文字情報データベースの開発とインターネット・リサーチによる音義未詳字の捜索

高田智和，横山詔一，米田純子
国立国語研究所

概要

国立国語研究所では，電子政府実現のために必要な漢字についての調査研究を行い，「文字情報データベース」の開発を進めている。調査研究のために収集・整理した文字数はおよそ60,000字に達する。その中には，漢和辞典に記載がなく，読みも使われ方も分からない音義未詳字が存在する。本発表では，インターネット・リサーチによる音義未詳字の調査方法について報告する。試験的に行った10文字の調査から，2文字について確例が得られ，文字を同定するに至った。

Construction of a kanji database and an internet-based search for information on unattested characters

TAKADA Tomokazu, YOKOYAMA Shoichi, YONEDA Junko
The National Institute for Japanese Language

Abstract

The National Institute for Japanese Language is in the process of constructing a kanji database containing information on all kanji that are necessary for implementing the Japanese e-government initiative. Approximately 60,000 kanji and kanji variants have been collected, some of which are unattested in kanji dictionaries and for which information on meaning and pronunciation is lacking. This presentation reports on the method we have adopted to search for information on these characters, which is being carried out with help from an internet research firm. Out of ten characters chosen for a pilot study, clear usage examples have been found for two characters, resulting in a conclusive identification.

1．はじめに

固有名詞の表記には，多種多様な文字がつかわれ，現代の漢和辞典に掲載されていない字種・字体も存する。現在行われている電子政府構築においても，住民の氏名や住所，事業所の名称や所在地を電算化して扱うために，行政情報処理用文字の調査研究が必要とされている。

本発表では，国立国語研究所が行なう電子政府の基盤整備にあたる漢字研究について紹介し，

文字同定に大きな問題となる，漢和辞典に掲載されていない音訓不明の文字（音義未詳字）のインターネット・リサーチによる調査方法の事例を報告する。

2．文字情報データベース

国立国語研究所では，経済産業省より「汎用電子情報交換環境整備プログラム」を受託し[1]，文字の整理・体系化に関する学術研究を行っている。総務省の住民基本台帳ネットワーク統一文字，法務省の戸籍統一文字から延べ約 80,000 字を収集し，部首，画数，読み，国語施策（常用漢字・印刷標準字体），JIS X 0213:2004 コード，ISO/IEC10646-1:2000 コード（UCS），大漢和辞典文字番号など，各種文字情報の付与を行い，情報処理学会とともに「文字情報データベース」の開発を進めている。

また，漢字には異体字が存在し，「文字情報データベース」のように大量の文字を扱う場合，異体字の処理が問題となる。このデータベースは，将来的に，行政文字情報交換の場での活用が期待されているため，異体字関係を明示した実現例として「異体字マップ」の制作も進めている。「異体字マップ」は，行政情報処理の実用面だけではなく，漢字字体研究などの学術面での応用も視野に入れて，整備・高度化していくことが課題である。

3．文字の同定と辞書非掲載字

3－1．辞書による文字同定

「文字情報データベース」に収録した文字は延べ約 80,000 字にのぼるが，住民基本台帳ネットワーク統一文字と戸籍統一文字とで共通する文字をまとめると，約 60,000 字に集約される。これらの文字に対して，『大漢和辞典』（諸橋轍次，大修館書店，修訂第 2 版 1990 年），『大字源』（尾崎雄二郎・都留春雄・西岡弘・山田勝美・山田俊雄編，角川書店，1992 年），『新大字典』（上田万年・岡田正之・飯島忠夫・栄田猛猪・飯田伝一編，講談社，1993 年），『増補改訂 JIS 漢字字典』（芝野耕司編，日本規格協会，2002 年）によって文字の同定を行なった。辞書に見えない文字（辞書非掲載字）は約 1,400 字である。

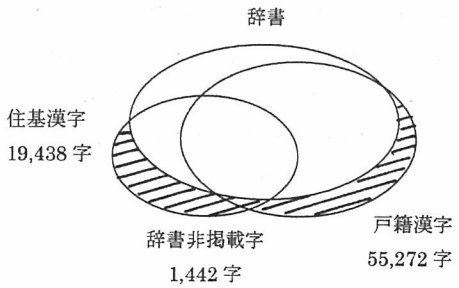

図 1:辞書による文字同定

[1] 国立国語研究所、情報処理学会、日本規格協会の三者で受託。情報処理学会は文字情報データベースの運営、日本規格協会はデザインを統一した平成明朝体グリフの作成を行なっている。

３－２．地名使用文字の現地調査

辞書非掲載字の多くは，地名や人名などの固有名詞の表記に使われる文字である。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。

