日本語処理システムにおける「外字」と「外字処理」

コンピュータ　システム　コンサルタント　　大倉　信治

1. はじめに

昭和40年代初頭、初期的な「漢字プリンタ」が実用化されて以来、古くから我が国でデータやメッセージの表記手段に用いられてきた漢字が、情報処理の対象として認知されるようになり、いわゆる「日本語情報処理」時代が開幕した。

今日、汎用コンピュータはもとより、ミニコン、パソコンの類にまで日本語ないしは漢字処理の機能が搭載され、さらには日本語ワードプロセッサが広く市場に浸透しつつあるのが現状である。

しかし、日本語の表記手段としての漢字は周知のとおり表意文字であり、仮名ローマ字、ギリシャ文字などの表音文字と違い、1個の漢字で表わせる概念の数だけ、それぞれ異った文字図形（graphics）で描ける漢字の種類が存在するのであり、その数は「無限」とはいえないにしても限定的ではない。

この点、表音文字集合が有限で閉じた体系であるのと対照的である。

そこで、我が国で社会生活を営むのに必要最少限の漢字集合を規定し、国や地方自治体、あるいは公共機関が民生上の情報を、この集合に属する漢字だけで書き表わすよう、一つの目安として当用漢字表（昭和20年内閣告示）が、さらに常用漢字表（昭和56年内閣告示）が制定された。

しかし、一般の日本人が実際の社会生活、文化生活を営むためには、常用漢字表に示された漢字集合の範囲内だけでは、表現したい概念を日本語として表記ができない場合もあり、あまつさえ人間の姓名を表記するための漢字までも考慮に入れるならば、常用漢字表におさめられたわずか2,000字種にも満たない漢字では表記しきれないのが実情である。

ここに有限な表音文字集合を処理対象とした、従来の情報処理技術を拡張して、漢字表記の情報に挑戦するとすれば、1個の情報処理アプリケーション・システムで取り扱い可能な漢字集合の範囲を何らかの方法で限定し、この範囲を逸脱した漢字の処理には、これに対処する特段の処理方法を別途講ぜざるを得ないというのが現実である。

そこで一般には、個々の情報処理アプリケーションで現実に扱わねばならなくなった漢字のうち「標準的でない」処理方式、別の言葉でいえば「例外処理」をせねばならない漢字を「外字」といい、その処理方法を「外字処理」といっているようである。

それでは、ここにいう「外字」または「外字処理」とは、どのように定義された概念かというと、人により意見がまちまちであったり、場合によって意味する内容が変ったりすることが多い。

この論文は多くの有識者や実務経験者が従来から述べており、かつ実務的に実践しているところを極力広範囲に見聞し、渉猟し、各人各様にいう「外字」および「外字処理」の実態をとらえ、これを整理して、これら概念の定義、ならびに「外字処理」方式に関する顕在的、潜在的な問題点をとりあげるものである。

2. 「外字」とは何か、何が「外字」ではないのか

「外字」という言葉は、古くから漢字符号を用いた漢字通信機（俗称「漢テレ」）や漢字モノタイプ（俗称「キャスター」）を運用する実務者の間で通用してきた。

昭和20年代後半、「活字印刷電信機」という名称で、俗にいう「漢テレ」が実用化され、昭和30年代に共同通信社が全国の加盟地方新聞社に対し、「漢テレ」による新聞記事原稿の通信サービスを、また一部の新聞社は漢字モノタイプによる刷版の作成を開始した。

「漢テレ」や「キャスター」への入力には、多段シフト・フルキー方式の「漢字鍵盤さん孔機」が用いられたが、鍵盤上には直接キータッチにより6単位符号化し得る文字、ならびに記号・約物が約2,300種収容されており、それ以外の「鍵盤上にない文字」を通信する必要が生じた場合は原稿中のその文字は「伏せ字」として何らかの記号（例えば〓など）代用しておき、文末で「伏せ字」となった文字は、いかような文字であるかを「解説する」よう、共同通信社発行のマニュアル『漢テレ　ハンドブック』に記載されている[1]。

ここに「解説」とは、例えば「鳩」という字は『九の右に鳥でハト』などと文字の形状や意味内容を言葉で記述することである。

ただし、この『漢テレ　ハンドブック』のいずれの箇所を見ても「外字」という用語は存在していないことは、「外字」という用語自体が『漢テレ　ハンドブック』中においてさえ、無定義術語であることに注目したい。

