マイクロコンピュータ　65－5
とワークステーション
（1991. 2. 1）

日本語電子文書処理の課題

安達　淳

学術情報センター研究開発部

ワープロやパソコンを利用した日本語文書作成におけるいろいろな問題を、標準化の観点から整理し、問題点をあげた。さらに、これからの発展の方向を踏まえての問題点への対処についての考え方を示した。重要な点としては、文字コード、文字のサイズやフォント、プリンタの規格、SGMLなどについて論じている。またより広い観点からの問題提起として、教育、リテラシー、標準の普及などについても言及した。

The issues on electronic processing of Japanese language documnets

Jun ADACHI

National Center for Science Information Systems

This paper surveys various problems related to Japanese language document preparation using word-processors and PCs from the viewpoint of standardization. Then, the future direction and the approaches to clear up the above problems are discussed. The major issues discussed here are concerned with character code sets, character sizes and fonts, the standards of printers and SGML. The related issues such as education, computer literacy and enforcement of standards are also mentioned and some approaches are proposed on those topics.

1. はじめに

本稿は、(財) 日本規格協会のもとに組織された「マルチメディア標準化調査研究委員会」の中に設けられた電子文書処理に関する分科会でのおよそ2年に渡る検討結果に基づいて、報告するものである。

この分科会では、現在および近い将来の日本語文書処理を考えた場合の様々な課題を網羅的に検討するという役割を担って、作業を開始した。このテーマは、きわめて広い話題を含んでおり、しかも多様な意見が出てくると考えられた。過去10年の急速なワープロの普及の中で典型的にみられるように、標準化や互換性の点でいくつかの解決すべき問題があることは事実である。また、「電子文書」という言葉がカバーする範囲についてであるが、狭く考えれば、いわゆるワープロやパソコンで行われている文書処理が該当すると思われる。一方、現在のこれらの機器の使い方を将来に外挿すると、当然電子出版、マルチメディア電子文書までを含めて検討する必要もあろう。従って、キーワードだけを列挙すると、通信、電子出版、卓上出版、用語用字、正書法などさまざまなものも含めて検討する方針とした。本稿では、パーソナルな電子文書処理を検討範囲として絞り、主に標準化の観点からの検討結果について、個人的な意見を含めて取りまとめたものである。従って、本稿の内容は前述の組織の検討結果と必ずしも一致するものではないことをまずお断りしておきたい。

2. 文書の標準化の範囲と問題

2.1 はじめに

一般に文書の作成に限らず標準化に関することを考えるには、規則や規範が存在することがが問題となろう。言い換えれば、まったく規則がなく、原則が自由であるなら、それを標準化しようと言うようなことは意味がないか、あるいは不可能に近いことであろう。英語の文書作成については、ある程度の規範があり、例えば有名なものとしてシカゴ大学の "The Chicago Manual of Style" が挙げられ、著者や編集者への参考書として定評がある。我国でも文書作成に関わる各種の本が市販されているが、機械を使って文書を作成する歴史がきわめて短いこともあり、まだ安定的な教科書のようなものはないと思われる。

2.2 用語と用字

漢字の使用範囲　日本語においては現在でも漢字制限についての議論が定まっていないことにより、漢字の使用範囲については社会の各部で大きなゆれがある。普通の文書作成については概ねJIS X0208と言う範囲で線が引かれているが、現実には各種の問題をはらんでいる。特に、人名については自由度が大きいことにより、なかなか統制が利かず、コンピュータ利用についても面倒な点が多い。また地名などの固有名詞も同様な性格を持っている。古典文学などにおいても漢字の範囲を限定してしまうのには本質的な難しさがある。

表記のゆれ　現在と今後に作られる文書に問題を限定しても、送り仮名や専門用語の表記についてはゆれがある。(例えば、「超電導」と「超伝導」) また、外来語のカタカナ表記について非常にゆれがあることも問題になろう。これらは、今後文書がそのままデータ

