朝日新聞における辞書非掲載漢字の出現状況

横山 詔一* 笹原 宏之* 野崎 浩成** 米田 純子*
YOKOYAMA Shoichi SASAHARA Hiroyuki NOZAKI Hironari YONEDA Junko

国立国語研究所*
The National Language Research Institute
115 東京都北区西が丘3丁目9番14号
3-9-14, Nishigaoka, Kita-ku, Tokyo, 115

名古屋市立大学**
Nagoya City University
467 名古屋市瑞穂区瑞穂町山の畑1
Yamanohata 1, Mizuho-cho, Mizuho-ku, Nagoya, 467

Email : [yokoyama, sasa, jyoneda]@kokken.go.jp, hnozaki@ccews1.cc.nagoya-cu.ac.jp

【1】はじめに

JIS 漢字には、読みや意味や出典がはっきりしない音義未詳字、通称「幽霊文字」が含まれているとの指摘が漢字・辞書研究者らの間でなされてきた。本研究は、1993年版朝日新聞記事全文データベース（以下、CD-HIASK'93(朝日新聞社・紀伊國屋書店・日外アソシエーツ, 1994)）を活用して、約11万件の記事のなかに「幽霊文字」や1978年以前の漢字辞書には掲載されていない「辞書非掲載漢字」がどの程度の頻度で出現するかを調査した。

今回、調査対象に新聞メディアを選択した理由は以下の3点による。

1. 現代日本において流通している文字の実態を包括的・網羅的に精査しようとする場合、大量データを含んだ電子化されたコーパスが必要となる。朝日・毎日・日経といった新聞社が発行している記事全文データベースは、比較的安価で市販されており、しかもCD-ROM化されている。
2. 字数が大量であり、さまざまな分野の用字を含むうえに、多くの人々の目に触れ、かつ影響力の大きいものは、新聞であろう。その発行部数は、『読売新聞』のように1000万を超えるものもある。また、日本国内に限らず、海外にも我が国の新聞記事はインターネット経由で流れており、日本語を解する世界中の人々に読まれているようである。
3. 新聞記事全文データベースに格納されているテキストデータを分析するには、著作権者である新聞社の使用許諾が必要である。今回、我々は朝日新聞社からCD-HIASK'93のテキストデータを光磁気ディスクに複写・蓄積する許諾を得たので、朝日新聞の記事を分析することにした。

文字調査にあたっては、上で述べたように、紙媒体上の記事だけではなく、今や新聞社各社のWebのホームページから発信されている記事も射程におさめたうえで議論すべき時代が到来したといっても過言ではない。その意味で、電子メディアとしての新聞記事CD-ROMを研究対象とすることは、世界中に流通している日本語の文字の実態を探る糸口としてたいへん重要であろう。

さて、『朝日新聞』は国内第2位の発行部数を誇っており、その紙面に、いわゆる「拡張新字体」（常用漢字表外字に表内字の新字体を準用した字体。当用漢字以前から存在した略体も含む）を大幅

に採用しているほか、固有名詞での異体字の使用を制限していたなど、他紙とは一線を画した漢字使用を行っている点で注目される（笹原・横山・米田・野崎，1997）。

CD-HIASK'93 に入っている新聞記事データは、通信社からの記事のほか、テレビ欄や天気予報欄、広告欄などを含んでおらず、紙面のすべてを包含するものではない。しかし、収録字数は、1 年間分で約 5500 万字と他の文字資料に比べて桁違いに多い。また、異なり字種も 4400 を超えており、3000 代に落ちつくことの多いサンプリングによる漢字頻度調査と比較して、異例の数を示している（笹原ほか，1997）。

本研究の目的は、JIS 漢字（「JIS X 0208」(1978 初版　1997 改正)に収める漢字）における「幽霊文字」を初めとする従前の漢和辞典に掲載されていなかった「辞書非掲載漢字」に焦点を絞ってCD-HIASK'93 と紙面（縮刷版，朝日新聞社：1993 年 1 月～12 月）との比較照合を行い、「幽霊文字」や「辞書非掲載漢字」の新聞紙面上での挙動を追跡するシステムを開発することである。

【2】調査対象

CD-HIASK'93 における 5462 万余字を対象とし、そのうち検索する漢字は次のような基準により選択した。

1．JIS 漢字における幽霊文字

「幽霊文字」とは、JIS 漢字（1978）選定の際に、原典から転記する際の誤写により生じたとみられる漢字字体のこと（笹原，1997）。なお，次の説明は，矢印記号を挟んで，左から順に「JIS 漢字の原典」「JIS 漢字表」「他の資料」を示すものである（笹原ほか，1997）。