次に国立国語研究所報告や文化庁、あるいは日本放送協会の出版物などに往々に見られる「表外字」という用語があるが、これは当用漢字表、人名漢字表、あるいは常用漢字表など、国の施策上、内閣告示などによって制定した「（漢字）表中に収載されていない文字」という意味で、前述の「鍵盤上にない文字」としこの「外字」の考え方に一脈通じるものであるが、ここにも「表外字」というものの厳密な定義を見出せない。

しかし実務家や著作物のいわんとする所を要約すれば、次のとおりとなる。

すなわち『何らかの施策や目的、方針のもとに、情報を表記するための文字集合を設定し通常の用字法は、その文字集合の範囲内で行なうことに規約するが、（広義の）情報処理過程中で、この文字集合に属さない文字の使用を余儀なくされたとき、その文字を「外字」と呼ぶ』ということのようである。

では、このように「外字」の使用を余儀なくされた場合、いかに対処するのかとなると情報処理の実務上では『運用上での逃げ』ともいうべき消極的な対処法しか説明されていないように見受けられる。

ここに問題なのは「外字」はあくまで、『何らかの施策や、目的、方針のもとに作られた標準的な文字集合』の存在を前提としていることである。

すなわち、「何が外字であるか？」との問い以前において、施策、目的、あるいは方針という人為的意図が働いて特定のアプリケーションで使用すべき文字の集合を限定したのであるから、その意志の存する所、すなわち「何は外字ではない」（換言すれば「これが意図的に集めた標準文字集合である」）という宣言が、あらかじめなされているはずなのである。

情報を表記する文字の種類を限定したということは、あたかもプログラム言語が限定されたコマンドやステートメントの集合からなっており、これらを含む有限個の文を組み合わせ、おびただしい種類のプログラミングとその実行が約束されているのに似ている。

目的や意図をもって使用文字種を限定した以上、それ以外の文字の使用を余儀なくされるということは、その目的や意図自体に限界があったことを示すものではなかろうか。故に「何が外字であるか」という問いは「何が外字でない」という主張によって答えられるべきであろう。[2)]

この問題は後で再び論ずることとする。

3. 日本文入力機における外字

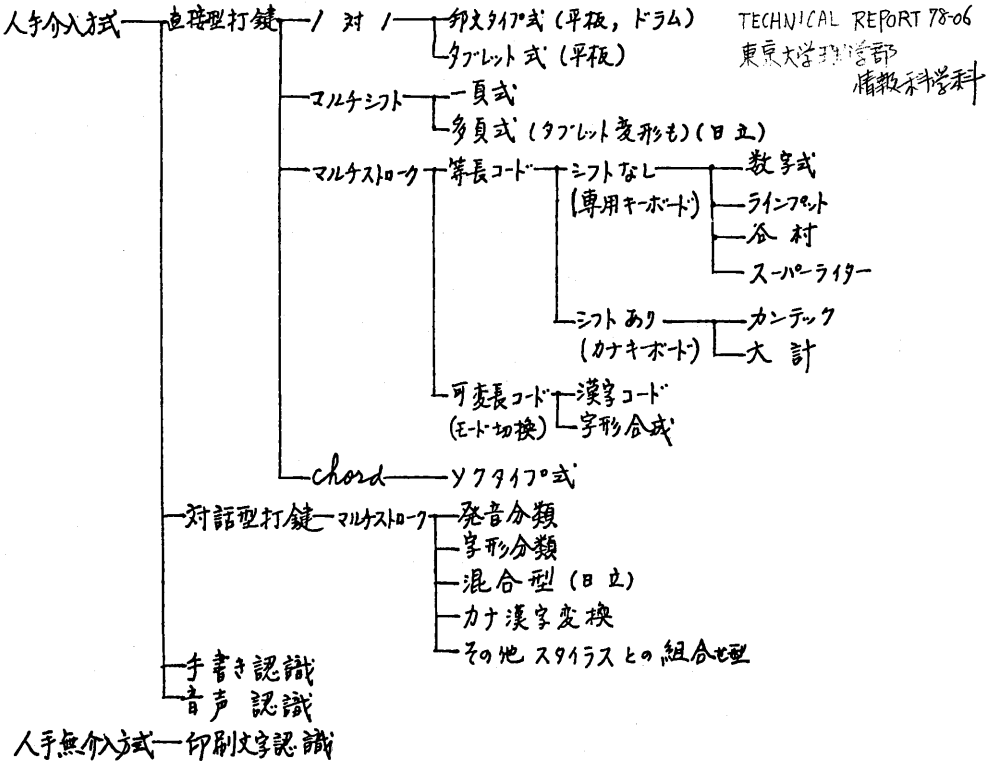
漢字仮名交りで表記されたデータやメッセージを情報処理用機器に入力するということは、データやメッセージを構成する個々の文字や記号を符号化（例えばJIS C 6226情報交換用漢字符号系など）することと見なし得る。

