新聞漢字調査の現状と将来

著者	横山 詔一, 笹原 宏之, エリク ロング, 谷本 玲大
雑誌名	日本語科学
巻	9
ページ	33-42
発行年	2001-04
URL	http://doi.org/10.15084/00002054

新聞漢字調査の現状と将来

付録：新聞記事データベースにおける「槙」の消失現象 (HTML 文書と外字 GIF)

横山 詔一
笹原 宏之
エリク・ロング
(国立国語研究所)
谷本 玲大
(茨城大学)

キーワード

文字計量，新聞，電子化テキスト，メディア変換

要 旨

新聞漢字調査について，豊島(1999)の論考を羅針盤としながら国内の状況を概観し，今後の調査に資する視点の設定を目指した。おもに新聞記事を電子化する際に原紙と電子化テキストの間で齟齬が生じる背景を考察し，メディア変換に伴って必然的に発生する諸問題の整理を試みた。そして，以下の提言を行った。将来，独立行政法人・国立国語研究所が新聞漢字調査を実施する場合は，調査精度と費用のバランスという観点から，大量の原紙を研究所側で電子化する作業は避けつつ，また外部から購入した電子化テキストを無批判に受け入れることもないよう，新聞社等の協力を得ながら原紙の組版に使用された文字データを分析するのが望ましいと考える。

以下のように略記することがある。

- 国語研選書1 ← 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――』(1998，三省堂刊)
- NTT データベース ← NTT データベースシリーズ『日本語の語彙特性』(2000，三省堂刊)

1．はじめに

近年，認知科学や心理言語学の領域では漢字の心的辞書 (mental lexicon) に関する研究が盛んになっており，海外からも注目を集めている (Kess & Miyamoto, 1994, 2000)。認知実験で漢字を被験者に呈示する場合は，漢字の出現頻度をあらかじめ統制しておくのが望ましい。このような背景も手伝ってか，電子化された漢字頻度表の必要性が国内外で高まりつつある。しかし，インターネット上での無償公開を目指している漢字頻度表はChikamatsu, Yokoyama, Nozaki, Long & Fukuda (2000) を除いてほとんど見当たらず，いささか寂しい感もする。単語頻度表については，日本語の場合は語の区切りについて定説が存在しないために整備が難しいことが広く理解されているが，文字単位の頻度表がなぜいまだにインターネット等で公開されないのか，疑問の声が聞かれるようになった。

この問題は認知科学などの領域だけで顕在化しているものではなく，日本語教育や国語学さらにはメディア政策の研究分野にも大いに関係があると言えよう。21世紀は電子メディアの時代である。もちろん紙の媒体は残るであろうが，従来型の印刷システムの圧倒的な優位性は次第に崩れつつある。押し寄せる印刷革命の波を越えて，精度の高い文字調査を経年的に展開するには何が必要となるのであろうか。かかる視点から，本稿は，文字計量研究に関する豊島(1999)の論考を羅針盤としながら国内の現状を概観し，「新聞記事全文データベース」を利用した漢字調査を中心に問題点を整理する。併せて，関連する文献の一つをHTML化し，付録のCD-ROMに収める。

2．原紙による調査

漢字使用率に関する標準的な資料として，『現代雑誌九十種の用語用字』(国立国語研究所，1962)と『現代新聞の漢字』(国立国語研究所，1976)がある。いずれの調査も，研究者が原紙を手元に置きながら作業を進めたので，ここでは「原紙による調査」と呼ぶ。(後者は世界で最初に大量の漢字仮名交じり文をコンピュータ処理して得た成果であるが，電子化テキストの作成において研究者は原紙に目を通している。)これらは，常用漢字表の制定やJIS漢字の選定，さらには漢字認知実験などにおいて幅広く活用され，諸学界から高い評価を受けてきたことは周知の事実と言えよう。しかし，認知科学の研究者からは，以下のような意見・要望を耳にすることも少なくはない。

(1)　使用率の低い漢字のデータも参照したい。

例えば，『現代新聞の漢字』は使用率0.001%以下の漢字を掲出していない。その理由は，文字資料の収集が紙面からのランダムサンプリングによるため，使用率の数値が推定値であることによる。使用率を統計的に区間推定すると，0.001%以下の場合は十分な精度が得られないことから，報告書から除外されている。これは，統計学的見地からすれば極めて妥当な方向である。しかし，認知研究においては，実験刺激選択の目的で使用率0.001%以下の漢字リストを参照したいという研究者が少なからず存在するようだ。

