

# 国立国語研究所学術情報リポジトリ

Promotion of the standardization of character encoding and the operation of a database

|       |                                                                                                                                                                     |
|-------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| メタデータ | 言語: jpn<br>出版者:<br>公開日: 2019-03-25<br>キーワード (Ja):<br>キーワード (En):<br>作成者: 高田, 智和, 井手, 順子, 虎岩, 千賀子, TAKADA, Tomokazu, IDE, Junko, TORAIWA, Chikako<br>メールアドレス:<br>所属: |
| URL   | <a href="https://doi.org/10.15084/00002197">https://doi.org/10.15084/00002197</a>                                                                                   |

# 行政用文字の調査研究

## ——汎用電子情報交換環境整備プログラム——

高田 智和      井手 順子      虎岩 千賀子  
(国立国語研究所)    (国立国語研究所)    (国立国語研究所)

### キーワード

電子政府, 文字コード, 外字, 地名

### 要 旨

さまざまな行政手続をインターネットで行う「電子政府」を構築するためには、氏名、住所、法人名などの固有名に使われる文字をも含め、行政情報処理で必要とされる文字をコンピュータで扱えるような環境を整えなければならない。国立国語研究所・情報処理学会・日本規格協会では、行政情報処理で必要とされる文字の調査研究（汎用電子情報交換環境整備プログラム）を実施している。この調査研究において、住民基本台帳ネットワーク統一文字、戸籍統一文字、登記統一文字を検討し、行政用文字の文字コード規格（JIS X 0213, ISO/IEC 10646）によるカバー率を明らかにした。また、漢和辞典に掲載されていない文字について、地名資料による文字同定を進めている。

### 1. 経済産業省委託「汎用電子情報交換環境整備プログラム」

1995年にマイクロソフト社からWindows95が発売されると、それまで業務用であったパソコンは家庭にまで普及し、広く社会に行きわたるようになった。それとともに、インターネットが急速に発達し、IT社会と呼ばれる世の中が到来した。これを背景に、全国の行政機関でも、さまざまな手続に使う文書の電算化を進め、インターネットを利用した電子申請のシステムが整えられつつある。コンピュータを用いて円滑な情報交換ができるような「電子政府」「電子自治体」を実現するためには、住民の氏名や住所、あるいは、法人の名称や所在地などを記載するために必要な文字についても、「電子政府」「電子自治体」を支える基盤の一つとして整備していく必要が生じる。

2002年に、世界最先端のIT国家を目指すとして、首相官邸は「e-Japan 重点計画－2002」を発表した。その中では、行政の情報化に関わる共通基盤の整備として、「文字情報・コードの整備」が挙げられている。行政の文字情報処理では、人名・地名・法人名等の固有名を扱うため、コンピュータに標準的に登載されている文字コード規格（JIS X 0208, JIS X 0212など）で定める文字セットでは文字が足りない、いわゆる「外字」問題が日常的に発生している。このような「外字」がどのくらいあるのか、そして、行政の文字情報処理を視野に入れて、将来的に文字コード規格を拡充するとしたら、どのような文字が必要となるのかという観点での検討は、「電子政府」「電子自治体」を実現するための基盤研究として重要である。そこで、「e-Japan 重点計画

-2002」に対応する行政用文字の調査研究として、経済産業省委託「汎用電子情報交換環境整備プログラム」（以下、本プログラム）が発足した。

本プログラムには、国立国語研究所・情報処理学会・日本規格協会が参加し、それぞれの特性を活かして共同研究を進めている。国立国語研究所は、ことばとしての文字を扱い、文字の属性や対応関係の記述、使用例の分析による文字同定など、基礎調査を課題とする。情報処理学会は、情報としての文字を扱い、国語研究所の基礎調査を経て整理体系化された文字情報データを元に、行政の文字情報処理に必要な文字を検討し、文字コード標準化のための活動（国際文字コード規格 ISO/IEC 10646 への追加提案）を行う。日本規格協会は、デザインとしての文字を扱い、字形上のデザイン的なゆれから生じる誤解を防ぐため、文字デザインの検討を行い、統一的なデザインコンセプトによる平成明朝体グリフの制作が課題である。

本プログラムは、平成 14 年度から 17 年度までをフェーズ 1 とし、総務省の住民基本台帳ネットワーク統一文字（以下、住基統一文字）約 21,000 字と、法務省の戸籍統一文字約 55,000 字を検討の対象とした。また、平成 18 年度からはフェーズ 2 とし（平成 20 年度末終了予定）、現在は、法務省の登記統一文字約 67,000 字を検討中である。

本プログラムの調査研究を進めるにあたり、国立国語研究所は、住基統一文字と戸籍統一文字の照合を行い、同字別字の判定を経た上で文字の目盛りを作り、目盛りとなるひとつひとつの文字に対して、部首・画数・読みなどの基本情報、国語施策・戸籍行政に関する行政の文字情報、JIS X 0213（国内規格）・ISO/IEC 10646（国際規格）の文字コード情報、大漢和辞典文字番号などの辞書情報を付与し、「漢字情報データベース」として蓄積している（図 1）。「漢字情報データベース」は、本プログラムでの文字台帳を意図して構築したものであるが、将来的には、漢字研究においての活用も期待できるものである。

