スペイン語テキスト・データとパーソナルコンピュータの使用環境

出口厚実

はじめに

自然言語のテキストがコンピュータで処理され始めて既に久しい。計算機を用いて大量のスペイン語文章資料体を検索したり集計する研究や調査も近年ますます盛んになって来ているようである。一方，扱い得るデータ量と処理速度には制限があるものの，個人の机上で比較的手軽に使用できるパーソナルコンピュータの普及により，上述のような作業に従事する一部の研究者だけでなく，一般の教師・学生もごく身近にパソコン上でスペイン語を扱う機会が急速に増加している。大阪外大イスパニア語科も1987年3月に普及型のパソコン1台と最少限度必要な周辺装置を購入し，共同研究室に設置している。

以下は，このような小規模なシステムでスペイン語言語データを扱おうとする際，いわばその入口の所で，i.e. パソコンにスペイン語を載せようとする初期段階で，コンピュータ操作経験のない筆者がぶつかった種々の問題点とその対策を経験に基いて記したものである。

1．使用環境

1.1．利用者と利用形態

パソコンの利用環境について善し悪しを論じるとき，どのような人が何の目的でどの特定ハードウェアを利用するかで千差万別の見解や評価が生じるだろう。ここでは，スペイン語語彙・形態・文法・意味など言語研究を直接の目的とする場合だけでなく，その他の人文科学・社会科学の各分野や学際領域においても，およそスペイン語文やその断片，語(句)を不可欠なデータとする人々にとって，多かれ少かれ基本的な部分であるスペイ

ン語テキストとパーソナルコンピュータの両立可能性を取り上げてみたい。

同時に，以下はコンピュータの利用経験がなく，ハードの技術的側面とＯＳ(オペレーティングシステム)を個人で改変することができない素人のユーザの立場から見た考察でもある。また，パソコン１基のみを１人が専用する，あるいは少人数が共用するスタンドアロンの利用形態においてはごく普通であると思われる，専門的な技術支援者の補助が受けられない状況を想定している。大・中型機や小型の高性能汎用機の中では，設備の供給者またはソフト開発業者による特製の環境が整えられることが多く，スペイン語に限らず特殊外国語の利用者はここで扱われるような雑多な細事に煩わされずに済むようである。勿論，パソコンに対してもスペイン語専用の統合的な環境を外注したり機器の１部を改造する可能性もあるだろうが，パーソナルな計算機という中に含められた大きな要素である“低コスト”を打ち砕くはずである。

１．２．機種

われわれの学科が導入したパソコンは，大学内で同時に購入された他の何台かと同じく，日本電気社製ＰＣ9800シリーズの１つ，ＰＣ9801ＶＸ２で，記憶装置として５インチフロッピーディスクドライブ２台を内蔵するタイプ（以下ＰＣ98と略す）である。現在のところ我が国で最も普及しており，学科の院生・学生を含む多数者の利便を始め，その他の諸条件を考慮すれば，この選定が特に不適切であったとは考えられない。

第２節以降で取り上げる不備・難点の多くはこの機種自身がもつ物理的な制約から来るが，ＰＣ98と同じくＯＳにＭＳ-ＤＯＳを採用している他の国内メーカー製のパソコンにも共通するものも少なくないと想像され，むしろ日本のパソコン上で日英語以外のデータを扱おうとする際に日常的に起り得る問題と言うべきかも知れない。[注1]　巷では異なるメーカー製のパソコン間に互換性が乏しいことに対する不満の声がくすぶっているけれども，一方で，例えばＰＣ98という単一種の装置においてすら非英外国語を満足にカバーし切れず，また異なる利用者間に互換性が保障され難い現状がある。極端な場合，同一利用者でもデータ形式を統一しておかなければ作成時期の異なる資料間に互換性が損われたり，誤処理を招来することもあり得るのである。

2．書字法とコンピュータの記号体系

2.1．実字と代字

一般のパソコンがスペイン語(及びその他の非英外国語)に対応しない，または容易にそれに適応しないという問題の中心は，前者の利用する記号体系と後者の書字(法)にズレが見られることに起因している。スペイン語の通常の正書法に用いられる各文字(図形素)の集合をここでは「実字」と呼ぶことにしよう。実字を単に表面的なフォントと見る限りでは，スペイン語の実字は英字の26文字と句読記号に(1) a，b，c，d を加えた記号セットと考えればよい。

