JIS漢字に於ける異字体とその変換システム

三　谷　和　史

1．はじめに

中国の長い歴史の中で様々な漢字が生まれて変化を遂げ，また，日本に伝えられてからも変化を続けている。そのため複数の漢字が同一の意味を持つ例が数多く存在し，このことが漢字をより複雑にしている。身近な例としては，文芸の「芸」と雑誌『文藝春秋』の「藝」などがある。しかし漢字が日本語の表現を豊かでバラエティに富んだものにしていることも事実である。

コンピュータで漢字を利用する場合，通常JIS [1]，[2]，[3]に基づいた漢字コードが使用されるため，日本の漢字と認められる5万字余りの漢字と比べると使用可能な漢字は制限される [4]。また，漢字の入力は通常仮名漢字変換システムを通じて行われるため，仮名漢字変換システムの辞書によっての制限も存在する。これは，「醫學」と記述したい場合に仮名漢字変換システムの辞書に登録されている語彙が「医学」だけであれば，記述したい「醫學」を入力するためにいちいち単漢字辞書を開いて1字ずつ入力するといった手間がかかり，結果使用する漢字は仮名漢字変換システムの辞書にある語彙となりがちになるという制限である。

コンピュータで漢字を使用する際も，深く豊かな文章表現を可能にするためには，できるだけ漢字を自由に使える環境が必要である。本研究ではその一歩として，漢字の異字体の整理を行い，異字体同士がコンピュータ上で変換できるようなシステムを構築し，これを用いてより自由な漢字使用を可能とすることを目的とする。このようなシステムによって，仮名漢字変換の辞書にある語

彙を異字体に変換した辞書の作成，一般的によく使われる漢字から古い漢字への置き換え，逆に古い文献を現行の漢字に置き換え読み易くするといったことが可能となる［5］，［6］，［7］。

2　漢字の字体変化の歴史

漢字はそもそも中国大陸で中国語を表すために生まれ，原則として一文字で一つの事象を表す表意文字であるため，字数が極めて多くその形態も実に豊富である。現在我が国で使用されている漢字は中国大陸で生まれたものと日本で作り出された国字と呼ばれる和製漢字（畑，峠，働）がある。

2．1　中国での変遷

中国では，時代ごとに実に多くの漢字が生まれ変化を重ねてきた。

殷代（紀元前1500年頃）

漢字の原始形として最も古いのは，殷代の契文（甲骨文）と青銅器に鋳込まれた金文と呼ばれるものである。漢字はこのときから既に2種類が存在した。

戦国時代（紀元前403年～）

その後戦国時代を迎えた中国は，混沌とした時代を反映し，籀文，篆文，古文等の新たな種類の漢字が生まれる。金文も変化しながら残り，この時代は主に4種類の漢字が存在した。

漢代（紀元前206年～）

漢代になると金文の流れを受け継ぎながら篆書を簡略化し，より直線的な隷書が生まれる。そしてさらに隷書から楷書が生まれる。これは後漢の王次仲が作ったもので，点画をくずさない現在最も標準的な漢字である。

2．2　日本での変遷

日本においても貴族社会や商人社会など，社会の状況が変化する度に様々な字体が使われてきたが，明治となり中央集権が強まってからはその氾濫を防ぐ

ため漢字使用に制限を加えるようになった。

国字政策

日本に伝えられてからの漢字は，人々の手で様々な字体に変化した。あまりにも字数が増えたため，国字改良論や漢字制限論が盛んになり，有識者の間で様々な議論がなされた。

そして明治33年小学校令施行規則の中で，初めて法令によって漢字規制が行われた。大正12年には常用漢字1962字，及び簡易字体154字を選定し，この後の漢字の使用に大きな影響を与えた。戦後には漢字そのものが廃止の危機に追い込まれるような出来事もあったが，昭和24年，当用漢字の字体の標準を示す「当用漢字字体表」が公布され新聞社もこれを用いることを決めた。

昭和56年には内閣告示によって常用漢字1945字が定められ，また同時に法務省令戸籍施行規則の附則によって人名用漢字別表が定められた。人名用漢字別表は平成2年に改正され，常用漢字に加えて284字が選定されている。また，平成4年施行の小学校学習指導要領の学年別配当表に，常用漢字のうち1006字が教育漢字として示されている。

このように日本でも，漢字は次々と改められたり統廃合されるなどして，様々な字体が生まれてきた。

JIS漢字の制定

昭和53年これまで不統一であったコンピュータの文字コードが日本工業規格（JIS）によって定められ，「情報交換用漢字符号系 JIS C 6226-1978」として制定された。第一水準の漢字が2965字，第二水準の漢字が3384字含まれ，この第一次規格がワープロや日本語処理システムに適用された。

これは国字政策とは別に産業界等の要請によって規格が求められた経緯があるため，国字政策で認められた漢字を全て含んだ上で，人名地名によく用いられる漢字等も相当数が含まれている。

この頃から，「日本語ワードプロセッサ」が企業を中心に普及し始め，1980年半ばには第一次規格にさらに追加，変更を行った第二次規格の「JIS C 6226-1983」が制定された。これは，後にJIS情報X部門が新設された際に

「JIS X 0208 1983」と改称された。

平成2年に定められた第3次規格の「JIS X 0208 1990」には，第一水準2965字，第二水準3390字の合計6335字が含まれている。また同年，日本語処理システムの普及，情報内容の多様化などの背景から，情報交換用の文字の追加が必要となり，「JIS X 0208 1990」の補助として「情報交換用漢字符号 補助漢字 JIS X 0212 1990」が制定された。これに含まれる漢字は5801字である。平成9年には最新の「JIS X 0208 1997」が制定され，より明確な規定が打ち出されている。ここでは，字体の追加，変更は行われていない。

