〈国語学会春季大会テーマ発表〉国語研究資料の「電子化」とその利用

電子化テキスト総論

——国語学の立場から——

當　山　日出夫

［0］ 要旨

国語学研究資料の電子化（国語学でのコンピュータ利用）について、学会および研究者が全体的に取り組むべき方向は、

1．研究者個人レベルでのパーソナルコンピュータ利用を前提に
2．電子化研究資料（データ）の公開・流通・共有をめざして
3．他の研究分野の動向にも配慮する学際的視野にたった

ものでなければならないことを述べる。

［1］ 情報化の流れのなかで

国語学におけるコンピュータ利用について述べる前に、国文学や歴史学などの周辺領域をふくめて、全体的にどのような状況にあるか、管見のおよぶかぎりで概観してみたい。人文科学の各研究分野においてコンピュータ利用は、年々さかんになってきており、一部では研究のための必需品とさえいえるが、全体としては、まだまだ普及の途上にあり、その利用方法をめぐって個別に試行錯誤を繰り返している状況である。

このような流れのなかで、最近、雑誌等でも、コンピュータ関係の特集が目立つようになってきた。最近のもので国語学関連としては、『日本語学』（1991年8月号、明治書院）の特集「新しいデータ・新しい研究」とか、『しにか』（1992年2月号、大修館書店）の特集「古典とコンピュータ」などがある。また、現在『日本語学』に連載の「私のパソコン言語学」などは、周知であろう。

また、いくつかの研究会・学会等が組織され、研究会の開催や機関誌の発行などを行っている。筆者が会員であったりするその主なものとしては、

1．情報処理語学文学研究会（JALLC）
2．テキスト・データベース研究会（JACH）
3．アート・ドキュメンテーション研究会
4．情報知識学会
5．情報処理学会内「人文科学とコンピュータ研究会」
6．「東洋学へのコンピュータ利用」（京都大学大型計算機センター）
7．「文献情報のデータベースとその利用に関する研究会」（統計数理研究所）
8．「国文学とコンピュータシンポジウム」（国文学研究資料館）

などがある。（これらの学会や研究会の簡単な活動紹介や連絡先等については、『しにか』1992年2月号所収の拙稿「関連研究会の紹介」を参照。）

このような全体的な動き（いわゆる「情報化」の流れ）に沿ったものとして、先般の国語学会テーマ発表「国語研究資料の〈電子化〉とその利用」も企画され、また、本稿の『国語学』掲載へと繋がっているのであると理解している。国語学においても、「情報化」は避けて通ることのできないものなってきている。

［2］ パーソナルコンピュータの普及

パソコン(パーソナルコンピュータ)が一般社会に普及しだしたのは、ここ数年の現象である。その普及に連動するかたちで、各方面でいわゆる「情報化」をめぐる諸問題が起ってきたのであり、また、それに対処するために前述した各種の学会や研究会の活動が行われてきたのである。それまでは、コンピュータといえば、大型計算機であった。だが、今では、コンピュータといえば、パソコンを意味するほどに、大きく社会全体の状況が変ってきている。国語学におけるコンピュータ利用を考えるにあたっても、パソコン利用を前提にしなければ、多くの国語学研究者の要望にこたえることはできない。

ここで、パソコン利用の特徴をまとめてみると、次のようになる。

1. 誰でも使える。個人で自由に自分の書斎や研究室で自分だけの機械として利用できる。一般に人文科学研究は、個人の研究者が自分の部屋の自分の机で考えたり書いたりという性格が強い。値段が安くて大きさも小さいパソコンは、このような利用にふさわしい。
2. 性能的にも十分実用にたえる。現在の段階でも、記憶容量とか処理速度など、文献目録の検索、語彙索引の作成などであれば、実用的に問題はない。

だが、だからといって、今後大型計算機が不用になるかといえば、そんなことはありえない。利用目的によっては、大型計算機の性能に頼らざるをえない仕事が多くある。これからも、大型計算機は国語学研究で利用されていくであろうが、その利用の前提として、研究者個人レベルでのパソコンの普及という現実をふまえたものであることが要求されよう。