(1)　a．アクセント付き文字
　　Á，É，Í，Ó，Ú，á，é，í，ó，ú
　b．diéresis付き文字　　Ü，ü
　c．tilde付き文字　　Ñ，ñ
　d．句読符号　　¿，¡，—(raya)

スペイン語に独特な ñ やアクセント付号付き母音字がなければ，英字以外の他の記号，例えば%，#や，2文字連続 n1，a/ などで代替する方法がすぐ思い浮ぶ。(1)の全部または一部に相当する価値をもつものとして代用された既存の記号を「代字」と呼ぶ。スペイン語テキストをパソコン上で扱う前にまず行なわなければならないのが，実字と代字の相互関係，翻字システムを確定することである。

実字／代字の関係が問題となるのは単にスクリーンに写し出される字形だけでなく，入力から内部処理・出力・保存までのパソコン処理の全体に及んでいる。

(2)

①キー入力　②画面表示　③内部処理　④記憶　⑤プリンタ出力
　実字/代字　実字/代字　実字/代字　実字/代字　実字/代字

すなわち各々の局面で二者択一と部分的／全体的実字(代字)の別に加え，それぞれの実現法に差異が生じ，それだけ多様な対処の仕方が生まれてくることがわかる。

キー入力をするときにボード上の実字を使用できれば理想的であるが，その実現は(2)の中でも最も見込みが乏しいであろう。(2)③④の演算や保

存における実字／代字は電気的な記憶形式の特徴を指すのではなく，スペイン語の書字体系におけるのと同等な価値をもつ数的順続で以て処理されるかどうかを意図している。言い換えれば実字コード体系が設定されているか，スペイン語の文字セットとしてのコードが無く，英字コード上で記号の振替えを行って処理するかの区別である。完全な実字はデータやプログラムの code が一貫してスペイン語用の特別なモードで実行されることを意味する。これは無論，望ましいことではあるが，ＰＣ98（MS-DOS）は多言語をこのように微調整して扱うように設計されていないため，[注2]ソフトによって対応するのが難しいようである。またこのモードで作成されたデータは通常の日英モードの中でそのまま使用できなくなるなど互換性に一部影響を及ぼす恐れも出てくる。とは言え，③を代字のまま放置すると，後述（§２．３）のようにスペイン語の正しいテキスト処理が不可能になるので，一部分にせよ実字をシミュレートせざるを得ない。結局，以下で具体的に考察するのは②ディスプレー表示における実字／代字の方法が主たる対象で，関連して§２．４で⑤についても少し触れるであろう〔*cf.* (3)〕。

(3)　①　②　③　④　⑤

代字　{実字／代字}　代字（＋実字）　代字　{実字／代字}

２．２．画面表示における実字・代字

２．２．１．実字方式

まず，コンピュータ側の文字表示能力とスペイン語書字との不整合の問題から取り上げよう。ＰＣ98がその画面に表示できる半角の欧文用記号は，ASCII コード表とも呼ばれるJIS C6220 符号表に含まれる a-z，A-Z の52字と若干の句読符号や記号で，これらはキーボードに刻印されている字形とほぼ一致する。そのため，正書法に則ってスペイン語を表記するには前出(1)の17記号が不足する。(1)d.の raya は形の似たハイフン(—)で間に合わせるとして，不足字リストから除いても構わないが，機能が異るのでやはり別記号と扱うのが筋であろう。

スペイン語のテキストを処理する以上はディスプレー上でもこの種の文字を表わすことができて当然で，むしろ不可欠だと考える人もいるであろ

う。ところが，これを実現するのはそれほど容易なことではない。第1に，ＮＥＣの供給するMS-DOSや，BASICを始めとする市販のＰＣ98用各種プログラミング言語は利用者がテキスト用半角文字を定義する命令を備えていない。文章中では稀にしか用いない記号，例えば#，$の代わりにáなり¿を定義して表示する，といったことができない。もしこの方式が可能ならば，キーのラベルと設定文字との対応づけに煩わしさが残るものの画面上で簡単に実字を扱うことができる。

この欠陥はＰＣ98が半角文字用のＰＣＧ（Programmable Character Generator）を持たないというハードの仕様に原因があるもので，ソフト的には克服するのが難しいらしい。ＰＣ98以前でも，他社製8 bit機のいくつかにこの機能が搭載されていたことを考えれば非常に惜しまれる設計である。どうしてもスクリーンにスペイン語の実字を見たいと望むならば，グラフィック画面をテキストの表示に転用しなければならないようである。しかし，このようなコンソール・デバイスの変更は一般ユーザには無理な操作と考えられているのであろうか，ハード・ソフト付属のマニュアル類にその具体的な実現法が示されていない。勿論，スペイン語を入力・表示でき，かつ他の付随する問題が調整された専用ソフトが別途に公開されたり販売されていれば，それに頼ればよいが，未だその種のパッケージは存在しないようである。次善の手段として，ビットマップ式のコンソール表示のできる市販ユーティリティソフトを使う手が考えられる。

