

国立国語研究所学術情報リポジトリ

言語研究資料としての電子媒体の問題点

メタデータ	言語: jpn 出版者: 公開日: 2020-06-29 キーワード (Ja): キーワード (En): 作成者: 横山, 詔一 メールアドレス: 所属:
URL	https://doi.org/10.15084/00002936

言語研究資料としての電子媒体の問題点

横山詔一

(情報資料研究部電子計算機システム開発研究室)

1. はじめに

21世紀は電子媒体(電子メディア)がさらに普及するものと思われる。もちろん紙による印刷媒体は今後も残るであろうが、従来型の印刷システムの圧倒的な優位性は次第に崩れていくに違いない。押し寄せる媒体革命の波を越えて、質の高い言語研究を展開するには何が必要なのか。ここではその手がかりの一端を探る。

電子媒体は言語研究の資料としてきわめて有用であることに疑いの余地はない。確かに電子媒体を利用すれば大量の資料を高速にさばくことができる。電子媒体をコンピュータで処理すれば、紙媒体を人間が処理した場合よりも「正確な」結果が得られると一般には信じられているようである。しかし果たして、この認識はどこまで正しいのであろうか。本稿は市場に流通している新聞記事全文データベース(朝日新聞1993年版コーパス;以下、CD-HIASK'93という)を対象にした国立国語研究所プロジェクト選書No.1『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—』(横山詔一・笹原宏之・野崎浩成・エリク=ロング, 1998, 三省堂刊;以下、国語研選書.1という)に掲出されている事実の一部を示しながら、どのような問題が生じているのか整理する。

さらに、日本語研究資料(研究成果を含む)を広く世界に電子媒体で提供するための先端技術の一つを後半部で簡単に紹介する。日本語研究資料へ海外からアクセスしようとする利用者を積極的に支援することは、国立国語研究所が今後も積極的に取り組むべき文化交流・情報発信事業の重要な柱の一つであり、海外からの期待も大きいと聞く。現時点では電子媒体上の日本語テキストは海外で文字化けするのが普通であり、「どこでも・いつでも・誰でも・簡単に利用できる」資料とは言い難い状況にあり、言語研究の発展を阻む要因となっている。

2. 新聞コーパスは原紙の「正確な」複製か

これまでの新聞漢字調査は原紙を第1次資料としてきた。近時インターネットや携帯電話でも新聞記事の配信が行われるようになったが、ここでは従来通り原紙に印刷された文字を第1次資料とする立場をとる。印刷媒体の情報を電子化した時に生じる問題、つまり紙媒体と電子媒体との間でなされるメディア変換において偶発的あるいは必然的に起こりうる齟齬に焦点を絞って論を進める。

2.1. 電子媒体と印刷媒体との比較方法

【方法1: 朝刊1日分の記事を対象としたもの】

CD-HIASK'93に収められている1993年2月1日朝刊の記事をすべてプリンタに出し、それらを縮刷版と逐一照合して、違いを検討した。

【方法 2：朝刊 1 ヶ月分の「一面記事」を対象としたもの】

CD-HIASK'93 から、1993 年 2 月 1 日～2 月 28 日までの朝刊「一面記事」をすべてプリンタに出力し、縮刷版と照合した。

【方法 3：朝夕刊 1 年分を対象としたもの】

ある文章と同じものが少し離れた場所に再び出現した場合は、そのような重複が原紙にも見られるかを縮刷版でチェックした。

2.2. 文章表現に関連する媒体間での齟齬の例

(1) 見出しに関すること

テキストデータを縮刷版と比較すると、見出し部に違いが多く発見された。その例を 1993 年 2 月 1 日朝日新聞の朝刊 1 面より 3 例示す。

【例 1】

縮刷版：政治改革関連法案 三塚氏が分離論 「一括論」の修正へ柔軟

CD-ROM：自民・三塚氏が分離論 「一括論」の修正へ柔軟 政治改革関連法案で

【例 2】

縮刷版：冬季五輪遠き照準 世界アルペン 3 日開幕の地元・岩手

CD-ROM：冬季五輪誘致に遠き照準 世界アルペン 3 日開幕の岩手

【例 3】

縮刷版：「21 世紀の日本」委員会 第 2 期委員会が発足 新たに 10 氏、留任は 6 氏

CD-ROM：「21 世紀の日本」委員会、第 2 期委員会が発足<社告>

朝日新聞社によると、CD-HIASK'93 作成時に、紙面の見出しの文言に編集をかけたうえで、そのテキストデータを手入力しているとのことである。その理由は、紙面の見出しで「カット部分（地紋の中に白抜きの見出しなど）」はイメージデータであること、CD-HIASK'93 では見出し部分の文字数に制限を設けていること、などによるという。なお、小見出しは、本文中に含まれる場合もあれば、そうでないこともあり、その扱いは CD-HIASK'93 では統一されていない。