ちなみに，『現代雑誌九十種の用語用字』は1997年からフロッピー版が市場に出ている(国立国語研究所，1997)。そこには雑誌から抽出した調査サンプルに一回でも出現した語がもれなく記載されており，表記のユレも網羅されている。例えば，「あう」の場合，「合う」と表記されることが最も多く143回，「会う」97回，「あう」86回，「逢う」46回，「遭う」7回，「遇う」3回，であることが分かるようになっている。

(2)　調査に使用した記事の発行から約30年の歳月が経過し，我が国の活字メディアにおける漢字使用の実態がかなり変化しているのではないか。

この30年間で社会情勢が大きく変化していることを考慮に入れれば，漢字使用の実態が変化しているかもしれない。もしそうであれば，国立国語研究所報告書に基づいて漢字刺激を選択することは今やあまり適切ではないのかもしれない。

経年的な文字調査の実施という観点からは，豊島(1999)の指摘を見逃してはならない。豊島によれば，これまでになされた国立国語研究所の文字調査は「文字の同定規準」が不明確である。『現代新聞の漢字』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が，そ

の他にどのような漢字を同一視したのかのリストが存在しない。そのため，国語研選書１のように「円・圓」「万・萬」の他にも多くの字体を区別する調査との異なり字数の比較さえ正確にはできなくなってしまった。この指摘は，結局のところ，現在の文字調査の学術的価値を判定するポイントを示しているように見える。文字同定規準をどう立てたのか，またその手続きを操作的にいかに定義しているのか，この２点をどの程度まで徹底して追究しているかが，当該の文字調査の将来における評価を左右するという。

（３） 漢字頻度表が紙の媒体でしか公開されておらず，利用者にとって不便なので電子化データを公開して欲しい。

先にふれたように『現代雑誌九十種の用語用字』についてはフロッピー版が公刊されている。しかし，『現代新聞の漢字』については今すぐに電子媒体で公開される見込みは薄い。紙媒体で公刊されたデータを電子化するにあたっては，著作権法にからむさまざまな問題を慎重に解決していかなければならず，時間がかかりそうに見える。

以上三つの意見は，国立国語研究所が実施した漢字調査の本来の目的を等閑視しており，その意味ではいささか的外れな側面がある。そもそも認知実験の刺激選択など眼中になかったからである。その一方で，国立国語研究所に対する期待・要望の一つとして耳をかたむけるべき内容を含んでいるようにも感じられる。

３．電子化テキストによる調査

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，本稿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。

原紙の文字は，新聞社内におけるコンピュータ組版の文字コード（以下，組版コードと呼ぶ）によって決定される。視点を逆転させて，原紙の文字が組版コードに「正確に」反映されている保証がある場合は，組版コードによる文字同定は妥当だということになる。

3.1．組版コードから変換されたJIS漢字コードによる調査

1990年代になってCD-ROM化された新聞記事全文データベース（以下，新聞CDと呼ぶ）が一般の市場に登場した。そこ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。新聞CDを対象にした文字調査の先駆は横山・野崎(1996)と野崎・横山・磯本・米田(1996)である。その後，Long & Yokoyama (1997)，久野(2000)，Chikamatsu et al.(2000)などが続いている。

最近は，企業体が文字・単語頻度表を販売する動きも出てきた。その初例がNTTデータベース（第７巻）である。冊子の解説によれば「これまでの文字・語彙調査で最も高い信頼性を有する」とある。果たして，この認識は正しいのであろうか。

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，コンピュータまかせで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，残念なことにNTTデータベースの解説にはそのような記述が一切ない。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である。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)が教えるように，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。外部から購入した電子化テキストをそのままコンピュータ処理するだけなら，学術的価値は一体どこに生じるのであろうか。

原紙の表記と電子化テキストの間には齟齬があるのが普通である(横山・笹原, 2000)。朝日新聞にもさまざまな問題が生じており，その実例の一端が国語研選書1や横山・笹原・ロング・野崎(1999：付録CD-ROMに収録）に報告されている。それらのうち，ここでは「異体字」の問題に関連する「JIS漢字のネジレ」を説明する。(異体字とは，読みも意味も同じで形だけが異なる「桧－檜」のようなものを指す。)

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「78 JIS」，1983年の第2次規格を「83 JIS」と呼ぶ)。その影響で，一つのコードポイントが二つの漢字字体を示すケースがある。83JISでは区点番号41-16で示される字は「桧」であるが，78 JISだと「檜」になる。もし，「文字頻度は，桧が50で檜は100」と83 JISで入力・作成し，78 JISで表示・印字したならば，「文字頻度は，檜が50で桧は100」となってしまい，「桧－檜」の文字が入れ替わる。つまり，情報が正確に伝達されない。