| 漢字情報データベース |                                                                                                                                                                                                                                                                         |
|------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 学          | 部首 39<br>総画数 8<br>読み ガク・カク・コウ・キョウ・まなぶ・おしえる<br>国語施策 常用漢字<br>戸籍番号 079270 戸籍行政 第2042号通達<br>住民基本台帳ネットワーク統一文字コード 5896 住民基本台帳ネットワーク使用情報 *****<br>JIS X 0213 2004 追加点 1-19-59<br>JIS X 0213 2004 追加情報<br>JIS X 0213 2004 追加区分 0<br>ISO/IEC 10646-1:2000 5896<br>大漢和辞典文字番号 6974 |
| 参照文字情報     |                                                                                                                                                                                                                                                                         |
| 戸籍         | 学 学 学 学 学 学 学 学 学 学<br>079270 079690 079910 150920 150960 150980 150740 146790                                                                                                                                                                                          |
| 住民         | 学 学 学 学 学 学 学 学 学 学<br>5896 5897 5898 5899 5900 5901 5902 5903                                                                                                                                                                                                          |
| JIS X 0213 | 学 学 学 学<br>1-19-56 1-53-60 1-53-61 2-13-72                                                                                                                                                                                                                              |
| 大漢和辞典      | 学 学 学 学 学 学 学 学 学 学<br>13124 13431 13453 13455 5696 6974 7003                                                                                                                                                                                                           |

図 1 「漢字情報データベース」の検索結果出力例

## 2. 行政用文字

### 2.1. 住民基本台帳ネットワーク統一文字

住民の居住関係を公証し、選挙人名簿の作成、国民健康保険や国民年金の被保険者としての資格の管理、学齢簿の作成など、地方自治体が行う各種サービスに用いられる住民基本台帳は、現在の IT 社会と呼ばれるコンピュータネットワークが出現する以前から電算化が始まり、2002 年から住民基本台帳ネットワークシステム（以下、住基ネット）が本格運用されている。住基ネットに登録されている統一文字は 21,039 字であり、仮名・ラテン文字・符号・記号などの非漢字を除くと、漢字は 19,435 字である<sup>1</sup>。これは、各地方自治体が住民基本台帳を電算化してコンピュータで扱うために必要であろうと考えられる文字の集まりである（図 2）。具体的には、日本国内のコンピュータに標準的に登録されている JIS 漢字（JIS X 0208 の第 1・第 2 水準漢字、JIS X 0213 の第 3・第 4 水準漢字、JIS X 0212 の補助漢字）、地名に用いられる文字（地名外字）、各地方自治体にシステムを納めている情報機器メーカー各社の拡張文字から成る。メーカー拡張文字は、情報機器メーカー各社が顧客からの要望に応じてシステムに取り込んだ文字であり、パソコンが普及する以前から蓄積されてきたものである。この点から、住基統一文字は、情報機器メーカー各社が電子計算機の時代から情報機器で扱ってきた文字の集合体と言ってもよい。

|                    |               |                      |
|--------------------|---------------|----------------------|
| JIS 漢字<br>13,141 字 | 地名外字<br>233 字 | メーカー拡張文字等<br>6,061 字 |
|--------------------|---------------|----------------------|

19,435 字

図 2 住基統一文字（漢字）の構成

また、住基統一文字については、総務省で行った住基ネットでの文字使用頻度報告から、使用実態を知ることができる<sup>2</sup>。それによると、

使用件数あり 11,643 字

使用件数なし 7,792 字

である。システム内で稼働している漢字は全体の 6 割程度であり、残り 4 割の漢字は、システムに登録されているだけで、実際には使われていないものである。このように、あるシステムを動かすために「必要とされる文字」と、実際に使われていて「必要な文字」とでは隔たりがある。住基統一文字に 19,435 字の漢字があるからといって、そのすべての文字が住民基本台帳で使われている文字ではない、ということの特記しておきたい。

### 2.2. 戸籍統一文字

戸籍の電算化とオンラインでの使用にあたり、法務省は戸籍統一文字を定め、戸籍簿に使うことができる文字を網羅している。その数は 56,040 字にのぼり、漢字は 55,267 字である<sup>3</sup>。戸籍簿に記載可能な文字については、戸籍行政における法務省民事局長通達によって取り決めがなされ

ており、現在、戸籍簿に記載することのできる主な漢字を示すと以下のようになる。

1. 常用漢字表の通用字体
2. 戸籍法施行規則別表第二「漢字の表」に掲げる字体
3. 漢和辞典に正字として記載されている漢字
4. 漢和辞典に同字・古字・本字として記載されている漢字
5. 漢和辞典に俗字として記載されている漢字
6. 第 5200 号通達別表に掲げる字体

ここで言う「漢和辞典」とは、現在市販されているものを指し、ゆるやかな定義である。戸籍行政において拠り所として重視されてきた『康熙字典』も含まれる。常用漢字と人名用漢字を合わせても 3,000 字に満たず、『康熙字典』の見出し字が 50,000 字程度であることから、戸籍統一文字の大部分が漢和辞典の見出し字であることは容易に想像できる。後述するように、現在市販されている中型規模以上の漢和辞典の見出し字が大部分を占めており、戸籍統一文字は現代の漢和辞典の集合体であると言ってよい（図 3）。

| 漢和辞典の見出し字       |                |  |
|-----------------|----------------|--|
| 常用漢字<br>1,945 字 | 人名用漢字<br>983 字 |  |
| 55,267 字        |                |  |

図 3 戸籍統一文字（漢字）の構成

住基統一文字が、実務に使用している道具である情報機器の登載文字に基づいているのに対し、戸籍統一文字は、実務のために設けられたルールの範囲内で、理論上可能となる臨界点を目指す網羅主義によって設計されている。住民基本台帳と戸籍とは、同じようなものを扱っているようであっても、電算化のためのシステムに登載する文字セットの設計思想には大きな違いがある。