一般に漢字入力機、および入力方式は、おおむね図1のように分類することが山田等により提称されているが、このうち現実に実用されているものは

1. 直接型打鍵　1対1　タブレット式
2. 　同　マルチシフト　（多段シフト・フルキー）
3. 　同　マルチストローク　等長コード
4. 対話型打鍵　マルチストローク　カナ漢字変換

の4方式である。

図1　入力法分類　山田尚勇 日本語テキスト入力法の人間工学的比較



これらいずれの入力方式でも実技的には漢字を含む1文字の入力につき1回または複数回の打鍵（またはペンタッチ）を行なうものである。

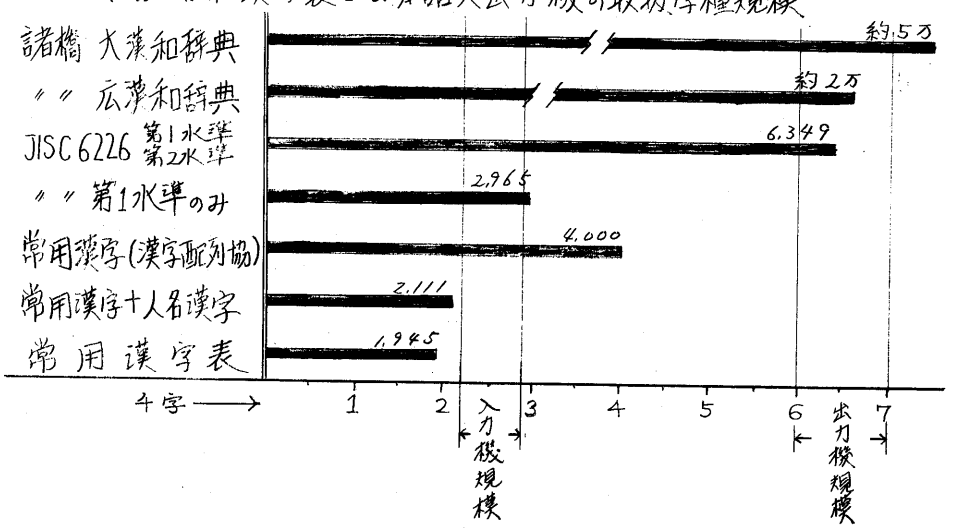
しかし、前述の3、マルチストローク法、および4、カナ漢字変換法では、カナまたはローマ字タイプライタ（44～48鍵）が用いられるが、1、タブレット（ペンタッチ）法や2、マルチシフト法では、鍵盤やタブレット盤面に直接入力可能な文字が配列されているので、入力オペレータは盤面上から必要な文字を探し出さねばならず（この字を探し出すことを「検字」という）、このため鍵盤やタブレットに収容される文字種はオペレータの入力作業能率を阻害しない範囲にまで制限しなければならない。

すなわち、あまりにも多種類にわたる文字配列の中から、所望する文字を1字だけ検字するということは、エントロピーの増大が人間の心理に影響を及ぼすためか、作業能率を阻外し、とくに鍵盤上に収容された文字種が3,000種を越えた場合、作業能率が甚だしく低下すると、経験的にいわれている。

従って、現実に存在する各社の日本語ワードプロセッサや和文タイプライタを見ても、鍵盤（またはタブレット）の収容文字種の規模は、N社の小型和文タイプライタの2,205字種が最低、O社の日本語ワードプロセッサの約3,000字種（JIS C 6226 第1水準の漢字全文字を収容）が最大であって、共同通信社コード（CO-59）入力用漢字鍵盤さん孔機の2,304字種ないし大型和文タイプライタの2,450字種程度のものが、最も多く市場に出回っている。

そこでタブレット法およびマルチシフト法による日本語入力機の文字収容規模と実際に我が国で公表されている代表的漢和辞典、または漢字表の収載字種の規模を図示すると、図2のとおりとなる。

図2. 各種漢字表と日本語入出力機の取扱字種規模



図で明らかなとおり、漢字の全数は不明であるにせよ、一応我が国で出版された最大規模の漢和辞典は、諸橋轍次「大漢和辞典」（全13巻）大修館であり、約5万字が収載されている。

また同辞典を集約した諸橋轍次他「広漢和字典」（全4巻）大修館は約2万字を収容しているという（昭和57年4月現在、第3巻以降未発行）。

しかし、周知のとおり、現代日本人の言語生活で実用される漢字はさほど多くなく、昭和56年内閣告示で示された常用漢字表収載の漢字は1945字、同じく人名

漢字に追加されたものを含め人名漢字174字（うち常用漢字と重複するもの8字）と併せ、内閣告示による漢字表の収容規模は2,111字種であり、一般に普及しているタブレットや鍵盤形式の日本語入力の漢字収容規模より小さい。