以上がJIS漢字のネジレである。多くの新聞社は記者が社内で過去の記事を検索できるよう記事テキストデータベースを構築・管理しているが，大抵はこの問題を抱えて混乱に陥っており，システム担当者の頭を悩ませている。社によっては，この問題が生じていることに気付いてさえいないところもあると聞く。

3.2. 組版コードによる調査

新聞社や書籍印刷会社の組版コードによると思われる調査もある。いずれも文化庁によるもので，『漢字出現頻度数調査』(1997)と『漢字出現頻度数調査（2）』(2000)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てい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。

4．電子化テキストと原紙を併用した調査

国語研選書1は「組版コード → JIS漢字コード」というメディア変換にともなう問題点を詳細

に検討した。組版コードのデータは入手不可能なので，原紙（縮刷版）と朝日新聞CD（CD-HIASK'93）の内容を比較照合し，両者の相違点を洗い出した。このような方法により，新しい知見がいくつかえられた。それらは以下の四つに集約できる。

（1） 原紙に高頻度で出現し，しかも83 JIS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「尭，槙，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，「作家森瑶子氏と槙原投手，遥かな旅に」と83 JISで入力し，78 JISで表示すると「作家森＝子氏と＝原投手，＝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。

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「＝」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(横山他，1999：付録CD-ROMに収録)。

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。〔注1〕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，「尭，槙，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）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。

（2） 83 JISにない漢字（83 JIS外漢字）のうち新聞によく使われるのは「補助漢字：JIS X 0212」に規定された160種である。

朝日新聞CDでは83 JIS外漢字の個所は「^」でマークしたうえ，仮名に開くというルールがある。そこで，電子化テキストから「^」でマークされた部分をすべて抜き出し，該当箇所を東京本社版の縮刷版で逐一字体を確認するという作業を行った。そうした調査の結果，83 JIS内漢字は4,300種を超え，延べ字数では99.99%以上を占めていることが判明した。

この残りが83 JIS外漢字である。延べ字数は少ないが，異なりでは240種を超えており，電子メディア上では無視しえないものである。この83 JIS外漢字は笹原・ロング・横山(1999)によって以下の四つに分類された。カッコ内に異なり字数と延べ字数をそれぞれ示す。

- 補助漢字でカバーできた字（異なり字数：160，延べ字数：1,308）
- 補助漢字で包摂できそうな字（23，29）
- IBM特殊漢字（12，40）
- その他（48，78）

ちなみに，「補助漢字」とは「JIS X 0212」(1990)の漢字（5,801字）の意である。これによって表現できる字は83 JIS外漢字全体において異なりで65.8%，延べで88.8%にのぼった。「補助漢字で包摂が可能かと思われるもの」を加えると，それぞれ75.3%，90.2%に達し，そのカバー率は

かなり高いことが明らかとなった。

「IBM 特殊漢字」は，ユーザーからの要望によってIBM 社で独自に追加した外字（279字）である。それは実際の需要の一部が蓄積されたものであるだけに，83 JIS 外漢字全体において補助漢字でカバーされた字を除いて，異なりで4.9%，延べで3.1%を補うことができた。IBM 特殊漢字は，補助漢字と重なる字を加えると，異なりで約３分の１，延べで半分以上の83 JIS 外漢字をカバーすることになる。なお，補助漢字にしかない83 JIS 外漢字は，83 JIS 外漢字全体の40%前後あった。

「その他」は補助漢字にもIBM 特殊漢字にもなかったもので，これが83 JIS 外漢字全体において，異なりで19.8%，延べで6.0%を占めた。

（３） 新聞には辞書に掲載されていない漢字（いわゆる辞書非掲載字）が出現することがあり，しかもそれが誤用ではないかと疑われるものが散見される。

漢字を大量に集めた資料として，我が国では漢和辞典や漢字辞典がある。しかし，それらの辞典に収められている漢字と，実際にさまざまなメディアで流通している漢字とを比較すると，互いに一致しないものがある。漢字の音訓や熟字以外では次の二つが挙げられる。

