

漢字教育と日本語処理¹⁾

矢澤真人 (YAZAWA Makoto)

筑波大学文芸・言語学系

Institute of Literature and Linguistics, TSUKUBA University

【要旨】 日本語ワードプロセッサにより、個人で手軽に漢字仮名交じり文の文書を活字体で作成することができるようになった。特に、その日本語処理の中心となる日本語入力システムは、言語情報を取り込むことによって、よりの確な表記を提供するように開発されてきたが、現在は、さらに、読みから表記を導くにとどまらず、表現を導くものへと発展しつつある。一方、日本語入力システムの普及にともない、その公共性が意識され、登録語彙や候補となる表記の配列順序などに関わるガイドライン策定も試みられた。1990年代は、ほぼ、理想的使用者の表記を導く方向で設定すればよかったが、2000年代に入り、携帯電話のメール機能の普及がきっかけになり、より個人的な表記を重視する方向に転換しつつある。今日の日本における言語政策や言語教育は、このような日本語入力システムの状況をふまえることが不可欠となっている。

1 日本語の表記システムと漢字

1-1 漢字と漢字仮名交り文

日本語では、漢字と仮名を組み合わせた「漢字仮名交り文」で文章を書き表す。漢字仮名交り文は、より実質的な意味を表す部分が漢字や片仮名で記され、補助的な部分や文法的な部分が平仮名で記されるため、自ずと分かち書きの効果が得られる。視覚的にも漢字や片仮名の部分が際だつ²⁾ことから、情報の伝達が容易であるとも言われる。しかし、この表記システムは、普通の言語生活で、約2000～3000字用いられるといわれる漢字によって支えられている。

漢字の読み方には、古代中国語の発音を元にした「音（おん）」と、漢字の意味から日本語をあてた「訓（くん）」とがある。さらに「音」には、借用した際の中国の時代的・地域的位相を反映して、唐代の中国語を元にした「漢音」と、それより古い比較的南方の中国語を元にした「呉音」とがあり、一部の漢字は、宋代から清代初期の中国語を元にした「唐音（唐宋音）」³⁾と呼ばれる「音」を持つ。例えば、現代の日本語でも、「頭」は、「ズ」（呉音）、「トウ」（漢音）、「チュウ」（唐音）という3種類の「音」を持ち、「あたま」と「かしら」という2種類の「訓」を持つ。そして、単語によってこれらの読み方のいずれを用いるかが定まっており、「頭蓋骨」（頭の骨）は「ズガイコツ」、「頭髮」（頭の毛）は「トウハツ」、「塔頭」（大寺院の中にある小さな寺院）は「タッチュウ」と読まれる。しかも、先行する漢字の音が[n]で終わる場合、後続の漢字音の無声音が有声音化する「連濁」と呼ばれる現象が生じる場合があり、同じく全体を司る立場を表す「頭」でありながら、「番頭」（店の支配人）は[n]の後でありながら「バントウ」と有聲化されないのに、「船頭」は「和船の船長」は「センドウ」と有聲化される⁴⁾。訓と音も同じ意味を表すとは限らず、「牧場」（まきば=ボクジョウ）のように、訓読みをした単語と音読みをした単語とがほぼ同じ意味を表すものもあれば、「風車」（かざぐるま≠フウシャ）のように

全く異なった意味を表す場合もある。

使用する文字の数だけを見れば、漢字のみを使う中国の方が圧倒的に多く、漢字とハングルを併用する韓国語⁵は日本語と大差ないといえるかもしれない。漢字は表語文字とされるが、音を表す部分（音符）と意味を表す部分（義符）とを組み合わせた形声の文字が多く、表音的な側面も併せ持っている。中国語や韓国語においては、一つの漢字が同じ意味を表す場合に異なった発音がなされることは原則としてない⁶。中国語や韓国語の漢字は、形態（発音）と意味とが対応する形態素であり、成り立ちや意味、発音を個別に教える漢字の単体指導も、語彙教育と比較的に結びつけることができる。これに対し、意味と発音とが対応しない日本語の漢字は、文字を介在させなければ、語彙教育と結びつけることができない。漢字に何種類もの発音があてられる日本語では、漢字の単体指導をそのまま語彙指導と結びつけることは難しく、文字教育と語彙教育との連携を阻害する要因となっているのである。

1-2 日本における出版形態と文字の規範

日本では、金属活字や木製活字を用いた活字印刷も部分的には行われていたが、一般的には、桜や杉の版木（はんぎ）に彫刻刀で文字や絵を彫り、紙をあてて馬棟（ばれん）でこすって刷る木版印刷が広く行われていた。木版印刷では、版面の摩耗のために大部数印刷は難しいが、印刷に必要な設備が小規模で安価であること、印刻の熟練は要求されるものの、複雑・巧妙な印刻から情報に応じた即時性を求められる印刻まで対応できること、印刷物を大量に運搬することが困難であっても、版木そのものを運搬することで遠方でも同じ印刷物を提供することができること、などの利点があった。木版印刷では、手書きの文字や図版を木版に彫りつけるところから、表記の面からすると、手書きと同様に考えてよい。

手書きの文字には、線の方向や線の末端の処理（はらい、止め、跳ねなど）、線の相対的な長短や配置といった、書き方のスタイルには一定の規範はあるものの、字形の許容範囲はかなり広い。美しい文字の「手本」はあっても、それは標準的な字形ではなかった。

1874年に東京築地に日本で最初の活版工場が開設され、新聞や書籍などの高速な大部数印刷が開始された。手書きや木版印刷から、金属活字による活版印刷へと移行したことで、表記の規範にもさまざまな変化が生じる。

その一つが、文字の切れ続きに関わる問題である。手書きには、文字の切れ続きがあり、そこから語句の切れ目がどこであるかが推測できる。右の図版は、紫日記所載の和歌であるが、これをそのまま活字に置き換えると、

きのくにのしららのはまにひろふてふ

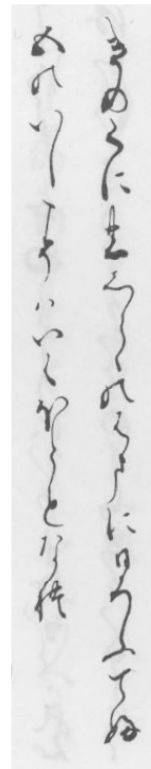
このいしこそはいはほともなれ

のように読みにくい文章になる。手書きの切れ続きは、

きのくにのしららのはまにひろふてふ

このいしこそはいはほともなれ

となっており、これを漢字仮名交り文に直すと、次のようになる。切れ続きがほぼ単語単位になっていることが知られるよう。



紀の国の白良の浜に拾ふてふ

この石こそは岩ほとも成れ

ここからさらに空白をとって、等間隔に並べたのが下の文である。漢字仮名交じり文でも語句の切れ続きが読み取れるのである。

紀の国の白良の浜に拾ふてふ

この石こそは岩ほとも成れ

漢字仮名交じり文は、そのまま活字体に置き換えることができるのに対し、仮名書きやローマ字書きでは、分かち書きが必要になる。活字印刷は、漢字仮名交じり文に潜在的なアドバンテージを与えるのである。