2.3 コンピュータ上での漢字使用の現状

以上を考えると，我々はJIS規格の制限，仮名漢字変換辞書の語彙による制限，さらには国字政策による制限という三つの制限の基にコンピュータ上で漢字を使用しているのが現状である。最後の制限は個人が自由な表現をする場合には問題としなければよいが，仮名漢字変換辞書の語彙による制限は快適に自由な表現を行う妨げとなり，JIS規格による制限は現状ではいかんともしがたい。つまり，大漢和辞典を通常の環境の人が持つ環境で表示可能な電子テキストとして表現する手段を持たないのが現状である。もちろん，専用のコードを制定して，それに対応するフォントを持てば表示可能であるが，そのテキストを加工するためには専用のソフトウェアが必要となる。また大漢和辞典には同じ漢字が重複して採録されているので，コードを制定する場合は注意が必要となる。

3 異字体の分類

本来異字体は日本の漢字5万字余りについて考えるべきではあるが，本研究ではコンピュータ上での異字体の取り扱いを考えているため，対象とする漢字をJISに存在する漢字に限定する。

そこで，JISに存在する第一水準2965字，第二水準3390字の合計6335字，補

助漢字の5801字の漢字について，異字体の調査を行った。

異字体とは，同一のルーツを持つ漢字であり，置き換えて使用しても差し支えがないものである。そのため，ここでは「花」と「華」のようなものは異字体とはせず，また数字の宛字についても異字体とはしていない。

3.1 分類のカテゴリー

JISの規格表にある漢字には，どの漢字がどの漢字の異字体であるという情報が記述されているが，どのような種類の異字体であるかの記述はない。しかし，一般の辞書では異字体に対して旧字や正字といった分類がなされ，それらの使用頻度も異なると考えられるので，ここでは異字体の分類を行う。これは，仮名漢字変換辞書では語彙に対して使用頻度が附随する場合が多く，このような辞書に対して異字体を追加する場合に分類によって頻度を変えるという操作を行いたいという要求があるからである。また，ある文章中の漢字を行ベースで異字体に自動変換するとき，複数の異字体の候補がある場合は1行が複数の候補行に変換されて，その中から希望の変換結果をユーザが選択する必要がある。このとき変換したい異字体の分類が決まっているのであれば，変換結果の候補行の数を減らすことができるためである。

異字体の分類を行うにあたって，字体に次の八つのカテゴリーを定義した。このカテゴリーは大漢和辞典（大修館書店）を中心として，他に大漢語林（大修館書店），漢和大辞典（小学館）を参照し，カテゴリー内の字数がグループとして成り立つ程度であるかを判断基準とした。分類は基本的には最も多くの漢字（48902字）が収録されている大漢和辞典全十二巻（大修館書店）を使用して字体を確認し，不明な場合は他の辞書を参考にした。以下に分類項目を示す。

・現行字体：現在広く一般的に用いられ，通常の出版物や新聞に使われている字体をさす。

・旧　　字：昭和24年に発表された『当用漢字字体表』で，当用漢字1850字のうち約400字に新しい字体が定められた。字画を簡略化したもの

が殆んどであるが，このときに簡略化されたものを新字体というのに対し，元の字体を旧字体という。例）藝（芸）　國（国）

・正　字：康熙字典を拠り所とし，そこで標準であると認められた点画を省略したり変えたりしない，正当と認められる漢字の字体をさす。例）冰（氷）　螢（蛍）

・俗　字：正字体ではないが，世間で通常用いられている字体をさす。例）舘（館）　做（作）

・古　字：特に古い起源をもつ文字のこと。辞書で「古文」と分類されている字体もこのカテゴリーに属する。例）龝（秋）　弍（二）

・本　字：漢字の元（ルーツ）となった漢字をさす。一般の正字よりもさらに字源的に忠実な形をしている。例）辦（弁）　鍼（針）

・略　字：漢字の点画などを省略して簡単にしたものや，その漢字と同意の漢字で字画を簡略化したものをさす。例）畄（留）　鼡（鼠）

・そ の 他：篆字や籀字，またはどのカテゴリーに属するか明確でない場合，その他に分類する。

同じ事象に対して二つ三つの漢字が作り出されるのは自然なことであって，その結果漢字には異なる漢字で同じ意味を持つものが存在する。例えば「幽玄」の「玄」と「夢幻」の「幻」は同じ事柄を表すといわれており，頻繁に目にする例としては「倉」と「蔵」がある。しかしほとんどの場合において後世に特有の慣用が定まり，日本では全く別の訓をつけているため，ここでは別の漢字として取り扱う。

3．2　異字体表の性質

異字体表を作成するとき，注意する点を考える。

AがBとCの異字体である場合に異字体表に両方を載せてしまうと，異字体表を使ってBからAという異字体に変換したのち変換をもう一度行って異字体を元の字体に戻す場合に，AからB，Cへの変換が可能となって誤ったCへの変換が含まれてしまう。

これを避けるためには，「一つの漢字はこの異字体表には高々1度しか出現してはならない」という性質を異字体表が持つ必要がある。

実際に二つの漢字が共通の異字体を持つ場合がある。このような異字体はこの性質を満たすため省略する必要がある。但しこのような場合でも明らかに一方の頻度が高いと思われるものは，頻度の少ないもののみを省略する。
この規則に従って，「冂」(「囲」の本字) と「冂」(「国」の古字または略字) の組み合わせ，「厂」(「庵」の略字)「厂」(「雁」の略字)「厂」(「歴」の略字) の3字の組み合わせなどは省略した。

一方のみを残した組み合わせは「×両 (輌の古字) ○両 (現行字体)」，「×嶌 (嶋の略字) ○嶌 (島の俗字)」，「×弍 (弐の略字) ○弍 (二の古字)」の三つである。

3.3 異字体表

調査の結果を，ピリオド (.) を区切り文字とする

現行字体．旧字．正字．俗字．古字．本字．略字．その他

という形式でテキストの表とした。当てはまる字がない場合，アスタリスク (*) を入れてある。第一・第二水準の漢字は「漢字：JISコード」もしくは「区／点」による表記，補助漢字は「区－点」による表記とした。また一つのフィールドに当てはまる漢字が複数ある場合，漢字と漢字の間を縦線 (|) で区切っている。その一部を表1に示す。

