ローマ字圏でない国の情報処理*

Information Processing in Non Latin-Alphabet Countries

植村俊亮
Shunsuke UEMURA

奈良先端科学技術大学院大学・情報科学研究科・教授
Professor,
Graduate School of Information Science
Nara Institute of Science and Technology

９６年６月２０日改訂

*電子情報通信学会誌「教養のページ」掲載予定．

１．はじめに

計算機の歴史は欧米から始まったので，ハードウェアもソフトウェアも最初は基本的にローマ字（ラテンアルファベット）文化圏だけを対象としていたのは，やむを得ないことかもしれない．ローマ字文化圏は世界総人口の４０パーセント強をしめ，近代文化の中心であり，牽引車でもあった．

しかし，世界中では3,000以上，細かく数えると10,000近くの言語が話されているという．実際に書かれている言語は100ほどを数える．人口5,000万人以上が使っている言語でも，中国語，英語，ロシア語，日本語など14もある．

漢字文化圏は世界総人口の３５パーセント強を占める．計算機が日本に導入されると，漢字を入出力可能にする努力が始まり，ワードプロセッサがうまれて，今日では日本語の文書のかなりの部分が計算機で作成されるに至っている．

インターネットの発展により，「地球規模の」情報資源の活用が議論される時代に，ローマ字を使わない国の情報処理の現状はどうなっているのであろうか．

２．東南アジアの計算機事情

東南アジア諸国の計算機の普及については，表１の統計がある(1)．日本だけではなくて，タイ，中国，韓国など，字種の多い国には，たいていその国の文字および符号に関する工業標準がすでにある．

例えば，タイでは，どこでもタイ語対応のWindowsを見かけることができる．マッキントッシュには，２０か国語におよぶ言語モジュールをもつWordPerfectのようなワープロソフトウェアがあるし，DOS/Vマシン，Windowsの世界でも各国の言語に対応する努力は，いろいろと進んでいる．文献（２）は，現状のよい入門書である．

中東の諸国については，残念ながらこうした資料は見あたらないが，マグレブ(北アフリカのモロッコ、アルジェリア、チュニジア）で人口7000万人に対してパソコンが35万台，リビアで500万人に3.5万台，エジプトで5500万人に11万台，湾岸諸国

（サウジアラビア、アラブ首長国連邦，イラク，シリア，　，，，）で合計4500万人に対して36万台くらいであるという(3)．人口100人当たりでは、0.2台～0.8台という計算になる。アラビア語の入出力も昔から研究開発が進んでいて，基本的には解決済みであるという．

３．国際化

Windows95の発売は社会現象として話題を呼んだ．Windows95は1995年8月にアメリカで発売され，日本語版の発売はその約３月後の11月であった．この期間のかなりの部分で，ソフトウェア技術者たちが自嘲を込めて「ジャパナイゼーション」と呼ぶ作業，すなわちシステムを日本語対応にする作業が行われたと考えられる．

日本語は，字種が多いだけではない．文章を横に書いたり，縦に書いたりする．アラビア語では，文章を右から左へ書き，行は上から下へ追う．文中に数字が現れると，その部分は，左から右に書く．モンゴル語は，文章を上から下へ書き，行は左から右に追う．数字があらわれても，書く順は変わらない．

情報技術の地球規模の拡散が始まると，従来のように，ソフトウェアをまずローマ字対応に作成しておいて，後でそれに手を加えて各国版を順次作っていくのは，合理的ではない．ソフトウェアは最初から，どのような文化環境にも容易に対応できる汎用系，いいかえると文化独立な系として作成しておくべきである．それに，特定の文化環境の名前を例えばパラメタとして与えるだけで，ソースプログラムの書き換えなしに，各国対応のソフトウェアができるようにすればよい．こうした文化独立な汎用系を作る過程を国際化(internationalizationあるいはi18n)，そこから特定の版を生成することを各国化(localizationあるいはl10n)という．i18nという名称は，internationalizationの先頭のiと最後のnとの間に文字が１８あるところからきている．オペレーティングシステムの世界では，Unixが先駆けで，localeという機能をもっている．国際標準化機構(ISO)もこのあたりに端を発して，国際化に関連する標準化を開始している．その第１歩は，文字符号系であった．