活字組の文書が規範的な文書形態となったことにより、活字の字体を規範的な字形と見なす素地が生み出された。手紙や日記、講義ノートなど、個人的な書記活動はともかく、公開を前提にした文書類では、活字体を用いるのがよいとみなされるようになったのである。日本では、和文タイプライターは、個人へは普及しなかったため、孔版印刷においても、タイプ製版ではなく手書き製版が普通であった。謄写版（ガリ版）と呼ばれる手書き製版においては、「沿溝ゴシック体」など、活字をまねた文字が多く用いられた。手書きでありながら書き手の個性は抑えられ、活字風の画一的な字体が選択されたのも、字体の規範意識が働いているのである。

文字種が多い日本語の文書は、活字を組むのに大変な手間がかかる。「活字になる」文書とは、その手間をとるだけの価値を持つものであると考えられるようになる。ここから、欧米の思想や学説を無批判に信用する「横文字信仰」や、書籍を無批判に信用する「書籍信仰」と並んで、営利活動である商業的な広告を除いて、活字で組まれたものの内容を無批判に信用する「活字信仰」といわれる風潮も生み出されている。

1-3 日本語の表記法と漢字制限

このような漢字の字種の多さにどう対処するかが、近代日本の言語政策の主要な課題であった。

明治の初期には、前島密や福沢諭吉が漢字の廃止を訴え、明治10年代には、漢字仮名交じり文のかわりに仮名を用いることを主張するグループが結集し、カナモジカイが組織された。同時期には、ローマ字を用いることを主張するグループによってローマ字会も組織されている。漢字全廃派と漢字擁護派、漢字制限派と漢字非制限派の激論を経て、漢字の数を制限するが漢字仮名交じり文を維持するという、漢字制限派の主張するゆるやかな改革案が取られることになった。

最初の漢字制限案は、1923年に出された「常用漢字表」である。新聞各社を含めた官民協力の下で、1962字の漢字が臨時国語調査会において定められたが、直後に関東大震災が起こったこともあり、告示には至らなかった。ほぼ10年後の1931年に「常用漢字表」の修正案（1858字）が国語審議会で議決され、さらに10年後の1942年には、漢字数を大幅に増やした「標準漢字表」（2669字）が閣議申し合わせまで至ったが、戦況の悪化もあり、やはり告示には至らなかった。

実際に内閣訓令による告示に至った、最初の漢字制限案は、戦後すぐの1946年に国語審議会が答申した「当用漢字表」（1850字）である。この「当用漢字表」は、漢字使用の

範囲を定めたものがあったが、1981年には、「一般の社会生活において、現代の国語を書き表す場合の漢字使用の目安を示す」ことを目的とした「常用漢字表」（1945字）が国語審議会から答申され、内閣訓令により告示された⁷。

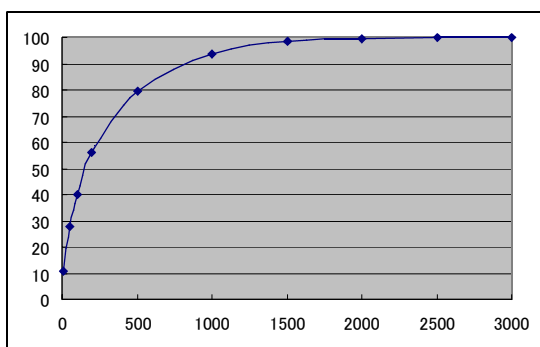
漢字の字種の多さは、言語の機械処理においても、記憶容量と処理能力が乏しい時代には、大きな負担となった。英語では、アルファベットの大文字と小文字や数字、記号を合わせても、7ビット（ $2^7 = 128$ ）に収めることができるが、日本語では、漢字を処理するためには、さらに大きな枠を準備しなくてはならず、それに応じた記憶容量と処理能力が必要になるのである。1970年代には、コンピュータでの日本語表記は、1バイトの枠内に組み入れた半角カタカナ文字を用いるのが普通であった。漢字処理に関わるJISの「第1水準漢字」や「第2水準漢字」などの制定も漢字制限の側面を持つが、これらについては3-1で触れる。

1-4 新聞における漢字の使用実態

現在の規準となっている「常用漢字表」の漢字は1945字、最も多くの漢字をあげた「標準漢字表」が2669字、「第1水準漢字」は2965字であり、ほぼ、日常使われる漢字は、2000字～3000字の枠の中に収まるという理解があると言える。一方、これらの漢字制限に対して、必要な漢字が入っていないという苦情も常に寄せられてきた。

では、日本で日常使われる漢字を2000字から3000字と想定することは本当に妥当なのであろうか。また、もし妥当だとしたら、なぜ、漢字が足りないと考えられるのだろうか。まず、最初の問題から考えていこう。

字	10	50	100	200	500	1000	1500	2000	2500	3000
%	10.6	27.7	40.2	56.1	79.4	93.9	98.4	99.6	99.9	99.9



上の表は、新聞に現れる漢字において、使用頻度順に漢字を並べ、その上位何字までで、全体の漢字数の何%を占めるかを表したものである⁸。使用頻度の高い上位1000位までの漢字で90%を越していることから、現在の「学年別漢字配当表漢字」（いわゆる「教育用漢字」；1006字）を習得していれば、日本の新聞は9割方読めることが知られる。義務

教育で習得が望まれている「常用漢字表漢字」（1945字）で、新聞に出てくる99%以上の漢字がカバーできることになる。日常使われる漢字は2000字～3000字であるという想定は、妥当であると言えよう。表をグラフにしたものを見ると、1500位から曲線が緩やかになり、2500位を越えると占有率の変化もなくなる。使用頻度の少ない漢字がずらっと横並びになっていることをうかがわせる。

さて、常用漢字表の漢字だけを習得した人は、新聞に出てくる0.4%の漢字は知らないことになる。毎日新聞の記事データを2000年度分をテキストデータにすると、約120MBになる。日本語の文字は2バイト文字であるから、これは約60,000,000字ということになる。

さらに、漢字と非漢字（仮名・英数字・記号など）の割合を計算すると、ほぼ1：1であることから⁹、1年間分の新聞に現れる漢字は、約30,000,000字となる。常用漢字表の漢字しか知らない人は、新聞に年間120,000回ほど、知らない漢字に出会うことになる。上位3000位までの漢字を習得したとしても、年間30,000回は、知らない漢字が出てくる計算になる。この数字をどのように解釈するのがよいか。

一般的には、次のように解釈されるだろう。2000字とか3000字とかは、読みも書きもできる漢字であり、書けなくても読める（意味のとれる）漢字はかなりあるので、このような数字が出たとしても、新聞を読むのには不自由はない、という解釈である。確かに、書けなくても読める漢字は相当ある。しかし、新聞のような公共性の高い出版物は、常用漢字表の枠内で書かれるのが基本である。大手の新聞社は独自の用字基準を設けているが、常用漢字表の枠を大きく出るものではない。とすると、常用漢字表を守ろうとしながらも、外れる漢字を年間120,000回も用いなければならなかったということになる。

2000字から3000字の漢字で99%以上の用が足りながら、それでも漢字が足りないという日本語の状態がこのデータからも見えてくる。

1-5 漢字と個人のアイデンティティ

我々は、使用する漢字の大多数が他の人の使用する漢字と共通するが、残りのわずかの漢字が共通しないことを経験的に知っている。「わたなべ」という姓を表すのに、「渡辺」「渡部」「渡邊」「渡邊」などの違いがあり、「さいとう」という姓を表すのにも、「斎藤」「齋藤」「齊藤」などの表記がある。地名でも、「かしま」市は、佐賀県は「鹿島」と書き、茨城県は「鹿嶋」と書いて区別する。渡邊さんは、「邊」という漢字は頻繁に用いても、「邊」はほとんど用いないであろうし、佐賀県鹿島市の市民は、「鹿嶋」という漢字を用いることはまずないであろう。しかし、地名や人名などは、個人個人のアイデンティティに深く関わるものであり、他の漢字で代替するのは、抵抗を覚えるだろう。