ベースとして蓄積されて行くことが予想されるので、情報を的確に検索する際に問題になると心配される。但し、日本語のカナによる表記は、外来語をのぞき、ほぼ安定していると思われる。(「ず」と「づ」などの使い分けにはゆれがある。)従って、現在的にはデータベースの問題はかなによってとりあえず回避する事ができているようである。さらに表記としては、「こと」と「事」のようにかなで書き下すかどうかに著者によるゆれがおおきい。

他国語の扱い　現在のJISでは、英字の他キリルとギリシャ文字が規格化されている。今後は、各国での規格化を次第に取り入れて行く方向であろうが、特に東アジアの漢字文化圏の文字の取扱については注意が必要である。例えば、中国の簡体字と日本語の漢字の扱いの考え方などである。

記号　JISには他国の文字標準と比べて多くの記号類が取り込まれている。やくものと呼ばれるものは、文書中で使われる記号類のことを指すと考えられるが、例えば数学記号等は文字と同列には取扱いにくい面もあり、標準作成にあたっては、文書作成における記号の考え方を再整理する必要があろう。また、丸で囲った数字のような対処しにくい(どこで制限するのか判断しにくい)ものも数多く記号として扱われていることにも注意すべきである。

2.3 文書の構造

字体とポイント数　印刷本来の考え方でよく使われるのは、明朝体とゴシック体といった字体とその大きさであるポイント数であろうが、ワープロでは、発達過程での技術的制約もあり、「半角、全角、倍角」とか「太文字による強調」という技法が用いられてきた。また、「網掛け」といった装飾が多用されること、また英文に比べて「罫線」を使うことが多いことから、結果として文書ファイルの互換性を維持することが難しいという結果となっている。

今後のひとつの方向としては、卓上出版のような技術的な進歩にも助けられて、より印刷的な技法を用いる形に収束して行くことが考えられる。例えば、英字を書くのに半角を用いるか全角を用いるかはまったく規範がなく、著者の審美的なその時々の判断によるものと思われる。これは今後、書体とそれに応じた字幅によって印刷的な観点からバランスのとれた形に版組されるという方向に行くのではないだろうか。また、和文中の英文の扱いには、ポイントサイズのバランスやハイフネーションの無視などの問題があるが、これもソフトウェアの高度化で期待を持てる課題である。

禁則処理　ワープロの出現以来広く意識されるようになって来たものに禁則処理がある。行頭、行末の禁則、改行禁止、改頁禁止、分離禁止などに分けられるが、これらの規則については、ゆれがあり、基本的には見た目の体裁の問題であるから、それほどはっきりとした規則として定着しているわけではなさそうである。禁則文字の設定とそれに伴う規則の標準がうまく設定できるかどうかについては定かではない。なお、これはプリンタの機能的な能力とも関係していると考えられる。

罫線　ワープロでは多種の罫線が用意されている点も英文の場合と異なる際だった特徴であろう。特に、商用では必要となる場合が多いと思われるが、実際の印刷物では図面やイラストを除きそれほど多くの種類は用いられていないのではないだろうか。

2.4 組版の要素

横組と縦組、段組　単に横のものを縦にするだけでは済まないのは言を待たない。例えば、美観上からは字体も縦と横では異なるということである。また、禁則の判断の基準も異なってくる。段組では、左右のバランス、マージンのとり方などが問題になる。

注釈　脚注が使われることが多いが、複雑なものとして割注、縦注がある。使い方の規則については一定していない。

ルビ　日本語に特徴的なものである。ポイントの下げ方、位置関係に注意する必要がある。

見出し　章、節などのインデクスの付け方、フォント、ポイントの選び方などについては、文章構造として応用別に一定の基準を設定できるであろう。また小見出しのようなものについても同様であろう。しかし、箇条書にする場合の見出し記号の使い方などについては、一定していない。また、右、左、中央揃えや均等割り付けについても留意する必要がある。

インデント　節が細かくなる場合や引用が割り込む場合にどの程度インデントするかについては、印刷の体裁上ある程度基準があるように見受けられるが、ワープロ作成文書ではあまり意識されず、すべて左つめが多い。専ら箇条書の場合に使われている。