Advanced Bits[注3]はそのようなツールの1つで，特に外国語文字を処理するべく設計されてはいないため，使い勝手は必ずしも良くないが，このソフトの自由フォントモード上で一応ディスプレー画面にスペイン語テキストを実字表示できることを確めることができた。パッケージにはイタリック・ゴシック・ＯＣＲ体など10種の字体フォントファイルが提供されている。ただしすべて基本英数字のセットで特殊な欧文文字はユーザ自身が付属のフォントエディタで作成しなければならない。この作業はそれほど手間がかからず，好みに合わせた字体を作り出せるが，(1)の16(または17)文字記号を既存の英数句読符号と代替する形でしか使えないのが難点である。すなわち，$，%，#などテキスト文で使用頻度の小さい記号を無効にした上で，その個数だけ新しい文字記号が利用できるようになる。

筆者は深く考えないままに試行的に(4)のような対応で読み替えを割り当

てて見た。

(4)

実字	代字	実字	代字
á	#	ü	^
é	%	Ü	/
í	,		
ó	&	ñ	~
ú	$	Ñ	—
Á	{		
É	}	¿	@
Í	<	¡	¦
Ó	=		
Ú	¥		

対応のキーを押し下げれば，スペイン語用の文字が表示されるし，またこうして作成保存されたファイルは，この補助プログラムを通して処理される限り，通常のスペイン語文のように書き出し，読むことができる。ただし，次のようないくつかの弊害を伴い，全面的に採用するには躊躇する。

(5) a) 16種類もの読み換えを一々記憶するのが面倒である。常時，スペイン語テキストのみを扱うわけではないので，普通のキー配列に慣れた指を，さらに(4)もスムーズに打鍵できるよう訓練する必要が生じる。

b) タイプライターや一部のワープロで採用されている Dead Key 方式によるアクセントと diéresis の入力ができない。

c) 市販のほとんどのアプリケーションソフトと共存できない。

d) テキスト編集に不可欠と思われる高機能エディタが利用できない。

上記4項のうち，a) b) は慣れと熟練によりある程度カバーできるであろうが，誰にでも手軽に使えないような方式はやはり好ましくないだろう。自作のスペイン語テキスト資料を多用途に開かれたデータとするためには，各種の応用ソフトやワープロ上でもそのまま使用できる柔軟性が望まれるが，c) はこの種のソフトの根本的な制約らしく，解決法が見当らない。市

販ソフトだけでなくグラフィック画面を用いる自作プログラムはすべて両立不能であるし，N88-日本語Basic（MS-DOS版も含めて）などのプログラミング言語も使えない。c) も d) も同一の原因から来るが，テキストデータを作成する道具としてユーザの好みに合った効率のよいエディタが選べないのは大きなデメリットと言える。MS-DOSシステム付属のエディタであるEDLINは使用可能とのことであるが，これは長文の言語データの編集にはおよそ不向きなラインエディタなので論外である。以上の難点のどれか1つでも我慢できない人はグラフィック画面利用のデバイスに頼る方式を諦めなければならないだろう。

しかし，使用環境が多少劣化し，操作性が落ちても，画面に実字表示がきちんとできる方を優先するという人々には(5)をいくらか改善する余地は残されている。高速スクロールや多彩なマクロ機能が売りもののエディタには及ばなくとも，一通りのスクリーンエディットが可能なエディタが利用できるからである。米国Mix社製の Split Screen Editor は Advanced Bits の自由フォントモードと両立する数少ないエディタの一つである。このソフトはIBM PC用にすぐ使える状態で出荷されているが，インストールをやり直すことでPC98のMS-DOS下で何とか使用可能になる。Split Screen Editor ではキーの割り当てがかなり自由に変更できるようになっていて，2ストローク＝1文字にも対応しているため，Dead Key方式の入力に簡単に切り換えることができる。例えば@あるいは〔の後に続けて母音を打つことにより強勢母音を入力・表示することもできる。また，diéresis をアクセント・キーの Shift に設定しておけば(4)のうち12文字の入力は非常に楽になり，残る4字のキー読み換えのみを記憶すればよく，a) b) はほぼ解決される。しかし(5) d) を埋め合わすには力不足である。編集可能なテキスト量にも問題はなく，同時編集・マクロ・スクリーン分割など一通りの機能は備わっているが，スクロールの遅さを特に気にするユーザには向かない。その反面，このエディタは機種に依存しないという利点を持っている。各社のパソコンのMS-DOS（PC-DOS）上で走らせることができるため，[注4] グループでの共同作業を統一基準で行なう際などには便利であろう。

