人文科学とコンピュータ　26－2
（１９９５．　５．　２５）

漢字検索システムの基礎的問題

當山　日出夫

花園大学(非常勤講師)

631　奈良市二名6-1492　王龍寺
PK8H-TOYM@asahi-net.or.jp

漢字とコンピュータの問題については、次の問題点がある。
（１）漢字の種類　（２）漢字の数　（３）フォント　（４）コード
しかし、これらの問題を考えるに時、漢字の検索のことを忘れてはならない。その理由は、検索できない漢字は、存在しないのと同じだからである。漢字の検索は、漢字の属性についての検索である。漢字の意味と形は、歴史と文化と深い関係がある。

Searching－system of KANJI

Hideo　TOUYAMA

HANAZONO　Unversity

Nara-shi Nimyou 6-1492, 631, Japan
PK8H-TOYM@asahi-net.or.jp

The problems between the computer and the kanji are as follows:
(1) The kind of kanji (2) The number of kanjis (3) Font (4) Code
However, thinking of these problems, we can not forget the reference to kanji. The reason is that kanji which can not be referred to is the same as the one which does not exist. The reference to kanji means the reference to attributes of kanji. The meaning and the form of the kanji are deeply related to history and culture.

【1】はじめに

漢字とコンピュータ(JIS漢字)をめぐっては、今までに多くの議論が展開されてきている(注1)。その論点としては、次のようなものがある。

(1). JIS漢字に収録する文字の字種・字数
(2). JIS漢字に収録する文字の字体・字形
(3). コード系

これらが主な論点である。これに加えて、最近では、

(4). ISO10646
(5). フォント

などが、新たな論点として登場してきている。漢字とコンピュータをめぐる議論は、ますます複雑な様相を呈しつつある。

しかし、これらの従来の議論で、意外と見過ごされてきた論点が残っているように思う。それは、漢字検索の問題である。本発表においては、漢字とコンピュータという問題を視野におきつつ、漢字検索にまつわる基礎的諸問題のいくつかを考えてみることにする。

【2】なぜ漢字検索が問題になるのか

なぜ漢字検索を問題にするのか……それは、仮にJIS漢字にある漢字であっても、検索できなければ存在しないに等しい、という現実があるからである。

近頃、目についたものから具体例をあげる。ワープロで書けない漢字について指摘したものである。

> 文句はある。日本文学に欠かせない文字がない。里見とんのとんがない。梶井基次郎の『檸もん』が書けない、永井荷風の『ぼく東綺譚』も「第二水準漢字」にはない。(注2)

これは、ロシア・ポーランド文学を専門にする工藤幸雄氏の発言である。ここで指摘されている、「里見とん・檸もん・ぼく東綺譚」のうち、少なくとも「檸檬」は書けるのである。つまり、「檬」はJIS漢字にある。(区点で6108、16進で5D28)。しかし、工藤氏はワープロで書けない(JIS漢字に無い)と誤解したままで、ワープロには「日本文学に欠かせない文字がない」と言っている。

この例は、たまたま目についたものであるが、個人的経験としても、ワープロで書かれた原稿で、JIS漢字にあるはずの漢字が手書きになっているものなど、決して珍しいことではない。

【3】漢字検索の重要性

既に述べたように、せっかくJIS漢字にある漢字であっても、検索できなければ存在しないに同じことになってしまう。これは単に個人的にワープロで文章を書く時に、欲しい漢字が出せる(出せない)の問題ではない。学術的データの利用において、大きな問題点をはらんでいるのである。

(1). すべての議論の前提である

既に述べたように、漢字コンピュータをめぐっては、字種・字体・フォントなどにかんして、実に様々な議論が展開されている。しかし、これらの議論は、その議論の対象としている漢字のすべてを網羅的に把握できていることを、大前提としている。