(2) 収録されていない記事・文字列：通信社からの記事、テレビ番組欄など

CD-HIASK'93 は、テレビ欄や天気予報欄、広告欄などを収録しておらず、紙面のすべてを包含するものではない（笹原・横山・野崎・米田，1998）。例えば、通信社から配信された記事（海外の動向に関する記事に多い）や評論家などによる署名記事（文化欄に多い）、連載小説などは、出典情報や見出し部分は検索できるが、本文は存在しない。これらは、朝日新聞社に著作権が帰属しないものである。

(3) 同じ文字列を2回以上重複して収録しているもの

【例1：用語欄，解説欄】

CD-HIASK'93において，記事本文の末尾に「<用語>」もしくは「<解説>」として示されている部分は，その部分だけがもう一度収録されている場合がある。これらは，紙面には1回しか出現しない。つまり，<用語>と<解説>の記事は，同じ文章がテキストデータ中に2回出現することがある。

【例2：記事が分割されたケース】（エリク＝ロングによる）

CD-HIASK'93は，紙面で1つの記事であるものを，複数の記事に分割している場合が珍しくない。そのようなケースにおいては，紙面で1回しか出現しない「注記」あるいは「ただし書き」のような記述が，分割された記事のそれぞれに付記されていることもある。

その典型例として，夏の全国高校野球大会関係の記事で甲子園出場校を紹介したものがあ。紙面では，各出場校のプロフィールが表の形にまとめられて1つの記事として掲載されており，その「《表の見方》」の説明が欄外に1回だけ登場する。ところが，CD-HIASK'93では「関東，中部，近畿，中国・四国，九州・沖縄」の5つに記事が分割されたうえで，記事のそれぞれに同じ《表の見方》が収録されている。つまり，その部分の文字列は重複収録ということになる。この5つの記事のうち，「関東」の例を以下に示す。

第75回全国高校野球 49代表校の横顔 関東
'93.8.4 朝刊 19頁 写図有 (全6727字) [930804119]
記事本文→省略

《表の見方》

校名の右の数字は出場回数，初は初出場。守備位置の左の数字は背番号。◎印は主将。選手名の次のカッコ内数字は学年。その後は投打の左，右（右左は両打ち），身長，体重，地方大会の打率の順。

学校所在地右側の数字は生徒数，カッコ内は野球部員数。所在地の上の数字は，地方大会のチーム成績で（1）勝利数（2）チーム打率（3）総得点（4）総失点。

(4) その他

【例1：訂正記事】（エリク＝ロングによる）

CD-HIASK'93の記事本文末尾に「<訂正>」と示されている部分の例を以下に示す。縮刷版により，テキストデータの本文部分は訂正の通りに直されていること，実際に訂正記事が掲載されたのは翌日の夕刊であったことなどを確認した。したがって，この記事の一部は実際の紙面と違っていることになる。また，末尾の2行「(8日夕刊に掲載)」と「●記事本文は，訂正の通りに直してあります。」の箇所は，ほとんどの<訂正>記事の末尾に出現するが，紙面には存在しない文字列である。

茶道習い，和食が好み プリンセス小和田雅子さん本人が語った「私」
'93.1.7 夕刊 7頁 写図有 (全2,367字) [930107185]

記事本文→省略

<訂正>

七日付「プリンセス小和田雅子さん」の記事で、「女性外交官第一号の楠田かおるさん」とあるのは誤りでした。第一号は山根敏子さん（故人）です。訂正します。

（8日夕刊に掲載）

●記事本文は、訂正の通りに直してあります。

【例2：電子メディアの注意点に言及した部分】（笹原宏之による）

CD-HIASK'93 の記事本文末尾に「<編注>」と示されている部分の例を以下に示す。縮刷版により、この部分は実際の紙面と違っていることを確認した。

厩廐廐…「きゅう」さまざま（赤えんぴつ）

'93.6.12 朝刊 29頁 写図無（全608字）[930612160]