日常用いる漢字はおおよそ2000字程度であるという理解があり、そのうち1800字程度は、だれもが選定する漢字となる。それなのに、漢字制限の見直しの他に、新聞社や出版社、研究者や教育者などによって、漢字の出入りに関する意見がかまびすしく交わされるのも、残りの数百字をどのようにするかで意見が大きく異なるからである。

これは、漢字制限を2000字から3000字に拡張したところでさほど状況は変わらない。第一水準漢字は、3000字近くまで選定しているが、「辺」(4A55)は入っているが、「邊」(6E34)は入っていない。旧漢字は第2水準にまわされているのである。戸籍に「渡邊」と旧漢字で記され、普段も手書きで「渡邊」とサインするのが普通になっている者にとっては、「渡辺」はあくまで略式表記に過ぎない。手書き文書と電子化文書は別物で、電子化文書には制限があるものだと割り切れば良いのだが、日常生活にコンピュータが浸透すればするほど、電子化文書にも正式な表記が求められて、第2水準漢字の「邊」が用いられることになる。第1水準漢字と第2水準漢字とを合わせた6000字あまりの漢字でも、十分というわけではない。最近の日本では、「韓流」がブームになっているが、韓国でかなり一般的な姓である「裴(9624)」は、第2水準にもない¹⁰。韓国や中国、台湾などとの交流が盛んになればなるほど、漢字が足りないという声も多くなる。

従来の漢字制限は、漢字の読み書きを考慮して定められてきた。個人の文書作成が手書

きで支えられていたから、これは当然だと言える。しかし、現在、個人の文書作成は、日本語ワードプロセッサ（日本語ワープロ）を用いるのが普通となっている。日本語ワードプロセッサで漢字表記を導くのは、その単語の「読み」（正確には仮名表記）である。このような環境において、漢字の制限がなお必要であるとしたら、それは従来のような、文字の側面からの漢字制限ではなく、単語とその漢字表記を同定するための、語彙＝表記的な側面からであろう。現在の日本語ワードプロセッサは、常用漢字の「渡辺」や「渡部」と変わらぬ手間で、「渡邊」や「渡邊」を導き出せる。「うっかりしたさま」を表す「ウカツ」と言う単語を知っていれば、日本語ワープロで「迂闊」を導くのも簡単である。

日本語ワードプロセッサを用いる場合に求められるのは、「迂」や「闊」の字の意味や成り立ちや部首などの漢字単体の知識ではなく、その単語を表記するのにどの形がふさわしいか（情報伝達に有効か）の見極めができることである。例えば、「当然であるさま」を表す「モチロン」は、「もちろん」と仮名で書かれることが多く、「勿論」と漢字で書かれることは少ないのに対し、同様の意味を表す「ムロン」は、仮名書きよりも「無論」と漢字で書かれる方が圧倒的に多い。この「ウカツ」の場合は、「うかつ」と仮名で書かれることも、「迂闊」と漢字で書かれることもともに多い。「迂」も「闊」も常用漢字表外の漢字であるが、「于」と「活」という音符から「ウカツ」という読みを導くことができ、文脈からこの漢字が「うっかりしたさま」の意味の「ウカツ」を表しているのだろうと読み取ることができるのである。これに対し、「勿論」の場合は、「勿」の読み方がわからず、「モチロン」という単語を導くことが困難なのである。

日本語ワープロは、個人レベルで手軽に通常の文章を活字体で作るという、日本語文書作成における夢を実現したものであった。しかし、その急激な発展が従来の手書きを中心とした習慣や教育システムとの間で、多少のずれや摩擦が生じているようだ。以下の章では、日本語ワープロの機能のうち、日本語の表記と最も深い関わりを持つ日本語入力システム（日本語IME）に注目し、その発展の方向と、現在の日本の言語生活や言語教育への影響について論じて行く。

2 日本語ワードプロセッサの発展

2-1 「漢字仮名交り文を 個人で 手軽に 活字体で」

欧文では、個人でも、タイプライターを使用して、活字体の文書を容易に作成することができる。しかし、文字種の多い日本語で、「個人」で「手軽」に「通常の文章」を「活字体」で印刷物にすることは、きわめて困難であった。活字組の凸版印刷は、漢字仮名交り文を活字体で印刷できるが、活字や印刷機などの施設が必要であり、活字組にも大きな手間がかかる¹⁾。とうてい、個人で手軽に使えるものではなかった。

欧文タイプライターや仮名タイプライターを用いれば活字が打てる。しかし、それで作成されるローマ字文書や仮名文書は、漢字仮名交り文という「通常の文章」から大きく外れるため、結局、限られた範囲でしか普及しなかった。

漢字と仮名とを打ち出せる和文タイプライターも開発され、漢字仮名交り文を活字体で印刷物にできたが、システムが高額で、かつ、文字を一字一字打ち込む必要があり、高速に文字を打ち込むにはかなりの熟練が必要であった。和文タイプライターは、学校や会社などの組織で、清書用に用いられるくらいであった。

謄写版（ガリ版）は、蠟を表面に塗った原紙を鉄筆でひっかいて、文字や線画の部分の蠟を削り、インクをしみ出させて紙に印刷する孔版印刷の一種である。設備は安価で、個人でも手軽に利用できるために、広く用いられた。手書きであったが、活字をまねた「沿溝ゴシック体」などが用いられたことは、1－2ですでに触れた。

これらの利点・欠点をまとめると、次のようになる。これらの方式は、すべての条件を満たす日本語ワープロに取って代わられたのである。

<p>欧文タイプ・仮名タイプ</p> <ul style="list-style-type: none"> ○設備は比較的安価 ○小規模 ○手軽に個人利用 ○活字体 ×ローマ字文・仮名文 	<p>活版</p> <ul style="list-style-type: none"> ×設備が非常に高価 ×大規模設備 ×手間・企業向き ○活字体 ○漢字仮名交じり文
<p>謄写版（孔版）</p> <ul style="list-style-type: none"> ○設備は安価 ○小規模 ○手軽に個人利用 ×手書き ○漢字仮名交り文 	<p>和文タイプ</p> <ul style="list-style-type: none"> ×設備は高価 ○比較的小規模 ×手間・小規模団体向き ○活字体 ○漢字仮名交り文

2-2 日本語ワードプロセッサの誕生

1978年のビジネスショーで東芝が日本語ワードプロセッサの試作機を発表し、同年内に、市販機も出された。この最初の日本語ワープロ専用機は、600万円を超える高額な機器であり、個人の文章作成用ではなく、和文タイプに変わる清書用機械として用いられた。

1980年代にはいると、富士通から100万円を切るワープロ専用機が出され、ついで、漢字を一字一字変換する単漢字変換のものながら10万円を切る機種や、文節変換タイプのビジネス用機種で40万円程度のもなど、より安価な機種が続々と出された。100万円を切る価格は、小規模の商店や学校でも導入することを可能とし、そこで用いられていた和文タイプライターと置き換わった。50万円を切る価格設定は、個人でもボーナスなどまとまった収入がある際に購入を検討させる値段であったし、10万円を切る価格設定は、月々の収入の中でやりくりすれば購入が可能だと思わせる値段であった。