2．2．2．代字表示

ビットマップ式による特殊字母使用については，入力と画面表示に限って見ても，少なからぬ困難が伴うことをみた。特に(5) c) d) を避けようとすれば，実字表示を諦めざるを得ないのが実状である。そこで á や ñ などの字形をそのまま表示しないで別の文字で代替表示する代字方式の可能性と効率を検討してみよう。既存の文字・記号から16字を選んで(1)の追加字母に充当する翻字案が満すべき要件として，恐らく(6)の各点があげられる。

(6) a) 表示(プリント)される図形記号から原テキストの書字が容易に復元し読解できること。
b) 代替のために使用禁止となる記号数がなるべく少ないこと
c) 翻字の対応に規則性があり，また記憶が容易なこと
d) キーボードからの入力が容易なこと

上記の各項に対する優先順位の与え方により，多数の解決案が可能であろう。最終的には各個人が最も使いやすいと思うものを採ればよいのであるが，筆者は次のような文字記号の割り当てを行なって試用している。

(7) 1) アクセント付き母音は当該母音字の後に ′ を続け 2 文字とする。
2) diéresis 付きの母音は当該母音字の後に ˆ を続けて 2 文字とする。
3) tilde 付きの N, n はそれぞれの後に～を続けて 2 文字とする。
4) ¿ は@で，¡ は¦でそれぞれ代替する。

以下，この代字システムをＳＴＸ方式と呼ぶことにする。強勢母音10文字を複文字に分解するやり方は(4)に較べるとずっと憶えやすく，予備知識がなくても誰でも瞬時に正書法へ翻字できるので，(6) a) の基準からは最善の解決法と思われる。また b) c) の面からも好都合である。もっとも，ＳＴＸは b) に関してはベストな方法ではない。例えば，Ñ, ñ に使用した tilde に対し母音後でアクセントの価値を与えるならば，シングルクォートを犠牲にしなくて済ませる。しかし，(6) a) c) の方を重視するならば，1 図形素を 2 価的または多価的に利用するのは好ましくないとも考えられる。シングルクォート(')はキーボード上では数字キー [7] の Shift に配

置され，指の休止位置から遠く，打ちづらいだけでなく，Shift Key を同時に押し下げなければならないというマイナス面がある。しかし，これは起動時に ' のキーを Shift の不要な打ちやすい位置，例えば / に入れ換えるよう予めキー・コマンドで設定できるので大した支障にならない。母音から独立して1字となる diéresis に字形上もっとも似ているのはダブルクォートであるが，これは引用句や文字列の区切りの標識になることがあり，むしろ circunflejo などの他記号に代替する方が無難であろう。また Ñ, ñ も個別の特定字に置き換えるよりは複文字化した方がわかりやすい。幸いキーボードには～が備わっているからそのまま流用すればよい。逆疑問符・逆感嘆符はそれぞれ1文字で(7) 4) のように対応してみたが，～と ¿ と比べると前者の方が頻度が高いにもかかわらず Shift Key 併用になるので，これも両者が逆になるよう予め Key 定義で変更しておけば，入力の手間をいくらか省くことができる。2文字→1文字の対応を含む(7)の翻字案は(6)の諸条件を満す点で，1字毎の読み換えを表現する代字体系，例えば(4)をそのまま流用するよりもかなり能率的ではないかと思われる。ただし後述のように，前節のビットマップ式や完全1文字翻字方式には起こらない難点を伴うことも見逃せない。[注5]

なお，ＳＴＸのような翻字を採用する際に，Dead Key 方式が利用できればさらに便利と考える人も多いことであろう。これはシステムプログラムを一部修正することで可能なはずだと思われるが，素人ユーザのわれわれにすぐできるのは，前述のMS-DOSの Key 定義と市販エディタのキーマクロを組み合わせる方法である。ある特定の Dead Key をまず押し，次に a を打てば a '（＝á）と等しくみなされるよう設定し直すことは可能である。ただ，これはワープロ・タイプライタで Dead Key のタッチに慣れている人にのみ有効で，その場合でも入力ストローク数の節約にはつながらないので，大幅にスピードを改善するのには役立たないかも知れない。