記事本文→省略

<編注> パソコンに表示されるフォントとプリントアウトしたフォントでは違った表示になることがあります。

この記事には、「きゅう（广に厩）」、「きゅう（厂に皂と死）」、「厩（厂と漑のサンズイを除いたもの）」、「きゅう（广に皂と受）」、「廐（广と漑の真ん中部分と受）」、「きゅう（厂に厩）」、「きゅう（广に皂と死）」といった記述が見られるが、これらはすべて紙面ではそれぞれ1つずつの漢字で表現されているものである。よって、これらの箇所は、CD-HIASK'93 と紙面に違いがある。ちなみに、「厩（厂と漑のサンズイを除いたもの）」、「廐（广と漑の真ん中部分と受）」は漢字と（）内の注記が一致しない。

【例3：横書きによる影響】（エリク＝ロングによる）

紙面で縦書きの記事は CD-HIASK'93 では横書きになる。そのため、将棋関係の記事などで、紙面で「上段が先手」とある箇所が、CD-HIASK'93 では「左が先手」と書き換えられていることがある。

また、紙面に掲載された写真のなかの文字列（紙面の本文には登場しないもの）が、CD-HIASK'93 では本文部に収録されているケースもあった。

さらに、CD-HIASK'93 に収録されているのに、紙面には見当たらない記事も発見された。おそらく、出典情報に誤りがあるものと思われる。

2.3. 文字表記に関連する媒体間での齟齬の例

ここでは JIS 漢字に関することからも取り上げる。JIS 漢字は、1978 年の第 1 次規格が出て以来、1983 年、1990 年、1997 年と改正を経てきた（以下、1978 年の第 1 次規格を「78JIS」、1983 年の第 2 次規格を「83JIS」という）。

(1)「^」が付された箇所の文字表記について

新聞記事には、当然のことながら 83JIS でカバーできない漢字（以下、JIS 外字と

いう)も出現する。朝日新聞社によると、CD-HIASK'93は、JIS外字を平仮名もしくは片仮名にひらき、その後ろに「^」マークを付けて、JIS漢字では表記できない漢字であることを示す、という方針をとってきた。JIS外字のうち、日本の地名・人名の場合は平仮名で正しい読みを、外国の地名・人名の場合は音読みを片仮名であることになっている。したがって、紙面でJIS外字の箇所は、CD-HIASK'93と表記が違う。

実際にCD-HIASK'93を調べてみると、JIS外字の漢字・熟字が同じ記事に複数回出現する場合は、2回目からマークしないようである。また、その記事で初めて出現したもののなかに、仮名にひらいただけで、「^」マークを付け落とした箇所も見られる。

「^」でマークされた箇所のCD-HIASK'93のKWICを以下に2例示す。下線部がJIS外字に相当する部分である。

ペット列伝 (新年特集・第5部)

'93.1.1 朝刊 76頁 写図無 (全1,359字) [930101139]
 をおそれず。しん恚 ^ =しんい (憎しみ・

中国の株式公開、全国規模で実験 (情報ファイル・国際)

'93.1.5 朝刊 11頁 写図無 (全296字) [930105067]
 海と深セン経済特区 ^ の二カ所に限っている

KWICを縮刷版ですべてチェックした結果、延べで814、異なりで236の漢字が同定された。なお、「^」マーク箇所の漢字は、すべてがJIS外字というわけではない。JIS漢字であるにもかかわらず、「^」でマークされて仮名化されたものが7文字「葛、俤、鴿、敵、祇、牀、瑰」、JIS記号が1文字「全」見つかった。また、「^」が屋号を示す記号として使われた箇所もあった。

(2)ゲタ文字(欠字・伏せ字)「=」について

朝日新聞社によると、いわゆるゲタ文字「=」はなくす方向でCD-HIASK'93を制作したという。しかし、我々は、JIS外字の一部が欠字・伏せ字として、ゲタ文字「=」に置換された可能性もあると考え、CD-HIASK'93のゲタ文字箇所が実際の紙面でどのように表記されているのかの一覧リストと頻度を求めた。ゲタ文字「=」は、83JISの区点番号で02-14である。CD-HIASK'93のKWICの例を以下に示す。

1軍復帰の=原、12奪三振初勝利 巨人2-1阪神 7回戦

'93.5.23 朝刊 22頁 写図有 (全566字) [930523104]

<勝> = 原1勝2敗
 点を先制。投げては = 原-石毛とつないで
 ムに落とされていた = 原が、ようやく戻っ
 地を通した」という = 原の表情にはのんび
 れ、ホッと一息。「 = 原の好投につきる。

