究で提案した符号化法を実用化するためには、属性符号を１バイトから２バイトに増加させることや間接指定法を標準化し、分類用レコードの作成を省略することが必要になる。また、属性と漢字の２項関係を符号化基準に使用する方法は、ハングルや漢字と異体字についても符号化できる可能性をもつものであり、多言語を異体字と同じ枠組みで符号化するための重要な課題となる。これは、今後、漢字の異体字と属性情報を一つの符号化法で処理する実験で確認する予定である。

［付記］本研究は、平成９年度・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 09NP0701）の分担として交付を受けた。

参考文献

[1]　斎藤秀紀　漢字清報と文例情報を結合した日本語データベースの構築・情報処理学会，人文科学とコンピュータ研究会,VoL96.No.42.pp.35.-40(1996).
[2]　斎藤秀紀　４バイトコードと対応文字の部分集合に対する利用者規定の方法，情報処理学会第 50 回全国大会講演資料集 pp.201-202(1995).
[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，第 19 巻５号,pp.223-233(1994).
[4]　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、 ,VoL35,No.9,pp.1119-1126(1994).
[5]　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（審査中）

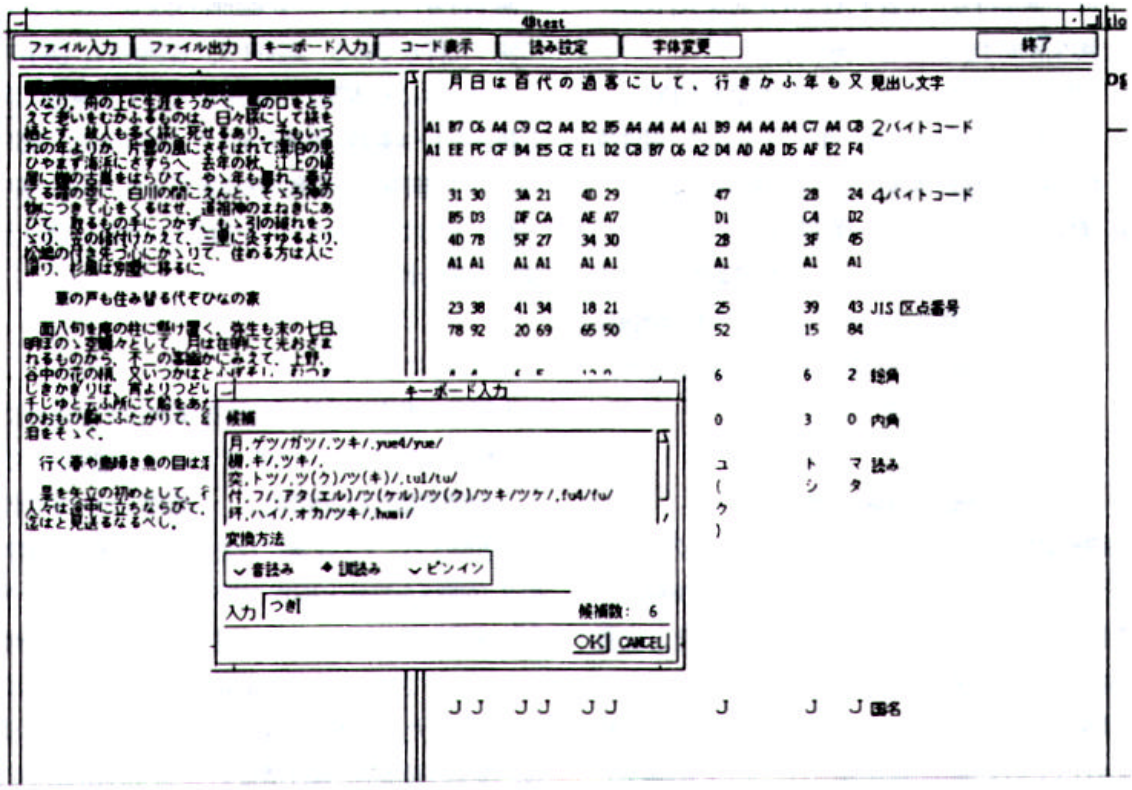

図２　属性情報を付加したデータ