段落　改段での一字下げは広く守られている規則である。しかし、改段時に行間をわずかに広く取るといったことについては、定まった習慣はないようである。

末端部の処理　改段、改頁の際にわずかにはみ出した数文字を前の部分に押し込めるようなことも考えられる。日本語文書中では基本的に等間隔に文字が並ぶので例外的な処理であろうが、ページ全体の空白の量による美観とも関係してくるため、考慮すべき課題であろう。

2.5 日本語入力

文字入力鍵盤　コンピュータへの入力については、メーカ各社から多種多様の方式が出回っていて、その互換性も薄く異機種を使いこなすのは困難な状況にある。現在、ASCII、JIS、新JIS、あいうえお配列、親指シフト、M式、トロン式、などが使われている。

入力方式　ローマ字入力では、“ん”の処置、促音、拗音の処理、撥音の後に母音を入れるときの処理、“てぃ”などの処理、などが統一されていない。キー操作では、シフト方法、機能キーの種類、コマンド類、操作方法など。

仮名漢字変換方式　辞書の保守、ユーザ辞書の互換性、思った通りに変換できなかった部分の指定の仕方、違った字種で入力したときのミスの修正の仕方、など。

2.6 関連する技術分野

ファイル交換　極めて必要性の高い問題で、まずはフォーマット標準を設ける必要がある。第1レベルとして、文字情報だけの交換、第2レベルとして編集指定、画像情報を含めての交換をするための形式の標準化が必要である。

情報検索　自然言語で書かれた文字情報のデータベースの検索の問題としては、文字コード体系では全角と半角の相違、異体字(国と國、口、圀など)がある。また表記の揺れ

としては、送り仮名(行う、行なう)、使用漢字の制限による揺れ(沈殿、沈澱、沈でん)などの例があげられる。

媒体の標準　小型化、大容量化が急速に進んでいる中で、積極的に標準化を勧める必要があろう。特にこの分野では日本が世界の最先端を走っており、比較的標準化がし易い分野でもあろう。また、標準化が遅れると利用者が被害をこうむることになるおそれがある。媒体としては、情報交換用DATのテープ、フロッピーディスク(2HDと呼ばれているものまでは、ほぼ標準が揃っているようであるが、より大きな容量のものについての標準化が遅れているようである)、ICカード、CD-ROM、CD-I等。

通信　標準化についての意識の高い分野である。プロトコル(異機種間のデータ交換)の標準化として、OSIが大前提となるが、実用上は全く貢献していないことは極めて残念である。

パーソナル利用では無手順の世界となり、いわゆる標準化とは無縁の混乱の状況にある。これとファイル形式、文字コードの混乱と合い待って、素人には極めて理解し難い状況にある。

3. 標準化の方向性

3.1 文書の交換性

全節の基本的な視点は、日本語において文書一般についての規範のようなものが曖昧であるため、標準化の作業では標準そのものの以前に規範についての検討が必要となる結果、すべてが困難な作業になりまたせっかく作った標準があまり使われない、というように見受けられる。

欧米では、タイプライターという機能の限定された機械による文書作成の長い伝統があるため、強い制限のもとでの文書作成の受容度が高いといえる。これのアナロジーで日本語の文書交換を考えると、記号や罫線を多用したワープロ文書の形での交換を行うことにもともと無理がある、という考え方も出てこよう。しかしこの場合でも、問題として残るのは文字コードである。

文字コードおよびセット

人名用の漢字に象徴的に表れる日本語の文字種の多様さは、閉じることのない開集合であり、それへの強い固執がある。標準化は一般に制限的に働くので、これは容易に解決し難い問題である。多様性を前提に考えると、電子文書処理では、これまでの情報交換用漢字符号系(X0208)で規定する図形文字の集合とその符号だけでは不十分であるのはいたしかたない。

文字コードには、1バイト系のX0202とX0208に加え、昨年JIS X0212(情報交換用漢字符合—補助漢字)が定められた。文字コードについては、現在のX0208の利用状況をどのように見るかにもよるが、基本的にX0202とX0208を基本文字セットとして使用し、それ以外は用途別に拡張された個別のセットを使用するという形態をまず考えることが出来る。大きく見ればシフトJISや企業毎の漢字コードもJISにならって定められているものが多いので、この考え方は受け入れ易いのではないだろうか。

