YOKOYAMA Shoichi, SASAHARA Hiroyuki, NOZAKI Hironari, Eric LONG "A Study of the Use of Kanji in Electronic Newspaper Media"

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—』

国立国語研究所プロジェクト選書1
A5判 304ページ　1998年7月10日　三省堂

豊島　正之
（東京外国語大学アジア・アフリカ言語文化研究所）

キーワード

文字統計，電子化テキスト，包摂規準，符号化文字集合，朝日文字

要　旨

本書は，「朝日新聞」原紙とそのCD-ROM版テキストファイルとを照合する事によって，成立過程から既にコード化されているテキストが，別のコード化規準の下でどの様に変容するかを克明に追究したものであり，その意味で，初の「電子メディア文献学的研究」と呼ぶべきものである。本書は，「朝日文字」を含む電子化テキストという特異な例を扱ってはいるが，文献学的な手法は，本書の扱う文字全般に徹底している。

本書の方法論が，それ以前の文字計量研究との対比に於て示唆するのは，明示的で操作可能な文字の同定規準が無い限り，文字計量研究の結果は扱い難い事である。本書が紙面照合を通じて文字同定に費やした労力は，本書の文字統計の将来の価値に大きく貢献している。これに比して，従来の文字統計研究は，本書が努力した様な文字同定の手続きが不分明で，他との統計的比較が無効になり兼ねないという問題を持つものもある。

本稿では，次の略称を用いる。

CD 93	「朝日新聞記事データベース CD-HIASK'93」
CD 96	「朝日新聞記事データベース CD-HIASK'96」
「雑誌九十種」	国立国語研究所（1962）「現代雑誌九十種の用語用字」
「戦後文字」	国立国語研究所（1966）「戦後の国民各層の文字生活」
「新聞漢字」	国立国語研究所（1976）「現代新聞の漢字」
「表外字試案」	文化庁（1998）「第21期国語審議会 新しい時代に応じた国語施策について（審議経過報告）」（冊子版）所収「表外漢字字体表試案」
「平成頻度調査」	文化庁（1997）「漢字出現頻度数調査」

1．本書の構成

本書は，著者らが追究してきた「朝日新聞記事データベース CD-HIASK'93」（以下CD 93）の文

字調査の方法論・研究成果の一部，調査の結果である漢字頻度表が一冊に纏められたものであり，本書によって，この有意義な一連の研究とその根本資料が容易に参照出来る様になった。

第１章は，野崎他(1996)の一部を含む。特に，第１章 3.1.は，野崎他(1996)では等閑視されていた1983年度版 JIS 漢字が行なった1978年度版との非互換変更問題の検証に詳しく，第１章 3.1.3.は，横山他(1997)が採り上げた外字の問題を含む。第３章は，笹原他(1998)（これは笹原他(1997)とほぼ同）である。

第２章が，野崎他(1996)では希望者に頒布するとされていた漢字頻度表（JIS 区点順・頻度順）である。この頻度表は MS 明朝で印刷されているから，所謂「朝日文字」を含む朝日新聞の紙面字体とは，時に大きく異なる。つまり，ここでの印刷字体は参考で，寧ろ掲出された JIS 漢字の区点番号の方が主であり，文字頻度表というよりは，コードポイント頻度表と見られる。

この頻度表は，「紙面頻度」「CD 頻度」を区別している。後者は，CD 93 が用いるコードポイントの頻度そのものであるが，前者は，その用例を紙面に当たり直して修正した頻度である。本書の研究の根幹は，メディア変換に伴う字体変容の文献学的検証にあるが，この２欄の並存は，本書に通底するこの厳密な姿勢を象徴している。

こうした，既発表の論の集大成としての成書の場合，この書一冊で全てが足る事が期待されるが，その点，既発表の文献の参照指示のみに終る事があるのは，やや残念である。例えば，CD 93 の記事の取捨選択に就ては，本書は笹原他(1998)に譲って明記しない。しかし，CD93 に収録されない記事は本書では計量していないのであるから，この情報は，本書に取っては不可欠の情報の筈である。因みに，笹原他(1998)は（上記の様に）本書第３章に再録されているが，そこにも，この詳細情報は明記されていないので，つまり，CD 93と原紙面との記事の出入りの詳細は，実はどこにも明記されていないのである。