例えば，住民基本台帳ネットワーク統一文字に，次の文字が採録されている（図 2)。この文字は，岩手県一関市（旧花泉町）の地名「＝輪田（ひしわだ)」に用いる。JIS X 0213 でも，この地名を典拠として採録された。ただし，規格票の例示字体は 2 点しんにょうである（図 3)。

図 2:住基文字

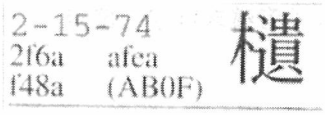

図 3:JIS 規格票

一関市花泉支所に残る「岩手縣西磐井郡金澤村字＝輪田絵圖」には，1 点しんにょうの文字が記載されている（図 4)。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 点しんにょうの文字を用いている（図 5)。このほかの文書でも 1 点しんにょうの文字を用い，役場では 1 点しんにょう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「櫃」が表記に用いられている。「匝」と「迊」のように，しんにょうがはこがまえの一部となる異体字の類型があり，「櫃」もしんにょうがはこがまえに変形したものと考えられる。ひし輪田在住の 80 代の男性は，住所の表記に「櫃」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。

図 4:絵図（明治期か）

図 5:法務局への回答（H15）

図 6:原野単価表

JIS X 0213 開発の際には，『国土行政区画総覧』から地名使用文字が採録されている。典拠資料か選定段階のいずれかにおいて，1 点しんにょうから 2 点しんにょうへの「拡張旧字体化」が行なわれたものと推測される。

３－３．「メーカー外字」

辞書非掲載字が「文字情報データベース」に採録された経緯をたどると，大部分が「メーカー外字」に由来する住民基本台帳ネットワーク統一文字であることが判明している。「メーカー外字」は，メーカー各社が規格外字としてのシステムに登載している文字であり，その中には意味も読みもわからない音義未詳字も存する。しかし，メーカー各社がシステムに取り込んだ以上，顧客の何らかの要求に応えた文字であり，使用の実績があったものであると考えられる。このような辞書非掲載字の使用例を求め，文字情報を記述していくことが，文字同定の次の段階であり，同

時に，文字生活の実態を解明する基礎研究として位置づけられる。

そこで，メーカー各社に対して，辞書非掲載字の採録経緯や使用例について質問を行なったところ，外字採録に関する記録は残っていないとの回答が寄せられた。「コードブック」に記載以上の情報は何も持っていないというのが実状のようである。よって，文字同定のための用例を求めて，独自の手段・方法を模索していくことが必要となった。

4．インターネット・リサーチによる音義未詳字の捜索

4－1．音義未詳字

音義未詳字を検討する前に，「文字情報データベース」に収集した文字の使用実態について述べておく。住民基本台帳ネットワークシステムは 2002 年から運用が開始されており，統一文字の使用件数は総務省より提供されている[2]。それによると，

使用件数あり　11,647 字

使用件数なし　7,791 字

である。実際に稼働している漢字は 6 割程度である。システムを構築するにあたって，多めに文字を収録したため，このような結果になったものと考えられる。

一方，戸籍統一文字は，戸籍に記載することが可能な文字の集まりである。戸籍に記載できる文字とは，子の名に使うことのできる常用漢字・人名用漢字をはじめとして，漢和辞典記載の「正字」や「俗字」などである。いわば，戸籍統一文字は現代の漢和辞典見出し字の集合体である。住民基本台帳ネットワーク統一文字の稼働状況から察するかぎり，戸籍統一文字は，日本国内における固有名詞の使用実態からかけ離れているものと見なされる。

辞書非掲載字を，住民基本台帳ネットワークシステムでの使用例の有無と，音訓の有無[3]とでクロス集計を行った結果が次の表 1 である。使用例のある音義未詳字 150 文字を検討していくことが，手順として妥当であろう。

	使用例あり	使用例なし	計
音訓あり	532 字	323 字	855 字
音訓なし	150字	437 字	587 字
計	682 字	760 字	1,442 字

表 1:辞書非掲載字の内訳

4－2．インターネット・リサーチ

音義未詳字は，住民基本台帳ネットワークシステムに使用が認められる文字であるため，人名・地名に用いられていることは容易に予測できるものの，現状において，固有名詞の表記を網羅的に扱った調査・研究は皆無といってよい。そこで，インフォプラント社のインターネット・リサ

---

[2] 住民基本台帳ネットワーク統一文字とその使用件数を記した使用頻度表による。個人情報に関わるため，使用例にまでさかのぼることはできないが，固有名詞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] 総務省より提供のあった『住民基本台帳ネットワーク統一文字属性辞書』による。

ーチを利用して，音義未詳字の捜索を行なうこととした。インフォプラント社に登録しているパネル（調査に協力してもらう会員）に対して，「この字を探しています」と質問を電子メールで配信し，目撃情報を求めるというものである。単純明快な質問であるが，文字使用例のデジタル画像付き回答など，文字同定に有益な判断材料となりうる回答もいくつか得られた。