また、戸籍は各地方自治体において電算化の途上にあるため、住民基本台帳のように、戸籍統一文字の使用実態を知ることにはできない。しかしながら、住基ネットで使用件数のある文字が 12,000 字程度であることから、これと同程度の水準であろうと推測される。戸籍統一文字もまた、戸籍簿で実際に使われている文字の集まりと同一視できないことを特記しておく。

### 2.3. 登記統一文字

登記事務の電算化を目的として、法務省は登記簿の記載事項に必要とされる文字を網羅した「登記統一文字」を定め、平成 20 年度から運用される次期登記システムへの登載準備を行っている。登記統一文字は、戸籍統一文字 56,040 字に、全国の登記所が保有する外字群 12,031 字を加えた 68,071 字の集合<sup>4</sup>である（図 4）。

|                    |                      |
|--------------------|----------------------|
| 戸籍統一文字<br>56,040 字 | 「登記固有文字」<br>12,031 字 |
|--------------------|----------------------|

68,071 字

図 4 登記統一文字の構成

「登記固有文字」は、全国約 600 ヶ所の登記所の現行登記システムにおいて、氏名・法人名・住所など登記簿の記載事項に用いられる外字（JIS X 0208 外字）を収集し、同字をまとめた上で戸籍統一文字と照合を行い、戸籍統一文字に含まれていない文字として抽出されたものである。したがって、すべての「登記固有文字」は、いずれかの登記所のいずれかの登記簿の記載事項を表記するために制作された文字、すなわち、使用実績がある文字ということになる。

法務省より提供された「登記固有文字」の一覧表によって、複数の登記所が制作した外字の上位 10 字を示すと表 1 のようになる。上位 10 字だけを見ても、「登記固有文字」の特徴を知るには十分である。「才・藤・真・均・雯・静・裴」はそれぞれ「第・藤・真（眞）・均・政・静（靜）・裴」の異体字である。「ㄣ」は「ど（登）」、「し」は「し（之）」の変体仮名である。また、「厶」は韓国国字とされる文字である。このように、「登記固有文字」には異体字・変体仮名・日本以外の地域で用いられている漢字が含まれており、「登記固有文字」の集合を特徴づけている。

表 1 複数の登記所が制作した外字の上位 10 字

| 順位 | 外字 | 制作登記所数 |
|----|----|--------|
| 1  | 才  | 142    |
| 2  | 藤  | 141    |
| 3  | ㄣ  | 138    |
| 4  | 真  | 96     |
| 4  | 均  | 96     |
| 6  | し  | 90     |
| 7  | 厶  | 84     |
| 8  | 雯  | 76     |
| 9  | 静  | 69     |
| 10 | 裴  | 56     |

本プログラムでは、漢字以外の文字は扱わないことを原則としてきた。しかし、「登記固有文字」に見るように、行政の文字処理では、主に人名で変体仮名を扱うことがあり、今後の課題として変体仮名の扱いが挙げられる。また、日本以外の地域で用いられている漢字としては、「龙（龍）・华（華）・郑（鄭）」などの中国簡化字も「登記固有文字」に見られる。国際化の時代

を迎え、日本国内に仕事や居住の場をもつ外国人が増えつつあることから、行政用文字を検討する場合であっても、日本国内で使われている漢字に限定することなく、漢字文化圏の漢字に視野を広げて調査研究を進めていくことが、今後ますます必要になってくる。

### 3. 文字の対応

次に、文字の対応関係の記述について述べる。文字の対応関係を記述する理由は、文字情報交換の信頼性と安定性を高め、システム間をつなぐ将来的な「電子政府」を構築するにあたり、現状行われている各府省庁の個別システム（住基ネットなど）に登載されている個々の文字同士の関係（同字であるか別字であるか）を検討し、基礎対応表を作り上げ、個別システム間でのデータ交換の可能性を拓くための基盤として整備することが、必要な基礎調査たりうると考えたからである。具体的には、住基統一文字・戸籍統一文字・登記統一文字それぞれに収録されている一々の文字について、それが3種集合に共通して収録されている文字なのか、どれか2種に共通し残り1種には存在しない文字なのか、あるいは、ただ1種集合にのみ存在する文字なのか、3種集合を文字という目盛りで横断的に貫く場合の対応関係を検討していくことになる。

一般に、文字同士の対応関係を捉えようとするとき、文字の形の単位をどのように設定するかには、さまざまな考え方がある。例えば、現代の「漢字使用の目安」を謳う常用漢字表には、「字種」「字体」の考え方が示されている。これを図示すると次のようになるであろう。

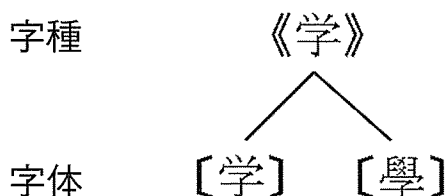


図5 「字種」と「字体」の関係

図5では、下段に新字体の「学」と旧字体の「學」を配置している。両者はそれぞれ「字体」が異なる。「学」「學」は字音ガク、意味は「マナブ」で、両者は文字の形が異なっても音義は共通している。また、「マナブ」という語に対して、「学ぶ」であっても「學ぶ」であっても語彙的な意味は変わらない。このように形は異なるが、音義が共通する「字体」同士の組を、常用漢字表では「字種」と呼んでいる。「字種」は図形化不可能なもので、上図では、便宜上「学」を代表させて《》で括って表し、また、字体は〔〕で括って表している。この考え方では、「字種」は「字体」の上位にあたる概念となり、同一の字種《学》には二つの字体〔学〕と〔學〕が所属していると捉えられる。すなわち、「学」と「學」は、字種としては同じ文字であるが、字体としては別な文字となる。