このような表記としたのは，補助漢字のフォントがまだ一般に整備されておらず，画面表示は可能な環境は存在するが，TeXやプリンタでの出力がまだ筆者の環境では不可能であるためである。

作成した表には2708字の漢字が含まれている。その内訳は第一・第二水準が1736文字，補助漢字が972文字である。字体別の文字数は表2の通りである。尚，現行字体を持たないものは149文字である。

表には亜・亞のような2文字の異字体の組み合わせから，剣・劍・劒・釼・劔・劒のように6文字が対応しているものまで存在する。漢字の異字体の組み

合わせは表3の通りである。

また，同一フィールドに複数の漢字があるものは，二つの漢字があるものが162組，三つの漢字があるものが11組である。

表1　異字体表（一部）

亜:3021.亞:5033.*.*.*.*.*.*.
悪:302d.惡:5828.*.*.*.*.*.*.
芦:3032.*.蘆:6943.*.*.*.*.*.
鯵:3033.鰺:724d.*.*.*.*.*.*.
圧:3035.壓:55a.*.*.*.*.*.*.
庵:3043.*.*.*.56-50.*.*.菴:683f
囲:304f.圍:5423.*.*.*.*.*.*.
為:3059.爲:602a.*.*.*.*.*.*.
医:3065.醫:6e50.*.*.*.*.*.*.
井:3066.*.*.*.*.井:5027.*.*.
育:3069.*.*.*.*.*.毓:5d5a
一:306c.*.*.*.弌:5021.*.*.*.
壱:306d.壹:5465.*.*.*.*.*.*.
稲:3070.稻:634b.*.*.*.*.*.*.
飲:307b.飮:5d3b.*.*.*.*.*.*.
淫:307c.*.*.*.*.*.*.婬:5535
隠:3123.隱:702c.*.*.16-20.*.*.*.
韻:3124.*.*.韵:7071.*.*.*.*.
卯:312c.*.*.*.*.*.*.夘:5249
欝:3135.*.鬱:5d35.*.*.*.*.*.
厩:3139.*.廐:567e.廄:567d.*.*.*.*.
叡:3143.*.*.*.24-59.*.*.睿:624f
営:3144.營:535b.*.*.*.*.*.*.
曳:3148.*.*.曵:5b2a.*.*.*.*.
栄:3149.榮:5c46.*.*.*.*.*.*.
衛:3152.*.衞:6a4c.*.*.*.*.*.
駅:3158.驛:7163.*.*.*.*.*.*.
円:315f.圓:5424.*.*.*.*.*.*.
煙:316c.*.*.*.*.*.*.烟:515d
艶:3170.*.*.艷:6766.62-76.*.*.62-77.
塩:3176.*.鹽:7345.*.*.*.*.*.
於:3177.*.*.*.*.于:5032.*.*.
奥:317c.奥:547c.*.*.*.*.*.27-84.
往:317d.*.*.徃:5748.*.*.*.*.
応:317e.應:5866.*.*.*.*.*.*.
欧:3224.歐:5d3f.*.*.*.*.*.*.
殴:3225.毆:5d58.*.*.*.*.*.*.
鴬:3229.鶯.7274.*.*.*.*.*.*.
岡:322c.*.*.崗:563e.*.*.*.*.
沖:322d.*.*.冲:5155.*.*.*.*.
穏:323a.穩:6353.*.*.*.*.*.*.
仮:323e.假:5071.*.*.*.*.*.*.
価:3241.價:512b.*.*.*.*.*.*.
歌:324e.*.*.*.哥:5327｜謌:6b68.*.*.*.
蝦:325c.*.鰕:7251.*.*.*.*.*.
嘩:325e.*.*.*.*.*.*.譁:6b76
峨:3265.*.*.峩:5636.*.*.*.*.
画:3268.畫:6141.*.*.*.*.*.*.
会:3271.會:5072.*.*.*.*.*.*.
解:3272.*.*.*.*.*.解:6b3b.*.
回:3273.*.囘:5145.*.*.*.*.*.
壊:3275.壞:5455.*.*.*.*.*.*.
:
蛯:6a48.*.*.*.*.*.*.:蝨6a49
袵:6a53.*.*.*.*.*.*.袵:6a54
祢:6a59.*.*.*.*.*.*.祢:472a
詑:6b42.*.*.譌:6b77.*.*.*.*.
諡:6b6a.*.*.謚:6b6b.*.*.*.*.
譖:6b7c.*.譖:6b7b.*.*.*.*.*.
貔:6c41.*.貙:6c40.*.*.*.*.*.
貳:6c48.*.*.貳:6c49.*.*.*.*.
躙:6d38.*.*.*.*.*.*.躪:6d39
輒:6d4c.*.*.輒:6d4d.*.*.*.*.
輌:6d52.輛:6d51.*.*.*.*.*.*.
迪:6d6c.*.*.廸:572f.*.*.*.*.
遖:6e24.*.*.遖:6e25.*.*.*.*.
邊:6e34.*.*.*.*.*.*.邉:6e35
鈩:6f4f.鑪:6f4e.*.*.*.*.*.*.
鑚:6f54.鑽:6f53.*.*.*.*.*.*.
閙:6f5e.*.鬧:7229.*.*.*.*.*.
濶:6f69.*.闊:6f68.*.*.*.*.*.
靱:7056.*.*.*.*.*.*.靭:7057