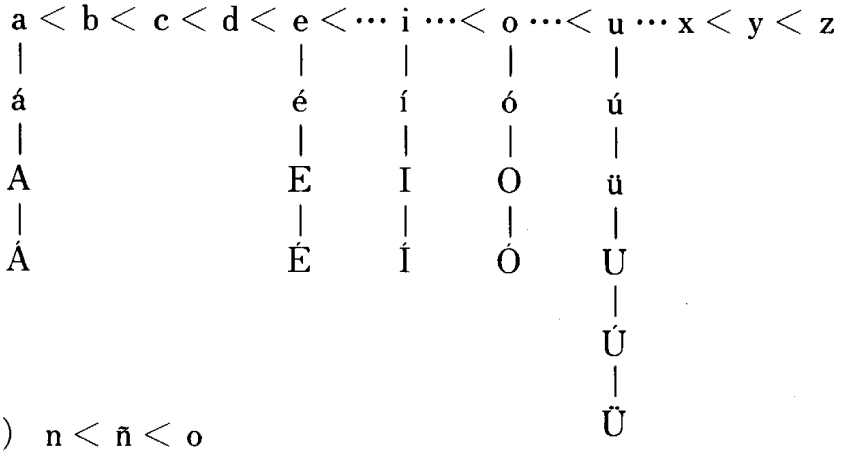
2.3. 内部処理における実字と代字

スペイン語の入力と画面表示に対処できたと仮定しても，なお解決しなければならない課題がある。一つは，テキスト処理の主要目的である形態素・単語・段落・文 etc. の単位を扱おうとするとき避けて通れない問題

で，コード内の記号配列順位を再編成する処理が中心となる。

一般にスペイン語のアルファベット順整列(辞書順)は次の大小原則に則って行われるのが慣習となっているようである。

(8) 1) 英字と同じ26種の単字(小・大文字)



2) n < ñ < o

N Ñ O

3) 複文字

c < ch < d　　l < ll < m

C CH D　　L LL M

N.B. i) 字母表では複文字とされる rr は整列上で単文字 r の2ヶ連続とみなされる。

ii) 同一字の大文字と小文字，区別符号のない母音字とそれの付いた母音字は同順位であるが，文字列の比較では，同一位置において1)～3)で上位にある字が小さいとされる。

実字方式・代字方式いずれを採るにせよ，通常のＪＩＳコード順と(8)の間にはかなり大幅なずれが存在し，ソフトウェアのどこかで各記号の順位を組み換えないと，文字列が正しくソートされなくなる。コードの変更手続きがし易いか否かの観点も，実は，(4)(6)での具体的な各字のコード割当てにおいて考慮に入れるべき要素であったのである。コンピュータ処理の全段階でコード体系をスペイン語に合致するようにMS-DOSシステムの基本部分を書き換えることが可能かどうか知らない。仮にそれが実現されたとしても，例えば，a → á → A → Á → b →…… のような一次的順

序づけをしたコードのみでは，現行のスペイン語辞書順序を自動的に生成しないことは(8)ii）の存在によって明らかである。つまり，ソート関数の呼び出しに使う，何らかのスペイン語専用の文字列比較関数を独自に用意する必要性があるということである。それならば，テキストの作成・表示やプログラムの全過程を通じて独自なコード系を持たなくても文字(列)の大小比較の場合にだけ(8)の基準が満たされれば一応の目的は達せられるであろう。このような関数処理はスピード低下という弊害を否めないが，最も簡便な対処法のように思われるので，筆者はそのような比較関数を自作してソートすることにしている。

もう一つ留意しなければならないのは，検索時に前提となる文・単語・形態素などの定義法である。文字列の同一性のみを判断するのであれば，特に支障はないが，形態素・単語など単位切り出しを含む手続きでは，標準アルファベットを変更した結果を，各単位の境界条件に反映させなければならない。前述のＳＴＸ形式を例にとれば，このテキストを走査するとき，語中に ’ ^ ～ が含まれていてもこれらを英字と同様に扱って「語」を切り出すよう設定されなければならない。また(4)に類する代字変換を行ったケースだと ＃，%，&,… など14種の非アルファベットを英字とみなして単語を再定義する必要が生じる。文切り出しの文末判断は一般の処理と異らないものの，文頭で raya，¿，¡ に代替されている記号が正しく文内先頭と解釈されるよう微調整されなければならないだろう。