基本的なコード以外に定められたいろいろのコード表の中で、将来的にどれが採用されていくかは市場の需要動向で定まっていくと考えられる。

文書の交換という観点から見た現在の問題点を列挙すると、(1) 仮名・英数字コードのセットに1バイトコードと2バイトコードがあるため混乱のもとになっている、(2) 漢字コードについては、1978年版と1983年版での変更の混乱がまだ残っている、(3) ローマ数字、丸付き数字、絵記号、業界特有の記号、約物(記号類や特殊記号)の使用についての考え方の調整、(4) 中途半端な罫線文字の規定、などがある。これらの諸点についての推奨方式を定めていけば、少なくとも話しをするための共通の基盤はできるわけで、プレインなテキストや通信による最低限の文書交換には問題がなくなると期待される。

文字コードに関して、さらに残された問題を挙げると次のようになる。

- ハングルなど他国語の文字セットへの対応や考え方。
- 外字(コード表にない字)、異体字(同一文字で書体の違うもの)の考え方に統一性を与えるためのシソーラスの必要性。
- 記号についての基本的な考え方の調整。
- やくものや記号の呼び方の調整。(〒や々をどのように呼ぶか)

3.2 ワープロに関係する問題

ワープロという短期間に急激に発展した装置による日本語の影響についてははかり知れないものがある。コードないし文字種については全節に述べたが、文字サイズ、書体(フォント)、プリンタなどの面において今後どのように発展していくかを見定め、標準化の方向を検討する必要がある。

文字サイズ　ポイントサイズ、倍角、全角、半角指定、縦横比指定、ルビ、画数けずりなど多くの寸法の考え方があり、文書処理上これらの処理基準を調整する必要がある。主としてプリンタの機能的制約から出てきたと考えられる半角などの考え方はすたれ、全体としての動向は卓上出版へ向かうとするならば、ポイントサイズによる考え方が組版との整合性もあり、最も自然であると考えられる。

書体(フォント)　ワープロでも、明朝体、ゴシック体、毛筆体など多種のものが使われている。表示装置用、ドットプリンタ用字形としてX9051、2があるが、文書処理上は写植機メーカの持つ各種字形や、文字サイズによる字形の変化もあり単なる拡大、縮小では対処できない。フォントの問題は、縦組と横組でもちがい芸術な側面もある。読み易い漢字書体を作るには縦長、偏平、正方形とか縦線、横線の太さの割合だけの問題ではなく、一画、一点に努力が込められていて、ここにフォントの著作権が発生する。今後は、目的別の各種フォントの体系的標準化が望まれるとともに、利用者により使い易い環境の確立が必要であろう。

プリンタの規格　現在のワープロやパーソナルコンピュータに接続するドットプリンタの制御コードの混乱については、まったく手のつけようがないと言ってもおかしくない。現状の低能力のドットプリンタを対象に標準化を改めて強力に押し進めても、すでに利用している消費者等に対して益となることは少ないと考えられる。

一方、今後に広く普及していくと考えられるのは、LBP(レーザプリンタ)であろうし、簡易型のプリンタもLBPの機能的影響のもとに発展していくことが予想される。これらプ

リンタの機能については、文字サイズの指定(ポイント、倍)、罫線の扱い、さらにはPDL (Page Description Language) の考え方等、使用目的、用途によって異なるものの、明確な規格等はまだ十分とはいえない。出力部に関して今後規格化すべき要素としては、記録形式や記録材料に関わる問題としてレーザ方式、フィルム、普通紙、写真植字、ドットとベクトル方式、インパクト方式などなど整理すべき問題がある。又出力精度についても、単位系(ミリ、インチ)の整理や、プリント、一般文書用、校正ゲラ用、印刷版下用など用途に応じた出力画線密度の整理、さらに新聞用、高級印刷、凸版用など印刷版下用高精細度出力装置の整理などが必要不可欠であろう。