２．文字の統計処理と文字の同定

2.1. 文字の統計処理の前提としての同定規準

野崎等(1996)は，「新聞漢字」では異なり字数3,213，CD 93 では4,476字であるとし，これは，本書 p.4にも繰り返されている。この「4,476」は CD 93 の生の値で，「紙面頻度」では4,583，「修正 CD 頻度」では4,488（p.218）であるが，とにかく，「新聞漢字」と本書では，漢字の種類に大きな違いがある様に見える。

しかし，「新聞漢字」は「円・圓」「万・萬」等を同一視してそれぞれ一つの字種としか数えないのに対し，本書はこれらを区別してそれぞれ二つと数えるので，単純に「異なり字数」として比べる事は出来ない。本書は，JIS 漢字（JIS X0208：1997）区点は原則として全て別字として区別している。「新聞漢字」が同一視した字の一覧は得られないが，JIS 漢字が参照区点ポインタを与えてグループ化した漢字は1,293字（616グループ）あり，これらはその候補となるであろう。しかも「新字源」見出し字に無い字でも[1]本書が区別しながら「新聞漢字」が同一視していた漢字は他にもあり得る。「簇・旗」，「寇・寇」は，「新聞漢字」はそれぞれ後者しか掲げないが，本書の計量は両者を区別してそれぞれ数えている。「簇」「寇」はたまたま「新聞漢字」の原紙面に現れな

かったのかも知れないが，或は原紙面に両方が現れたものを「新聞漢字」が資料作成時に同一視した結果かも知れず，事実は知りようが無い。

つまり，「新聞漢字」の文字同定規準が不明確である以上，本書と「新聞漢字」とは，実は異なり字数の比較すら困難なのである。

これを厳密に調査するには，「新聞漢字」の元データ（翻刻結果としてのカード）に遡り，それを更に原紙面に遡って両者を逐一照合して，「新聞漢字」の同定規準を再構成する以外に方法は無い。「新聞漢字」の元データが残っていなければ，「新聞漢字」同定規準は不明のままで，本書との比較はもはや不可能になる。又，元データが残っていたとしても，この様な再構成・再調査を行う位なら，原紙面から再度（本書と同じ規準で）サンプルを取り直す方が，まだしも現実的であろう。

「新聞漢字」は，他の文字統計との比較検討を十分意識して行なわれた研究なのであるが，文字の同定規準は，現在の JIS 漢字（及びそれに則ってなされた文字調査）のそれよりも遥に緩やかなものであり，且つその同定規準が，操作可能な形では明示されなかった。

本書の厳密な文字の扱いは，文字の統計は，明示された同定規準が無い限りその意味を喪失し兼ねない事を，暗に警告している。

2.2. 文字同定の前提としての用例調査

文字を先ず実際の用例から切り離して調査したり一覧したりする事には，意味が無い。文字を，語から切り離して同定する事は出来ないからである。いきなり文字を用例から切り出してしまっては，同定自体が困難になる事は，本書の苦闘（書き方は如何にもさりげないが）が十分に語っている。

因みに，文字の符号化は，文字を文脈から切り離した形で行なわれるが，それは，こうした文字同定が成立した上で初めて可能になるのであって，その逆ではない。用例・文脈の調査を伴わない符号化等あり得ないのである。文字蒐集は即ち文字用例蒐集なのであって，文脈は無視して取り敢えず文字（文字表）だけを集める，等というのでは，本書が獲得した様な精度を維持する事が殆ど不可能であるのは，本書から容易に伺える。

新聞は，（依頼原稿・外注欄・広告等多少例外はあるが）殆どの原稿の作成から整版・印刷・出荷に至る迄を全て自社の制御下に置き，紙面に一貫して責任を負うシステムなので，字体が印刷所の都合で勝手に置き換えられる，ページ毎に違った字体が出て来る，等の問題は殆ど生じない筈であるが，それでも，本書の目指した精度の調査を維持するには，本書第１章・第３章が詳説する用例調査が必須となったのである。

2.3. 用例調査が文字調査に必須である例

電子メディアに限らず，データには様々なエラーが混入するが，電子メディア変換が介在する故に生じた新たな問題もあり，その結果，メディア上の文字情報が混乱する事もある。

この典型は誤植，類形異字の代用である。

1. 誤植

原紙面自体にある誤植は，それ自体が文字使用の実態と言えなくもないから，必ずしもデータのエラー（ノイズ）とは言えない。しかし，電子メディアへの変換に際して，誤植が悪影響を及ぼし，結果的に文字情報が混乱する事もある。