［3］ 国語学と情報化

国語学における「情報化」とか国語研究資料の「電子化」などというが、そもそも何をさしているのだろうか。厳密に考えればキリのない、ある意味では些末な議論になりかねない。本稿の段階では、とりあえず、国語学研究におけるコンピュータ利用の可能性とその問題点、ぐらいの認識としておきたい。

ここで確認しておきたいのは、国語学研究（あるいは一般に人文科学研究）におけるコンピュータ利用が何をもたらすか、2つの視点についてである。それは、

1. 既存の研究方法を踏襲するなかで、コンピュータがどのように利用可能か
2. コンピュータの利用によって、新たな研究方法・領域が拓かれていく可能性

の2つである。

1の立場は、単純に道具としてのコンピュータ利用である。

例えば、次のような例である。ある文献からある語の用例を検索するのに、人間が目で読んで探したり語彙索引を利用したりするのに較べて、電子化されたテキストを利用すれば、より速くより正確に目的を達することができる。語彙索引の作成にも、手作業でカードを並べ変えてそれを原稿に写すのにくらべて、コンピュータを利用すれば、はるかに簡単であり確実に作業をおこなえる。

これに対して、2の立場は、コンピュータが人間の言語とどうかかわるかという点に着目したものである。

コンピュータは、日本語をどのように変えていくだろうか。例えばワープロがある。ワープロといえども、文章の作成・編集・印刷専用に開発された一種のコンピュータである。少なくとも近現代の日本語の歴史において、ワープロほど文字や表記に多大な影響を与えたものは他にない。ワープロと日本語の関係は、新たな国語学の研究課題として浮び上がってこよう。文字に限っても、現在では、工業的な印刷をふくめて、コンピュータ(ワープロ)で使える文字こそが文字であって、コンピュータ(ワープロ)で利用不可能な字は、使えなくなりつつあるし、また、実際に使わなくなりつつある。

コンピュータ(ワープロ)の文字つまりJIS漢字について、「欲しい字がない」「字体が気にいらない」と不満をよく耳にする。だが、不満を言うだけでは解決にはならないし、また、あきらめてしまってもいけないだろう。そうではなく、日本語の表記史の観点から、現在のJISコードや他の漢字コード案について、厳密な検証作業が必要なのではなかろうか。そして、それをふまえてのコンピュータ利用を開拓していくことこそ、国語学研究者に課せられた課題であるのかもしれない。

コンピュータは大量の文書をきわめて簡単に保存し伝達することができる。「源氏物語」のような大部の文学作品であっても、ごく普通のフロッピーディスクの1枚か2枚で収ってしまう。最近では、パソコン通信などの新しいコミュニケーション手段も広まりつつある。このような新しい記録と通信の手段を手に入れることによって、人間はどのような言語行動をとるのか、既存の研究領域では対処できない新しい言語研究のテーマであるといえるかもしれない。

[4] 得るものと失うもの

社会全体の流れとして「情報化」があり、国語学におけるコンピュータ利用は、必然である。

だが、安易にこの必然に身をまかせて流れに流されるだけではいけない。たちどまって、これが国語学研究に何をもたらし、同時に、何を失わせるのか、反省する視点も確保しておくべきではなかろうか。

ところで、筆者自身の専門はあくまでも国語史・訓点語である(決してコンピュータではない、念のため)。その立場から省みてであるが、現在では写本ということをしなくなってい

る。コピー（複写機）が普及する以前、資料の収集は、写真によるか筆写によるかしかなかった。写真さえ普及しない以前は、当然ながら、すべて筆写によっていた。近代の学問史において、研究者が文献の筆写をしなくなりコピーを使うようになったことによって、資料収集が便利で正確になったことは否定できない事実であろう。だが、ここであえていえば(個人的感想にすぎないかもしれないが)、筆写からコピーへの変化のなかで、はたして便利さ正確さを得ただけであろうか、ひょっとして、失ってしまったものがあるのではなかろうか、そんな気がしてならないのである。

現在、コピーなしの文献研究とかテープレコーダーなしの音声研究など、想像もできないことかもしれない。これと同じように、国語学研究においてコンピュータも必需品になるにちがいない。いや、必需品であるという意識さえも存在しなくなるかもしれない。そうなった時、我々は、コンピュータによって多大の恩恵をこうむると同時に、喪失してしまったものの価値をいとおしむであろうか。