さらに、常用漢字表が描く文字の形に関する考え方は、「字体」の下位概念として「字形」を

設定する。「字形」「字体」の考え方は、常用漢字表に限るものではなく、文字研究者から概ね支持されているものである。「字形」「字体」の枠組みは、音声学・音韻論で言う「音声」「音韻」の關係に相似し、具体的・個別的に、目に見える形で実現された文字の形を「字形」、人間の脳裏にあって、社会的な約束事として存在する抽象字形を「字体」と呼ぶ。これを図示すると次のようになる。

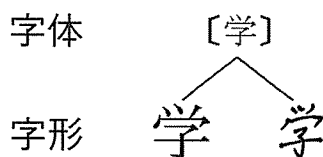


図6 「字体」と「字形」の關係

図6では、明朝体で表現された「学」も、ペンで手書きされた「学」も、「字体」としては同じ「学」であるが、実現形としての「字形」が異なっている。「字体」は抽象字形であるから、厳密には図形化不能なものであると考えられる。「字体」の概念については、骨組み説、代表字形説、社会共通の基準説などさまざまであるが、実現形である個々の「字形」によって文字コミュニケーションが成り立つ以上、我々の脳裏に「字体」が共有されていることは疑いないであろう。上図では、便宜上、明朝体の「学」を代表させて〔〕で括って表している。図5でも同様である。

ここまで見てくると、文字の形は3段階に階層化することができる。

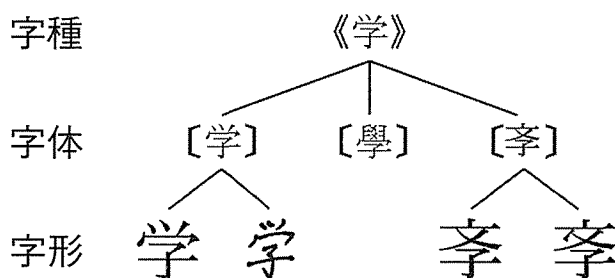


図7 「字種」「字体」「字形」の階層構造

図7の階層構造からわかるように、「同じ文字」と言う場合には、「字種」「字体」「字形」のどのレベルで「同じ」なのか、文字の形の単位を考慮する必要が生じる。「学」と「学」が「同じ文字」というのは、「字種」のレベルでのことであり、「字体」のレベルでは「別な文字」である。同様に、「字形」のレベルでも「別な文字」である。

文字の対応を考えると、「字体」のレベルで行うことが理想的であろう。実現形である「字形」



のレベルで行うと、無限のパターンを扱わなくてはならない。漢和辞典における見出し字の設定や、日本での文字コード標準化における例示文字の設定でも、明記されてはいないが、概ね「字体」レベルを想定して、これまで検討が行われてきたようである。

しかし、住基統一文字・戸籍統一文字・登記統一文字の3種集合には、同一「字体」内における「字形」の差を持つような複数の「字形」が、集合の要素として収録されている。例えば、図7の右下にある「𪛗」と「𪛗」とは、住基統一文字では、「𪛗」はB00B,「𪛗」は6588と、それぞれ別コードを与えられ、住基ネットのシステムでは「別な文字」として運用されている。

「𪛗」と「𪛗」には、筆押さえの有無という実現形としての図形上の差が存する。この筆押さえは、明朝体活字のデザイン上との差と捉えられることもある。このように、行政の実務で用いられているシステム内では、微差とも言える違いを持つ複数の「字形」が、それぞれ独立した「別な文字」として扱われていることがあるため、いわゆるデザイン差をも考慮して、文字の対応関係を検討し、基礎対応表を記述することとした。

この基準による同字別字の判定を経て得られた基礎対応表は、行政の実務で必要とされている「字形」の和集合を取り出したものである。住基統一文字と戸籍統一文字との対応では、「字形」のレベルで58,700字に集約される。現在検討中の登記統一文字を加えると、69,000字程度になる見込みである（図8）。

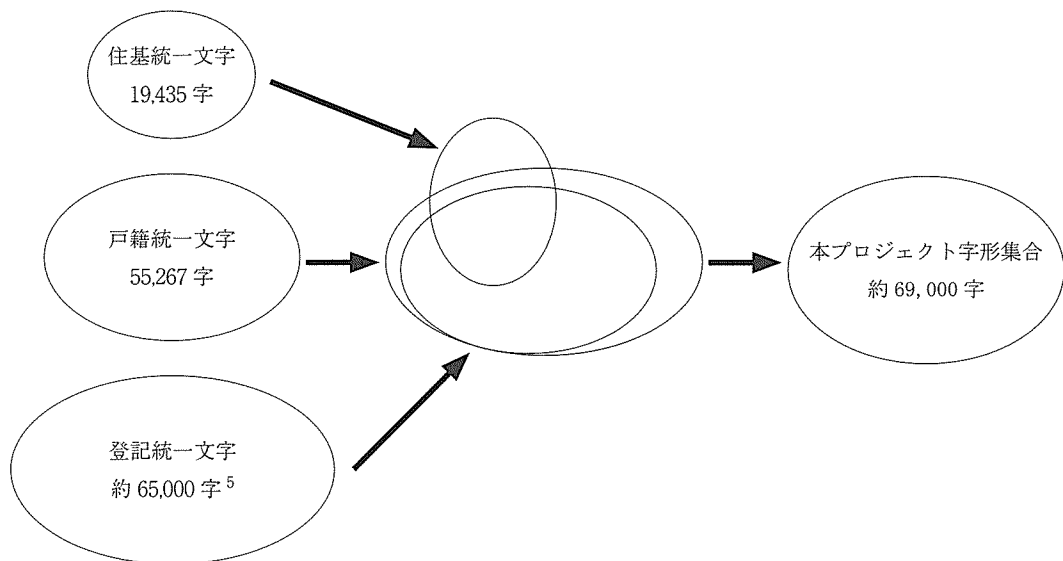


図8 行政用文字（漢字）の和集合

#### 4. 行政用文字と文字コード規格

本プログラムの課題の一つに、行政の文字情報処理において必要な文字の中で、現行の文字コード規格に採録されていない「外字」がどのくらいあるのか、というものがある。そこで本節では、戸籍統一文字と住基統一文字が、国内規格のJIS X 0213:2004を用いてどの程度表現できる