４．文字符号の国際化

世界中の文字を表現できる統一的な計算機符号系を作りたいという夢は古くからあり，国際標準化機構でも，１９８４年ごろから作業を始めていた．１９９０年に入って，Unicodeという提案が参入して作業は難航したが，ようやく１９９３年に国際規格，１９９５年には対応する日本工業規格 (JIS) が制定された．国際規格の番号や略称から，１０６４６（イチマルロクヨンロク）あるいは UCS と通称されるが，JIS でいえば，X0221「 国際符号化文字集合（UCS） - 第 1 部　体系及び基本多言語面」(4) である．技術的には，世界の文字（目で見える文字，図形文字という）を３２ビットで表現する壮大な規格である．しかし，実質はそのなかの基本多言語面という部分に，世界各国で標準化されている具体的な文字集合が集約されている．基本多言語面は，1 文字を１６ビットで表現する．これで，世界の主要文字を全部表現できるのは，中国，日本，韓国で使われている漢字をひとまとめに統合したためで，この作業が関係各国では大きい反響を呼んだ．例を図 1 に示す．基本多言語面は Unicodeと基本的に同じである． Unicode開発グループは，通常１０６４６を Unicodeと呼んでいる．

１６ビット符号系は，7 ビット 2 列や，8 ビット 2 列の符号系とは根本的に異なる．基本多言語面は，すべての文字を１６ビットで表現するので，ローマ字圏にとっては，重い．WindowsNT は Unicodeを採用していたが，Windows95で はやめてしまった．JAVA　は Unicodeによっているが，情報転送のためにUTF-8 という形式を採用している．UTF-8 は，国際符号化文字集合を基本としながら，ローマ字は従来どおりの8(7)ビット，基本多言語面の文字は 8 ビット 3 列というふうに表現する可変長の文字符号系である．シフト JIS や，Unix 系の EUC の考え方を国際的に拡張したものといえる．中国では，シフト JIS の考え方を取り入れた新たな文字符号系が国の規格として登場場していて，Windows95 の中国語版は，これによるという．情報通信の世界はなかなか 8 ビット系を抜け出せない．

5．文化要素

ある国対応のソフトウェアが実現するというのは，その国の文字が使えるだけでは十分ではない．文章を書く方向，文字列の整列順，日付の表現などの文化要素がきちんと反映されていてほしい．それをどこまで標準化するか，できるか，大きい問題であるが，とにかく情報処理の国際化を推進するにあたっては，各国の文化事情をよく理解する必要がある．ヨーロッパでは，この作業はかなり進んでいる．東南アジア，中近東諸国については，そんなにすすんでいない．わが国では，国際情報化協力センター（Centerof theInternationalCooperationforComputerization,CICC）という財団があって，東南アジア諸国と協力して，こうした文化要素を収集，整理する作業を進めている．

地球規模の情報ネットワーク時代を迎えて，その上を流れる情報がローマ字だけを基本としていてよいはずがない．世界各国の人々がそれぞれの文化を反映する言語を使って互いに情報を交換できて始めて，真の情報社会が誕生する．

執筆にあたり，貴重な情報を提供していただいた Rafik Belhadjさん，佐藤敬幸さん，木戸彰夫さんに感謝します．

６．参考文献

（１）"アジア地域における自然言語処理環境の現状に関する資料集"，（財）国際情報化協力センターシンガポール事務所，１９９６年２月

（２）三上吉彦，池田巧，山口真也，"電脳外国語大学," 技術評論社，１９９３年

（３）フランスBull社のR. Belhadj氏による。

（4）JIS X0221-1995(ISO/IEC 10646-1: 1993)国際符号化文字集合（UCS）- 第１部 体系及び基本多言語面，日本規格協会，１９９５年

国名	人口（千人）	パソコン台数（千台）	人口１００人当りの台数	主に使う文字
日本	130,000	12,500	9.62	漢字，かな
韓国				