1980年代は、個人にワープロ専用機が普及した時代である。多くのメーカーから工夫を凝らしたワープロ専用機が出され、ハードウェア・ソフトウェアともに大きく発展した。

1983年には、パーソナルコンピュータ（パソコン）で動くワープロソフトも相次いで出された¹²。初期のパソコン用ワープロソフトは、パソコン自体の性能も低く、また、フロッピーベースでプログラムをインストールする必要上、プログラムや変換用辞書に制限があった。パソコンの低価格化と高性能化、ハードディスクやCD-ROMといった記憶装置の低価格化と普及により、1990年代に入ると、徐々にワープロ専用機を駆逐し、パ

ソコンのワープロソフトが主流となった。

1978	日本語ワープロ誕生
1982	100万円を切る専用機発売
1983	パソコン用ワープロソフト誕生
1984	10万円以下のワープロ専用機発売 ワープロ専用機のパーソナル化が加速
1990年代	パソコン用ワープロソフトの隆盛
1990年末	携帯電話の爆発的普及・携帯用変換辞書の開発
2000年初	ワープロ専用機の終焉 2000年 松下・東芝がワープロ専用機から撤退 2001年 NEC・富士通がワープロ専用機から撤退 2002年 シャープがワープロ専用機から撤退

3 日本語入力システムの発展

3-1 文字コード

まず、コンピュータで用いる文字コードについて触れておこう。文字コードとは、コンピュータ上で文字や記号を扱うために、文字や記号に割り振った一定の符号をいう。

まず、アメリカ規格協会(ANSI ; American National Standard Institute)が1963年に制定した「ASCII (American Standard Code for Information Interchange)」と呼ばれるコードがある。これは、英数字や記号を7ビット ($2^7 = 128$) の枠に入れたもので、16進法で示すと、「21」(!) から「7E」(~)までの2桁で表される。「A」は「41」、「?」は「3F」となる。

このASCIIのコードを拡張して、半角片仮名を組み込んだのが、1976年に定められた日本工業規格(JIS ; Japanese Industrial Standards)の「JIS-X0201」である。ASCIIでは、英数字を00から7Fまでの128の枠の中に収めるが、これを8ビット($2^8=256$)に拡張して、A1以降の128の枠に句読点やカギ括弧、半角片仮名、濁音符や半濁音符などを入れている。「。」(句点)は「A1」、「ア」は「B1」、「ン」は「D0」となる。

漢字を処理するためには、1バイト(8ビット、 $2^8 = 256$)では足りず、2バイト($256 * 256=65536$)の枠が必要になり、16進法では4桁のコードが振られることになる。

1978年に制定された「JIS X0208-1978」(旧JIS)は、全角の平仮名や片仮名、記号類など「非漢字」524字を「2121」(; 全角スペース)から「2840」(+ ; 罫線)まで割り当て、さらに、使用頻度を元に、基本的な漢字2965字を第1水準漢字として、「3021」

(亜)から「4F53」(腕)までに割り当て、次いで使用頻度の高い3384字を第2水準漢字として、「5021」(弍)から「7426」(熙)までに割り当てている。

「JIS X0208-1978」の不備を修正するために、1983年に「JIS X0208-1983」、1990年に「JIS X0208-1990」、1997年には「JIS X0208-1997」が制定された。「第1水準」文字2965字、「第2水準」文字3390字を定めている。先の「JIS X0208-1978」を「旧JIS」、「JIS X0208-1987」以降の修正版を「新JIS」とも言う。

「JIS X0208」で定めた第1水準漢字や第2水準漢字をさらに補うものとして、1990年

に「JIS X0212-1990」によって「補助漢字」5801字が制定された。また、これとは別に、2000年に「JIS X0213」によって、「第3水準漢字」1910字、「第4水準漢字」2436字が制定されている。

この他、世界各国の文字体系に対応する統一文字コード体系として、Unicode コンソーシアムの提唱する「Unicode」がある。

3-2 コード入力と単漢字変換、熟語変換

コード入力は、「4A38」を入れて「文」を呼び出し、「3E4F」を入れて「章」を呼び出すように、直接コードを指定して文字や記号を入力する方式である。「文章を書いた」という文字列をコードで入力すると、「4A38、3E4F、2472、3D71、2424、243F」といったように、それぞれの文字のコードを一字一字入力しなくてはならない。機械の性能が劣り、十分な記憶容量もない時期には、これも仕方がなかったが、あまりに非効率的である。

そこで、それぞれの漢字の読み方と漢字との対照表を作り、漢字の音訓をインデックスとして、同音字の一覧から漢字を選び出す「単漢字変換」と呼ばれる方式が考え出された。例えば、「ブン」と読みを入力すると、「文・分・聞・蚊」などの漢字が変換候補としてあがり、人間がその中から求める漢字を選び出すというものである。「コード入力」では言語情報は関与しないが、「単漢字変換」では文字レベルの情報が組み込まれているのである。もちろん、漢字一字一字に対し、音訓を登録しておくのであるから、仮に「望」のように、常用漢字表レベルで、訓「のぞむ」、呉音「モウ」・漢音「ボウ」の音訓を持つ漢字は、一文字につき3つのインデックスを持つことになる。3000字程度の漢字でもインデックスはその何倍かの数になり、それを記憶する容量と検索する能力が必要になる。

先の「文章を書いた」を単漢字変換すると、「ブン」と入力して「文・分・聞・蚊」などの候補から「文」を選択し、「ショウ」と入力して「小・省・将・章・生・・・」などの候補から「章」を選択し、「ヲ」と入力して平仮名変換し、「カク」と入れて、「各・画・角（音読候補）・欠・書・描（訓読候補）・…」などから「書」を選択し、残りの「イ」と「タ」を平仮名変換するのである。

コンピュータの能力と記憶容量に余裕ができると、単語の読みを索引として同音語の一覧から単語を選び出す「熟語変換」という方式が用いられるようになった。「ブンショウ」と入力して、「文章・文相・文正・分掌」などの候補から「文章」を選択する方式である。小型国語辞典には、おおよそ5万語から7万語の見出しが立っている。一つの見出しの元に、「あし；足・脚」のように複数の表記が示されているものもかなりあり、さらに国語辞典では集録されない人名や地名などもあるので、日常用いる単語の表記を導くには、10万を超えるインデックスが必要になるのである。

3-3 文節変換

「熟語変換」では、「を」や「た」のような付属語は、個々に処理しなくてはならず、付属語や活用変化などの文法情報を加えて文節単位で変換する「単文節変換」という方式が考えられ、さらに、2文節以上の文章を自動的に文節に区切る「連文節（自動）変換」という方式が考え出された。「単文節変換」では、「ブンショウヲ」という文節で一端入力を止め、変換キーを押して「文章を、文相を、文正を、分掌を」などの候補から「文

章を」を選択し、再び「カイト」と入力して「颯田・海田・開田・貝田・書いた・描いた・欠いた」などの候補から「書いた」を選択するのである。

「連文節（自動）変換」では、「ブンショウウヲカイト」とまとめて入力すると、自動的に「ブンショウウノカイト」のように文節区切りをし、それぞれの文節で単文節変換と同様の候補を提示する。擬似的な構文解析を行っているのである。