図 7:「この字を探しています」と質問をした 10 文字

今回は，住民基本台帳ネットワークシステムで使用件数の多い上記 10 文字をサンプルとして選び，個々の文字に対して，以下の【Q1】から【Q4】までの質問に順次回答してもらう手順を設定した。調査期間は 2005 年 8 月 11 日から 9 月 30 日までの約 50 日間である。

【Q1】この漢字が使われているのを見たことがありますか。＜選択＞
　　使われているのを見たことがある／見たことがない
【Q2】「使われているのを見たことがある」とお答えになった漢字について，どこで見たか，使われていたモノなどについて詳細をご記入ください。＜自由記述＞
【Q3】「使われているのを見たことがある」とお答えになった漢字について，実際にその漢字を使用している画像を撮った方は以下によりアップロードしてください。＜依頼＞
【Q4】調査終了後に，今回の調査について「国立国語研究所」より詳細をお聞きするために，連絡をさせていただいてもよろしいでしょうか。＜選択＞
　　はい／いいえ

調査期間内に配信したメール数は 38,494 件である。回答者数は 876 件であり，回答率 2.28% であった。次の表 2 に，個々の文字に対する回答の内訳をまとめて示す。

文字	回答数（写真付）	回答例
A 杢	73 件（1 件）	・名刺で見た ・中華料理屋
B 尭	28 件（0 件）	・お酒の名前でみたような
C 乗	7 件（1 件）	・お客さんの下の名前に使われていた
D 辰	27 件（1 件）	・浙江省蕭山市北山公園内の石板 ・戸籍謄本
E 桟	527 件（14 件）	・材木屋で見た ・「くい」を打つとかいうときに使っていたと思う

F 枡	21 件（0 件）	・住宅の表札にあったように思う
G 裄	343 件（18 件）	・「ゆき」と読み着物の寸法表示に使用している ・桁という文字かと思ったら違った
H 斎	468 件（6 件）	・斎藤という知人の苗字で知っている ・葬儀屋
I 膳	131 件（0 件）	・小学校時代の同級生の苗字の一部 ・医療関係
J 蓑	158 件（0 件）	・職場の同僚の苗字 ・みのという字？

表 2:回答の内訳

G ではころもへんの「裄」，J では「蓑」に見誤った回答が多く寄せられた。本調査で提示した 10 文字は，何かの異体字であると推定されるものであり，逆に考えると，回答にあるような見誤りによって字体が変容し，異体字として成立・定着していく過程が想定される。

4－3. 杁

次にあげる図は，いずれも回答のデジタル画像である。注意深く街を歩いていると見かけることのある文字であるが，現代の漢和辞典に記載されていない。「材」の異体字と同定する。図 8 は木材店のトラック，図 9 は包装材料を扱う企業の看板を撮影したものである。図 9 の看板では，「包装材料」と普通名詞の場合と，「○○包材株式会社」と企業名の場合とで字体が異なる。こだわりや固有性が主張されている事例と見なされる。

図 10 は石材店の看板であるが，最終画の点がない。この字体は，鎌倉時代の辞書である観智院本類聚名義抄にも見え，楷書の筆写字体として古くから使われてきたものである。E の字体は，「丈」や「土」に見られる，書法上の安定性を持たせるために点が付加したものか，あるいは，「戈」や「伐」からの類推によって発生した字体であると考えられる。

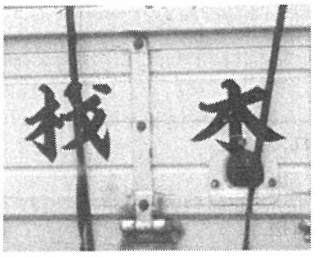

図 8:「木材」

図 9:「包材」

図 10:「石材」

4－4. 髂

Ⅰに関する回答に次のようなものがあった。

タウンガイド誌で仙台の多賀城近くにある「あらはばき神社」のはばきがこの字で紹介されていたが、ネットで検索してもこの字ではなかった。（おそらく字自体変換できないからでは。）

下図は宮城県多賀城市の現地調査によって得た「あらはばき神社」の表記である。図 11 は JR 国府多賀城駅前の観光案内図，図 12 は多賀城遺跡脇に多賀城市教育委員会が設置したと見られる案内図，図 13 は多賀城市設置の住居表示案内図である。それぞれ表記が異なり，図 11 は「脛」と「巾」を合字したもの，図 12 はⅠと同じ字体，図 13 は 2 文字で「脛巾」となっている。神社での表記は図 11 と同じく，「脛」と「巾」とを縦に重ねた字体である。神社を管理されている方[4]によれば，以前からこの文字を使っているとのことである。