例えば「この字が無いのは問題だ」と言ってはみたが、実はあったのだが探せなかっただけ、というのでは、学術的な議論として問題外である。

漢字とコンピュータを考えるとき、本来的には、コード表(JIS X 0208)によるべきであろう。

しかし、単に漢字をコード順に配列しただけのコード表だけでは、探す漢字がはたしてあるのか無いのか確認するだけでも困難であるし、その全貌を把握することは容易ではない。これまでのJIS漢字についての議論では、JISにある漢字(および無い漢字)を、総合的に検索して把握するシステムについて本格的に考察されてこなかったのではなかろうか。

(2). データ入力・検索の能率

学術的データについて、実用的レベルでいえば、難しい漢字を多用する、古典作品のデータ入力において、使いたい漢字を即座に入力できなければ、効率的ではない。これは、データ入力だけではなく、既に作成されたデータの検索においても同様である。(この場合、無い文字について無いことの確認が素早くとれることも重要である。これについては後述。)

(3). データ共有の阻害要因になる

データを他の研究者と共有するとき、作成者が無いと勘違いして仮名表記にして しまった場合、検索が困難になる。前述の例でいえば、「檸檬」が無いと思い違いしたまま、「檸もん／檸 モン」のように表記されてしまったテキストで、「檸檬／れもん／レモン」を検索しても、見つからない。常識的に考えて、検索用シソーラスで「檸檬／れもん／レモン」はあり得ても、「檸もん／檸モン」まで含んだものが、作られるとは思えない。

【4】コンピュータによる検索

漢字とコンピュータという視点から漢字検索を問題にする以上、まず考えるべきは、コンピュータの機能を使った漢字検索である。一般的には、ワープロの文字入力機能と言い換えることもできる。

(1). 仮名漢字変換
(2). 部首検索
(3). 総画検索
(4). コード入力

これらコンピュータによる漢字検索方法については、伝統的な漢字検索方法に依拠したものであることを指摘しておきたい。

このうち、(1)(2)(3)は、漢字辞典の音訓索引・部首索引・総画索引などを、コンピュータ化したものである。(4)のコード入力は、実際には後述のJIS漢字コード辞典類を併用することになるが、これも基本的には、(1)(2)(3)の音訓・部首・総画などによって、求める漢字のコードを探すことになる。

したがって、コンピュータによる漢字検索は、基本的には、伝統的な漢字辞典の手法を超えるものではないことになる。また、実用的な観点からは、一般の利用者の漢字についての知識(漢字辞典の使い方)に依存するものである。

【5】JIS漢字コード辞典による検索

JIS漢字コード辞典類が数多く市販されている(後掲のリスト参照)。そして、実際のデータ入力においては、これらのJIS漢字コード辞典類によって、仮名漢字変換で出せないような漢字を入力することが多い。

これらは、JIS漢字を、

(1). よみ(音訓)
(2). 部首
(3). 画数

などに従って配列し、求める漢字のコード(区点・16進・シフトJISなど)を表示したものである。中には、JIS漢字内部での異体字などを参照できるようにしたものもある。これも基本的には、既存の漢字辞典の漢字検索方法を踏襲したものであるといえる。

【6】斬新な漢字検索法

漢字検索は、その多くが既存の漢字辞典にならったものであるが、中にわずかではあるが、斬新な方法を採用したものもある。

コンピュータによる検索では、市販の日本語入力フロントプロセッサ「VJE-Delta」の複合部首入力などが、コンピュータならでは機能を使った漢字検索法として注目される。また、JIS漢字コード字典類では、リストの(11)(12)(17)などが、特徴のある検索法を採用している。

【7】無い字を探す必要

漢字検索の実用的ポイントは次の2点である。

(1). ある字を探す。

JISにある漢字を漏れなく網羅的に検索が可能であるかどうか。そのためには、どのような方法が適当かが課題である。各種の索引の整備や、コンピュータならでは機能を活かした漢字検索方法を開発する必要がある。