CD96では「陝西省」1件に対し「陝西省」23件が検索されるが，それらの文脈「〇西省西安市」等からすれば，当然「陝」が期待される。本書によればCD93のCD頻度も「陝」(80-01) 4対「陝」(79-93) 27で，こちらも「陝西省」を多く含むかと疑われる。正誤が件数上逆転している訳だが，これは，「陝」を「陝」とした誤植が，CD編集過程で更に誤ってコード変換された結果であるらしく，数少ない誤植が変換テーブル作成のエラーを誘発し，更に正誤を逆転させた可能性も考えられる。

2. 代用

CD 96では「柿（こけら）落とし[2]」2件に対し「こけら落とし」31件である。CD 93で「柿」がどの程度「こけら」に用いられているかは本書からは不明であるが，こうした計量結果が，そのまま文字の同定を表すと見てよいかは疑問で，寧ろ代用（或は誤用）の可能性がある。CD 96には，人名の韓侘胄（長子）を冑（武具）とした例もある。

こうした代用には，JIS X0208：1997が「区点位置詳説」で取り上げた「叱」・「姫」・「籾」等もあるが，さすがにこれらの字の「本来の」用法は極く稀なので，これらは，完全に同定されていると見てよかろう[3]。

CD 93には，紙面に仞 (48-32) 仭 (48-33) に関連するゆれがあり，CDにゆれが無く (p.285)，CD 93には，紙面に鴎のゆれがあり，CDに無い (p.286)。この様に，紙面には無くCD化で初めて生じた「代用」もある。

以上の様に，文字調査は，必然的に文脈・語の調査を伴わねばならない。1983年度版JIS漢字の非互換変更（第1章 3.1.1.）と所謂「幽霊字」（第3章）に就て，これを徹底して行なったのは，本書の眼目の一つであり，そこには，一々の用例を原紙面と照合して文脈を把握した上で，文字の同定を再検討するという厳密な姿勢が貫かれている。

3．文字調査での文字の同定

これまで行なわれた文字調査での同定の方法には，次の様なものがある。

1. 語による同定

（a）「雑誌九十種」（第2分冊p.15注）

「円・圓」「台・臺」「予・豫」を同一視し（これは，97JISの包摂規準とは比較にならない程緩やかな同定規準である），「歳・才」「連・聯」「裏・裡」を区別する。「連・聯」「裏・裡」を区別しながら「予・豫」（ワレ・アラカジメ）を区別しないとする理由は，明記されていない。この為，例えば「余」「餘」が区別されたのか否か（「餘」の立項が無いのは「余」に同一視されたのか，それともたまたま「餘」の用例が無かったのか），判然としない。

（b）「新聞漢字」(p.11)

「万・萬」，「餅・餠」を同定する点で，やはり極めて緩やかな同定であるが，同定規準

が「雑誌九十種」と一致しているのか否か，不明である。

例えば，「雑誌九十種」が区別する「峰・峯」は「新聞漢字」には「峰」しか立項が無いが，「峯」が「峰」に同定された結果か，それとも実際に「峯」の用例が無かったのかは，分からない。「雑誌九十種」には「萠」のみが立項され，「新聞漢字」には「萌」のみがあるが，これも，実は両者同一字体を計量した可能性もある。

（c）「戦後文字」

「新字体」「旧字体」という区分があるが，「旧字体」が具体的に何を指すか不明なので，計量結果の解釈に問題が残る。

例えば，既を新字体，常用漢字表の「いわゆる康熙字典体」旣を「旧字体」とするには恐らく異論が無かろうが，この時，康熙字典内府本の旣[4]は「旧字体」ではないのか，もしこれも「旧字体」とするならば，既（康熙字典字体）・旣（「いわゆる康熙字典体」）の相互は同定されるのか，等の疑問に対しては，「新字体」「旧字体」という区分は，何も答えて呉れない。

語による同定は，これを推し進めれば一字多義のものは別字となり，極端には音訓の別（例「為」）すら別字と判断される事になり兼ねない。更に，語としての意味を欠く[5]固有名詞は更に同定方法が必要な筈だが，この辺りの判断規準は，不明というより他はない。