これをもとに、ゲタ文字箇所が実際の紙面ではどのように表記されていたかを縮刷版に当たりながらすべてチェックした。その結果、縮刷版のゲタ文字箇所の文字・記号は延べで 876 に達し、その内訳は以下の通りになった。

漢字：延べ 856, 異なり 43

記号：延べ 20, 異なり 6

上記の漢字の一覧を調べた結果、先に述べた「^」箇所の漢字頻度表に含まれているものが発見された。つまり、それらの漢字は、「^」でマークされて仮名にひらかれている場合もあれば、ゲタ文字になっていることもあるということで、CD-HIASK'93 制作における JIS 外字処理の不統一を示唆している。

「^」マーク箇所の漢字の場合と同様、ゲタ文字箇所の漢字もすべてが JIS 外字ではなく、以下の 2 タイプに分類できる。それぞれの典型例を示す。

【JIS 外字のケース】

鄧小平 → 二小平 (ゲタ文字のなかで最多。頻度 426。)

【JIS 漢字のケース】

槇原投手 → 二原投手 (ゲタ文字のなかで順位が第 2 位。頻度 230。)

遙かな → 二かな

ここで注目すべき点は、CD-HIASK'93 のテキストデータ内を検索して頻度ゼロの「堯, 槇, 遙, 瑤」が含まれていることである。この 4 文字は、CD-HIASK'93 の制作過程において、何らかの事情により、誰も気づかないうちにすべてゲタ文字になった可能性がある。あるいは、この 4 文字が 78JIS では扱えないことに配慮して、制作者側であらかじめ意図的にゲタ文字化したのかもしれない (前に述べたように CD-HIASK'93 は 78JIS にも対応している)。

ちなみに、「堯, 槇, 遙, 瑤」を『CD-毎日新聞'93』(毎日新聞社・日外アソシエーツ, 1994) で検索したところ、やはり頻度ゼロであることが判明した。次に、それらの正字体である「堯, 槇, 遙, 瑤」を検索したところ、数多くヒットした。例えば、「槇原投手 (読売巨人軍)」は、毎日新聞 CD-ROM ではすべて「槇原」と表記されている。

(3) JIS 外字の JIS 内字への置換

CD-HIASK'93 の漢字コードは 83JIS を基本としており、90JIS (1990 年改正の JIS 規格) で追加された「熙(熙の中は口)」と「凜(凜の示は禾)」は検索できない。紙面で細川護「熙(熙の中は口)」元首相と表記されている箇所が、CD-ROM ではすべて「熙」になっている。このように、83JIS 外字は 83JIS の類似した異体字に置換されたケースが少なくないようである。

(4) 朝日文字の JIS 内字への置換(笹原宏之による)

新聞で使われる漢字は、戦後、当用漢字を用いるようになったが、表外字については新聞社によって対応が異なった。読売新聞社・毎日新聞社などでは、部首の「食偏」

「しんにょう」「示偏」の部分に限っては、当用漢字・常用漢字の新字体を準用している。一方、朝日新聞社は、より積極的に表外字の字体の整理を行い、部首だけでなく旁に対しても、当用漢字・常用漢字新字体を準用した「拡張新字体」を用いている。「壺」→「壺」がその例で、中には「檜」→「桧」,「摺(正)」→「摺(新)」,のように、ずいぶん字体が違うものを含んでいる。

これらのなかには、83JIS 第 1 水準の字体と一致するものがあるほか、83JIS で字体が変更されなかった字でも字体が大幅に簡略化されたものが存在する。これらは、「朝日文字」「朝日字体」とも呼ばれ、JIS 漢字よりも以前の昭和 30 年代から使われ始めている。朝日文字は、朝日新聞の紙面のほかに、『日本経済新聞』の紙面などでも使用されている。

3. 電子媒体を用いた言語研究の精度について

以上、文字表記研究の側面から言語研究資料としての電子媒体の問題点を検討した。その結果は以下の 4 点に整理できる。

- (1) 組版コードから変換された JIS 漢字コードを用いた文字調査は精度を高めようがない。

新聞 CD に納められている電子化テキストは、原紙の印刷に用いた組版コードをパソコンで扱えるように JIS 漢字コードへ変換したものである。最近、国立国語研究所以外の研究機関でも文字・単語頻度表を公刊する動きが出てきた。その初例が NTT データベースシリーズ『日本語の語彙特性：第 7 巻』（NTT コミュニケーション科学基礎研究所、2000；以下、NTT データベースという）である。NTT データベースの文字頻度表は、朝日新聞社内で「組版コード → JIS 漢字コード」の変換を経た電子化テキストを対象に、文字頻度を計数したものである。文字調査においては調査対象のデータとしての精度に関する記述が必要不可欠であるが、NTT データベースの解説にはそのような記述が欠落しており、残念である。調査対象のコーパスが「文字化け」のようなエラーデータを一定の割合で含むようなものであれば、コーパスの規模をいくら大きくしたところで調査精度はまったく向上しないのは自明の理と言えよう。