3.3 情報処理上の課題

電子化されたより高度な文書の交換性が期待されるのは、例えばデータベース化した場合の有効性、可用性が期待されるからであろう。そのためには、以上に述べた文字コードレベルでの交換性のみならず、図や表を含む文書レベルでの交換性が必要となり、現状では多様なアプローチがしのぎを削っている状態で、なかなか将来動向を見定めにくい。この中で最近注目されるのはSGMLを中心とするアプローチである。

SGML　SGML (Standard Generalized Mark-up Language; 標準一般マーク付け言語) とは、文書データの記述仕様の国際標準規格として、1986年にISOにおいて定められた(ISO8879; JCT SC18/WG8)。既にECの多言語による官報の出版、米国国防省での文書作成などで使われ始めている。これは元々印刷・出版関係の規格として考案されたもので、AAP(米国出版協会)が積極的に関与している。

SGMLは、ある特定の目的に合致した文書の構造を定め、その機械可読テキストを全文データベースとして利用できるような形とすることを主たる目的としている。

すなわち、与えられた文書の仕様に沿って「標題」、「著者」、「章立て」など、文書のとるべき構造を任意に定義できるような文法記述言語にSGMLは相当する。その仕様定義のことをDTD(Document Type Definition)と言っている。また、文書の中にDTDに従って、例えば「表題」を指示するようなコマンドを挿入しつつ、文書を書いていくことになるが、このようなコマンドを“タグ”と呼んでいる。そこで、SGMLを使うと、編集者は今まで赤鉛筆で書き込んでいた各種の指示の代わりに、このタグを打ち込んでいくことになるが、これを“タグ付け”と言っている。

一般的に「文書」といってもその汎用構造は定めにくくなるが、例えば商用文、学術論文などと範囲を限定すると、日本語文書にも一定の形式があり、規格化し得る要素となると考えられる。現実には理工学系の学会では学術論文の体裁をかなり厳しく規定しているところもある。

SGMLの利点は、使用する文字コードの範囲を含めて個々の文書タイプ毎に指定するということが必要なため、自然に文字や記号の使用制限を取り込まざるを得ないという点である。現在の日本語化の作業でも、禁則処理の明確化など本稿前半で指摘したいくつかの項目の規格化を内包している。

もう一つの特徴は、最終的な版組出力を記述するためのものとしてSPDL (Standard Page Description Language) と連結しており、これはほぼPostScript様のものとして標準化される見通しである。このため、プリンタや版組要素関連の互換性を維持していくのに好都合であるという特徴を持つ。

現時点では、SGMLの利用については様々の周辺部の規格化やシステムの作成などを必要としており、具体的な普及の予測はし難いが、標準化推進の方向性としては、最も有望なものであると期待される。

3.4 その他の検討項目

教育　電子文書に関する能力は、これからは基本的なリテラシーの一つとなることは当然である。しかし、前述したように様々な論理的な整合性を欠いているのが現状である。特に、小学校における情報処理教育のなかでの文書処理をどのように規格化していくかは重要で、本稿で指摘した多くの問題についての見解を求めることにもなると思われる。

標準の普及　迅速的確な標準の制定も重要であるが、標準を定めてもそれを普及させるような方向に力が働かなければなかなか広まらない。このような標準の強制力を持っているものは、一つには学校教育での利用があげられる。また、官公庁などのなかでの利用も重要であろう。完全に市場での動向に任せるのではなく、標準を実現し使用していくという態度が重要であり、またそうでなければ真に有用な標準が作られないのではないだろうか。標準の制定だけではなく、常にプラクティスとのギャップを埋めるように、標準の利用を重視した政策が必要であろう。

4. むすび

本稿では、パーソナルな情報処理の中核である日本語文書処理の持つ問題点について概観し、その整理を試みた。現状のパソコン、ワープロの置かれた状況を問題視するか否かについて、まず議論する必要があろうが、ここでは今後重要となるより高度な文書処理への円滑な発展を念頭に置いて議論を展開した。

最後に、本稿をまとめるには、「マルチメディア標準化調査研究委員会」、なかでも特に詳細な検討に携わった「電子文書処理第一分科会」の委員の方々の活動がなければ不可能であったことを記し、感謝の意を表します。