デジタル・ドキュメント 13－3
（1998．7．17）

JIS 漢字異字体変換システムとその応用

三谷和史
mit@mit-s.otaru-uc.ac.jp
小樽商科大学・商学部・社会情報学科

概 要

日本語を表すために用いられる漢字は，大漢和辞典にある 48902 字の漢字と，それ以外の若干の漢字で，おおよそ 5 万語程度といわれている．そのうち，日本工業規格 (JIS) によって規定された漢字は 12136 字である．通常我々が電子的にテキストを記述する場合はこの JIS に従った漢字を使用している．漢字の中には異字体と呼ばれる，同じ起原を持つが異なった表記をもち，同じ意味を表すものが多数含まれる．

自由に漢字を駆使するためには，異字体を自由に駆使できる必要がある．また古い文献を電子化する場合はできる限り原文にある字をそのまま使って保存すべきである．また場合によっては現在使われている字体に変換する必要もある．さらに，電子化されたテキスト中を検索する場合も，異字体を含めて検索したいという要求も高まっている．

本論文では異字体の分類，JIS で規定された漢字の異字体調査の結果，異字体間の変換を行なうシステムおよびその応用について述べる．

Translating System for JIS KANJI Character Alternatives

Kazufumi MITANI
mit@mit-s.otaru-uc.ac.jp
Department of Information and Management Science
Faculty of Commerce, Otaru University of Commerce

Abstract

In Japanese, we have 48902 KANJI characters according to DAIKANWA dictionary and some more are not in it. The total is over 50000 KANJI characters.

We usually use JIS KANJI code that defines 12136 KANJI characters for writing/exchanging electronic texts. However, there are many alternative KANJI characters (same roots, different character but same meaning) in our KANJI characters land.

To express ourselves freely in Japanese, we need to have a good command of alternative KANJI characters.

When we convert old texts to electronic form, we ought to preserve the characters in original text. However, sometimes we need to convert the characters into present style. And there are more demands to searching electronic text in regardless of alternative KANJI characters.

In this paper, we define and classify alternative KANJI character, describe the results of our examination of alternative KANJI characters in JIS KANJI character code, and introduce translating system we developed to handle alternative KANJI characters and its application.

1 はじめに

漢字は中国からもたらされたものである．そして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，通常は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5万余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なわれるため，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,「辭書」と書きたいときに「辞書」しか仮名漢字変換システムの辞書に登録されていなければ，使いたい「辭書」を探し出すためにいちいち単漢字辞書を開いて一字ずつ入力する手間がかかり，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漢字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，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，漢字の異字体を整理し，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にすることを目的とする．このようなシステムがあれば，仮名漢字変換の辞書にある漢字を異字体に変換した辞書を作成したり，一般的によく使われる漢字から古い漢字に置き換えたり，逆に古い文献を現在よく使われる漢字に直し，簡単に読めるものに変換することが可能となる[5, 6]．

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たものである．原則として一文字で一つの事象を表す表意文字なので，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．漢字はこの時から既に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なわれた．大正12年には常用漢字1962字，および簡易字体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，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たのである．

JIS 漢字の制定

昭和53年これまで不統一だ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」として制定される．第一水準の漢字が2965字，第二水準の漢字が3384字含まれ，これが第一次規格でワープロや日本語処理システムに適用された．

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，変更した第二次規格の「JIS C 6226-1983」が制定される．これは，後にJIS情報X部門が新設された際に「JIS X 0208 1983」と改称される．

平成2年に定められた第三次規格の「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年に最新の「JIS X 0208 1997」が出て，より明確な規定が打ち出されている．

3 異字体の分類

本来異字体は日本の漢字5万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して考える．

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，異字体の調査を行なった．

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支えはないものである．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は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なう．これは，仮名漢字変換辞書では単語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な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．

異字体の分類を行なうにあたって，字体に次の 8 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 文字) 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．

- 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』では，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ほとんどであるが，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だと認められた，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) 倣 (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) 穐 (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) 当 (留) 単 (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，その他に区分することにした．

また，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ある．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ので，別の漢字として取り扱う．

3.2 異字体表の性質

ＡがＢとＣの異字体である場合に表に両方を載せてしまうと，表を使ってＢからＡという異字体に変換したのち，変換をもう一度行なって異字体を元の字体に戻すときに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てばよい.