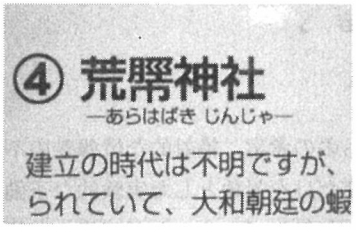

図 11:国府多賀城駅前の案内図

図 12:多賀城遺跡脇の案内図

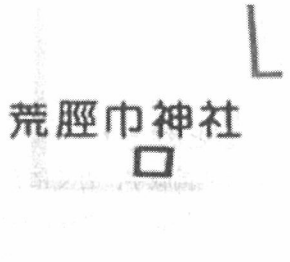

図 13:多賀城市住居表示案内図

図 14:神社鳥居

図 15:神社額

図 16:戸籍文字

図 17: 実用難読奇姓辞典

また，戸籍統一文字には，「はばき」と読む上のような文字が採録されている（図 16）。この文字は，『実用難読奇姓辞典』（篠崎晃雄，日本加除出版，1981 年），『国字の字典』（飛田良文監修・菅原義三編，東京堂出版，1985 年），『大字源』国字一覧の順を経て戸籍統一文字となった。Ⅰの異体字と考えられ，「はばき」字の成り立ちと変容は，以下の図 18 のように示すことができよう。

脚絆・脛あてを意味する 2 文字で 1 語の「脛巾」が，縦に重ねて合字されて 1 文字の字体①と

[4] あらはばき神社は個人宅の敷地内にある。

なる。「麿」や「粂」など，古来より行われている合字法である。字体①が動用を起こした字体②が住民基本台帳ネットワーク統一文字に採録され，さらに，字体②を省画した字体③が戸籍統一文字に採録されたと考えられえる。

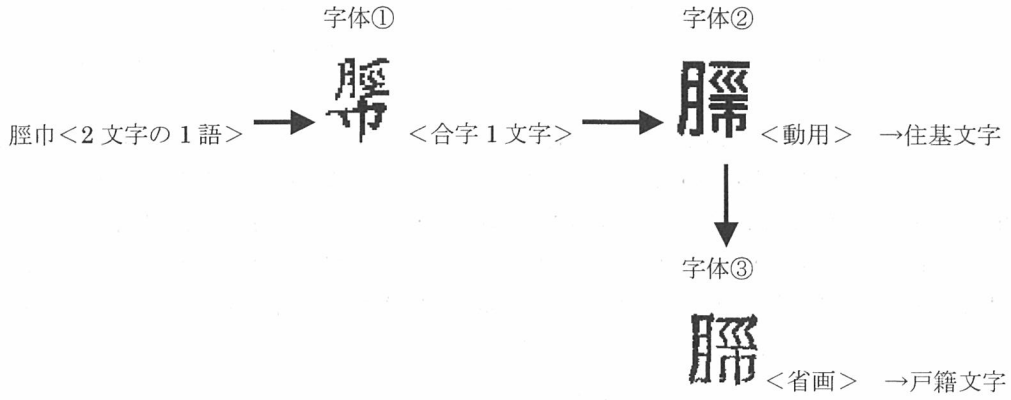

図 18:「はばき」の成り立ちと変容

5．まとめ

辞書非掲載字，とくに音義未詳字の調査では，調査方法そのものの考案，絶対的な情報量の不足など，克服すべき課題が山積している。今回は，わずか 10 文字だけを対象とした実験的な予備調査に過ぎないが，この調査手法によって，有効性のある情報を収集していくことが可能であると見通しが立てられた。より大きな規模で調査が行なえるよう調査手法の洗練を模索し，インターネットを介した文字や言葉の調査・研究方法の確立に寄与することを目指すものである。

参考文献

［1］江守賢治：解説字体辞典，三省堂，1986
［2］笹原宏之，横山詔一，エリク=ロング：国立国語研究所プロジェクト選書 2 現代日本の異体字—漢字環境学序説—，三省堂，2003
［3］高田智和：合成字，北海道大学大学院文学研究科研究論集創刊号，2001
［4］高田智和：漢字字体研究と文字情報データベース，Proceedings of the 3th International Conference for Kugyol Studies ”How to Read Written Chinese and Asian Writings”，2005
［5］高田智和：公共サービスと漢字，日本語学第 24 巻第 13 号，2005
［6］豊島正之：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ー利用，2000
［7］丹羽基二：人名と漢字，朝倉漢字講座 3 現代の漢字，朝倉書店，2003
［8］横山詔一，笹原宏之，黒田信二郎，澤田照一郎，野島伸一，石岡俊明：漢字ユビキタスを支える文字情報集積体の開発，情報処理学会研究報告 2004-CH-64，2004