のか、JIS X 0213:2004 の規則に則って厳密に対応づけを行った場合の結果を示す。

JIS X 0213:2004 規格票の「6.6.2 字体の実現としての字形」では、「一つの字体の図形的実現としては、デザインの差に基づく複数の字形が考えられるが、この規格はそれらを互いに区別しない」とし、デザインの差の例として常用漢字表の「(付) 字体についての解説 第1 明朝体活字のデザインについて」から用例を引用している。前述のように、住基統一文字にはデザインの差に該当する複数の字形が採録されており、JIS X 0213:2004 の規則では、それらは規格票の例示字体内におけるデザインのゆれとして同一の区点位置で表現することになる。例えば、住基統一文字では別コードを与えられ、「別な文字」として扱われている「班 (73ED)」と「班 (B5A1)」は、双方とも 1-40-41 (班) で表現され、互いに区別されない (図 9)。図 7 で示した階層構造にしたがうと、「字形」レベルのゆれに該当する。

班 (73ED) → 1-40-41 班  
班 (B5A1) → 1-40-41 班

図 9 デザインのゆれ

また、JIS X 0213:2004 には 199 の包摂規準<sup>6</sup>が規定されている。これは、符号化を行う対象文字の字体と規格票の例示字体との差が、包摂規準に記述された範囲内のものであれば、例示字体と同一の区点位置で表現できるようにするというものである。例えば、戸籍統一文字には、それぞれ別番号を与えられ、「別な文字」とされている「葛 (352760)」と「葛 (349920)」とがある。これらを JIS X 0213:2004 で表現しようとする、規格票の例示字体と同じ「葛 (352760)」はそのままに 1-19-75 (葛) で表現できるが、規格票の例示字体と異なる「葛 (349920)」の方は、包摂規準連番 150 を適用することで 1-19-75 (葛) に「包摂」して表現することになる。結果として、「葛 (352760)」も「葛 (349920)」も同一の区点位置 1-19-75 (葛) で表され、互いに区別されない (図 10)。JIS 包摂規準が記述する差は、図 7 の階層構造では「字体」レベルのゆれに該当するものと見られる。

葛 (352760) → 1-19-75 葛  
連番 150 曷 曷 葛 (349920) → 1-19-75 葛

図 10 包摂規準適用例

このような操作を経て、住基統一文字・戸籍統一文字と JIS X 0213:2004 との対応をまとめる

と、表2のようになる。JIS X 0213:2004の規則に則ると、住基統一文字の64.9%、戸籍統一文字の21.9%が表現できる。この結果だけを見るかぎり、行政の実務で「必要とされる文字」を扱いきることを目的とすれば、国内規格では文字が大幅に足りないと思なされる。また、前述の住基ネットで使用件数のある11,643字を範囲としても、JIS X 0213:2004のカバー率は77.0%であり、「必要な文字」を扱う際にも不足が生じる。

表2 JIS X 0213:2004 との対応

|                     | 住基統一文字 | 戸籍統一文字 |
|---------------------|--------|--------|
| 例示字体の字形と同じ字形        | 10,050 | 9,755  |
| 例示字体の字形とデザインの差のある字形 | 654    | 336    |
| JIS 包摂規準が適用される字体の字形 | 1,900  | 2,003  |
| JIS X 0213:2004 外字  | 6,831  | 43,173 |
| 計                   | 19,435 | 55,267 |

では、同様に、国際規格のISO/IEC 10646:2003を当てはめてみるとどうなるであろうか。ISO/IEC 10646には、CJK(中日韓)統合漢字・CJK 統合漢字拡張A・CJK 統合漢字拡張Bが収録され、実装されれば約70,000字の多漢字環境が実現される。住基統一文字・戸籍統一文字とISO/IEC 10646:2003との対応をまとめたものが表3である。ISO/IEC 10646:2003では、住基統一文字の93.1%、戸籍統一文字の95.5%が表現可能である。住基ネットで使用件数のある11,643字の範囲では、94.4%にあたる10,995字が表現できる。一方、ISO/IEC 10646:2003で表現できない648字は、国際規格への優先順位の高い追加候補となり得る。

また、ISO/IEC 10646:2003にも、JIS X 0213:2004ほど明確ではないが、「包摂」の概念が存在する。ISO/IEC 10646:2003でも、「葛」と「葛」とは互いに区別されず、同一の符号位置で表現される。「葛飾区」「葛城市」のように、固有名で区別して用いられることがあるため、行政の文字処理で不足する文字と言う場合には、現行の文字コード規格において「包摂」されるものにむしろ大きな問題があると言える。ひと口に足りない文字と言っても、現行の文字コード規格の枠組みを当てはめてながめてみると、既存の文字セットに収録されていない「外字」を指す場合と、既存の文字セットの収録文字に「包摂」される繰り返しのバリエーションを指す場合と、二通りがある。特に後者は、無限ともいえる字形パターンの記述につながる可能性があり、現行の文字コード規格と、それに基づく文字情報交換の根幹に関わる問題である。それと同時に、明らかに異なる字体同士を同じものと見なすことは、使用する人間の側の感覚の問題として、抵抗感なり違和感なりが生じることが考えられる。文字包摂の許容範囲について、認知上の観点から検討していくことも、異体字に関わる研究の一領域として行われるべきであろう。