齏:706d.*.*.齏:706e.*.*.*.*.
飄:7128.*.*.飇:7129.*.*.*.*.
鬨:722a.*.*.鬩:6f62.*.*.*.*.
鰛:725c.*.鱚:725b.*.*.*.*.*.
鳧:726a.*.*.鳬:726b.*.*.*.*.
鴈:726e.*.*.鳫:726f.*.*.*.*.
鵄:7276.*.*.*.*.*.鵐:7277
鶫:7329.鶇:732a.*.*.*.*.*.
鷆:7337.鷁:7336.*.*.*.*.*.
麪:734f.*.*.76-74.*.*.*.麩:7350
鼇:7367.*.*.鰲:7262.*.*.*.*.
齧:7376.*.*.囓:5377.*.*.*.*.
30/69.*.*.*.16-02.*.*.*.
18/28.*.*.*.16-03.*.*.*.
34/22.*.*.*.16-04.*.*.*.
17/15.*.*.16-05.*.*.*.*.
38/51.*.*.*.16-07.*.*.*.
27/47.*.*.*.16-23 | 18-75.*.*.*.
20/05.*.*.16-27.*.*.*.*.
27/45.*.*.*.*.*.*.16-30
26/48.*.*.*.16-31.*.*.*.
21/93.*.*.*.16-33.*.*.*.
27/17.*.*.*.*.*.16-36.*.
36/67.*.*.16-48.*.*.*.*.
54/70.*.*.*.*.*.*.16-58
18/48.*.*.16-60.*.*.*.*.
37/56.*.*.*.*.*.*.16-61
70/84.*.*.*.*.*.*.16-66
27/87.*.*.*.*.*.*.16-70
55/43.*.*.*.*.*.*.16-82
43/01.*.*.*.16-85$ | $41-28.*.*.*.
55/44.*.*.*.*.*.*.16-86
17/02.*.*.70-40.17-04.*.*.70-65.
:
..*.*.*.*.*.51-71 | 52-23
..*.*.*.*.*.52-45 | 73-03
..*.*.*.*.*.52-59 | 71-56
..*.53-10.*.*.*.53-09
..*.*.*.*.*.53-40 | 53-41
..*.*.*.*.*.54-58 | 73-73
..*.*.*.*.*.55-47 | 55-89
..*.*.*.*.*.55-86 | 56-10
..*.*.*.56-56.*.56-27
..*.*.*.*.*.56-35 | 56-47
..*.*.*.*.*.56-60 | 57-63
..*.*.*.*.*.56-79 | 57-29
..*.*.*.*.*.57-28 | 57-75
..*.*.*.*.*.57-87 | 58-21 | 58-24
..*.*.*.*.*.58-38 | 58-39
..*.*.*.*.*.58-46 | 59-29
..*.*.*.*.*.60-56 | 60-57
..*.61-18.*.*.*.61-14
..*.61-19.*.*.*.61-17
..*.*.*.*.*.61-23 | 61-31
..*.*.*.*.*.61-32 | 61-33 | 61-34
..*.*.*.*.*.62-32 | 62-54
..*.*.*.*.*.62-84 | 63-01
..*.63-53.*.*.*.63-48
..*.*.*.*.*.63-55 | 64-31
..*.*.*.*.*.63-61 | 64-08
..*.*.*.*.*.64-73 | 64-90
..*.*.*.*.*.65-19 | 65-20
..*.*.*.*.*.65-25 | 65-35
..*.*.*.*.*.65-91 | 65-92
..*.66-49.*.*.*.66-52
..*.*.*.*.*.66-72 | 72-58
..*.*.*.*.*.67-04 | 67-08 | 67-22
..*.*.*.*.*.67-44 | 69-26
..*.*.*.*.*.67-45 | 67-74
..*.*.*.*.*.67-70 | 68-73
..*.*.*.*.*.68-38 | 69-88
..*.*.*.*.*.68-41 | 69-05
..*.68-89.*.*.*.69-38
..*.69-17.*.*.*.77-09
..*.69-20.*.*.*.69-53
..*.*.*.*.*.70-84 | 70-89
..*.*.*.*.*.71-07 | 71-08
..*.*.*.*.*.71-24 | 72-02
..*.*.*.*.*.71-34 | 71-35 | 77-12
..*.*.*.*.71-42.71-51
..*.*.*.*.*.71-47 | 71-71
..*.*.*.*.*.71-77 | 71-78
..*.*.*.*.*.72-47 | 72-88
..*.*.*.*.*.72-66 | 72-81
..*.*.*.*.*.72-93 | 73-53
..*.*.*.*.*.73-19 | 73-38 | 73-50

表2　字体別文字数

現行字体	旧字	正字	俗字
1095	288	89	262
古字	本字	略字	その他
100	55	25	794

表3　字体組合せ数

2文字	3文字	4文字	5文字	6文字
1057	153	22	7	2

4　異字体変換プログラム ktr

この表を使って，EUC コードで書かれたプレーンテキスト中の漢字の字体を行ベースで変換するプログラムの一つとして“ktr”を作成した。

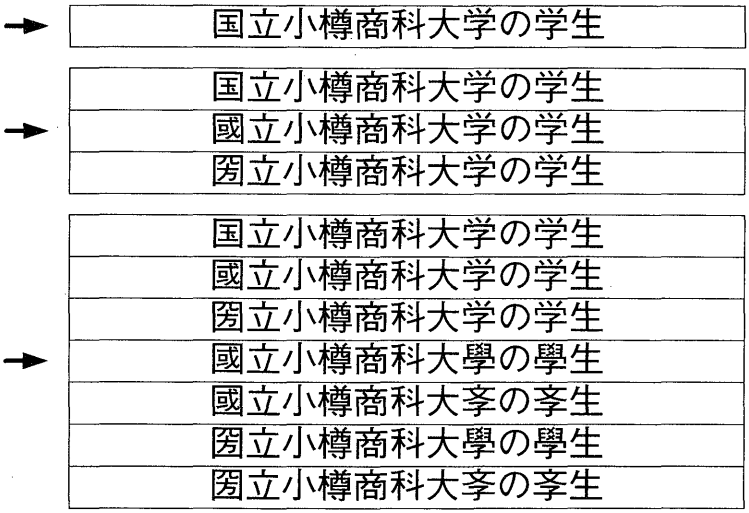
ktr は基本的に ktr －[数字] ＋[数字] {＋[数字]}*の形式で起動される。数字には1から8までが一つ以上入り，各々が順に1：現行字体，2：旧字，3：正字，4：俗字，5：古字，6：本字，7：略字，8：その他を意味する。ktr を実行すると入力行は－で指定した字体から＋で指定した字体に変換される。複数の変換がある場合は複数の行に分かれて出力がなされる。尚，＋[数字]の数字で示される字体は同じ優先度となり，＋[数字]が複数ある場合は右側の＋[数字]の優先度が高くなり，最も優先度が高い行が出力される。例えば，ktr －1 ＋23は現行字体を旧字体と正字体へ変換せよということを示し，ktr －1 ＋2 ＋3 は現行字体を旧字体または正字体へ変換せよ但し旧字体と正字体の両方がある場合は正字体のみに変換せよ，ということを示す。