2. 既に為されたコード化に依拠するもの

この場合の文字の同定規準は，元のコード化に際しての同定規準に全面的に依拠する事になる。

（a）活字使用実績調査によるもの

これは，原則として一つのタイプフェイス（時にサイズも）で通した，特定分野の印刷物に限って行なわれて来たが，その為，結果が偏る事もある。

i．「本邦常用漢字の研究」(1941)（内閣印刷局）

内閣印刷局（当時）が「従來…關係方面にのみ配付し來たるも，今後之を廣く一般に頒布することとし」たもので，「官報號外帝國議會議事速記録」の組版に使用した活字の使用度（発注伝票によるか）調査である。この為，使用第１位が「議」，第２位「第」，第３位「君」と，結果はかなり偏っている。第１位・第２位は官報の「柱」に用いられた結果であろう。

ii．毎日新聞漢字調査（1957）

全国共通の組版システムの為，外字を一意に指示する外字表の策定が必要になり，その基礎研究として行われたもの。

前文によれば，毎日新聞社は何度か漢字調査を行なっており，古くは手組版からモノタイプ文字盤設計への基礎研究として「活字鋳造日票」による活字使用度数調査を行なっていたが，鋳造日票は活字ストックを発注するものなので，ストック中の活字が確実に全て印刷に回ったとは思われず，「回転率の悪い低使用度のも

の」が実際の使用例よりもかなり大目に出て仕舞う。この為，「外字ケース補充日票」による活字使用度数調査に切り替えて調査を進めたもの。これでも，「途中に誤ってコボしたような事故［の結果必要となった補充字］もそのまま記録され，また文選者によって誤って拾われ，校閲で除外されるようなものも記録されてしまう欠点」がある。

新聞社による文字調査の目標が漢字制限である事は，標榜されている。「新聞社がこのように漢字の使用制限を提唱したのは当然のことである。新聞はまず広く読まれることが絶対の条件であり，広く読まれるためには難解な文字の使用を避けなければならない。」(p.1)。昭和30年代には，こうした考え方に基づく漢字制限は当然の事とされており，

> 「氏名の名の方は…［漢字制限で］ともかく解決したと思われる。…氏の方は今後も残されるが…氏は夫婦のいずれをとってもよいことになっている。むずかしい漢字の氏の人はなるべく当用漢字，人名漢字の範囲内の氏の人と婚姻し，氏を変えてほしいものと思う。」(p.15)

とあるには驚かされる。冗談なのかとも思うが，その直後に，

> 「人名の方は…極めて徐々にではあるが改善の方向に進んでいるように思われる。残ったのは地名である。閖上町のように仙台付近の町名は全くのローカル性の字でこれが読めるのは国民の何パーセントであろうか。昭和28年（1953年）10月１日に施行された町村合併促進法により多くの町村が合併し新しい市ができた。何とか手は打たれてなかったかと期待したが全く期待は外れ，むずかしい字が使われていた。あとで文部省…から伺ったところによると文部省国語課から自治庁あてに新市名に当用漢字励行方を勧告したが，これを定めるのは地方議会の議員さんたちで徹底することができなかったとのことであった。」(p.15)

と続く。いずれにせよ，現在の目からは驚くばかりである。

（b）漢字コードによるもの

ｉ．JIS漢字コード（JIS X0208）によるもの

本書を始め，多数ある。

ii．印刷会社内部コード（CTSコード）によるもの

「平成頻度調査」。調査結果自体が，更に印刷される事による問題（後述）と，サイズ，タイプフェイスの抽象化規準・同定規準が公開されていないため，計量の対象が明確にならない事がある。

（c）独自の同定規準を明示するもの

ｉ．「太陽」調査

木村他(1999)が，自ら雑誌「太陽」を電子化データにする際に，文字の同定規準を策定し，操作可能な形で明記して，文字の同定に就て詳細に論じたもの。同定規

準は，JIS X0208：1997の「包摂規準」を批判的に独自拡張したもので，電子化データでの文字同定に関しては，本書と並んで必見の文献である。

4．新聞 CD-ROM と新聞紙面

4.1．朝日文字の問題

CD 93 には，他の電子化テキストにはない独特の問題がある。

朝日新聞社は，戦前の「字体整理案」や「当用漢字字体表」が採った様な部分字体の統一的簡素化を積極的に推進した所謂「朝日文字」を利用する事で名高い。

CD 93 は，紙面では「朝日文字」を印字しながらも，それを CD-ROM に変換する時には JIS 漢字字体の区点位置に変換している。仮にこれを逆転して，紙面の文字列が CD 93 の電子化テキストの印字出力であると見るならば，「朝日文字」による印字は，出力装置としては規格非合致である，とも見る事が出来る。