しかし、これこそが時代の流れなのであろうと、思うのである。筆者とてコンピュータを使わない手作業の価値を認めるにやぶさかではないが、それにこだわるあまり、コンピュータ利用のもたらす多大な利点を評価しそこなってはならないだろう。あるいは、手作業の価値を知るものこそが、その利点を欠点と共に正しく評価できるとさえもいえようか。

［５］ データは公開され流通しなければならない

データの公開・流通についての諸問題を考える前に、なぜデータの公開が課題となるのかについて簡単にふれておきたい。筆者は、研究のためコンピュータでつかうデータは、原則的に公開され流通させなければいけないと考えている。それは次の理由による。

1. パソコンの個人レベルでの普及という現実
2. 一般に、近代の学問は資料の公開と共有を前提に成立している

このうち２について説明すると、例えば、古代語の研究に、「万葉集」「源氏物語」などは必須の文献資料である。現在、研究者が「万葉集」「源氏物語」を研究しようと思えば、通常は市販の校訂本や影印本を利用する。これら一般に使用される通行の校訂本などは、研究者・学生・一般市民の区別なく誰でもごく普通に書店で購入することができるし、図書館などでの閲覧も自由である。すべての日本語研究者にとって、古代語資料である「万葉集」「源氏物語」は、公開され共有されており、その基盤の上に現在までの種々の研究が積み重ねられてきているのである。資料の公開と共有こそが近代的な学問の前提条件である。

かつてパソコンの普及以前の大型計算機主流の時代には、大型機を使えるのはごく一部の限られた人だけであった。この段階では、データの公開ということは、特に考える必要はなかった。一部の人だけが使えるデータであったとしても、特に問題になることはなくそれでことは済んだのである。

だが、今日では、コンピュータ利用は一部の人だけのものではなくなっている。研究者個人レベルでのコンピュータの普及という現実をふまえて、研究資料の公開と共有はどう

あるべきだろうか。

先に例としてあげた（書物としての）「万葉集」「源氏物語」の公開と共有、これをコンピュータの普及のなかでとらえなおせば、「万葉集」「源氏物語」の本文データの公開と共有に他ならない。

俗に、コンピュータについて「ソフトウエアが無ければただの箱」という言い方をするが、国語学での学術的な利用という観点からすれば、「データが無ければただの箱」である。国語学におけるコンピュータ利用の推進は、個人レベルでのコンピュータの普及という現実をふまえるならば、何よりもまず、データの公開・流通・共有を基盤とすると同時にそれを目的としたものでなければならない。

［6］ データの公開・流通の問題点

データの公開・流通・共有が課題であるとしても、現実には、いくつかの問題点がある。その主なものをあげると次のごとくである。

（1）使えるデータが少ない

公的な研究機関等において、データの入力は進行しつつある。国語史研究の分野にかかわるものとしては、国文学研究資料館で行われている「日本古典文学大系」（岩波書店・旧版全100巻）の全文データベース化の企画がある。が、全体としては、これら一部の大規模な事業を除くと、個々の研究者個人レベルで自分の専門分野の資料についてデータ入力が行われつつある状態である。

国語学研究全体から見た場合、まだまだデータ入力の量的な不足が問題である。もちろん、研究資料のデータである以上、量があればよいというものではなく、最終的には質が重要である。だが、質の向上のためには、まず、ある程度の量の確保と、それを様々に利用し学問的に検証する過程が不可欠である。質的により良いデータを作るためにとりあえず試用してみる基礎的なデータが必要なのだが、それさえ不足しているのが現状である。

（2）データ入力情報の不足

使えるデータが少ないのみならず、その少ないデータについて、何処で、誰が、何の文献について、どのような方式で、データを入力しているのか、また、そのデータは公開されているかいないか、もし公開であればその利用条件はどんなものか、等々についての情報流通のシステムが整っていない。現在では、知っていそうな人を個人的に探して聞き出すぐらいがせいぜいである。これを、学会全体（あるいは周辺の関連領域までふくめて学際的）規模で、データ入力情報の流通を図らねばならない。