先に示した通り、たとえ大手の新聞社が作成した電子化テキストであろうとも、そこにエラーが混入していないという保証はない。豊島(1999)は「新聞社が作成した電子化テキストを信頼性のあるテキストデータと無批判に受け入れてはならない」と述べている。この豊島の論考は文字の計量的研究に携わる者にとっては必見の文献である。

- (2) 新聞社内の組版コードを直接調査対象とする方法はコストや調査精度の面で有望であるが、注意を払わねばならない点もある。また、組版コードに対応する文字の字体を人間の目で確認できるようにするシステムも必要であろう。

新聞社や書籍印刷会社の組版コードを利用したと思われる調査が文化庁によってなされている。一つは読売新聞の漢字を調査した『漢字出現頻度数調査(Ⅱ)』(2000)、もう一つは凸版印刷による書籍を対象にした『漢字出現頻度数調査』(1997)である。新聞社や印刷会社の協力を得ながら実施されたという点と、JIS 漢字コードへの変換を経ない電子化テキストを分析した点で、調査の精度は高いと思われる。ただし、

原紙の文字を完全には補足できていない懸念もない訳ではない（豊島，1999，p.98）。たとえ組版コードを直接対象とした場合であっても，新聞社や印刷会社の内部でシステムに変更があるとコード体系が変化する可能性がある。その結果，文字情報の管理に混乱が起きて，結果的に文字統計の数値が原紙の数値と一致しないケースも生じてしまうようである。

(3) 原紙に高頻度で出現し，しかも 83JIS にある漢字の一部が，電子化テキストでは失われていることがある。

83JIS に含まれている「堯，楨，遥，瑶」の 4 文字は，78JIS にコードポイントが存在しない。そのため，「作家森瑶子氏と楨原投手，遥かな旅に」と 83JIS で入力し，78JIS で表示すると「作家森ニ子氏とニ原投手，ニかな旅に」などと欠字（ゲタ文字）だらけとなり，意味不明となるのが普通である。

先に示した通り朝日新聞 CD では，この 4 文字は見出し部分を除く記事本文において例外なく「ニ」に置き換えられている。おそらく，組版システムからテキストデータを抽出する際のコード変換テーブルに乱れがあったのだろう。

この問題は，文字調査のみならず単語調査にも影響を及ぼすので注意が必要である。NTT データベースには「遥かだ」の出現頻度が掲出されているが，原紙照合を経た国語研選書 1 のデータと比較するとその数値は異常に低い（NTT データベースでは「遥かな」「遥かに」などの活用形は「遥かだ」の終止形にまとめられている）。NTT データベースは CD 化される前段階の朝日新聞記事テキストデータを分析対象としており，その点で CD データと一線を画すると言われている。それにもかかわらず，「堯，楨，遥，瑶」の 4 文字については，朝日新聞 CD と同様，原紙と大きな齟齬が見られるようである。

ちなみに，市販の毎日新聞 CD（CD-毎日'93：毎日新聞社，1994）ならびに言語処理学会員に販売されている毎日新聞テキストデータには，ここで述べたゲタ文字化の問題は見当たらないようである。

(4) 朝日新聞は 1993 年秋から「葛 {旧} 飾区」と印字するが，83JIS 漢字コードによる調査では「葛 {旧}」を頻度表に掲出しない。また，読売新聞や書籍では「摺 {旧}」「頬 {旧}」「剝 {旧}」と印字するのが一般的であるが，83JIS 漢字コードによる調査はそれらの字体を掲出しない。これらは旧字体が 83JIS 外漢字となる例である。

国語研選書 1 をはじめ NTT データベースや Chikamatsu, Yokoyama, Nozaki, Long & Fukuda (2000) は，83JIS に含まれない「葛 {旧}」「摺 {旧}」などを分析の対象からカットしている。これらの字種は 83JIS において拡張新字体のみが採用され，旧字体は 83JIS 外漢字となってしまったものである。同じあるいは類似の理由で，高頻度漢字が頻度表に掲出されないケースが珍しくない。