2．4．プリンタにおける実字と代字

コンソールでの実字表示の難しさに比べて，プリンタ上の非標準文字の扱いは早い時期から考慮され，任意の半角ダウンロード文字が使用可能になっていた。前述の Advanced Bits でもユーザ定義の独自文字がそのまま印刷文字として展開できるよう配慮されている。一方，代字方式の場合，2文字分解を含むＳＴＸ型式のテキストはそのまま印刷されても可読性がかなり高いものの，プリンタ用のフォントを定義しそのためのプログラムを作成する労を厭わなければ，やはり画面上で a ’，n ～ であるものを á，ñ と書き出した方が見やすいだろう。各ユーザが用いるプリンタの仕様に合わせてプリンタ・ドライバを作り，差し替えるのが本筋だがこれは簡単にできない。そこで次善の策として，代替記号のコードをプリンタに送

って定義文字を印字するダウンロードモードを利用することになろう。すなわち，p. ej. 外字 Á の代わりに { を使用するとすれば，プリンタ側に Á のドットイメージを所定の方法で { のコード番号に登録しておくと，実際に { のコードが送られたときに，Á が印刷される。

もっとも，たいていプリンタは国際文字をサポートしているので，ユーザがスペイン語の特殊文字をすべて登録する必要はないようである。エプソン社製の普及型プリンタ（v. gr. VP135K）でも「スペインⅠ」「スペインⅡ」「ラテンアメリカ」の3種の国際文字セットがそれぞれ12字ずつ用意されている。このうち「ラテンアメリカ」を使用すれば，小文字のアクセント付き母音，ü，¡，¿ がプリントアウトできるが，なぜか Á など大文字の区別符号付き母音は欠落している。プリンタの中にはＩＢＭ拡張グラフィックスモードに対応するものもあるが，この時でもこれらの特殊大文字が使用不能である。用紙への印字を単に補助的手段とみなし実用的な範囲の精度しか要求しないのであれば，上述の文字セットへコンピュータの出力を合致させるだけで十分である。この調整は MS-DOS のフィルタ・プログラムとして準備しておけば，特にスペイン語字母への印字が必要な時のみ，臨機にフィルタを通過させるという便法がとれるだろう。

テキスト画面内でＳＴＸのように1対2文字の字数増加の対応をした場合，スペイン語用フォントを用いて印刷出力しないと1行の桁数が原テキストと合わなくなる点は注意しなければならない。オリジナルでは適正な改行やハイフォネーションを忠実に再現してテキスト化すると行末が不揃いになってしまうことがある。データが単語・語句や数値から成る表形式のとき，代字表示はさらに厄介である。特に字数が増え設定セル幅を超える可能性が予想されれば，予め代字の枠にその余裕を設けておくしか方法がないだろう。逆にプリントの際に代字から実字表示へ変換するケースでは，桁の不整列が起らないよう各語(句)の区切り毎に実質字数に入らない識別記号の数だけスペースを挿入する処置が必要となる。

3．おわりに

普通のパソコンでスペイン語テキストをどのようにしたら扱えるかというテーマを個人的な体験を中心にまとめて見た。実際の作業の中で気づいた細かな問題や改善の余地など触れられなかった点を未だ多く残している。

また“不可能である”とか“難しい”と述べた個所は，ただ筆者の知識不足によりそう思い込んだだけで，実は案外簡単な対応策があるのかも知れない。御教示いただければ幸いである。拙文を書くに当っては多くの方々から直接・間接に各自の使用経験やパソコンのハード・ソフトに関する情報を得ることができ，それらを利用させていただいたことも付記しておきたい。

(1988年12月5日)

〔注〕

1．外国製及び外国系の機種においてはほぼ完全にあるいは大部分克服されていると言われている。それらを使用した経験がないので確認していない。

2．新しい (MS-) OS/2 ではコードページの管理により入出力装置の文字セットを切り換えられるようになったそうであるが，未見である。

3．現在 Version 5.0 が入手可能で，オンメモリ・エディタや電卓機能，ＤＯＳの拡張などが付属している。発売元イシガキM.E.S。

4．全く同じ操作性をもつ CP/M80 版も発売されているから，各メーカーの 8 bit 機上でも同等に使え，ファイル変換も可能である。

5．編集時には代字，最終の表示・出力に対しては実字を利用する折衷案も考えられる。ＳＴＸでテキストを作成し完成後に，一括して文字置換を行ない(4)の対応関係に変換する方法である。