- **広義**：字体が過去に存在しなかったもの
  ？　→「冕」
- **狭義**：字体が過去に存在したもの
  同字の場合…暗合
  閏　→「閏」　←　閏
  別字の場合…衝突
  □(木＋品)　→「櫃」　←　櫃

2．幽霊文字と辞書非掲載漢字

- 本研究では、幽霊文字を「JIS 漢字選定時に、原典から転記する際に誤写されて変化・派生した、とみられる字体」と定義する。ただし、その字種が漢和辞典にある字体と暗合・衝突した字体については調査を継続中である（笹原ほか，1997）。
- このような事情に配慮して、やや広めに『新字源』や『大漢和辞典』にない字体をも視野に入れて、129 字を選び、それを調査対象とした。その結果、純然たる広義・狭義の幽霊文字のほかに、漢和辞典に載せられることのほとんどなかった「辞書非掲載漢字」（国字・異体字を含む）や、頻度の低い漢字をも対象に含めた（笹原ほか，1997）。

【3】方　法

縮刷版で辞書非掲載漢字を検索するには、当該の漢字を含むＫＷＩＣとその出典情報が必要である。ＫＷＩＣプログラムの開発は、awk、sed、Perl などのスクリプト言語を用いて行い、データ処理はワークステーションを活用した（野崎・笹原・米田・横山，1997）。

本プログラムは，調査対象となる文字

（以下，「ターゲット文字」と呼ぶ）を読み込み，入力されるテキストファイル中に出現するターゲット文字について，ＫＷＩＣを作成する。ＫＷＩＣには，記事番号，見出し，行番号などの出典情報が付加されており，新聞紙面上の記事との照会が可能である。入力データとなるテキストファイルはテキスト形式でありさえすれば，任意の文書の取り扱いが可能である。ターゲット文字には，先に述べた幽霊文字と辞書非掲載漢字の候補を併せて 129 字用いた。図１にＫＷＩＣ作成プログラムの処理手順を示す（野崎ほか，1997）。

【4】結果と考察

１．非出現字

上記の 129 字について、CD-HIASK'93 を検索したところ、出現しなかった字が 95 種ある。これらは、いずれも国立国語研究所が 1966 年に実施した新聞に対する漢字調査（1976）でも出現していない。むろん、他の資料には用例があるものがほとんどである（笹原ほか，1997）。

冩 丗 听 啣 嘛 嚠 噌 圸 垈 垉
垳 埖 塗 堤 墹 墸 壗 壢 佞 寉
妛 岼 岾 嵶 幣 廰 弉 弖 彁 悠
抻 挧 掄 播 暃 晳 朷 枩 椢 榜
椨 椪 椣 椡 樢 梗 椦 樶 楅 樢
櫑 汢 溂 畁 眩 煇 硴 碵 穃 笂
筺 笋 篏 粐 粭 粫 糘 糢 緕 脣
艝 舮 藕 蟐 蝍 袮 貳 賍 軅 叓
轌 逎 逧 釛 鈒 鈬 鍄 鈩 閖 閠
駇 駲 鮴 鵈 鶇

２．出現字

一方、CD-HIASK'93 に出現した字は 34 種（353 回）ある。国立国語研究所の 1966 年調査でも出現していた字は「粱」「礦」(石＋廣)「莵」「峅」「帋」の 5 種（笹原ほか，1997）。

亊 伋 伜 劔 劒 垪 岃 峅 嵜 廸
杁 枦 档 梛 珎 畳 砿 礦 聨 筝
籑 粱 簱 莵 萢 帋 袰 逹 邉 鐵
鏤 鞆 鼓 甬

ちなみに、第１水準の 2560「砿」と、第２水準の 6672「礦」とは、1978JIS では正字体と拡張新字体の関係であったが、1983 年の JIS 漢字改正において、入れ替えがあった字体である。そのため、「砿」(石＋広)が調査の対象であったが、「礦」(石＋廣)も確認のために合わせて調査した。

３．ゲタ文字

本研究では JIS 外字の文字・記号を表すのに用いられることが多い「ゲタ文字：＝」も検索してみた。ゲタ文字の総度数は 876 あり、紙面と対照したところ以下のようになった。