例えば、筆者の知る限りでも、「万葉集」「源氏物語」などについては、別個に独立していくつかのデータ入力がなされている。底本の違いとか入力形式の違いなどの差異はあるにしても、全体的視野から見て、相互の連絡の不行き届きの弊害を感じざるをえない。

データ入力情報の流通するシステムが無い結果として、次のような問題が生じる。

1. 利用者の立場からは、公開されているものであればそのデータを利用したいと思っても利用できない。その存在を知ることさえ容易ではない。

2. データ入力者の立場からは、自分の入力したデータを公開し広く学会に提供したいと思っても、知らせるすべがない。せっかく作ったデータをより多くの研究者に使ってもらえない。

さらに、あるテキストの本文データは、現段階での主な利用法に限ったとしても、その文献の総語彙索引作成に匹敵する価値を持つ。書物として総語彙索引を印刷・製本して刊行することは学問的業績になるが、はたして、コンピュータによる本文データの作成は、業績として認知されるであろうか。もし業績となるべきならば、データの公開・流通を前提としなければならない。

（３）共有と個別利用、標準化と再加工

データの公開・流通のためには、ある程度の標準化・規格化が必要である。が、それには、次のような問題点がある。

1. データの互換性。一般に、ワープロ専用機の文書データは、そのままの形式ではコンピュータであつかうことができない。また、コンピュータに入力されたデータであっても、利用するソフトウエアによっては、直接の互換性がないこともある。
2. 文字コード。国語学研究に利用するとなると、通常、文字はJISコードに依存することになるが、これだけでは不十分である。したがって、JIS外の漢字とか音声記号など、外字として、個々のデータ入力者ごとに作成されることになる。JISコード外に利用者が個別に作成した文字（外字）の互換性をとることは、困難な場合が多い。
3. 付加情報。ある文献資料をそのまま入力しただけでは、非常に使いづらい。これは日本語の表記の特性（漢字表記と仮名表記・異体字・ふり仮名など、また英語のように単語ごとに空白を置かない）に起因する。効率的に語彙検索などをおこなうためには、単語や形態素単位に印をつけるとか、品詞などの文法情報を付加する必要がある。この付加情報も、個々の研究者の興味や関心、資料の性格によってかなり変化し、統一は難しい。文法的な情報の他にも、テキストの書誌情報なども必要である。

今後2・3がこれからの大きな課題であろう。特に3についてみれば、電子化資料の公開と流通という発想、その原点に帰って考えてみた場合、研究者がみんなで共有するためのデータに必要とされるものと、個々の研究者が個別の研究で必要とするものとは、自ずと違うのであると、認識しておくべきではなかろうか。データの公開・流通にとって最低限の標準化・規格化は必要であるが、個々の研究者が、それをそのまま使わねばならない義務はない。必要に応じてさらに文法情報を付加するなりして（あるいは、削除するなりして）、自分の研究目的に応じて再加工し利用していけばよいのである。

では、個別の研究目的に応じた再加工を許容するものして、全体としてどのような標準化が適当であるのか。それは、実際の利用経験の蓄積としてのみ可能なのである。

（４）著作権の問題

各種資料の電子化、特にその公開・流通をめぐる問題として、著作権（知的所有権）がある。これについては、現在までのところ、どのような行為についていったい何が問題となるのかということさえ、未解明であるのが、現状である。これについては、国語学研究者

の判断だけで決められることではない。同じように電子化資料の流通・公開を課題とする他の学問分野と歩調を合せたものでなければならないし、また、何よりも出版・印刷関係の意向は尊重されなければならないであろう。

［7］ 使ってみたいがどうすればよいか

最初に記したコンピュータ利用をめぐる学会・研究会等にかかわっていて感じることとしてであるが、すでにコンピュータをある程度つかいこなしている人からコンピュータ利用のメリット(場合によってはデメリット)についての発表がなされる一方で、これから自分もコンピュータを使ってみたいがどうしたらよいか、という相談が寄せられることがある。具体的には、コンピュータを買いたいがどの機種が適当か、プリンタなどの機器は何がいいか、どんなソフトウエアが使いやすいか、などである。ざっくばらんにいえば、「初心者」への適切なアドバイスを要求されるのである。