一つは，字種すなわち字の種類である。辞書には掲載されていない漢字（辞書非掲載字）については，笹原・横山・野崎・米田(1997)や国語研選書１に先行研究がある。〔注２〕それによると，大規模な漢和辞典にもない漢字のうちで，最も多いのは日本人の固有名詞，つまり姓名や各地の地名などに使われている字である。姓の「彌」，地名の「峅」のような国字（日本製漢字）や人名の「廸」のような漢字の異体字がその大部分を占めている。これらの頻度そのものは高いとはいえないが，日本語の文章を表記する際に欠くことのできない文字として，日常的に使用されている漢字である。

もう一つは，字体である。とくに手書きに基づく略字体には辞書に掲載されていないものが少なくない。例えば，常用漢字表にある「事」「畳」であっても，「爭」「疂」のような常用漢字表にない字体が朝日新聞CD と原紙の両者に使用されているケースもあった。特に「爭」は「〇務局」という文脈で使用されており，「事」と誤って用いられたものと確認できた。

ところで，常用漢字の表外字については，漢字そのものは旧字体で辞書に載っていても，その略字体（新字体）までは載っていないものが少なくない。辞書とは違って，朝日新聞は「鷗｛旧｝」を紙面に「鴎」のように印刷することが多い。このいわゆる「朝日字体」は，常用漢字の字体を常用漢字でない漢字にも拡張して準用した「拡張新字体」の一種とされ，JIS 漢字と一致するものもある。〔注３〕そのため，朝日新聞CD をパソコン画面上で見ると「鴎」と表示される。ただし，朝日新聞CD の「鴎」が本当に紙面でも「鴎」となっているかどうかは，原紙で確認する必要がある。

このような紙面照合の調査は，国語研選書１にその成果が示されていたが，調査対象をJIS 漢字のネジレに関係する字，「=」字，「^」字などに限定した調査であった。そこで，笹原・横山・ロングは，朝日新聞社の協力を得て，調査対象とする字の範囲を300種余り，延べ3000箇所以上へと広げ，朝日新聞における漢字字体の使用頻度のほぼ全貌を捉えつつある。朝日新聞でも拡張新字体ばかりでなく正字体を使っているケースがあることのほか，朝日新聞CD では一つの字体とさ

れているのに原紙では二つ，極端な例では三つの字体が出現しているケースなども確認されている。

（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「摑〔旧〕」「頬〔旧〕」「剝〔旧〕」と印字するのが一般的であるが，83 JIS 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 83 JIS 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 et al. (2000) は，83 JIS に含まれない「葛〔旧〕」「摑〔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83 JIS において拡張新字体のみが採用され，旧字体は83 JIS 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「掴む」について「摑〔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「摑〔旧〕む」と表記する場合が圧倒的に多く，「掴む」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，「掴」よりも「摑〔旧〕」，「頬」よりも「頬〔旧〕」，「剥」よりも「剝〔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，1998，2000；横山・笹原，1999)。ここでの好みとは，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・野崎，1998)。

以上の事実は，たとえ認知科学の研究であろうとも，漢字刺激を扱う際は異体字や83 JIS 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。被験者が見慣れてい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の精度を低下させるおそれがあるからである。

5．まとめ

将来，諸研究機関や研究者あるいは独立行政法人・国立国語研究所が大規模な新聞漢字調査を実施する機会に恵まれたならば，コストと調査精度のバランスという側面を重視して，以下の点を考慮に入れた計画を立案するのが望ましいと考える。

（1） 新聞社内の組版コードに基づく電子化テキストを調査対象とするのが合理的である。従来のように，研究所側で原紙から電子化テキストを作成することは避ける。原紙を電子化して蓄積するには，著作権者である新聞社から著作権使用許諾を得なければならない。そのため，現時点ではかなり高額な対価を新聞社から要求される。〔注4〕大量の原紙を入力す

るための人件費・役務費などのコストも膨大なものとなるだろう。そもそも新聞社内には組版コードによる電子化テキストが存在するのだから，それを利用しない手はない。

（2） ただし，組版コードによる電子化テキストを無批判に受け入れてはならない。JIS 漢字のネジレに関係する文字と組版コードの対応関係などについて，原紙との照合を必ず行うべきである。〔注５〕

（3） 原紙との照合作業においては，組版コードに対応する漢字を表示・検索できるシステムが必要である。それがない限り，組版コードと漢字との対応テーブルを見ながら作業しなければならず，現実的ではない。原紙に登場した JIS 外字を JIS 漢字コードやユニコードに包摂するのは避ける。〔注6〕できるだけ原紙の字体と１対１に対応する電子化テキストを作成することを目指す。このような高度な技術は，産官学の協同で開発に取り組まないと成功はおぼつかないであろう。21世紀の新聞漢字調査は，産業界の資本と人材でシステム開発を行い，学界が原紙の表記を字体レベルで峻別し，官界が調査結果をメディア政策に反映させる，という役割分担が求められるようになるであろう。