また，同じ行中に同じ文字が出現した場合は，同じ文字への変換のみを出力することとした。例えば「小樽商科大学の学生」という行を ktr －1 ＋27によって変換する場合「学」を変換させる候補は「學」と「孝」の二つあるが，もし「大学」の「学」を「學」に変換したのであれば，次の「学生」の「学」も同じく「學」に変換させるという仕組みである。これによって，変換の不統一を防いでいる。

以下，プログラムの開発の順を追って，プログラムの設計方法について説明を行う。

4.1 ktr Version.0

このプログラムはktrの原形となった初期のバージョンで，現行字体を第一・第二水準の他の字体全てに変換したものを出力する目的で作られた。ここでは表をif( ) else if( )文の並びに変換し，行中の漢字を行の先頭から一つ一つif文の並びで処理を行い，変換すべき漢字であれば変換した行を作成していく方法をとる。変換が複数ある場合は複数の行が作成され，続く処理はこの変換した行全ての変換が済んだ位置以降に対して同様に行われる。また変換すべき漢字が見つかった場合は，行の最後まで同じ漢字が存在するかをチェックして，存在すれば同時に変換を行うことによって変換の不統一を防いでいる（図1参照)。



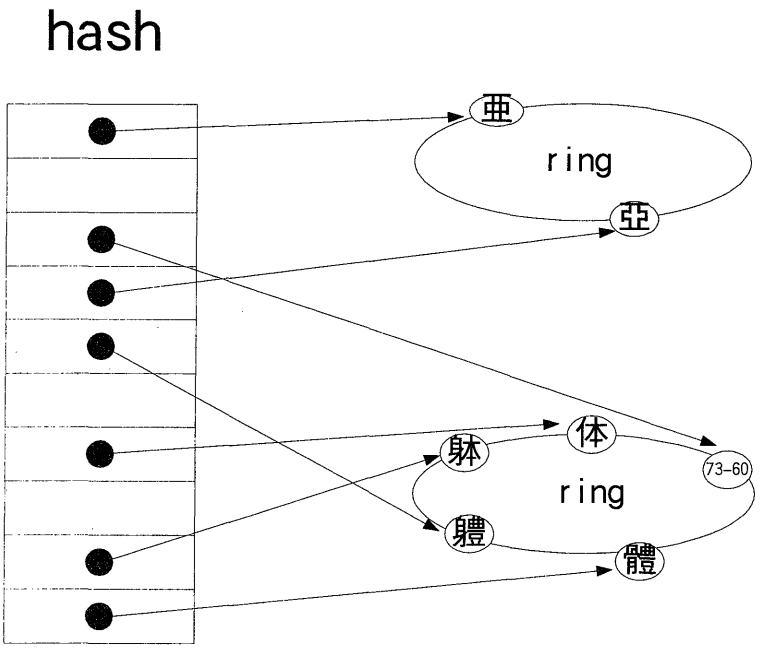
図－1

本手法の欠点は，変換対象かどうかの決定のために，行中の漢字1字ごとに多数のif文の判断を行うため処理速度が遅いこと，及び変換結果を行として積み重ねていくため記憶域が無駄になり，長い行に対しては実行ができないことである。しかしながら，仮名漢字辞書の語彙を変換するためには十分利用可能であった。

4．2 ktr Version.1

Version.0の欠点を補うために，表を起動時に読み込みハッシュ表の形での内部表現を行って任意のカテゴリー間での変換を可能としたものがVersion.1である。また，記憶域の管理のため変換結果はリストの形で保持することとして，リストの上でロックシステムを用いて変換の不統一を防ぐこととした。