文節区切りの仕組みには、様々な方式が用いられているが、有意義な最も長い2文節を切り出して候補とする「2文節最長一致法」という方式を例に、「ハナコニハタラキカケルトモダチハ」という文字列の文節わけを示そう。まず、この文字列の先頭から1文字をとり、それと変換辞書にある語と引き比べる。「ハ」には、「葉・歯・刃・羽・・・」など文節となりうる有意義な候補がある¹³ことから、まず、この1文字を仮に第1文節とする。さらに、それに続く「ナコニハ・・・」から第2文節の候補を探すと、「ナ」は「名・葉・・・」などの有意義な文節候補があがるが、「ナコ」「ナコニ」「ナコニハ」など、2文字以上では、有意義な文節候補が設定できない。そこで、「ハ」という最初の1文字を取った場合、2つの文節の最長文字数は「ハノナ」の2文字ということになる。

続いて、「ハナ」と2文字取ると、「花・鼻・・・」などの文節候補が上がり、第2文節も「コ」（子、個・・・）、「コニ」（子に、個に）、「コニハ」（子には、個には・・・）などの候補が取り出せるが、「コニハタ」以上では有意義な候補が取り出せない。第1文節を「ハナ」で取った場合の2文節の最長文字数は、「ハナノコニハ」の5文字となる。

第1文節は、「ハナコ」、「ハナコニ」、「ハナコニハ」のように5文字までは取ることができるが、「ハナコニハタ」以上はもはや有意義な文節にはならない。「ハナコ（花子・華子）」という3文字を第1文節にすると、第2文節は「ニ（荷・2・二・・・）」と「ニハ（荷は、2は、二は・・・）」が取れ、2文節の最長文字数は「ハナコノニハ」の5文字になり、「ハナコニ（花子に、華子に）」だと、第2文節は「ハ（葉・歯・・・）」「ハタ（旗・畑・端・・・）」「ハタラ（幡羅；地名）」「ハタラキ（働き）」「ハタカキカ（働きか；「働き」＋疑問「か）」、「ハタラキカケ（働きかけ・働き掛け）」、「ハタラキカケル（働きかける・働き掛ける）」「ハタラキカケルト（働きかけると、働き掛けると）」のように取れるが、「ハタラキカケルトキ」以上は1文節では解釈できないので、2文節の最長は、「ハナコニノハタラキカケルト」の12文字となる。第1文節を最も長く5文字取って、「ハナコニハ（花子には、華子には）」にした場合、第2文節は「タ（田、他）」、「タラ（鱈、太良、多良）」と取れるが、「タラキ」以上は有意義な文節にならない。2文節の最長文字数は「ハナコニハノタラ」の7文字となる。以上の中で、最も長く2文節の文字数が取れたのは、「ハナコニノハタラキカケル」であるから、第1文節は、「ハナコニ」で区切られると仮定するのである。次いで、残った「ハタラキカケルトモダチハ」でまた「ハ」「ハタ」「ハタラ」・・・と文節を取っていくと、最も長く取った「ハタラキカケルト」では、続く文節が「モ（藻・裳・・・）」しかとれないが、「ハタラキカケル」だと「トモダチハ（友達は、友だちは）」まで取れる。そこで、全体の第2文節は「ハタラキカケル」で区切られると仮定する。

上の例では、2文節最長一致法がうまく機能しているが、これがうまくいかない場合もある。有名なのは、「オカノウエニハナガサキマシタ」という文字列であり、2文節最長一致法だと、「オカノノウエニハノナガサキノマシタ」と分析され、期待する「オカノノ

ウエニ／ハナガ／サキマシタ」という分析が得られない¹⁴。これは、切り取った文字列を変換辞書に機械的に当てはめていくからで、文字列から「岡の」「上には」「長崎」「真下」という有意義な単語を導けても、その単語の連続が有意義であるかどうかは全く考慮されていないからである。意味の接続を考慮しない分析の限界がここにある。

そこで、「花」は「咲く」と共起しやすいという意味的な情報を組み込むことが考えられた。いわゆる「AI変換」である。「花」と「咲く」とを優先的に結びつけることにより、「花が／咲きました」という2文節が抽出でき、「岡の／上に／花が／咲きました」という正しい文節区切りが可能となるのである。単語の共起情報を組み込むことは、「暑い」「熱い」「厚い」「篤い」などの同音異義語を処理にも役立っている。「夏・気温」などは「暑い」、「湯・スープ」などは「熱い」、「本・板」などは「厚い」、「志・病気」などは「篤い」と結びつきやすいという情報を与えることで、「病気が篤い」「厚い板」などを的確に出せるのである。さらに、「割る・切る・蹴る」などの動詞は、「(人)が(物)を」のような型を取りやすく、「行く・帰る・乗る」などの動詞は「(人)が(場所)に」のような型を、「運ぶ・移す・積み上げる」などの動詞は「(人)が(物)を(場所)に」のような型を取りやすいといった動詞の格体制の情報を組み込むことで、さらに的確な変換を可能にしている。

3-4 日本語入力システムの可能性

現在では、シソーラスを準備し、前の部分で「火星」や「宇宙」などが話題になっていたなら「キヤイ」を「金星」に変換し、「江戸」や「中世」などが話題になっていたなら「近世」に変換するような文脈上の情報を考慮した「文脈変換」や、思いつく単語、例えば、「持って行くのを忘れた」と仮に入力して、そこから「忘れ(た)」の部分で「失念し(た)」や「うっかりし(た)」などに変換し、また「持って行く(のを)」の部分も「失念した」に合わせて、「持参する(のを)」に交換する「類義表現変換」といった、より意味を盛り込んだ変換技術の開発も進められている。

コード入力から単漢字変換までは、対象の漢字を知らなければ利用できず、所与の漢字を導く「漢字」変換システムであった。手書きで言えば、漢字索引を用いて漢字を調べる作業と同等である。次いで、熟語変換から連文節変換までは、対象の単語を知らなければ利用できないのであり、所与の単語について、読みから標準的な日本語表記を導く「読み-表記」変換システムである。これは、手書きで言えば、国語辞典の見出し語と漢字表記欄とを利用する「字引き」の作業と同等である。さらに、AI変換では、利用者が対象となる単語を知らなくても正しい単語とその表記が提供される、「読み-単語」変換システムである。これは、国語辞典の意味を調べて、適当な漢字表記を探す作業と同等である。最後の類義表現変換になると、読みからも離れてしまい、利用者の知っている単語から利用者の知らない表現を導く、「単語-表現」変換システムになっている。手書きで言えば、類義語辞典を用いるのと同様であると言える。

「類義表現変換」は、提示するものと得られるものは、類義語辞典を用いるのと同様であるが、作業の意識の面からすると、全く異なったものと言える。現在、日本語入力システムから国語辞典や類義語辞典などの電子化辞典をスムーズに引く仕組みは、いくつか存在している¹⁵。しかし、これらは、「文書を作る」という作業とは別に、「辞書を引く」と