本書は，ここに大きな問題がある事を自覚しているが，「朝日文字は文字種が多く…頻度も高いため」（p.16）この点に就ては原紙照合を行わなかった。このままでは，コード化済のテキストが紙面に優先する逆転状態のままで，本書の厳密な姿勢は，ここでやや一貫を欠いている。

この点，本書は不徹底ともいえるが，著者チームは，この問題を解決すべく，既に朝日文字の原本照合に乗り出しており，その成果の発表が待たれる事である。

4.2．CD-ROM 不採録部分

原紙の「見出し」部分には，文字列としての電子化データが無く，「絵」として紙面に貼り込まれるものもあると聞く。CD 93 の見出し部分は，実際に紙面製版に使われたデータではなく別途入力されたもので，原紙と大きく異なる事がある事から，本書ではデータから捨てている。

又，CD 93 は画像データを収めない為，CD 化に当たって画像を文章に「翻訳」したケースがある（p.17）。これらも，本書ではデータから捨てている。

本書の様な電子化データ自身に対する徹底した研究に於ては，こうした CD-ROM 不採録部分に就ては，原紙のどこが不採録であるかの詳細な記述が望ましかった様にも思われる。

5．コード化済のテキストを調査する意味

5.1．暗黙の文字包摂

文字調査は，暗黙の文字包摂を排除出来ない。

文化庁国語教科書調査（文化庁 1975）は，表外字の洗い出しが目的なので，（「万・萬」を同一視する「新聞漢字」の様な）「語による同定」を行っていない事は，明白である。

しかし，調査対象と思われる教科書[6]から「逆引き」を行なって見ると，実際には，草冠３・４画の同一視，しんにょう１・２点（辻等）の同一視[7]は，当然の如くに行われている。この他，暗黙に同一視されている例をいくつか挙げれば，

1．[目]関連の同一視（「表外字試案」は認めない）。

報告書は，何れも 𠦂 に作る。

- 𩙿 応 三省513 p.246
- 𩙿 応 学図505 p.261

2.「食へん」新旧の同一視（「表外字試案」は，特定字以外は認めない）。

報告書は，旧字体に作る。

- 餞 別 学図505 p.264

3.「殻」の一画増減

調査は当用漢字時代で，𣏌 は当用漢字補正案字。報告書は 殻 に作る。

- 𣏌 を 三省513 p.126（他に旧字体の教科書用例あり）

等がある。これらは，（ここに用例を摘記した様に）実際の教科書にはいずれも両様に現れており，調査・集計に当って暗黙に同一視されたと覚しい。この調査が，表外字の調査というよりも表外字を必要とする語の調査の性格が強い為に必然的に現れたものであろう。

5.2. 印刷に伴う暗黙の包摂と見掛けの多様

印刷会社は，それぞれ代表字体を持っており，要求された字体は，別段の指定が無い限りその代表字体に置換して植字する。この場合，印刷会社によって包摂された字体は，実際には統一が行なわれている様に見える。

この統一が，複数の出力元で共通しない場合は，その統一は見掛け上のもので，実際には，出力された文字の背後に多くの字体が包摂されている事になる。時には，その「代表字体」が，同一社内でも入れ替わる事（規準変遷）があり，この結果，実態は変わらないのに逆に見掛け上の変化が生ずる事もある。

「平成頻度調査」は，凸版印刷の内部コードに依拠しており，この見掛け上の多様現象が生じている。「表外字試案」124番「卉」は，凸版印刷4,500位以内に草４画体 卉 しか無い事を理由に，草３画体 卉 が認められない[8]のであるが，実際には

	４画体	３画体
凸版印刷 昭和50年度報告書		27回 (2,645位)
凸版印刷 昭和51年度報告書	27回 (2,767位)	
凸版印刷 平成９年度調査	184回 (2,865位)	

と，昭和50年/51年の間に字体が全く入れ替わっている。昭和50年→51年の一年で印刷物に一斉に４画体へのシフトが起こった可能性もあろうが，寧ろ，調査報告書が昭和50年度には３画体で，昭和51年度には４画体で印刷されただけで，頻度統計としては実は同じものであると見る方が，自然の様に思われる。