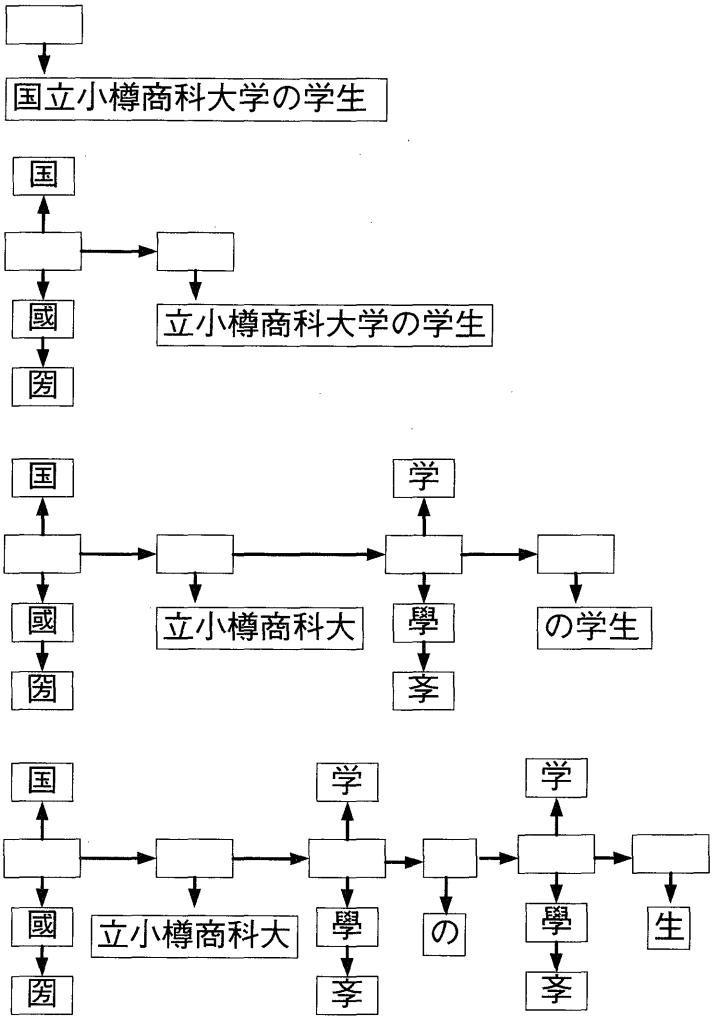
ハッシュは図2のような形をしており，ハッシュ関数H( )に対して漢字Aを与えたH(A)が左のハッシュテーブルの位置(添字)を返し，そこにring形式で繋がっている漢字へのポインターが格納されている。ポイントされた先の漢字がAであれば，あとはringを辿って変換すべきカテゴリーの漢字が存在するかを調べればよい。また，ハッシュであるからH(A)が異なる漢字に対して同一の値を返す場合があるので，その場合のオーバフロー処理も行う。こ



図－2

れによって，多数の if 文による比較とは桁違いの検索速度が得られる。

また，変換結果のリストは図 3 に示すような形で変化していく。図中の四角い空白がリストのノードであり，そこから変換文字列がポイントされている。

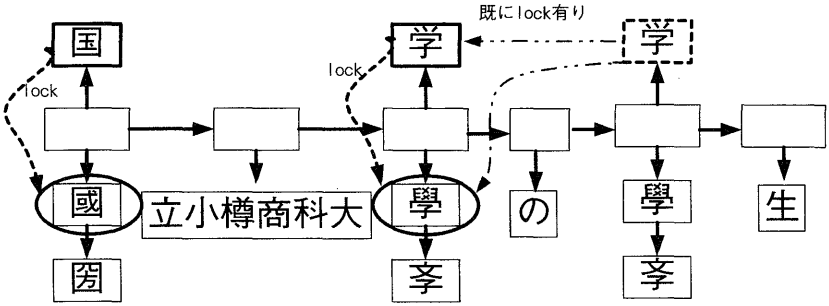


図－3

最初は変換前の文字列がポイントされており,最初の変換によって,「国」が「國」と「圀」に変換されると，次のリストの形になる。これは，変換前の文字がリストの上に，変換された結果の文字が下にこれもリストの形で付加されて，残った変換対象の文字列が次の四角い空白のリストに繋がれる。そして，この残った変換対象の文字列のリストに同様の操作を行う。Version.0とは異なり，同じ変換文字列があるかを一度に調べることは行わない。そして，変換結果として最後のリストが得られる。

このリストをトラバースして最終出力の行を生成するわけであるが，トラバースの際は変換結果の不統一を防ぐための方法が必要となる。この方法を図4に示す。これは，変換されたリストをトラバースするときに，同じ漢字が既に使用されているかどうかを最初にチェックして，そうでない場合に限って元の漢字がどの漢字に変換されているかをロックとして記憶しておき，同じ漢字が既に使用されていればロックされた漢字をそのまま出力する方法である。これによって変換の不統一を避けることができる。この例では，前から順番にトラバースが進み，「国」が「國」にロックされ，「学」が「學」にロックされ，最後の「学」の時点でのチェックで「学」が既に使用されているので，ロックされた「學」をそのまま使用することとしている。

本手法の欠点は，ハッシュ表をプログラム起動時に作成するため表のファイルが実行時に必要である点と,ハッシュの記憶域に多少の無駄がある点である。



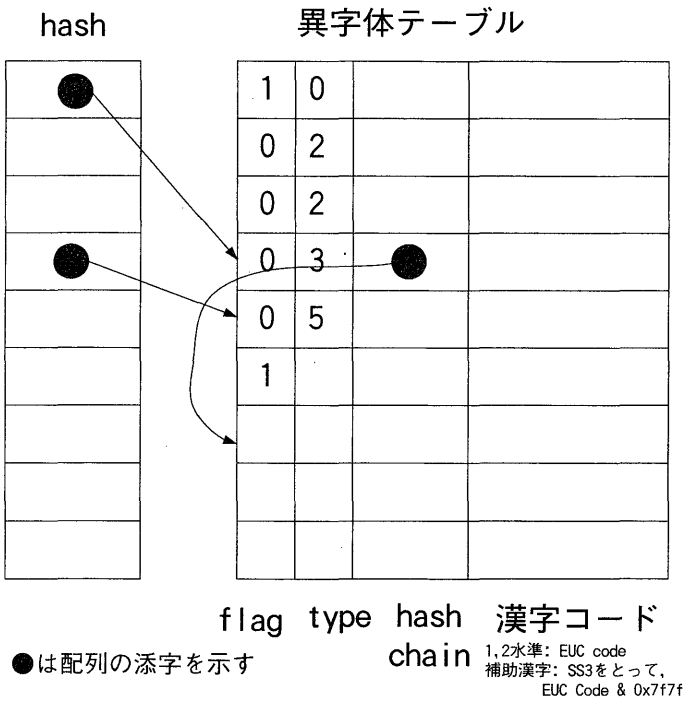
図－4

4.3 ktr Version.2

Version.1の欠点を補うため，表をコンパクトにして，プログラムのコンパイル時に読み込める形式にした。異字体テーブルは，

```
struct ktable {
unsigned short flag: 1;
unsigned short type: 3;
unsigned short chain: 12;
unsigned char code [2];
} ;
```