この問題は単語調査にも波及する。NTT データベースは，動詞「摺む」について「摺 {旧} む」の表記は掲出しない。ところが，一般の書籍や読売新聞などでは「摺 {旧} む」と表記する場が圧倒的に多く，「摺む」は相対的にまれである。笹原・横山 (1998) を中心とする一連の研究によれば，大学生に「葛」と「葛 {旧}」のペアでより「なじみ」深い方を直観的に選択させたところ，「葛 {旧}」を選択した人数が統計的に有意

に多くなった。つまり、この異体字ペアにおいては旧字体の方が新字体よりもなじみ深いのである。同様に、「搦」よりも「搦 {旧}」、「頬」よりも「頬 {旧}」、「剥」よりも「剥 {旧}」の方が、それぞれなじみ深いと受け取られている。

なじみ調査と同様の傾向は「好み」調査においても見られる（笹原・横山，1998；横山・笹原，1999）。好みとは、ワープロやパソコンで文字を打っている時に異体字ペアの一方を選択しなければならないとしたらどちらを選ぶか、というものである。心理学や経済学の用語では「選好 (preference)」とも呼ぶ。なじみと好みの相関を算出すると.95に達し、両者には強い正の相関関係がある（笹原・横山，1998）。

以上の事実は、例えば認知科学などの研究で漢字刺激を扱う際は、異体字や 83JIS 外漢字の問題を等閑視するのは危険であることを示唆している。被験者になじみのない表記で単語刺激を呈示するのは、特別な目的がある場合に限られる（浮田・杉島・井上・皆川・賀集，1996；横山，1997）。なじみの薄い表記は被験者に違和感を生じさせ、その効果が攪乱要因として実験データの精度を低下させるおそれがあるからである。

4. 日本語研究資料は電子媒体で海外に提供可能か

次に視点を少し変えて、電子媒体による言語研究資料の普及・共有化の問題を考えてみよう。英国の国立研究機関は、英語に関する言語研究資料をインターネットで世界に無償公開している (<http://www.itd.clrc.ac.uk/Projects/Psych/index.htm>)。このような状況を考えると、国立国語研究所が中心になって蓄積してきた日本語研究の成果も、今やインターネットを介して広く世界に提供する時期が到来したと言えよう。日本語研究資源へ海外からアクセスしようとする利用者を積極的に支援することは、我が国が行うべき文化交流・情報発信事業の一つでもある。

4.1. 日本語研究資源を世界に開くために

海外には約 210 万人もの日本語学習者が存在し、確実にしかも急激に増加している。日本語学習者の手元にインターネットを介して日本語教材が円滑に届く情報基盤の必要性が叫ばれているのであるが、その仕組みは諸学界で作成・蓄積された日本語データベースを世界に提供するシステムの整備にも応用が可能であろう。

かかる視点から、『現代雑誌九十種の用語用字：全語彙・表記』（国立国語研究所言語処理データ集 No.7, 1996, 三省堂；以下、「雑誌九十種」という）を例として、そのデータを海外の WWW ブラウザ（いわゆるホームページ閲覧ソフト）で検索できるシステムの第 1 版を開発した（以下、このシステムを「雑誌九十種 WWW 検索システム」という；横山詔一・エリク＝ロング・江川清・笹原宏之・古家時雄，2000）。

現在のところ、海外の一般的な WWW ブラウザは日本語を簡単に表示することができない。日本語処理ができるようにするためのソフトや文字フォントをクライアントに組み込む必要があるが、海外の大学などでは学内 LAN に接続されている端末のパソコンに使用者が勝手に日本語処理ソフトを組み込むことを許していないケースも多い。ちなみに、一般には Java のアプレットに日本語フォントを畳み込んでクライアントに送り込むという方法がよく知られている。しかし、アプレットが開くまで時間

がかかり過ぎるという評判も耳にする。また PDF ファイル形式で文書をやり取りする方法も有効だが、アドビ社の Acrobat Reader というソフトをダウンロードしたり、文書ファイルをダウンロードする手間が必要である。

このような壁を少しでも突き崩すには、ネットワークの端末装置やソフトの性能に左右されることなく、インターネット回線の状態が安定していれば常に世界中の一般的な WWW ブラウザから日本語データベースを検索可能な、軽快な動作のシステムが是非とも必要である。

4.2. 雑誌九十種WWW検索システムの概要

雑誌九十種 WWW 検索システムの使用法は以下のようにいたって簡単である。インターネットに接続されたパソコンで WWW ブラウザを起動し、本システムにアクセスすると図 1 に示す画面が表示される。