表3 ISO/IEC 10646:2003 との対応

|                          | 住基統一文字 | 戸籍統一文字 |
|--------------------------|--------|--------|
| ISO/IEC 10646:2003 で表現可能 | 18,093 | 52,756 |
| ISO/IEC 10646:2003 外字    | 1,342  | 2,511  |
| 計                        | 19,435 | 55,267 |

## 5. 辞書にない文字

### 5.1. 辞書による文字同定

本節では、住基統一文字・戸籍統一文字と漢和辞典・漢字字典の見出し字との対応について述べる。文字の集合について、特徴や傾向を見出したり、評価を行ったりする場合には、辞書との照合が基礎調査として行われる。また、個々の文字に関しては、辞書の掲出字との照合は、それが辞書に採録されているような典拠のある文字か否かによって、文字の属性や素性を明らかにする文字同定の第一段階である。

明治以降、漢和辞典・漢字字典は数多く生み出されている。初学者のための学習用辞書を除けば、一般的な辞書は小学の流れを汲み、漢籍に用いられる漢字を土台にし、若干の「国字（和製漢字）」を加えて見出し字としている。本プログラムでは、以下の4種の辞書の掲出字と住基統一文字・戸籍統一文字との文字同定を行っている。また、これら4種の辞書を総称して同定辞書と呼ぶ。

1. 諸橋轍次『大漢和辞典』修訂第2版、大修館書店、2001年
2. 尾崎雄二郎・都留春雄・西岡弘・山田勝美・山田俊雄『大字源』、角川書店、1992年
3. 上田万年・岡田正之・飯島忠夫・栄田猛猪・飯田伝一『新大字典』、講談社、1993年
4. 芝野耕司『増補改訂 JIS 漢字字典』、日本規格協会、2002年

『大漢和辞典』は収録字数50,000字の国内最大の漢和辞典である。利用は国内外に及び、学術的権威と信頼の高さには定評がある。『大字源』は収録字数12,000字の中型規模の漢和辞典であるが、「国字一覧」を付録として掲載するなどの特色がある。『新大字典』は戦前から広く一般に普及していた『大字典』の改訂版である。収録字数21,000字の中型規模の漢和辞典にあたるが、多様な異体字を収録する特長をもつ。これらの現在市販されていて、それぞれの特色が異なる中型規模以上の漢和辞典に加えて、『増補改訂 JIS 漢字字典』も同定辞書に選定した。本プログラムで検討する文字群は、電子情報交換の場で扱われるものであることと、いわゆる JIS 漢字が明治以降の実用の漢字に関する調査研究の蓄積を基盤としていることによる。

同定辞書との照合結果を表4に示す。住基統一文字の92.2%、戸籍統一文字の97.9%が同定辞書に見出され、辞書典拠を有する文字である。戸籍統一文字に至っては、大部分を辞書掲載字に結びつけることができる。しかし、一方で、行政の文字処理において必要とされる文字の中には辞書にない文字が含まれており、これらの辞書非掲載字について文字情報を収集・蓄積していくことが、文字同定の第二段階である。

表 4 同定辞書との照合結果

|         | 住基統一文字 | 戸籍統一文字 |
|---------|--------|--------|
| 同定辞書にあり | 17,914 | 54,094 |
| 同定辞書になし | 1,521  | 1,173  |
| 計       | 19,435 | 55,267 |

## 5.2. 地名使用文字

地名・人名など固有名の表記に、辞書にない文字が使用されていることは、従来から知られている。地名使用文字については、現地調査によっていくつかの文字を同定するに至っている。ここでは二つの事例を紹介する。

〔櫓〕

「櫓」は岩手県一関市（旧花泉町）の地名「ひし輪田（ひしわだ）」の表記に用いる文字である。JIS X 0213:2004 においても、この地名を典拠として 2-15-74 の区点位置に採録されている。ただし、JIS 規格票の例示字体は 2 点しんによろである（図 11）。

2-15-74  
2f6a    afea  
f48a    (AB0F)    櫓

図 11 JIS 規格票

一関市花泉支所に残る「岩手縣西磐井郡金澤村字櫓輪田絵圖」には、1 点しんによろの文字が記載されている（図 12）。法務局の書類電算化にあたって、平成 15 年 1 月の盛岡地方法務局一関支局から旧花泉町への確認文書には、字（あざ）名の正式名称に 1 点しんによろの文字を用いている（図 13）。このほかの文書でも 1 点しんによろの文字を用い、役場では 1 点しんによろの文字を安定的に使用していたものと見られる。

また、「原野単価表」には「櫓」が用いられている（図 14）。「ひし輪田」在住の 80 代の男性は、住所の表記に「櫓」を用いていた。行政上の「正式な」表記と、現地に住んでいる方の「日常的な」表記とが乖離することは、往々にして起こる現象である。

櫓

図 12 絵圖（明治期）

櫓輪田

図 13 法務局への回答（H15）

櫓輪田

図 14 原野単価表

「𪛗」と「𪛗」のように、はこがまえとしんによろとが交替する異体字の類型があり、明治期

の絵図から確認できるしんによろにつくる文字を定点とするならば、「櫃」はしんによろがはこがまえに變形したものと考えられる。この場合、「米櫃」の「櫃」とは同形異字である。また、JIS 規格票の例示字体は、常用漢字表外字が活字字体として慣習的に2点しんによろにつくることから、それに準じて「拡張旧字体」化が行われた結果と見られる。