という形式とし，漢字1文字のために4バイトを使うだけで済ませることができた。図5にハッシュ表と異字体テーブルを，表4に実際の異字体テーブルの



図－5

一部を示す。異字体テーブルではflagが1の行から次にflagが1となる行の手前，もしくはそのような行がない場合はテーブルの終りまでが一つの漢字の異字体ファミリーを表している。これはVersoin.1のringと同様のものである。typeはカテゴリーを示す数値が入る。chainはハッシュ表におけるオーバフロー処理で用いられる。codeは漢字コードが第一・第二水準についてはEUCコードそのものを，補助漢字についてはECUコードに0x7f7fを論理積演算したもの（各バイトのMSBを0とする）を入れている。実行時にこのテーブル上でハッシュ表を作り，構成されたハッシュ表を使うこととした。またこの表をアクセスするルーチンを独立させ，他のプログラムからも使用可能とした。これは，EUC漢字1文字と変換対象のカテゴリーを引数として変換後の文字を返すものであり，汎用的な利用が可能となった。プログラム自体もこのインタフェースを用いて書き直された。

表4　ktr Ver.2の表異字体テーブル（一部）

```
{1,0,4095,0xb0,0xed} ,/* 壱:16/77*/
{0,1,4095,0xd4,0xe5} ,/* 壹:52/69*/
{1,0,4095,0xb0,0xec} ,/* 一:16/76*/
{0,4,4095,0xd0,0xa1} ,/* 弐:48/01*/
{1,0,4095,0xb0,0xe9} ,/* 育:16/73*/
{0,7,4095,0xdd,0xda} ,/* 毓:61/58*/
{1,0,4095,0xb0,0xe6} ,/* 井:16/70*/
{0,5,4095,0xd0,0xa7} ,/* 丼:48/07*/
{1,0,4095,0xb0,0xe5} ,/* 医:16/69*/
{0,1,4095,0xee,0xd0} ,/* 醫:78/48*/
{1,0,4095,0xb0,0xd9} ,/* 為:16/57*/
{0,1,4095,0xe0,0xaa} ,/* 爲:64/10*/
{1,0,4095,0xb0,0xcf} ,/* 囲:16/47*/
{0,1,4095,0xd4,0xa3} ,/* 圍:52/03*/
{1,0,4095,0xb0,0xc3} ,/* 庵:16/35*/
{0,4,4095,0x58,0x52} ,/*:56-50*/
{0,7,4095,0xe8,0xbf} ,/* 菴:72/31*/
{1,0,4095,0xb0,0xb5} ,/* 圧:16/21*/
{0,1,4095,0xd4,0xda} ,/* 壓:52/58*/
{1,0,4095,0xb0,0xb3} ,/* 鯵:16/19*/
{0,1,4095,0xf2,0xcd} ,/* 鰺:82/45*/
{1,0,4095,0xb0,0xb2} ,/* 芦:16/18*/
{0,2,4095,0xe9,0xc3} ,/* 蘆:73/35*/
{1,0,4095,0xb0,0xad} ,/* 悪:16/13*/
{0,1,4095,0xd8,0xa8} ,/* 惡:56/08*/
{1,0,4095,0xb0,0xa1} ,/* 亜:16/01*/
{0,1,4095,0xd0,0xb3} ,/* 亞:48/19*/
```

更なる変更として，リストの文字列の持ち方を工夫して，より少ない記憶域で動作可能とする予定である。

4.4　変換例

現行字体への変換は1種類しかないので，ktr　−2345678　+1とすることで，

任意の文章を現行字体へ変換することが可能である。

例えば，昔の落語のテキストを現行字体へ変換させてみると，「それは氣の毒だなァ。おめえにそんな散財をさせようと思つて，俺は呼込んだ譯ぢやァねえ。(三遊亭圓生 「駱駝」より)」が，

「それは気の毒だなァ。おめえにそんな散財をさせようと思つて、俺は呼込んだ訳ぢやァねえ。(三遊亭円生 「駱駝」より)」へと変換される。

5 異字体表の応用

5.1 仮名漢字変換辞書への適応

ktr を用いて仮名漢字変換システム wnn の辞書にある語彙の変換を行い，異字体の辞書を作成することが可能である。これによって wnn で利用できる字体の幅が増加する（本文もこの辭書を使って書かれている)。また，辞書の頻度をカテゴリーによって変更することができるので，旧字体の語彙は現行字体の半分の頻度に，正字体の語彙は 1 / 4 にといった細かい制御が可能となる。他の仮名漢字変換システムの辞書に対しても同様の手順を踏むことにより，利用できる字体を増やしかつ頻度情報の制御も行える。但し，ここで増えた異字体語彙が全て有用であるとは限らない。

ここでは pubdic [†1] の辞書にある語彙について ktr を使った変換を行い，辞書の語彙の字体を増やす実験を行った結果を示す。尚，symbol.u，tankan.u はここでは除いて考える。基本辞書 kihon.u の一部を表 5 に，それを変換した結果の一部を表 6 に示す。また，語彙増加数の結果を表 7 に示す。この結果から，異字体によって60% 程度の異字体語彙の増加が見込めると考えられる。もっと多くの語彙をもった辞書について調べてみると，253605個の語彙に対して第一・第二水準での異字体語彙の増加が101141個，補助漢字での異字体語彙の増加が72073個得られ，68.3% の増加となった。異字体語彙の増加は真の語