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しなければいけない.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,「口」(「囲」の本字)と「口」(「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)の三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(輛の古字)○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)若しくは(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3に示す.

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1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1: 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・劔・釼・劒・劍のように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2のようになっている.

表2: 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

また,同一字体に複数の漢字があるものは,二つの漢字があるものが162個,三つの漢字があるものが11個である.

4 異字体変換プログラム ktr

この表を使って,EUC で書かれた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.プログラムでは表を漢字をキーとするハッシュかつ異字体同士をリング状にリンク

表 3: 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.*.*.*.*. 丼: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. 廏:567d.*.*.*.*.
叡:3143.*.*.*.24-59.*.*. 睿:624f
営:3144. 營:535b.*.*.*.*.*.*.
曳:3148.*.*. 曵:5b2a.*.*.*.*.
栄:3149. 榮:5c46.*.*.*.*.*.*.
衛:3152.*. 衞:6a4c.*.*.*.*.*.
詠:3153.*.*.*.*.*.*. 咏:5269
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. 奧: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.*.*.*.*.
:
:

濶:6f69.*. 闊:6f68.*.*.*.*.*.
靱:7056.*.*.*.*.*.*. 靫:7057
齏:706d.*.*. 齑:706e.*.*.*.*.
飄:7128.*.*. 颪:7129.*.*.*.*.
閧:722a.*.*. 閙:6f62.*.*.*.*.
鯉:725c.*. 鱚:725b.*.*.*.*.*.
鳧:726a.*.*. 鳬:726b.*.*.*.*.
鴈:726e.*.*. 鳫:726f.*.*.*.*.
鴟:7276.*.*.*.*.*.*. 鵄:7277
鵜:7321.*.*.*.*.*.*. 鵞:7322
鶇:7329. 鶫:732a.*.*.*.*.*.*.
鷆:7337. 鷁:7336.*.*.*.*.*.*.
麪:734f.*.*.76-74.*.*.*. 麩:7350
鼈:7367.*.*. 鰲:7262.*.*.*.*.
齧:7376.*.*. 嚙: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.
35/43.*.*.17-20.*.*.*.*.
42/69.*.*.*.*.*.*.17-29
22/02.*.*.*.*.*.*.17-34
29/04.*.*.*.17-35.*.*.*.
:
:

..*.*.*.*.*.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
..*.*.*.*.*.64-13|64-30|64-43
..*.*.*.*.*.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
..*.*.*.*.*.73-79|73-87
..*.*.*.*.*.74-35|75-51
..*.*.*.*.*.74-37|74-74
..*.*.*.*.*.74-42|77-02
..*.*.*.*.*.74-59|75-01
..*.*.*.*.*.74-66|74-72|74-93
..*.*.*.*.*.74-76|74-84

した形式で格納した．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，ということを意味する．

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．

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ない，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 (本文もこの辞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なえる．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 の辞書にある語彙について ktr を使った変換を行ない，辞書の語彙の字体を増やす実験を行なった結果を示す．尚，symbol.u，tankan.u はここでは除いて考える．結果を表 4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 1，2 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．

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なう手法がある．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 の処理と比べて手間がかかる上，現行字体を持たないと分類されたものが存在

表 4: pubdic への適応結果

辞書名	辞書中の語彙数	1,2 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%

するため，その分の処理も組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」も検索対象とさせて，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なえるようになる．

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を付加して異字体での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をせず，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な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なえるものもあり，文字検索への応用は既に実用化している．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．

本研究では異字体同士の変換を行な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 への適用を行ない，仮名漢字変換辞書の持つ語彙からその異字体語彙へ機械的に 6 割程度増すことを確認した．

今後も JIS 第 3 水準，第 4 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組みを行なっていきたい．

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.