はこがまえとしんによろが交替するならば、逆に、しんによろの文字は「櫃」が變形したものである可能性もある。「ひつ」と「ひし」は音が類似することから、「ひしわだ」は「ひつわだ」の転訛であると推論することも許されよう。つまり、「櫃」を定点とするならば、しんによろの文字の方がはこがまえの「櫃」から變形を経たものと見なすこともできる。

「櫃」の字体の變容についての推定をまとめると、図15・図16のようになる。

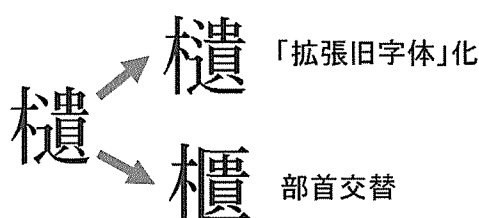


図15 「櫃」の字体の變容（パターンA）

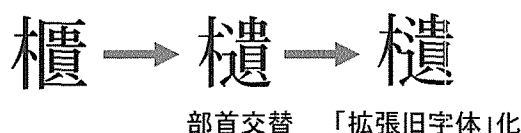


図16 「櫃」の字体の變容（パターンB）

## 〔盈・盈〕

「盈」「盈」はいずれも福島市の地名「こぼれ田（こぼれた）」の表記に使用し、住基統一文字・戸籍統一文字に採録されている。易林本節用集では、「盈」に「コボル」と訓をあてていることから、また、字形が類似していることから、この二つの文字は「盈」の異体字であると考えられる。

地名「こぼれ田」がどのように表記されているのか、明治期から近年の電算化以前までの手書きによる土地関係資料の用例を表5に示す。一見して、さまざまな字体が歴史的に使われてきたことがわかる。類型をまとめると、おおよそAからFまでの六種である。

⑤土地台帳は、明治期から昭和40年代まで、継ぎ足しながら使い続けられてきたものである。よって、その時どきの役場担当者によって字体が異なり、大変興味深い資料を提供している。康熙字典体にあたる「盈」を書くのは、昭和23年ごろの担当者だけである。

行政用文字「盈」の起源は、明治期の③地籍から使用が認められるDであると考えられる。また「盈」は、⑤土地台帳の昭和41年ごろの担当者が書いたEを反映しているものと推測される。『書道大字典』（伏見冲敬，角川書店，1974年）では、陳智永千字文谷氏本や唐温泉銘などにDが見え、康熙字典体よりも古くから使用されている字体であることが確認できる。Eも同様に、隋宋玉艶墓誌に見える。「盈」「盈」とも、個人の書き癖として結論付けてしまうことはできない。

現在福島市では、戸籍や住民基本台帳に、「𪛗」(D),「𪛗」(E)の双方が使われているため、二つながら地名表記の文字として許容している。「こぼれ田」在住の50代の女性は、住所の表記に「𪛗」(D)を用いるが、極めて類似する字体「𪛗」との関係に思い至ることはほとんどないようである。当面「𪛗」(D)も「𪛗」(E)も使い続けられることであろう。

このように、行政用文字に含まれる辞書非掲載字には、従来活字として一般的に存在しなかったもの、つまり、筆写文字として存在していた文字が、行政処理の電算化によって、コンピュータで扱うことのできる「デジタル文字」になったものが多数あると推測される。「こぼれ田」のような地名表記の文字であれば、公共性があるため、現地調査によって文字同定の根拠を示すことができる。しかし、人名表記に由来する文字の場合、個人情報であるため、全く情報が得られず、用例に基づく文字同定が行き届かないのが現状である。

表5 「こぼれ田」の表記

|                         | A         | B         | C         | D         | E         | F        |
|-------------------------|-----------|-----------|-----------|-----------|-----------|----------|
| ①丈量帳<br>(明治17)          |           | 𪛗         |           |           |           |          |
| ②地籍図<br>(明治期)           |           | 𪛗         |           |           |           |          |
| ③地籍<br>(明治期)            |           |           |           | 𪛗         |           |          |
| ④字切図<br>(明治～戦前)         |           |           |           | 𪛗         |           |          |
| ⑤土地台帳<br>(明治～<br>昭和40代) | 𪛗<br>明治21 | 𪛗<br>明治26 | 𪛗<br>昭和23 | 𪛗<br>昭和31 | 𪛗<br>昭和41 | 𪛗<br>昭和? |

## 6. むすびにかえて

小論では、汎用電子情報交換環境整備プログラムにおける行政用文字の基礎調査について述べてきた。最後に、繰り返しになるが、文字の対応を考える場合、これまでのように「字体」のレベルで行うことが、やはり理想的であろう。デザインの違いのような「字形」の微差に差異を認め、「別な文字」として扱うことは、運用上においても、処理上においても、多くの問題を引き起こす要因となるであろう。しかし、一方で、行政処理の実務では、人名・地名・法人名などの固有名を扱うため、「字形」の微差のようなものであっても、それが文字の弁別に機能し続け、コンピュータによる電算化の時代になってなお維持されていることは、紛れもない事実である。微差によって「別な文字」とされるものが増え続けるとしても、常用漢字表など現今の漢字政策