６．付録 CD-ROM に収録したファイルについて

6.1. HTML ファイルについて

朝日新聞 CD（CD-HIASK '93）における「槙」など４文字のゲタ文字化に関する論考「新聞記事データベースにおける「槙」の消失現象」を HTML 化し，付録の CD-ROM に全文を転載した。（転載許可を勉誠出版から取得済み。）ファイル名は「MAKI. html」である。この HTML ファイルのうち，下記の「文字鏡フォント」を除く部分については，営利目的でない限り，インターネット等での自由な配布を許可する。ただし，その際は国立国語研究所『日本語科学』編集委員会宛に連絡されたい。（なお，この HTML ファイルの作成は谷本玲大が担当した。）

6.2. 文字鏡 GIF ファイルについて

ゲタ文字の箇所には JIS 外字も含まれている。その箇所を WWW 閲覧ソフトで表示するために「文字鏡フォント」の96 dot サイズ GIF 形式ファイルを CD-ROM に納めた。（この画像形式ファイルの CD-ROM 収録に当たっては、権利者の許可を取得済み。）文字鏡に関する詳細は、http://www.mojikyo.org/を参照されたい。

注

1 NTT データベースでは原則として「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。

2 ここでの辞書非掲載字とは，早くから漢字情報処理に利用されてきた『新字源』や国内最大規模の漢和辞典である『大漢和辞典』に掲載されていない漢字やその字体を指す。

3 「桧－檜」「事－事」「尭－堯」は83 JIS に新旧両字体が存在する例。「鴎」「掴」「葛」は83 JIS に新字体のみが掲出されている例で，朝日字体と一致するものが多い。83 JIS に旧字体のみが存在するのは「鱸」など（拡張新字体ならば「鮖」となる）。

4　伊藤(1994)によれば，米国のジョージタウン大学ではおもに著作権上の理由により1990年代前半から電子化テキストの作成はなるべく少なくし，市販の電子化テキストをできるだけ多く購入するようにしている。

5　最近の新聞記事はJIS漢字の字体そのものを取り上げたものも珍しくはないが，これを通常の他の記事と一緒に扱うことはできないであろう。紙面照合の際に研究者がチェックすべきポイントの一つである。

6　「包摂」はその規準の設定こそが重要であり，規準を明示しない包摂の適用は問題である。また，JIS規格の包摂以外にも別の規準の設定が可能である（研究者ごとに異なる規準が成立しうる）ことは言うまでもない。

引用文献（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）pp.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

伊藤 雅光（1994）「海外のテキスト・アーカイヴにおける管理・運営上の問題点についてーアンケート調査報告ー」『国立国語研究所研究報告集15』（国立国語研究所報告107），秀英出版

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ー日本語教育支援の観点からー」『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・横山 詔一（2000）「異体字に対するなじみと好みー接触印象・使用頻度との関係ー」『日本語科学』８号 pp.110-125，国立国語研究所〔編〕，国書刊行会
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 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漢字ー「JIS X 0208」の漢字を中心にー」『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 浩成・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 浩成（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 浩成・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 浩成（1999）「新聞記事データベースにおける「槙」の消失現象」『人文学と情報処理』No.20 pp.57-63，勉誠出版

付 記

本稿の執筆にあたり，森内豊四氏（日経広告研究所・前専務理事）と軽部能彦氏・松居秀記氏（毎日新聞社・編集局・編集総センター）には新聞制作や校閲に関する貴重な情報をいただいた。また，張元哉（チャン・ウォンゼ）氏（東京都立大学大学院博士課程）には韓国の新聞コーパスについてご教示いただいた。さらに，査読者の方々からは的確なコメントをいただいた。ここに記して感謝の意を表する次第である。

（投稿受理日：2000年７月31日）

横山 詔一（よこやま しょういち）
国立国語研究所 115-8620 東京都北区西が丘3-9-14 yokoyama@kokken.go.jp
笹原 宏之（ささはら ひろゆき）
国立国語研究所 sasa@kokken.go.jp
エリク・ロング（Eric Long）
国立国語研究所 KGD03011@nifty.ne.jp
谷本 玲大（たにもと さちひろ）
茨城大学 s_tanimoto@amy.hi-ho.ne.jp