いう作業を並行して行わせるものである。コンピュータ上で、どんなに容易に国語辞典や類義語辞典が立ち上げられるにせよ、国語辞典の画面が現れた時点で、文書作成の画面が部分的に遮られるばかりでなく、「文書を作る」から「辞書を引く」へと意識も変わらざるを得ない。文書の作成者は、漢字を調べたいのでも、単語の意味を知りたいのでもない。ただ、その文書にふさわしい表現を求めているのである。従来の辞書引きは、よりふさわしい表現を得るために、漢字や単語を引くという迂遠な道を経なければならなかった、といっても良いだろう。単語の意味や用法を調べることが主たる目的であるときには、電子国語辞典を繰り広げればよい。しかし、文書作成時には、辞書引きは夾雑物になるのである¹⁶。

日本語入力システムでは、ソフトウェアは変換辞書を参照しているが、使用者はそれを意識しない。ワープロの文書作成の際に、日本語入力システムで単語の読みから漢字表記を導くのは、手書きの文書作成の際に、書けない漢字を「よみ」から国語辞典で引くのと、似ているようで全く違った作業である。日本語入力システムの漢字変換は、文書作成の一部であって、文書作成を妨げる夾雑物にはならない。従来の日本語入力システムが文書作成を妨げることなく、単語のよみから漢字表記を導くのと同様に、「忘れた」から「失念した」にシームレスに変換する「類義表現変換」も、文書作成の一部として、単語の意味から、よりふさわしい表現を導く。

3-5 標準表記と表記の公共性

日本語入力システムは、おおよそ、常用漢字音訓表に準拠した、標準的な表記を第1候補として提供する。日本語では、漢字の使用法にはかなり厳しいが、単語の正書法は緩やかである。小学校の国語の漢字のテストでは、「(重さを)ハカル」に「図る」や「諮る」と書くのはもちろん、「測る」や「計る」と書いても点はもらえない。「体重の測定」とか「釣った魚の重量の計測」とか言うのではないかと反論しても、「重さ」や「容量」は「量る」と書かなくてはならない。一方で、「大根」という野菜は「ダイコン」と片仮名で表記しても「だいこん」と平仮名で表記しても許容される。多くの国語辞典で、動物の「いぬ」の項目は、見出し語では「いぬ」と平仮名で示され、漢字表記欄では「犬」と「狗」が並べて掲げられ、語釈では「イヌ科の小動物」と片仮名表記が見られる。

もちろん、国語辞典に載っている表記が普通に使われているのかというと、そうでもない。昆虫の「かぶとむし」や「くわがたむし」は、平仮名や片仮名で書かれるが、「甲虫」や「鍬形虫」のように漢字で書かれることは滅多にない。一方、「ちょう」や「が」「せみ」などは、「蝶」「蛾」「蟬」のように漢字で書かれることも多く、分類学的なカテゴリーと表記とは一致しない。

このような多様な表記に対し、国語辞典では、それを併記すればすむが、日本語入力システムでは、どれがより標準的なのか、変換候補の順序をつけなければならない。そして、これは上に見たように、それぞれの語によってかなり違ってくる。以下の表は、『新潮文庫の100冊』中の日本作品（文学作品）と『毎日新聞』1999年・2000年の2年間分（新聞）、2004年12月時点のinfoseekによる検索ヒット件数（インターネット）の3種をデータベースとして、「だいこん」と「ごぼう」の表記を調査したものである¹⁷。

	文学作品	新聞	インターネット
だいこん	0	20	7,012
ダイコン	3	169	8,137
大根	84	419	93,980
ごぼう	3	27	18,866
ゴボウ	4	72	11,218
牛蒡	18	0	3,169

「だいこん」の表記は、3種とも「大根」>「ダイコン」>「だいこん」の順になっており、この順で変換候補として出せばよいことがわかる。ところが、「ごぼう」は、文学作品では漢字表記が最も多く、新聞では片仮名表記が最も多く、インターネットでは平仮名表記が最も多くなっている。ここで調べた文学作品は、明治時代から昭和40年代のものなので、時代を反映していると解釈でき、「牛蒡」という表記候補を3番目にするには問題ないだろうが、新聞とインターネットのどちらを優先させるか、悩ましいところである。googleのサーチでは、「ごぼう」約325,000、「ゴボウ」約165,000、「牛蒡」約80,800、BIGLOBEの「Attayo」でも、「ごぼう」約49,300、「ゴボウ」約26,000、「牛蒡」8,960と、infoseekと同様の順番となっている。ホームページ上の表記は、ほとんど日本語入力システムを用いたものと考えられるので、現在の使用傾向に合わせるならば、「ごぼう」>「ゴボウ」>「牛蒡」の順番で候補にあげればよいことがわかる。

AI変換や文脈変換など、自動変換の実用性が高まることは、そのまま第一候補がそのまま確定されやすくなることを表すのであり、上のような調査は、日本語入力システムの変換辞書作成には欠かせないものとなっている¹⁸。かつては、日本語入力システムの変換辞書には、「水飯器」のような誤登録も少なくなかった。日本語入力システムの「表記の公共性」が意識されるようになり、どのような基準で登録語彙を選定するか、また、何を第一候補をし、他の変換候補もどのような順番で並べるかといったガイドライン策定のための監修委員会を設置するところも出てきた¹⁹。この監修委員会の検討内容を伝える資料によると、1990年代の日本語入力システムは、理想的な使用者を想定して開発が進められ、「中年」の「文系」の「男性」が「仕事」で書く「標準語」の文書作成を支援することを想定していたという²⁰。父親が書斎やオフィスで文書を作成するイメージであると言われている。

「理想的使用者」を想定することは、一面では「表記の公共性」を支えるが、他方、その「理想的な教養に裏打ちされた達意の表現を実現するために、「(茶を) 淹れる」や「趨る・奔る」など、従来は、特殊な学習や辞書引きの手間をかけて得られた、高コストな表記についても、同音語の変換の選択肢として第二候補以下で提供することになった。実際の使用者の側も、これらの表記が提供されることは、表記の自由が保証されると共に、もともと、和語をどの漢字で書き表すかという知識は、日本語の教養の一つと見なされてきたこともあり、自分の日本語の運用能力を高めるのに役立つと、好意的に捉えられたようだ。そして、高コストな表現も手軽に利用できるようになったことから、「お茶ならば『淹れる』を使うべきだ」「勢いよく走る場合には、『奔る』がふさわしい」といったよう

な、漢字の書き分けを偏重する傾向が強まった面さえ見られたのである。

3-6 携帯電話の普及と表記の個人性

ところが、近年、若年層における携帯電話の普及と携帯メールの利用にともない、話し言葉への対応が必要となり、若者言葉や地域方言など、言語の位相差に対処することが求められた。使用者の表記を「理想的使用者」に近づけさせるという方向から、個人個人の出したい表記をそのまま出すという個別化への転換が求められたのである²¹。

若者やビジネスマン、主婦など、使用者の職業や年齢などによった使用語彙の調査のほか、最初にインストールする時点で、個人個人が既に作成した文書をまとめて読み込ませて、個人の書き癖や表記法の好みを探り、変換に反映させるシステムも取り入れられている。

日本語変換システムには、学習機能が付いており、表記毎に確定させた数を登録している。これをもとに、確定数が多い表記や直前に確定された表記を、確定数が少ない候補や以前に確定された表記よりも優先させることで、個人の趣向や今書いている文章に対応させてきたのである。しかし、学習機能は、場面を分けず一律に働く。そのために、硬い文章を書いた後で、私的な e-mail の文章を書くような場合、前の硬い言葉遣いが学習機能によって後にも反映されることになる²²。学習機能を一律ではなく、場面や文脈に応じて、細かく学習を行わせることで、逆に、学習情報の場面に応じた利用が可能となる。今後、日本語入力システムは、個性化を進め、手書きと同様の感覚で用いられるようになることが予想されるのである。