しかし、昭和20年、当用漢字表が制定された当時とは事情が変り、常用漢字表は、いわゆる漢字制限的意図はなくなり、日本語表記における漢字使用の「目安」とされているのであって、漢字使用による学術的または文芸的表現が、常用漢字表によって拘束をうけるものではない。

そこで現代日本人の言語生活で必要とする漢字集合の最低規模を一応2,000字程度と考えるならば、タブレットまたは鍵盤式入力機の漢字収容能力は、その約1～2割増程度と考えた方がよいであろう。

従って、この程度の規模の基本的な漢字集合を適当に選んで入力機に収容すれば、おおむね日本語情報処理システムが扱わねばならないであろう漢字のうち、約95%程度は1回の打鍵、またはペンタッチで直接入力可能と考えられる。

しかし逆にいえば、きわめて大量の日本文をタブレットまたは鍵盤式入力機を用いて入力した場合、5%程度の確率で「鍵盤上にない」漢字の入力処理を余儀なくされる場合が生じることを、あらかじめ覚悟しなければならない。

このことはタブレットや鍵盤式でない入力機、すなわち3、マルチストローク等長コード法においてもいえることで、例えば48種類の仮名文字2個を組み合わせて、1個の漢字のニーモニック コード（*mnemonic code*）を定義した場合、$48^2=2304$字種の漢字コードが設定できるわけである。

そしてこのコード体系とその漢字集合の規模は、前述のタブレットや鍵盤のそれと、ほぼ等しく、結果的に「コード体系にない」漢字の入力を余儀なくされる場合が想定できる。

また、4、カナ漢字変換方式の場合でも、この方式がカナ文字列で表現された語句を、漢字や漢語に置き換えるための、いわゆる「カナ漢字変換辞書」の能力によって拘束されている以上、「辞書にない漢字」の入力を余儀なくされることが考えられるであろう。

以上のように、日本文入力機や入力方式には、機械自体の機構的な制約からばかりではなく、オペレータの作業能率維持のうえからも、入力機本来の操作要領で入力可能な漢字の種類には限界があり、これを越えた文字入力を余儀なくされる場合、その文字を「外字」、またこれを入力する方法を「外字入力」と呼ぶ。

そこで、「外字入力」を必要とする場合を想定すると、次の二つが考えられる。

1、日本文入力機では入力できないが、その漢字のパターンは出力機の文字発生装置が持っている場合

2、入力機はもとより、出力機の文字発生機にもパターン登録されていない漢字を入力する場合

1、の場合は何らかの手段でその漢字固有のコード（例えば、JIS C 6226における区点番号など）を数字キーなどを利用して入力し、出力の際、強制的に所望の文字パターン記憶をアクセスする方法がとられる。

ただこの場合、問題の文字図形を認知してから、そのコードを知るまでの人間（オペレータ）の知的労働時間を最短化することが重要な課題となる。

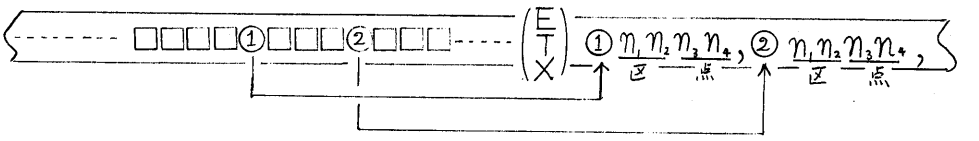
一般に、このような場合、一種の漢字辞書を引いて問題の漢字を見出し、そのコードを読みとるのであるが、漢字の図形をパターン認識して辞書を引くこと自

体がある程度の訓練を必要とするし、また技能上の個人差があるばかりでなく、漢字によって辞書から見出しやすいものと難しいものとがあるので、漢字1字あたりの検索時間を管理するには多くの困難がある。

従って、最も多く用いられている慣行では、日本文の入力の際、入力不能または検字困難な漢字に遭遇した場合、遠慮なくこれを伏せ字扱いとし、全文入力終了後あらためて伏せ字を漢字コードに翻訳する方法が、入力作業のリズムを阻害するおそれがないので、最も望ましい方法であるとされている。