(2). 無い字を探す。

JISに無い漢字を確認できることが必要である。どれほどの字数を収録するにせよ、JIS漢字は有限である。しかし、実際の研究活動で必要とする漢字は、無限であると言っても過言では ない。少なくとも、古典作品の研究では、非JIS漢字を数多く必要とする。このとき、ある漢字がJISにある漢字であるのか、それとも無い漢字であるのか、的確に判断することが求められる。そのためには、非JIS漢字までをも含んだ漢字検索システムが必要になる。

コンピュータを使えば、確かに手作業では困難な作業を容易にこなすことができる。では、コンピュータを使って完璧な漢字検索システムが構築可能であろうか……答えは否である。当たり前のことだが、コンピュータを使った漢字検索の致命的欠陥として、コンピュータで使えない漢字は検索不可能、ということがある。コンピュータで検索できる漢字は、あくまでもコンピュータで使える漢字の範囲内(つまりJIS漢字)でしかないのである。

これは、JIS漢字コード辞典についても同様である。JIS漢字の範囲しか収録していない辞典では、非JIS漢字の確認が難しい。

一般的に言っても、ある範囲内にあるものが「存在する」ことの確認よりも、「存在しない」ことの確認の方が、より困難である。したがって、実用的な観点から、JIS漢字の全貌を把握しようと思えば、非JIS漢字まで含めた検索システムが必要になる。これは、コンピュータでは原理的に不可能であり、書物としてのJIS漢字コード辞典類に頼らざるをえなくなる。

この意味で、現在、補助集合の漢字まで収録してあるJIS漢字コード辞典(リストの(14)(18)(20)の3冊)が、比較的これに近い機能を持っていることになる。今のところ、補助漢字は一般に使える状態にはないので、補助漢字であることが確認できれば、非JIS漢字であると判断が下せる。

【8】漢字検索と属性

これまで、漢字検索と一括して述べてきたが、これを別の角度から言い換えるならば、漢字の属性の検索と言うことができる。つまり、漢字検索とは、漢字について、その読み方は何か(音・訓)、ど

の部首に所属するか、総画数はいくつか、という漢字の属性ごとに、ある一定の配列(五十音順・画数順)を与えて、検索の手段とする……このように定義可能である。

では、漢字の属性(形音義)は、すべての漢字について、一義的に決定可能でかつ網羅的に適用可能であろうか。この問題については以前に発表したことがある。結論を先に述べれば、私見では不可能である、ということになる。(注3)

具体例をあげる。「芸」は現在の常用漢字では「藝」の新字体である。音は「ゲイ」。しかし、この字は本来「藝」とは別の字で、「ウン」とよむ草の名称である。この字は、どうやって検索すればいいのだろうか。もちろん、JIS漢字には、「芸」は一つの文字に対して一つのコードしか与えられていない。

JIS漢字制定者の意図としては、「芸」は「ゲイ」と読むことになっている。

芸　頴鶏芸迎鯨(第1水準)

藝　藐藕藝藥藜(第2水準)

であり、「芸」には明らかに「ゲイ」の読みが与えてある。では、JIS漢字には、「芸」(ウン)の文字は存在しないと考えてよいのだろうか。利用者が、「芸」を「ウン」の文字として使うことは不当なのだろうか。

この文字のあつかいは、かなり複雑である。

(1). 現在、通行の漢字辞典では、「藝」の新字体(ゲイ)と、草の名称(ウン)とは、別見出しで区別するのが普通である。

(2). JIS漢字コード辞典類を見ると、「ゲイ」では出ていても「ウン」では出てこない。

(3). パソコンの日本語入力システムの場合、ATOK9では、「ウン」「ゲイ」両方の読みで仮名漢字変換が可能。VJE-Deltaでは、「ゲイ」しかない。

「芸」にどのような属性を与えるか(どんな意味の漢字として使うか)は、最終的には利用者の判断にゆだねられていると考えるべきだろうか。そうであるならば、すでに社会的存在となっているJIS漢字について、個々の漢字にどのような属性を与え、どような検索システムを用意すれば、円滑な利用が可能になるのか、狭義の情報処理技術の枠を超えて、文化的社会的に各方面から多面的に考察されるべきである。