Complete Vocabulary from a Survey of Ninety Contemporary Japanese Magazines

Enter word to search for:

図 1 入力画面の例

入力エリアに検索したい語のローマ字表記を半角英文字で入力し<Find>のボタンをクリックすると図 2 に示す検索結果が表示される。ちなみにこれは「いし」という読みの語を検索した例である。Written form (表記) は当然のことながら日本語表示であるが、Origin (語種: 和語, 漢語, 混種語, 外来語など) や Part of Speech (品詞) などの情報は英語で表示される。

このシステムの最大のポイントは日本語の表示方法にある。日本語環境を持たない海外の WWW ブラウザに日本語を表示するために、ここでは文字の画像 (GIF ファイル) を利用する方法を選択した。

雑誌九十種 WWW 検索システムは以下のような仕組みになっている。WWW ブラウザから入力された文字列は「データベース検索サーバ」へ送られ、検索完了後その結果が WWW ブラウザに返送される。日本語の部分は GIF リンクで表示されるようになっているので、WWW ブラウザはリンク先の「文字 GIF 配信サーバ」にある GIF ファイルにアクセスする (データベース検索サーバは国立国語研究所内に、文字 GIF 配信サーバは紀伊國屋書店内に設置されている)。

いし			
Written form	Form count	Total count	Origin /Part of Speech
石	28	28	Native noun
意志	28	40	Sino noun
意思	12		
遺志	1	1	Sino noun
医師	33	33	Sino noun
絵死	1	1	Sino noun

図 2 検索結果の例

5. まとめ

本稿の前半部では、「新聞コーパスは原紙の正確な複製か」という問いについてさまざまな側面から検討を行った。その結果「No」という答えが導かれた。電子媒体による言語研究は紙媒体を用いたそれよりもいつも格段に優れているとは言えないようである。例えば漢字調査について言えば、コンピュータによって張られた電子の網にはところどころポツカリと穴のあいた箇所があり、JIS 漢字の範囲といえども全部を捕捉できる訳ではない。新聞紙面における漢字の実物と、それを模写したはずのパーティクルな電子メディア空間における漢字とでは、さまざまな違いが見られる。紙面上の漢字と電子媒体の関係は、航空機と航空管制レーダーのそれに似ている。レーダーは、人間の視力をはるかに超えた広い範囲を素早く探索する。しかし、レーダー画面に映し出される機影だけに頼って、それがどの航空会社の旅客機なのかを判断することは危険きわまりない行為である。システムに死角がないとも限らない。一見手間がかかりそうに思えるかもしれないが、紙媒体を併用しながら、電子の網にかかった漢字の実物を専門家が目で確認する必要がある。同様のことが文字表記研究以外の言語研究全般にも言えるのではないか。

本稿の後半部は、日本語研究資料を「どこでも・いつでも・誰でも・簡単に利用できる」ようにするための最新技術の一つを紹介した。このような情報技術を取り込んだ言語研究を推進するには産業界の協力がどうしても必要である。情報通信研究と国語学・言語学の間には大きな距離があるように見えるかもしれない。しかし、文字調査の歴史は日本語情報処理の工学的研究の歩みと軌を一にする部分が少なくないのであって、両者は車の両輪である。漢字仮名交じりで書かれた大量の新聞記事テキストを世界で初めてコンピュータ処理して漢字頻度調査を行ったのは国立国語研究所であり（国立国語研究所，1976），その成果の上に我が国の情報基盤が築かれてきたと言っても過言ではなかろう。ただし，その成功はコンピュータメーカ各社の旺盛な技術開発意欲に支えられていたことも周知の事実である。

高品質の日本語研究資料は研究・教育機関のみならず，出版社各社でも積極的に作成されるようになりつつある。三省堂はインターネット閲覧ソフトを介して『大辞林』『新明解国語辞典』『新グローバル英和辞典』『クラウン独和辞典』など 16 タイトル 120 万語の辞書検索が可能な辞書検索サービス（『三省堂 Web Dictionary』）を 2001 年 1 月上旬から開始する（<http://www.sanseido.net>）。また，紀伊國屋書店は 9 万字以上の漢字フォントを WWW に配信する『文字鏡 URL 文字配信サービス』をすでに行っている（<http://font.mojikyo.com/>）。国立国語研究所はこのような「日本語資源」も視野に入れながら，独立行政法人化の後も，これまで以上に日本語研究を推進する使命があると考えられる。