2、の場合については、いわゆる「出力文字パターンにおける外字」の問題とも密接に関連し合うので、次節において詳述するものとする。

図3. 日本文入力テキスト中での外字コードの扱い方



□ 標準文字
ⓝ 外字に対する伏せ字

ETX後、ⓝで伏せ字とした外字のコードを数字キーなどで入力する

4、日本文出力過程での「外字」

日本文出力の場合、いわゆる漢字プリンタや漢字ディスプレイの文字発生装置内に、出力されるべき文字のパターン(字母)が存在しないとき、一般にその文字を「外字」といっている。

従って日本文入力の場合とは大きくその趣きが異ることに注目したい。入力の場合は、使用する入力機、または入力方式の標準的手続により入力し得ない文字が「外字」であり、「外字」のコードを何らかの別方法で強制的に入力するのが「外字処理」であった。

また標準的手続により入力可能な文字種は、入力機の機構上の規模により制約をうけるばかりでなく、検字能力という人間の知能や技術上の限界も、その制約条件になったのである。

出力の場合では、出力すべき文字のコードが与えられたにもかかわらず、出力機の文字発生装置に所要の文字パターンが記憶されていなかったなどということは、実際問題として考えられない。

およそ、ある目的で漢字出力機を用いる場合、出力させるべき文字集合は、あらかじめその目的によって定まっているわけであり、その文字集合に含まれない文字の出力を余儀なくされるならばその出力機を選定し、使用すべき文字集合を決定した、当初の機能仕様設定が不完全であったと考えざるを得ない。この問題は後節で詳述する。

現在、市販されている多くの漢字出力機では、記憶素子の大容量化と原価低減により、図2に示したとおり、約6,000～8,000種類程度の文字パターンが文字発生装置に記憶されているのが常である。

従って現代日本語の表記用漢字は数のうえからはほとんど網羅的に収容可能といってよいほどの記憶容量をもっているといってよいであろう。

そのため漢字出力機が保持する文字集合を、入力機の扱う文字集合が真部分集合として含まれるように当初から設計しておけば、入力機から標準的手続で入力された文字はもとより、「外字入力」としてそのコードを強制的に入力した文字でも出力が可能となる。

では漢字出力機が文字発生装置に記憶されていないような文字の出力を余儀なくされる場合とは、いかなる場合であろうか。次の三つのケースが想定できる。

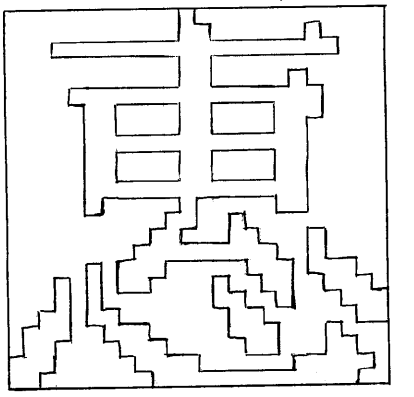
1. 地人名など、固有名詞を表記する場合
2. 古典的典籍復刻など文化財を研究あるいは管理する場合
3. 日本以外の漢字使用国民の文化と接触する場合

1. の場合は現実に戸籍業務、住民情報業務、人事関係業務などの実務上、常に直面している問題であり、2.や3.の場合は、今後情報処理対象業務の拡大により重要となってくることが予想される。

そこで具体的に上記のような事態が発生した場合、一般に漢字出力機のエンドユーザが文字パターンを作成して使用する場合、メーカまたはユーザ システムのセンターから文字パターンの供給を受ける場合が考えられる。

図4は、ある日本語ワードプロセッサにより[3]新規文字パターンを作成して登録するため、下書きとして作成した文字パターンであるが、この日本語ワードプロセッサでは、機械を文字パターン登録モードに設定すると、ディスプレイ表示面に24×24の目盛をもった枠が表示されるので、オペレータがカーソルを用いて表示された枠内に字を書くとカーソルの軌跡どおりに□印が表示され書体が作られる。ただし、実務的にはディスプレイ表示面にいきなりフリーハンドで字を書いても形の整った書体を作るのは難しいので、結局図4のような下書きを作成せざるを得ないだろう。

図4. 24×24画素文字パターン下画



また電算写植システムではDEGISETが早くから回転ドラム式走査法により、手書きでデザインした文字や模様を走査し、ランレングス法でパターンし、圧縮を行なって写植機本体へ伝送する方法を採用した。

サンケイ新聞社の新聞製作システムSUCSESでは[4]、あらかじめ約数百種類の字体素を記憶しておき、新規に文字パターンを作成するときは既存の字体素中から適当なものを選び、グラフィック ディスプレイ上で天地、左右にパターンを移動、または拡大／縮小し、字体素で漢字を合成する方式をとった。