もしこれが正しければ，「卉」の変化は，実は印刷所の「代表字体」の置換えによって生じた見掛け上のもので，実際の字体は両者弁別不能であり，どの字体が計量されたのか不明なのである。

5.3. コード化済テキストの調査

既存のコード系（必ずしも符号化文字集合ではない）に依拠する調査では，こうした暗黙の包摂が顕在化しない。包摂に関する不安定部分は，全てコード化の段階で切り捨てられている。

しかし，この切り捨てに無自覚にテキストを扱うならば，それはテキストをデータとして調査した事にはならない。

コード化済テキストにも，印刷所のそれと同じく規準変遷がある。これに，1983年度版 JIS の行なった非互換な字体変更の問題が絡むと，話は大変厄介になる。長い期間維持されている大規模な「外字表」（メインフレーム各社の外字表等）は大抵この問題を抱えており，1978版 JIS 時代に作成した「外字」（例えば 麺 ）が1983版 JIS で「内字」になって仕舞い，その結果「 麺 」（1978版 JIS 時代の「内字」）を更にもう一つ「外字」として登録した結果，都合３つの「麺」区点を持つデータが入り乱れて，どれがどの字を意図して入力されたか分からなくなったという大混乱も，稀ではない。

5.4. 本書の原本照合の意味

本書は，新聞紙面が如何に CD-ROM 化されたかの調査である。しかも，朝日新聞の制作過程としては，紙面自体が既にコード化された形で作成される。つまり，本書は，コード化されたテキストが，更に別のコード化規準の下で如何に変容するかの調査と見る事が出来る。

新聞紙面の編集データがそのまま CD-ROM に焼かれる訳ではなく，そこには変換・整形過程があり，「紙面のコード化」過程が存在している。新聞 CD-ROM を新聞紙面に対して原本照合するという，本書が採った文献学的な研究手法は，この紙面のコード化の検証なのである。本書第１章・第３章を倉卒に読むと，恰も朝日文字・83 JIS 非互換字・外字だけが問題であるかの様に見えるが，これらは最も目立つ形で顕れているに過ぎず，本書の綿密な研究手法は，全ての字に及んでいるのである。

5.5. サンプリングに拠らない事の意義

本書の特徴の一つに，サンプリングに拠らず，全数調査を行った事がある。

従来の国立国語研究所による文字・語彙調査は，サンプリングによるものが殆どである。これは，統計処理の段階では調査対象（90種雑誌等）そのものを母集団とみなすにしても，研究の真の対象は，調査対象（universe）の背後にある現代日本語そのものであって，調査対象である雑誌などは，既にその言語のサンプルである，という暗黙の前提に立っている。これに対し，本書は，所与のテキスト全体を文献として扱って，そこからサンプルを抜いて母集団推測を行なう等はしない。これは，コンピュータの発達で全数調査が可能になった事もあるが，寧ろ，電子化テキストをあくまで文献学的に扱うという本書に通底する徹底した姿勢に由来するものである。

5.6. 本書の採った研究手法の意義

本書は，漢字コードという既存の同定規準に強く影響されるものに依存している事を自覚し，

原紙面に立ち返って，その同定規準を文献学的に検証しようとした点で，空前といってもよい位置を占めている。

近時市場に又はネットワーク上に溢れている電子化テキストは，言語研究にも盛んに利用される様になっているが，本書の様に，そのテキストのデータとしての性質から吟味する研究は極めて稀である。電子化テキストそのものを便利なデータとして受入れ，「テキストデータベースを作成する手間が…なくな」り，且つ「信頼性のあるテキストデータ」であると評価するもの（荻野他1994）もあるが，本書の実践する様なデータの性質の吟味抜きでデータの信頼性云々は虚しい。本書がCD 93を採用したのは，単に楽にデータを採れるから等という安直な姿勢によるものではない。本書の様な紙面照合手続きを踏んで電子化テキストデータを扱うのは，寧ろ大変な労力を要する。

本書の研究対象は，メディア変換とコード化に伴って生ずる文字区別の変容自体なのである。著者達の努力が教える事は，文字はメディアから自由にあるものではなく，そのコード化にも，予め定められた「原本に忠実」なコード化がある訳ではない事，メディアに応じて変化する文字は，メディアの上に於て検討されねばならない事である。