街の書店に行ってコンピュータ関係の本の売場を見れば、初心者向けのコンピュータ入門書や雑誌があふれている。とりあえず、機械を買いそろえてコード類をつないでスイッチをいれて、ワープロなどのソフトを使って文章を書いてみる、というぐらいまでは、市販の入門書や雑誌類で対応できる。が、これから一歩進んで、国語学研究のためには、どの資料をどのようにあつかえばいいのかとなると、全くといってよいほど役にたたない。例えば、コンピュータで利用可能な漢字の問題とか、方言研究で必要な音声記号をどうやって使えばいいのか、どのような形式でデータを入力すればいいのか、語彙の検索にはどんなソフトを使うのがいいのか、などである。

一番いい解決法は、身近にコンピュータに詳しい人がいれば、その人に聞くことである。もし、その人が同じ国語学研究者でしかも比較的専門分野が近ければ、これにまさる幸運はない。だが、すべての人がそのような幸運にめぐまれているわけではない。このような幸運にみはなされた(おそらく)大部分の人は、いくぶん極端にいえば、絶望してあきらめるか、さもなくば無駄を覚悟で試行錯誤を積み重ねていくしか、道は残されていない。

今は、「国語学研究におけるコンピュータ入門」とでもいうような書物・雑誌あるいは講習会などが、本当に必要とされている状況である。コンピュータ利用について、機械翻訳や自然言語処理などの先端的な問題点を考えることももちろん重要であるが、その一方で、いかにすみやかにコンピュータを国語学研究者全体に普及させるかということが現実的課題としてある。

［8］ 解決のために、私的提言として

解決のためには何をなすべきであろうか。

基本的姿勢としては、本稿の冒頭にかかげた3箇条になる。これまで述べてきた筋道にしたがって要約して繰り返せば……研究者個人レベルでのパソコンの普及という現実をふまえての電子化資料の公開・流通・共有を達成するためには、より広い学際的視野からの取り組みが必要である、ということである。

では、何を実際に行っていけばよいであろうか。

個々の研究者については、コンピュータ利用により積極的になってもらう。まだ使っていない人はどんどん使うようになってもらいたい。もはや、コンピュータによってしか利用できない国語学研究資料が登場しつつある。例えば、先般の学会テーマ発表で紹介された音声資料についてのデータなどがそうである。コンピュータ利用に抵抗したり逡巡している余裕などない時代である。すでに使っている人は利用経験やデータ入力情報について、公開を原則に、よりオープンな姿勢で臨んでもらいたい。

そして学会全体としては、これを支援するシステムを考えるべきである。

具体的には（実現可能性の高いものとしては）、例えば、コンピュータ利用についての講習会・研究会などを開催し、この時のテキストや発表予稿などをもとに情報誌を発行する、などであろうか。これは、春秋の学会大会や『国語学』などとは別に、考えられてもよいかもしれない。このような講習会・研究会の開催や情報誌の発行の主体として何処が適当か、今後の現実的課題であろう。必ずしも国語学会を始めとする日本語研究にかかわる諸学会や専門研究機関などにこだわる必要はなく、総合的視野から、実現可能なものを模索しつつ柔軟に対処していくことが現実的ではなかろうか。そして、これらの活動も、前述したコンピュータ利用についての各種の学会・研究会、さらには国語学の周辺学問分野での動きを視野にいれたものでなければならないことを最後に強調しておきたい。

付　記

本稿は、国語学会テーマ発表で、同じ「電子化テキスト総論――国語学の立場から――」と題して話したことの内容・質疑応答・予稿等を総合的にまとめて再編して新たに書いたものである。紙数の都合もあり、予稿の内容からは、「【1】個人的経験から……なぜパソコンを使い始めたのか」および「【4】電子化テキストの具体例……「源氏物語」」の大部分を割愛した。「【8】付記」として、「TEI」と「ISO 10646」について若干の言及があるが、これも本稿では省いた。また、「注」としたことがらについても原則的に省略してある。発表時の予稿に無く本稿で追加した主なものは（主に当日寄せられた質問票・質疑応答をふまえて書いたものである）は、［４］、［５］、［６］の（3）、［８］で述べたことがらである。

（1992年７月１日）

――円満寺――