現在、我々は、携帯電話で、国語辞典や漢和辞典、和英・英和辞典をはじめとした各国語辞典、百科事典などを検索することができる。書齋でしか実現し得なかった国語辞典や百科事典が常に脇にある環境を、ほぼ常時手に入れているのである。現在の検索サービスは、利用者が指定した語をインデックスにして情報を得るもので、思いついた語が不適切であれば検索はそこで止まってしまい、それ以上の展開は望めなくなる。

しかし、現在、パーソナルコンピュータの日本語入力システムで進められている「類義表現変換」は、いわば、手元の簡単なきっかけを元にさらに広く深く考えを展開させるための道具である。パーソナルコンピュータで実現されるものは、携帯電話でもすぐに利用可能となる。携帯電話の日本語入力システムを用いて、一つの言葉から自分では思いつかなかったような様々な言葉を導き出し、それをキーワードにして、種々のデータベースで検索することも、ここ数年のうちにできるようになるだろう。我々は、手書きでは考えられなかったインターフェースを獲得しつつあるのである。

4 日本語入力システムと言語教育

4-1 日本語変換システムを理解することの意味

現在、日本語入力システムを使った文書作成は、現代日本社会において、最も標準的な方式となっている。言語政策や言語教育について論じるには、日本語入力システムの現状をふまえることが不可欠であり、国語辞典もこのような言語環境において使われることを想定して編纂されることが望まれるし、言語教育もこのような環境において言語運用能力が高まるように指導されることが望まれる。しかし、実態はどうであろうか。

かつて、「情報化時代に適合した」ことを謳った国語辞典があったが、その内容は、五十音順に並べられた漢字項目にコード番号を付すというものであった。その国語辞典に漢字単独の項目として掲出されたのは、常用漢字と人名用漢字を含む 2663 字であり、その程度の漢字であれば、当時の日本語変換システムでも当然登録されており、国語辞典で引くことのできる「読み」ならば、日本語入力システムの通常の変換でも対応している。この当時、ワープロの使用者がわざわざ「ワープロ漢字辞典」を調べなくてはならなかったのは、漢字の読みがわからない場合か、変換辞書に登録されていない読みしかわからない場合、ワープロが第 2 水準漢字や補助漢字に対応していないなど、その漢字を出すこと自体をあきらめなければならないか確かめる場合、等であった。このような要求には、五十音順に配列された 2663 字程度の漢字では、所詮応じることができないのである。この程度の処置で「情報化時代に適合した」と言ってしまうとすれば、それは、日本語入力システムを全く利用していないか、理解が極めて浅いかである。

学校教育でも、しばしば、「ワープロ時代にちゃんとした漢字の知識を与えることが必要である」といった説明もなされる。しかし、その「ちゃんとした漢字の教育」は、日本語入力システムのあり方を踏まえて設定されたものではなく、「伝統的な漢字教育」を示すに過ぎない場合も見受けられる。漢字の成り立ちや変遷は、古代中国文化の理解に必要であるし、日本における漢字の受容の歴史は日本文化をより深く知る上で欠かせない。伝統的な漢字教育は、それ相応の意義を持つ。ただ、現在の日本語の表記システムを考えると、手書きを前提にした伝統的な表記観（漢字観）では処理しきれないことが少なくない。原稿用紙の使い方の指導以上にワープロでの文書作成の指導が必要であり、国語辞典や漢和辞典を使いこなせるよう指導することと並んで、日本語入力システムを使いこなせるよう指導することが必要になっている。当然、それを指導する教員は、国語辞典や漢和辞典について正しく理解しているように、日本語入力システムについても正しく理解することが望まれているのである。

4-2 情報化時代の言語教育

最後に、日本語入力システムの発展と実現国語教育との関わりについて、まとめておく。

- 1) 日本語入力システムが日常化した現在においては、「漢字を正しく書く」という文字教育から、「語の適切な表記を知る」という表記教育へと移行する必要がある。「漢字を正しく書く」ことは主たる目的ではなく、「語の適切な表記を知る」ための補助的な手段と位置づけるのが適当である。
- 2) 言語教育のための情報端末の利用法をもっと追求しなめればならない。情報端末は胡散臭いものだという教師の直感は、8割方正しい。しかし、その胡散臭さは、技術の進展によりすぐに解消される。「水飯器」のような胡散臭い変換をしていた日本語入力システムにもうそのレベルの胡散臭さはない（すぐにまた別の胡散臭さがつきまとうが、それも解消されていく）。最初の胡散臭さから敬遠し、子供たちを守ろうとするのだろうが、どうせ、子供たちはそこに入っていく。先に、携帯電話の情報端末としての可能性について述べたが、現在、初等・中等教育に関わる多くの学校では、携帯電話の携行を制限している。そこには相応の理由があることはわかるが、携帯電話の情報端末としての機能を十分に理解させ、適切な使用法を指導することは、情報教育・言語教育の面から必要である。

コンピュータを用いた情報検索と並んで、携帯電話を介した情報検索も、社会生活に必要な技術となりつつあるし、電話での対話と対面して対話との比較や、携帯メールと e-mail や手紙文との比較など、自己の言語生活を自覚させる多くのきっかけが潜んでいる。道具を道具としての的確に効率よく使う技術を子供と一緒に磨けばよい。

3) 手書きの意味づけをする必要がある。現在、ワープロを用いた文書作成が主流になっていることは、だれもが認めるだろう。このような時代に、あえて手書きをすることの意味を積極的に問いかけ、子供とともに考えていくことは大きな意義を持つだろう。教師の答えを復唱させてはならない。現在、我々は、従来の言語生活や言語意識を根本から変える大きな変革の中にある。子供たちは、この変革を生き抜き、次のステージに進む存在であるが、我々は、半ばこの変革から取り残された存在である。一世代前の考えを理解し、自分たちのあり方の中でそれを位置づけることが、文化の継承となる。

4) グローバル化の中で、日本というローカルな言語がどのような意味を持つのか、世界共通語となりつつある英語に対し、どのようなアドバンテージを見いだせるか、考えていかなければならない。キーボードで入力した文字がそのまま日常の表記にならない日本語だからこそ、よりふさわしい表記を導くための様々な工夫がなされてきた。そこで行われた言語知識と言語処理と融合は、このような処理を必要としない他言語の言語処理に対して、大きなアドバンテージとなっている。グローバル化の中で、日本語も中国語もフランス語も、地域方言の位置に納まることになる。その方言の豊かさや暖かさ、細やかさなど、良い点を自覚し、共通語話者や他の方言話者に示すことができなければ、また、共通語に対して劣等感を感じるだけであれば、その方言は共通語に徐々に飲み込まれるだけのものとなる。伝統に安穩として、縮小再生産をするような教育であってはならない。

【参考文献】

- 小林龍生(2002)「漢字・日本語処理技術の発展：仮名漢字変換技術」『情報処理』43-10
- 篠原一(2003)『電脳日本語論』 作品社
- 矢澤真人(1993)「国語電子辞書引き比べ」『月刊言語』22-5 大修館書店
- 矢澤真人(1994)「道具としての国語辞書」『日本語論』2-4 山本書房
- 矢澤真人(1995)「辞書の記述と利用—言語系—」『日本語学』14-4 明治書院
- 矢澤真人(1998)「より豊かな表現を目指そう」沖森卓也・半沢幹一編『日本語表現法』三省堂書店
- 矢澤真人(2000)「情報化社会と国語科教育」2000年度筑波大学公開講座「学校教育のための情報教育講座」資料；<http://quoniam.social.tsukuba.ac.jp/yamane/article/intel20000801/>
- 箭内敏夫(1994)『電脳辞書の国語学』おうふう