† 1 Wnn バージョン 3 の時代（1987年～1989年）に元 ASTEC の橘 浩志氏が世話人となって行われた pubdic プロジェクトの成果物。

彙の増加といえるかというと疑問の向きもあるが，仮名漢字変換辞書の語彙に対する機械的操作によって異字体語彙が6割程度増加することが判明した。

表5　pubdic（一部）

よみ	語彙	品詞	頻度
あ	あ	接続詞，感動詞	33
あ	あ	ワ行五段	16
あ	あ	ラ行五段	1163
あ	あ	カ行五段	1
あ	逢	ワ行五段	10
あ	会	ワ行五段	204
あ	開	カ行五段	2
あ	空	カ行五段	166
あ	合	ワ行五段	603
あ	在	ラ行五段	1
あ	遭	ワ行五段	1
あ	編	マ行五段	29
あ	飽	カ行五段	1
あ	明	カ行五段	1
あ	有	ラ行五段	29
あーく	アーク	名詞	1
あーす	アース	名詞	1
あーち	アーチ	名詞	4
あーちすと	アーチスト	名詞	3
あーてぃすと	アーティスト	名詞	3
あーびとれーしょん	アービトレーション	サ行(する)＆名詞	2
ああ	ああ	サ行(する)＆名詞	29
ああいう	ああいう	連体詞	2
あい	愛	サ行(する)＆名詞	33
あい	藍	名詞	1

表6　pubdicに第一・第二水準で変換を行った結果（一部）

よ　み	語　彙	品　詞	頻度
あ	會	ワ行五段	136
あいかわらず	相變わらず	副詞	6
あいこく	愛國	名詞	0
あいしょう	愛稱	名詞	1
あいじ	愛兒	名詞	0
あいず	合圖	サ行(する) & 名詞	0
あいたい	相對	サ行(する) & 名詞	0
あいつ	相繼	ガ行五段	0
あいどく	愛讀	サ行(する) & 名詞	0
あいらく	哀樂	名詞	0
あえ	會え	一段	0
あえん	亞鉛	名詞	0
あかし	證	名詞	0
あかし	燈	名詞	0
あかつき	曉	名詞	1
あかぬけ	垢抜け	一段	0
あかり	燈り	名詞	0
あかり	燈	名詞	0
あが	擧が	ラ行五段	4
あき	穐	名詞	28
あきさめ	穐雨	名詞	0
あきばれ	穐晴れ	名詞	0
あく	惡	名詞	1
あくしつ	惡質	名詞	0

表7　pubdic への適応結果

辞書名	辞書中の語彙数	第一・第二水準での増加数	補助漢字での増加数	総増加数	増加分の%
bio.u	465	162	129	291	62.6%
chimei.u	4693	2041	1686	3727	79.4%
computer.u	900	109	62	171	19.0%
jinmei.u	2493	731	902	1633	65.5%
kihon.u	22707	7891	5122	13013	57.3%
koyuu.u	252	107	81	188	74.6%
setsuji.u	879	196	111	307	34.9%
special.u	26	6	1	7	26.9%
合　計	32415	11243	8096	19337	59.7%

5.2　文字検索への応用

さらなる応用としては，漢字の異字体をアルファベットの大文字と小文字の関係のような扱いとすることが考えられる。例えば，通常文章内の「真実」という単語を検索すると，「眞實」と旧字で表記されているものは検索の対象外となる。アルファベットで大文字小文字の区別をせずに検索する場合の一つの手法として，大文字に全てを変換して，それに対して大文字で検索を行う手法がある。しかし日本語の場合，大文字を現行字体と考えて同様の手法を適用すると，現行字体に変換する手間がアルファベットを大文字に変換するための toupper ( )の処理と比べてかかる上，現行字体を持たないと分類されたものが存在するため，その分の処理も組み入れる必要が生じる。

そこで，正規表現を使った検索システムに異字体を取り込み，「真実」の検索を「(真｜眞)(実｜實)」と正規表現の検索に自動的に置き換えれば，「真実」と共に「眞實」も検索対象となり，異字体同士を同じ文字として扱うことが可能となる。このような機構を日本語処理プログラムに組み込むことで，異字体の利用を妨げることなく日本語の処理が行えるようになる。

6 まとめ

本システム以外でも異字体への取り組みは沢山ある。例えば wnn6では，仮名漢字変換に異字体へ変換するという枠組みを付加して異字体入力の便宜をはかっている。これは異字体を持つ語彙を仮名漢字変換辞書に取り込むことはせず，変換した結果の語彙中の文字に対して異字体への変換をさらに行うというものである。また，市販のワープロソフトでは“あいまい検索”によって異字体を含めた検索が行えるものもあり，文字検索への応用は既に一部実用化している。このように，日本語を扱う環境では異字体への取り組みが不可欠であるという認識が徐々に広まっている。

本研究では異字体同士の変換を行うシステムのための異字体表が持つべき性質を定め，実際に表を作成し，それを使った変換プログラムを作成して pubdic への適用を行い，仮名漢字変換辞書の持つ語彙からその異字体語彙が機械的に 6 割程度生成されることを確認した。

今後も JIS 第三水準，第四水準の出現に合わせて異字体表を update すると共に，漢字の自由な使用に向けての取り組みを行っていきたい。

謝辞

ktr Version.0 で用いた漢字データを提供して頂いた，塚本高之さん（元東京理科大学工学部），表の作成をお手伝い頂いた田中智子さん（元本学三谷ゼミ学生，現第一生命）を始めとする三谷ゼミの学生諸君に感謝します。

参　考　文　献

[1] 日本工業標準調査会，JIS X 0208:1997，日本規格協会，1997.
[2] 日本工業標準調査会，JIS X 0212:1990，日本規格協会，1990.
[3] 芝野耕司，JIS漢字辞典，日本規格協会，1997.
[4] 太田昌孝，いま日本語が危ない，丸山学芸図書，1997.
[5] 田中智子，漢字異字体変換システムの開発，小樽商科大学卒業論文，1998.
[6] 三谷和史，田中智子，漢字異字体変換システムの開発，情報処理北海道シンポジウム'98，pp.69－70，1998.
[7] 三谷和史，JIS漢字異字体変換システムとその応用，情報処理学会デジタルドキュメント研究会，DDS-13-3，pp.17－24，1998.