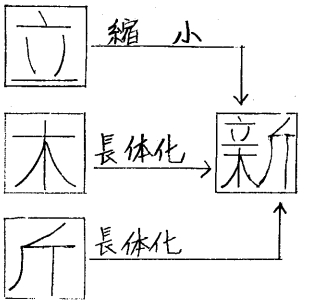
朝日新聞社もNELSONの「HCS方式」[5]により新規文字パターンの作成を行なっている。この方法はSUCSESが字体素から漢字を合成するのに対し、既存の漢字パターンから必要とする字体素を抽出し、グラフィック ディスプレイ上で適当に拡大／縮小し、新しい漢字パターンを組み立てるものである。

図5、はSUCSESおよびNELSONの両方式を原理的に図示、比較したものである。

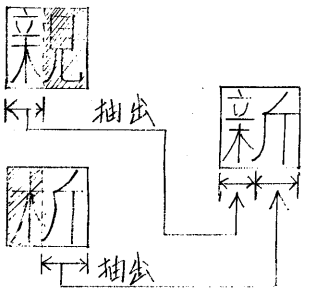
さて問題は、各種の図形入力法による新規漢字パターンの作成では、実際の作業を行なった場合、どの程度の期間と工数を要するかである。

それも、印刷や製版を本業とする新聞社や印刷業者のように、レタリング技能

図5. 文字パタン合成と組立



SUCSESによる文字合成



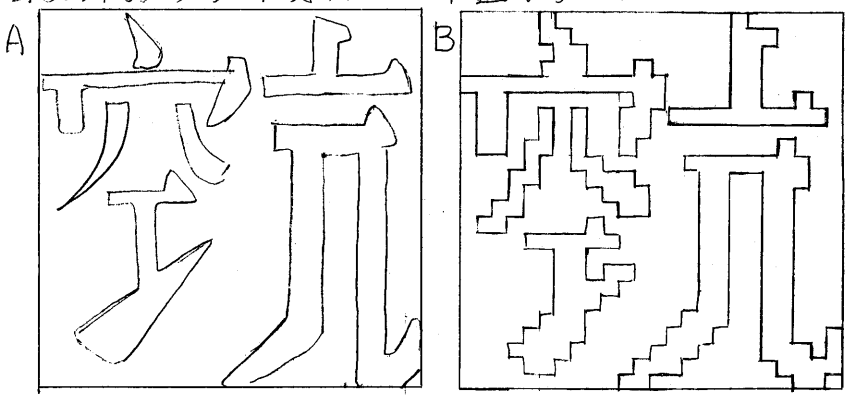
NELSONによる文字組立

者が常時勤務している職場ならば問題はないだろうが、一般の事務職員の操作する日本語ワードプロセッサや、EDP要員が管理一切をあずかっている漢字プリンタなどが設置されている職場では、新規の文字パターン作成は事実上、実務的にみて不可能ではないだろうか。

図6.は24×24 ドットパターンで作った全く新規な文字であるが、この程度のやや複雑な文字図形になると、多少「画心」のある人材でも、図6.Aのようなラフ デッサンを作るのに5分これをドットパターンに展開して下画を作るのに15分、さらに実際の日本語ワードプロセッサを用いて図形入力するのに15分を要している。

この点、NELSONの文字パターン作成のように、メーカが作成した既存の文字パターンをドット展開して、ディスプレイ上に表示し、これを手直ししたり、字体素を部分的に抽出し、組み合わせするような方式で新規文字パターンの作成ができれば、より実用的に価値のある方式となるであろう。

図6. 外字パターン作成のための下画デザイン



5. 文字パターンと文字コード体系の管理

～ アプリケーション システムとしての文字管理システム ～

さて、ここで問題となるのは、日本文入力の際の「外字」と、その入力処理、および日本文出力の際の「外字」と、文字パターン入力処理とは全く別の問題であるのに対し、従来、これらを一括して「外字」あるいは「外字処理」といわれてきたことである。

また、新規に文字パターンを作成したからといって、そのパターンを標準の文字パターンに対して、どのように管理するのか一考を要するところである。

そこで、入力に際しての「外字」と、出力の際の「外字」とを分けて考え、さらにあるアプリケーション システムが保有する文字パターンおよびそのコード体系についての考え方を整理しておかねばならないだろう。

図7.は、これらの問題を整理するために描いた概念図である。

前述のとおり、およそ漢字に限らず、一般に人類の用いている文字の種類は有限ではない。従って、ある自然言語の処理対象とするアプリケーション システムでは、その言語を表記するため、極力、必要かつ十分な種類の文字を収容した文字集合を用意し、その文字パターンと、これに対応するコード体系を用意してから言語処理を開始するのであるが、当初の方針やシステム設計上の機能仕様の枠を越えた言語処理の必要があった場合は、当然、システム完成時点では予想し得なかった新規文字の取り扱いを余儀なくされるであろう。