引用・参考文献(アルファベット順)

- 朝日新聞社 (1994) 『CD-HIASK'93 朝日新聞記事データベース』, 紀伊國屋書店・日外アソシエーツ
文化庁国語課 (1997) 『漢字出現頻度数調査』 漢字字体関係参考資料集, 文化庁
文化庁国語課 (2000) 『漢字出現頻度数調査(2)』 漢字字体関係参考資料集, 文化庁
CHIKAMATSU Nobuko, YOKOYAMA Shoichi, NOZAKI Hironari, Eric LONG, & FUKUDA Sachio

- (2000) 「A Japanese Logographic Character Frequency List for Cognitive Science Research」
『Behavior Research Methods, Instruments, and Computers』,32(3), 482-500, Psychonomic Society
- 久野雅樹 (2000) 「新聞の用字の面による変動と時系列変動」『自然言語処理』7 巻 2 号 pp45-61, 言語処理学会
- KESS Joseph F & MIYAMOTO Tadao (1994) 『Japanese Psycholinguistics : A Classified and Annotated Research Bibliography』, John Benjamins
- KESS Joseph F & MIYAMOTO Tadao (2000) 『Japanese Mental Lexicon : Psycholinguistic Studies of Kana and Kanji Processing』, John Benjamins
- 国立国語研究所 (1962) 『現代雑誌九十種の用字用語』(国立国語研究所報告 21,22,25), 秀英出版
- 国立国語研究所 (1976) 『現代新聞の漢字』(国立国語研究所報告 56), 秀英出版
- 国立国語研究所 (1997) 『現代雑誌九十種の用語用字 全語彙・表記【FD 版】』(国立国語研究所言語処理データ集 7), 三省堂
- LONG Eric T & YOKOYAMA Shoichi (1997) 「An Analysis of Kanji Strings in the CD-HIASK'93 Data Base」『人文科学における数量的分析 (2)』シンポジウム報告書 pp.15-20, 文部省統計数理研究所
- 毎日新聞社 (1994) 『CD-毎日新聞'93』, 日外アソシエーツ
- NTT コミュニケーション科学基礎研究所〔監修〕天野成昭・近藤公久〔編著〕(2000) 『日本語の語彙特性』NTT データベースシリーズ, 三省堂
- 野崎造成・横山詔一・磯本征雄・米田純子 (1996) 「文字使用に関する計量的研究—日本語教育支援の観点から—」『日本教育工学雑誌』20 巻 3 号 pp.141-149, 日本教育工学会
- 笹原宏之・横山詔一 (1998) 「異体字選択に影響する要因」『計量国語学』21 巻 7 号 pp.291-310, 計量国語学会
- 笹原宏之・エリク＝ロング・横山詔一 (1998) 『朝日新聞』における JIS 外漢字 (計量国語学会第 42 回大会) 『計量国語学』21 巻 7 号 pp.336-337, 計量国語学会
- 笹原宏之・横山詔一・野崎造成・米田純子 (1998) 『朝日新聞』の CD-ROM と紙面における幽霊文字と辞書非掲載漢字—「JIS X 0208」の漢字を中心に— 『計量国語学』21 巻 4 号 pp.145-161, 計量国語学会
- 豊島正之 (1999) 「書評 横山詔一・笹原宏之・野崎造成・エリク＝ロング〔編著〕『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——』国立国語研究所プロジェクト選書 1」『日本語科学』6 号 pp.91-102, 国立国語研究所〔編〕, 国書刊行会
- 浮田潤・杉島一郎・井上道雄・皆川直凡・賀集寛 (1996) 『日本語の表記形態に関する心理学的研究』心理学モノグラフNo.25, 日本心理学会
- 横山詔一 (1997) 『表記と記憶』心理学モノグラフ No.26, 日本心理学会
- 横山詔一・野崎造成 (1996) 「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』シンポジウム報告書 pp.11-14, 文部省統計数理研究所
- 横山詔一・笹原宏之 (2000) 「文字と暮らし」『豊かな言語生活のために』(新「ことば」シリーズ 11) pp52-63, 国立国語研究所〔編〕, 大蔵省印刷局
- 横山詔一・笹原宏之・野崎造成・エリク＝ロング〔編著〕(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——』国立国語研究所プロジェクト選書No.1, 三省堂
- 横山詔一・笹原宏之・エリク＝ロング・野崎造成 (1999) 「新聞記事データベースにおける「楨」の消失現象」『人文学と情報処理』No. 20pp. 57-63, 勉誠出版