1 本稿は、クリスティアン・ガラン氏によるフランス語訳の元となった日本語原稿に一部手を加えたものである。この日本語版は、フランス語版で削除された要旨を復活させたほか、一部の図を削除するとともに、2006年において再調査した結果を注で補ってある。

2 平仮名と片仮名にも使い分けがあり、基本的に片仮名はmarkedを表す（和語に対する外来語やオノマトペ、個々の動植物名など）。

3 （広義の）「唐音」を、鎌倉時代から室町時代に、宋・元の時代の中国語から伝わった「宋音」と、江戸時代に、明・清の時代の中国語から伝わった（狭義の）「唐音」とに分けることもある。

4 「船尾（船の後部）」に対する「船の前部」の意味なら「セントウ」と発音される。

5 言語的には、北朝鮮の言語も韓国の言語も「朝鮮語」としてくれるが、ここでは、表記面に注目し、ハングルを専用する北朝鮮の朝鮮語と区別して、漢字とハングルを併用する韓国の朝鮮語を「韓国語」と呼ぶ。

6 中国語や韓国語においても、意味の違いに応じて、同じ漢字が二つ以上の発音を持つ。例えば、「率」は、中国語（北京官話）では、「率いる」の意味では「shuai3」、「割合」の意味では「lü3」と発音される。日本漢字音でも、この区別は保持され、前者の意味では「ソチ」（呉音）・「シュツ」（漢音）、後者の意味では「リチ」（呉音）・「リツ」（漢音）と発音され、また、前者には「ソツ」という慣用音も持つ。中国語では、さらに声調の区別が意味の区別に大きく関わるが、日本語の漢字音や韓国語の漢字音では、中国語の声調の区別を反映していない。

7 この他に、小学校で学習すべき漢字を定めた「当用漢字別表」（教育漢字）が1948年に881字制定され、その後、「学年別漢字配当表漢字」と名称が変わり、1999年には1006字となっている。また、「当用漢字」を補うものとして、1951年に、法務省の管轄のもと、「人名用漢字」92字が制定され、その後の増補を受け、2004年には287字が制定されている。

8 『日本語百科大事典』（大修館書店；1988）p342による。

※2003年1年分の毎日新聞データによる再調査の結果は、以下の通り。

10	50	100	200	500	1000	1500	2000	2500	3000
8.7	24.5	37.3	53.7	78.4	93.7	98.4	99.7	99.9	99.9

9 同じデータから約2万字のデータ（分野の違う2つの抽出データ）で文字数を比べると、漢字が9521字で48.1%、非漢字が10269字で51.9%となった。ただし、これは新聞の比率であって、一般の文章ではもっと非漢字の比率が高くなる。例えば、『照葉樹林文化』（上山春平、中公新書）では、漢字25440字（28.4%）：非漢字64093字（72.6%）であり、『新潮文庫の百冊』所収の中島敦の小説（『山月記』『名人伝』『弟子』『李陵』）では、漢字24983字（36.9%）：非漢字42791字（63.1%）である。本論文は、漢字34.0%：非漢字66.0%の比率となっている。

※後日、2003年1年分の毎日新聞データを集計したところ、以下のような数値になった。

年間の文字数約5300万字 漢字2250万字（42.4%）：非漢字（3050万字）57.6%

第1水準（約3000字）以外の漢字の出現回数 36700回

また、「新潮文庫の100冊」および「明治の文豪」所収の何人かの作家について、同様の調査を行ったところ、以下のような数値が得られた。

	漢字	非漢字
夏目漱石	33.6%	66.4%
森鷗外	32.3%	67.7%
尾崎紅葉	33.3%	66.7%
芥川龍之介	31.6%	68.4%
島崎藤村	34.2%	65.8%
野坂昭如	25.9%	74.1%
三島由紀夫	30.0%	70.0%
井上ひさし	20.9%	79.1%
司馬遼太郎	28.7%	71.3%
村上春樹	22.1%	77.9%

10 「裵」は、第2水準漢字である「裴」(6A6A)の異体字。先の「辺」と「邊」と同様に、「裵」さんを「裴」と書いてすませしてしまうかどうかである。

11 詩人・童話作家の宮沢賢治(1896-1933)は、童話『銀河鉄道の夜』において、活版所で働く少年の姿を描いている。少年は、まだ昼なのに電灯がたくさんついている活版所で、一枚の紙切れを渡され、虫眼鏡とピンセットを用いて、粟粒くらいの活字を何べんも眼を拭いながら拾い、6時過ぎにようやく拾い終える。

12 当時のパーソナルコンピュータは高価であり、日本語ワードプロセッサのソフトウェアも10万円を越す価格であった。さらにプリンタも別途用意する必要があった。これに対し、プリンタ内蔵のワープロ専用機は、1980年代中頃には、文節変換のできるビジネスモデルで40万円前後であり、個人用の機種は20万円を切っていた。さらに、年々、性能の向上と価格の引き下げとが続けられ、各家庭に1台、そして各個人に1台と、ワープロ専用機が普及していった。パソコンがこれを挽回するのは、機体やワープロソフトの値段が下がると共に、ワープロ以外の表計算ソフトやデータベースソフトの価格も下がり、それらとの連携が可能になるという、パーソナルコンピュータの汎用性が発揮されてからのことである。

13 自立語1語でも文節になると見なす。

14 2番目の文節と3番目の文節で、「ウエニハ／ナガサキ」は8文字になるのに対し、「ウエニ／ハナガ」は7文字になる。

15 株式会社JUSTSYSTEMの日本語変換システム「ATOK」とATOK連携電子辞典「明鏡国語辞典・ジーニアス英和／和英辞典」(いずれも書籍版は大修館書店)など

16 もちろん、わざわざ辞書を引くことで、あらたな言葉をお言葉覚えたり、自分の誤りを知ったりする利点はある。しかし、これは、文書作成そのものとは関わらない、副次的な利点に過ぎない。

17 ※2006年5月の時点での再調査の結果は以下の通り(小説:「新潮文庫の百冊」no 日本人作家、新聞:毎日新聞2003年1年分、web:yahoo Japan2006/5/23)。

小説 新聞 yahoo

小説 新聞 yahoo

だいこん	0	3	647000	ごぼう	1	16	1510000
ダイコン	3	42	649000	ゴボウ	1	49	826000
大根	84	286	6440000	牛蒡	11	0	531000

18 この一方で、変換辞書が上位にあげればあげるほど、その表記が多く確定される可能性が増すのであり、いたちごっこになる怖れもある。

19 株式会社JUSTSYSTEMの「ATOK監修委員会」。

20 篠原一(2003)「第一章 ATOK監修委員会の設立とその問題意識」参照。

21 篠原一(2003)「第二章 第二期監修委員会の議論」参照。

22 筆者は、よく明治時代の漢文訓読体の文章を入力するが、その後は「之を」や「其れ」「或いは」などの表記が学習されて、なかなか元に戻らず、苛立ちを覚えることも少なくない。