このように、あるアプリケーション上の目的によって、特定システムのため策定された条件により構成した文字集合の範囲内では扱うことのできない、新しい文字の取り扱いを余儀なくされた場合、その新しい文字を、そのシステムにおける、「システム レベル外字」と呼ぶことにしよう。

前節に述べたとおり、あるアプリケーション システムで新たに必要となった「システム外字」は、何らかの方法を用いて、その文字パターンを作成しなければならないし、必要とあれば、そのコードも設定されるであろう。

問題は、システムが新たに取り扱うことに決定した、新規文字概念としての「外字」と、その文字パターンおよびそのコードとを、いかに管理するかということであり、ここに日本語処理システム(広くいえば、自然言語処理システム)のサブ システムとして、「文字管理システム」が必要となるのである。

さて、あるアプリケーション システムが現在までに、たとえ1回でも使用したことのある文字ならば、それがいかに使用頻度の少ないものでも、実務上常時使用する使用頻度の高い文字と混じて用いるのは得策ではない。

実際、多くの日本語ワードプロセッサでも、JIS C 6226の第1水準の文字パターンは始業時にRAMへロードして用いるが、第2水準のそれは、必要に応じて、そのつどフロッピイ ディスクからアクセスして用いるなどの方法がとられているであろう。

そのため、あるアプリケーション システム内で用いられる入出力デバイスのレベルで、あらゆる文字パターンを保持し、そのコード体系を管理することは得策ではない。「文字管理システム」は、アプリケーション上でのCPUが分担すべきものであって、入出力デバイス レベルでは、使用頻度の高いアプリケーション システムとして標準的な常用文字を保持すべきであろう。

そこで日本語入出力デバイス レベル(たとえば周辺装置としての漢字入出力機、インテリジェント日本語ターミナル、または日本語ワードプロセッサ)で保持すべき文字の種類を限定し、その範囲を逸脱した種類の文字は、たとえばホストコンピュータや計算センター マシンなど、上位システムに管理を委ね、必要のつど、入出力デバイス レベルから見てホストコンピュータやセンター マシンに委ねた文字の範囲は、すべて入出力デバイス レベルでの「外字」ということになるわけで、これを「出力機レベル外字」と呼ぼう。

さらに、入出力デバイス　レベルで見た処理対象文字でも、前述のとおり、日本語入力機のレベルになると、機構そのものの規模よりも、オペレータの知能・技能上の限界から、能率よく入力できる文字集合の規模に限界が生じ、出力機レベルで取り扱える字種の約1/3程度しか、入力機では取り扱えない。

従って、日本語入力機のレベルから見た場合は、ホストコンピュータに管理を委ねられた文字集合はもとより、出力機レベルで取り扱うレベルの文字集合の過半数までが、「外字」となってしまうので、これらを「入力機レベル外字」と呼ぶことにしよう。

図7. 各種レベルでの「外字」概念図

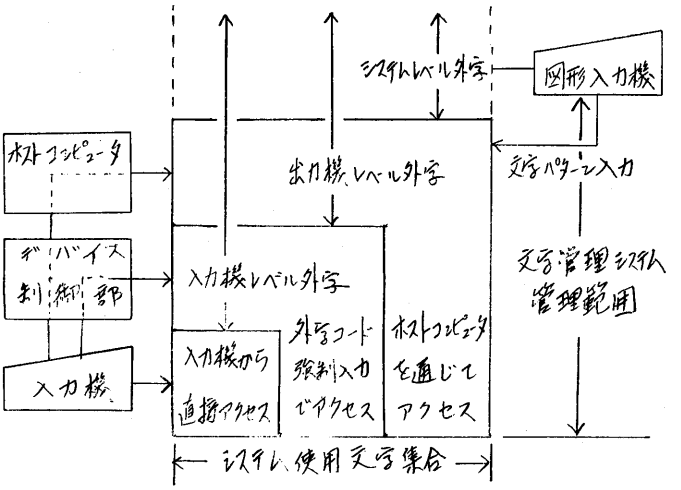


図7.で明らかなとおり、そのアプリケーション　システムで使用される、最も使用頻度の高い、高々2,500字種程度の文字集合のみが、そのシステムで採用した入力機の標準的入力方法で直接コード入力ができ、かつ出力可能な文字であり、「入力機レベル外字」は一般に、文字コードの強制入力で文字パターンをアクセスできるもののその範囲は、標準的には出力機の文字発生装置に記憶されている文字集合の範囲に限られる。