が固有名を対象外としていることや、文字の形を捉えるための学術的理論の深化が近年あまり進んでいないことなどを鑑みると、「字体」のレベルだけに限定して文字の対応を組み、本プログラムにおいて漢字整理や漢字制限に類することを行うのは極めて困難である。今後、本プログラムによって得られた知見や成果をもとに、各方面で更なる検討や議論が行われることを期待するものである。

#### 注

- 1 本プログラムで検討を行った住基統一文字は、「住民基本台帳ネットワークシステム統一文字 確定版（2002年5月8日版）」に基づく。
- 2 総務省より提供のあった住基統一文字の使用頻度表による。個人情報に関わるため、人名の使用例まで知ることはできないが、固有名の使用文字の実態を記述した国内唯一の頻度表である。
- 3 本プログラムで検討を行った戸籍統一文字は、「平成16年4月1日付け法務省民一第928号(改正平成16年12月6日付け法務省民一第3464号) 民事局長通達」に基づく。
- 4 本プログラムで検討を行った登記統一文字は、平成19年8月現在の「暫定版」に基づく。平成20年度からの次期登記システム登載に向け、法務省において最終調整を行うため、小論で示す登記統一文字の字数等の情報には、今後若干の変動が起こりうる。
- 5 登記統一文字の漢字は、65,000字程度と見込まれる。
- 6 JIS X 0213:2004 規格票「6.6.3 漢字の字体の包摂規準」参照。

#### 参考文献

- 池田証寿（2001）「『標準コード用漢字表（試案）』と JIS 漢字」『国語文字史の研究 6』259-289, 和泉書院
- 石塚晴通（1984）『圖書寮本日本書紀 研究篇』汲古書院
- 江守賢治（1986）『解説字体辞典』三省堂
- 樺島忠夫（1975）「文字体系の構造」『計量国語学』75, 13-22, 計量国語学会
- 小池和夫・府川充男・直井靖・永瀬唯（1999）『漢字問題と文字コード』太田出版
- 小宮山博史（2000）「書体設計と JIS 包摂規準」『人文学と情報処理』26, 48-66, 勉誠出版
- 笹原宏之（2007）『国字の位相と展開』三省堂
- 笹原宏之・横山詔一・エリク＝ロング（2003）『現代日本の異体字—漢字環境学序説—』三省堂
- 佐藤栄作（2002）「字形のゆれ・変化と字体のゆれ・変化の相互関係について—字体・字体単位体のほりあい・収斂に着目して—」『日本語の文字・表記—研究会報告論集—』33-48, 国立国語研究所
- 佐藤稔（1987）「異体字」『漢字講座 3 漢字と日本語』183-210, 明治書院
- 高田智和（2002）「『大字典』と漢字処理」『訓点語と訓点資料』109, 99-107, 訓点語学会
- 高田智和（2005）「公共サービスと漢字」『日本語学』24 (13), 58-66, 明治書院
- 高田智和（2006a）「『新しい文字』の発生とその要因」『日本語学』25 (9), 28-36, 明治書院
- 高田智和（2006b）「漢字字体研究의 文字情報 데이터 베이스」『漢文讀法과 東아시아의 文字』621-630, 태학사



- 高田智和・横山詔一・米田純子（2007）「インターネット・リサーチで文字を探す」『京都大学学術メディアセンター第18回研究セミナー報告 東洋学へのコンピュータ利用』175-182, 京都大学人文科学研究所附属漢字情報センター
- 當山日出夫（1999）「コンピュータの文字に対する意識について—錯綜する JIS 漢字論の根底にあるもの—」『国語と国文学』76 (5), 122-130, 東京大学国語国文学会
- 豊島正之（1999）「書評 横山詔一・笹原宏之・野崎浩成・エリク＝ロング『新聞電子メディアの漢字—朝日新聞 CD-ROM による漢字頻度表—』」『日本語科学』6, 91-102, 国書刊行会
- 豊島正之（2000）「文字の符号化—新 JIS 漢字第3・第4水準の開発から見た—」『京都大学大型計算機センター第64回研究セミナー報告 東洋学へのコンピュータ利用』京都大学人文科学研究所附属漢字情報センター
- 中田祝夫（1982）『日本語の世界4 日本の漢字』中央公論社
- 日本加除出版企画部編（2005）『最新人名用漢字と誤字俗字関係通達の解説』日本加除出版
- 野村雅昭（1984）「同字と別字のあいだ」『日本語学』3 (3), 23-31, 明治書院
- 林大（1984）「字体・字形・書体をめぐって」『日本語学』3 (3), 10-15, 明治書院
- 前田富祺（1992）「国語文字史研究の課題」『国語文字史の研究1』1-30, 和泉書院
- 山田忠雄（1981）『国語辞書の歩み—その模倣と創意と—』三省堂
- 山田俊雄（1968）「漢字字形の史的研究の問題とその一方向」『国語学』72, 110-126, 国語学会
- 横山詔一・笹原宏之・野崎浩成・エリク＝ロング（1998）『新聞電子メディアの漢字—朝日新聞 CD-ROM による漢字頻度表—』三省堂
- 横山詔一・當山日出夫・高田智和・米田純子（2008）「台湾日本語学習者は日本人の字体選好をいかに推論するのか」『情報処理学会研究報告』2008-CH-77, 43-50, 情報処理学会
- 

高田 智和（たかだ ともかず）

国立国語研究所研究開発部門  
190-8561 東京都立川市緑町 10-2  
ttakada@kokken.go.jp

井手 順子（いで じゅんこ）

国立国語研究所研究開発部門  
ishiyama@kokken.go.jp

虎岩 千賀子（とらいわ ちかこ）

国立国語研究所研究開発部門  
ctoraiwa@kokken.go.jp