この意味で，本書は，「電子メディアの文献学的研究」の嚆矢と言うべく，その厳密な姿勢は（若干徹底を欠く憾みはあるものの），単に「入力する手間が省ける」といった安易な姿勢の電子化テキスト利用の対極に位置し，以て範とすべきものである。

注

1　JIS X0208：1997は，「新字源」の見出し字に限って参照区点ポインタを新たに追加した。このため「薗・園」，「競・競」，「剳・箚」等のポインタは（それぞれ左が「新字源」見出し字に見えない為），張られていない。

2　括弧書きは原本のまま。

3　「姫」は当用漢字字体表の時点で，既に「代用」ではなくなっている。一方，「叱」は，表外字試案は認めない事としている。

4　図は書香文庫蔵本による。

5　太田さんが太っている保証は無く，良子さんが良い子だとは限らない。又，meaning と referent を混同したものは措く。

6　報告書には，年季以外は調査対象の教科書の明記が無いので，当該年季である昭和49・50年高校「現代国語」の教科書から，東京書籍，学校図書出版，教育図書研究会，三省堂，筑摩書房発行の文部省検定済教科書を検した。

7　「表外字試案」は特定字以外認めない。

8　表外字試案が，草４画・３画を区別するのは，この字のみ。

引用文献

荻野綱男・塩田雄大（1994）「朝日新聞データベースを利用した言語研究」『日本語学』13-5
木村睦子・田中牧郎・飯島満・笹原宏之（1999）『『太陽』コーパスの漢字処理』科学研究費研究報告書・新プロ「日本語」研究班４・木村チーム
国立国語研究所（1962-64）『現代雑誌九十種の用語用字』（国立国語研究所報告21,22,25）秀英出版
国立国語研究所（1966）『戦後の国民各層の文字生活』（国立国語研究所報告29）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
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」『シンポジウム 人文科学における数量的分析（2)』予稿集,「人文科学とコンピュータ」数量的分析計画研究班
笹原宏之・横山詔一・野崎浩成・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字」『計量国語学』21-4
内閣印刷局（1941）『本邦常用漢字の研究』内閣印刷局
日外アソシエーツ（1994）『朝日新聞記事データベースCD-HIASK'93』紀伊國屋書店
日外アソシエーツ（1997）『朝日新聞記事データベースCD-HIASK'96』紀伊國屋書店
日本工業規格JISX 0208：1997『７ビット及び８ビットの２バイト情報交換用符号化漢字集合』日本規格協会
野崎浩成・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」『日本教育工学雑誌』20-3
文化庁国語課（1975）『国語教科書（中学校・高等学校用）における表外漢字の出現状況―高等学校国語　現代国語」編―』
文化庁国語課（1997）『漢字出現頻度数調査』
文化庁（1998）『第21期国語審議会 新しい時代に応じた国語施策について（審議経過報告)』
毎日新聞社東京本社印刷局（1960）『外字調査報告書』（毎日新聞社）
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)』予稿集,「人文科学とコンピュータ」数量的分析計画研究班

謝　辞

康熙字典内府本の利用をお許し戴いた書香文庫境田稔信氏，CD-HIASKに就て種々御教示戴いた比留間直和氏（朝日新聞社用語幹事室），御蒐集の教科書の閲覧を許された東京書籍印刷株式会社東書文庫，過去の文字調査資料各種の閲覧を許された国立国語研究所図書館に深謝申し上げます。

Review

YOKOYAMA Shoichi, SASAHARA Hiroyuki, NOZAKI Hironari, Eric LONG

A Study of the Use of Kanji in Electronic Newspaper Media

TOYOSHIMA Masayuki
Institute for the Study of Languages and Cultures of Asia and Africa
Tokyo University of Foreign Studies

Keywords

character census, machine readable text, unification principle, coded character set, character identification

Abstract

The authors' study, based on an exhaustive listing of the kanji in the articles of *Asahi Shinbun*, one of the major newspapers of Japan, suggests that for valid analyses of character statistics, an explicit and stable procedure for character identification is crucial. The most typical case of this is the problem of the so-called "Asahi characters" (simplified versions of kanji used in *Asahi Shinbun* which are not sanctioned by government decree).

There are numerous discrepancies between the CD-ROM text of the newspaper and the printed one, especially because the character identification principles used in production of the CD-ROM version were sometimes inconsistent. The authors dedicated a great deal of work to the verification of the CD-ROM text based on the printed text, which justifies characterizing this work as the first philological study of machinereadable Japanese texts.