さらに「出力機レベル外字」については、ホストコンピュータからの文字パターン伝送によりバックアップされねばならない。これには様々にして図8.に示すようにテキスト中に文字パターンを記述する形態で伝送する方法が提案されている。

さて最後に、「システム　レベル外字」として新規に図形入力された文字パターンの問題であるが、これに対しては、いわば、実験用ないしは練習用入力としてそのつど「使い捨て」される文字パターンと、戸籍上の人名や文化遺産として残され、後々までの利用、活用を考慮しなければならないものがある。

後者の場合、何らかの標準的方式により文字コードが定義され、文字パターンは登録されるであろうが、ここに案外に見逃されている問題点があるので、これを列挙して今後の論議のいとぐちとしておこう。

すなわち、

1. 登録される文字の登録期限の問題

文化遺産としての文字はともかく、ただ1人だけの人名に用いられ、戸籍に登録されたが、当事者が死亡した場合、いかに処置するか。

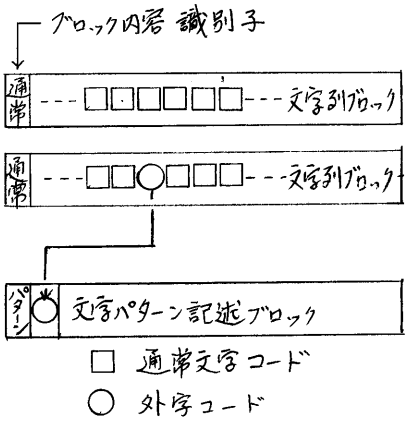
2、同一パターンの文字が重複登録されることを、いかに防止するか
かつて「システムレベル外字」として登録された文字パターンを、いかに参照するか
とくにその文字の意味や読み方が不明であったり、違った読み方がなされる場合

3、過去に登録された文字で、その出所原典が分からなくなった場合
諸橋轍次「大漢和辞典」の中に収載された漢字の中でも、出所原典、読み方、意味の不明なものが数多く存在する。

などが考えられる。

よって、前述の自然言語処理アプリケーション システムのサブ システムとして必要となる「文字管理システム」は、単に「文字パターンの倉庫」であってはならず、十分に完備された「文字データベース」でなければならないであろう。

図8. 外字パターンの伝送



6、おわりに

自然言語処理アプリケーションの中での日本語処理アプリケーションでは、きわめて数多い、不特定多数の表音・表意文字を扱わねばならない。

従って、文字の入力、出力さらに管理にあたっては、それぞれのデバイス、システムの能力や規模はもとより、それらの作業に従事する人間の知能・技能のうえからも制約をうけるであろう。

そのため、いわゆる「外字」や、「外字処理」が問題になるわけであるが、ここに述べた小論は、一貫して、日本語処理を担うアプリケーション システムは、その本来の目的に応じて、あらかじめ「システム レベル」「出力機レベル」さらに「入力機レベル」で扱うべき文字およびそのコードと文字パターンとをひとつのシステム設計の範畴において設定すべきものであることをとなえている。

そして、ひとたびそのアプリケーション システムが成立した以上、これが扱う文字集合の文字パターンおよびそのコード体系はシステムの一部であり、もしこの文字集合の範囲を越えた、「特異な新しい文字」の取り扱いが余儀なくされたとすれば、これを一種の「システム障害」であると見なすことを提案する。

すなわち、いわゆる「外字処理」を、「システム レベル」「出力機レベル」および「入力機レベル」でのシステム障害であると見たとき、これを復旧させるのが、それぞれのレベルにおける「外字処理」と見なし得るのではなかろうか。もし、この見方に妥当性が認められれば、「入力機レベル外字」を人間が漢字辞書を検索してコード入力する作業も、「出力機レベル外字」に対し、これを文字パターン入力する作業も、システム保守の問題として、各種の外字処理方式やそ

の活用技能を評価できるのではあるまいか。
「外字」の存在は、アプリケーション　システム設計の段階におけるシステム信頼性の限界であるとも考えられるのではあるまいか。

終

〔参考文献〕

1) 共同通信社　『漢テレ　ハンドブック』
2) 大倉　信治　漢字情報処理における特殊漢字　「情報処理」 Vol.16 NO.6
3) シャープビジネス 株式会社　書院 WD-3500 操作説明書
4) 長谷川実邦　パターン合成による漢字入出力処理「情報処理」Vol.16 NO.9
5) 細川　淳一　HCSによる見出しカットの作成「新聞印刷技術」1981-2. No.96