【9】おわりに

本稿は、さる3月10日に統計数理研究所で行われた研究会(統計数理研究所共同研究/(第7回)文献情報のデータベースとその利用に関する研究・総合大学院大学共同研究/日本語テキストデータベースの利用法に関する研究・平成6年度合同研究会)において、「漢字コードと漢字検索システム」と題して発表した時の資料を、要約・整理したものである。

漢字検索という問題は、JIS漢字の有効利用のために実用的観点から重要なテーマであるばかりでなく、漢字属性の再検討にも新たな視点を提供してくれるものである。今や、活版印刷が滅亡し、文字は、手で書くか、コンピュータで使うか、この二者択一的状況に追い込まれている。

人文系学術研究は、文字で記された文献を多用する。文字そのものが研究対象であったりもする。本研究会のタイトルである「人文科学とコンピュータ」について考えるとき、その基底をささえるものは、ミクロ的に見れば、文献を表記する一つ一つの文字である。この意味で文字についての考察は、研究者自身によってなされなければならない、重要な研究改題であると考える次第である。

(注1)最近のものとしては、『月刊しにか』1995.5《特集:"正しい漢字"とは何か》などがある。

(注2)『私のワープロ考』(リテレール・ブックス7)安原顯編　メタローグ　1994年8月刊

(注3)當山日出夫『漢字コードをめぐる諸概念について』情報処理学会・人文科学とコンピュータ研究会(94-CH-23)1994/09/16

JIS漢字コード辞典リスト(架蔵)

(1)パソコンワープロ漢字辞書
　柳澤章喜編　オーム社　1983年9月刊
(2)ＯＡ漢字辞典
　吉野敏也著　蕗出版　1985年12月刊
(3)ワープロ・パソコン用漢字コード辞典
　九段コンピュータサービス編　九段コンピュータサービス　1985年9月刊
(4)ワープロ・パソコンの漢字辞典
　黒須重彦監修　日本実業出版社編　日本実業出版社　1986年1月刊
(5)ＯＡ時代の新漢字辞典
　野本菊雄監修　ぎょうせい　1986年4月刊
(6)パソコンワープロJIS第2水準漢字辞書
　柳澤章喜編　オーム社　1986年5月刊
(7)三省堂ワープロ漢字辞典
　三省堂編修所編　三省堂　1986年8月刊
(8)ワープロ・パソコンのための漢字コード辞典
　富士通編　工学図書　1986年9月刊
(9)ワープロ・パソコン漢字辞典
　西東社編集部編　西東社　1987年3月刊
(10)パソコンワープロ漢字辞典
　上柿力編　ナツメ社　1987年9月刊
(11)パソコン・ワープロ漢字スピード入力辞典
　石川譲編　東洋堂企画出版社　1988年6月刊
(12)直感！ワープロ漢字辞典
　増田忠編　日本経済新聞社　1990年3月刊　1993年1月刊（増補改訂版）
(13)最新ＯＡに強くなる／ワープロ漢字辞典
　学研語学ソフトウェア開発部編　学研　1990年8月刊
(14)最新JIS漢字辞典
　田島一夫監修・(財)日本規格協会編集協力　講談社　1990年11月刊
(15)三省堂実用ワープロ漢字の辞典
　三省堂編修所編　三省堂　1991年7月刊
(16)ユーザー必携！／早引きパソコンワープロ漢字辞典
　上柿力監修　ナツメ社　1993年4月刊
(17)JIS第1・2水準漢字索引辞書／漢べき君
　高田任康著　ソフトバンク　1993年8月刊
(18)早引きワープロ漢字辞典
　小和田顯監修・旺文社編　旺文社　1993年9月刊
(19)「見る」「引く」簡単／パソコンワープロ漢字コード辞典
　実教出版出版部編　実教出版　1993年12月刊
(20)ワープロ・パソコン／最新漢字辞典
　小学館辞典編集部編　小学館　1994年4月刊
(21)コンサイスワープロ漢字辞典
　三省堂編修所編　三省堂　1994年9月刊