漢字：延べ 856、異なり 43

記号：延べ 20、異なり 6

ゲタ文字に置換された漢字の一覧表から、CD-HIASK'93 には不正ゲタ文字とでも言うべきものが浮き彫りになった。

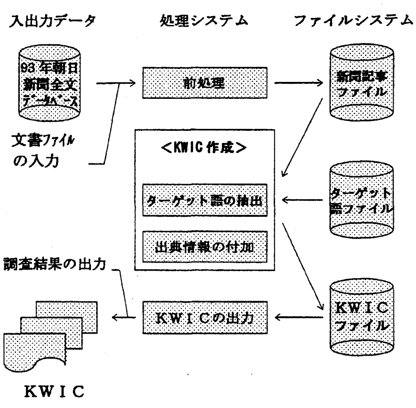

図１．ＫＷＩＣ作成プログラムの処理手順

その典型例を示す（横山・笹原・米田，1997）。

● **不正ゲタ文字の例**

83JIS に存在するのに、なぜかゲタ文字になっているケース。

槙原投手　→　＝原投手
遥かな　→　＝かな

「槙」は度数 230 で相当目立つ。ゲタ文字に置換された文字のなかで出現順位が第２位である。この事実は CD-HIASK'93 を扱う際に注意すべき側面を如実に示したものである（横山ほか，1997）。

【5】まとめ

本研究は、辞書非掲載漢字（幽霊文字を含む）が新聞記事のなかでどのような振る舞いをしているのかを悉皆的に精査した。

その結果、国立国語研究所の 1966 年調査で出現していた５種の漢字（「簗」「礦」(石＋廣)「菟」「峅」「朖」）以外に、29 種の漢字が実際の新聞紙面上に登場したことが明らかになった。これら併せて 34 種の漢字の出現頻度は 350 を越えており、これまでの文字調査では得られなかった新たな資料を手にすることができた。

またゲタ文字の調査により、電子メディアを資料として扱う際に注意を払うべきことがらの一端が明らかになった。

本研究の知見は、今後の漢字符号化研究の基礎資料として役立つことが期待される。

引用文献

1. 朝日新聞社・紀伊國屋書店・日外アソシエーツ（1994）『CD-HIASK'93』
2. 国立国語研究所（1976）『現代新聞の漢字』，集英出版
3. 笹原宏之（印刷中）「字体に生じる偶然の一致」『日本語科学』創刊号，国立国語研究所
3. 笹原宏之・横山詔一・米田純子・野崎浩成（1997）「文字資料としての『朝日新聞』紙面と CD-ROM －「JIS X0208」における辞書非掲載漢字を中心に－」，シンポジウム人文科学における数量的分析(2)，文部省統計数理研究所
4. 野崎浩成・笹原宏之・米田純子・横山詔一（1997）「幽霊文字調査のための KWIC 作成プログラムの開発」，シンポジウム人文科学における数量的分析(2)，文部省統計数理研究所
5. 横山詔一・笹原宏之・米田純子（1997）「朝日新聞 CD-ROM に出現するゲタ文字の分析」，シンポジウム人文科学における数量的分析(2)，文部省統計数理研究所

参考文献

1. 横山詔一・野崎浩成（1996 年３月）「朝日新聞 CD-ROM による漢字頻度基準表の作成と数量分析」，人文科学における数量的分析シンポジウム，文部省統計数理研究所，pp. 11-14
2. 横山詔一・野崎浩成（1996 年９月）「コーパスを利用した日本語環境の分析」，日本行動計量学会第 24 回大会特別セッション，pp. 138-139
3. 野崎浩成・横山詔一（1996 年９月）「新聞と雑誌における漢字使用頻度の分析 －心理学での材料統制の観点から－」，日本行動計量学会第 24 回大会，pp. 266-267
4. 横山詔一・野崎浩成・米田純子（1996 年９月）「新聞の漢字使用順位に影響する要因の分析」，計量国語学会第 40 回大会，p. 7
5. 野崎浩成・横山詔一ほか（1996 年９月）「漢字使用頻度の時代的変化に関する考察」，計量国語学会第 40 回大会，p. 8

＜謝辞＞

本研究は、横山・野崎・米田に対する以下の文部省科研費の援助を受けた。①「国際社会における日本語についての総合的研究」（創成的基礎研究、代表者：水谷修）、②「インターネットにおける学術漢字の符号化に関する基礎的研究」（重点領域研究、代表者：斎藤秀紀）、③「海外日本語学習リソース提供システムの実験研究」（国際学術研究、代表者：柳澤好昭）。

また、研究の推進に当たり、朝日新聞社電子電波メディア局著作権チームマネジャーの杉野信雄氏には多大なるご理解と貴重なご助言をいただいた。

紙面と CD-ROM の綿密な照合作業は佐藤朋子氏と藤川美穂氏に、83JIS 漢字符号については太田幸代氏にお手伝いいただいた。ここに記して厚く感謝申し上げる。