自然言語処理 35－1
（1983. 1. 25）

アジア・アフリカ言語のコンピュータ処理について

杉田繁治　江口一久　中谷英明
（国立民族学博物館）（神戸学院大学）

1．はしがき

ここ数年来コンピュータ・ハードウエア、とりわけＶＬＳＩの急激な発達により、安価で能力の高い機器が身近に利用出来るようになり、パーソナル化が進み、人文科学の分野でも、単に統計処理のみならず言語、画像、音響処理など広範囲の問題にも使われるようになってきている。

国立民族学博物館では、それらの多様な要求に応えるとともに新たなニーズを引だすために、昭和５４年より本格的なコンピュータ・システムと各種の入出力装置を導入し、さまざまな応用プログラムを開発してきている。とくに人文系ではエンド・ユーザであることを強く意識して、道具性が高く使い勝手の良いソフトウエアを開発するよう留意している。それは出来るだけもとのデータのまま扱えるようにすることである。

言語の問題に関していえば、コンピュータの発達経過からして、その取扱える文字はラテン文字（英語アルファベッド）が中心であった。我国においてさえ最近やっと数千字種を持つ漢字混りの日本語が英文タイプライター並の簡略さで入出力処理出来るようになったにすぎない。開発途上国においてはまだラテン文字ではなく自国の文字を直接扱うことはそれほど進んでいない。しかし現在の情報処理技術からすれば、アラビア語でもハングルでもシュメールやマヤ文字でさえ、形式的には何等問題無く取扱う事が出来る。アジアやアフリカにおけるそれら言語のコンピュータ化の遅れは、そこではまだ情報化社会が進んでおらず、コンピュータ需要が少なく、商業的関心が薄いためにすぎないのであろう。

アジア・アフリカ諸国の文盲率は低くなく、ましてや固有の文字でなくラテン文字化された形を読みとれる人は極めて少数のインテリ層に限られてしまう。『情報量の南北格差』ということがいわれて久しいが、もし『南』の国々が将来それぞれの文字を自由に出し入れ出来、また相互に翻訳出来るコンピュータ・システムを持つことになれば、それはこの格差是正に大いに役立つものと思われる。

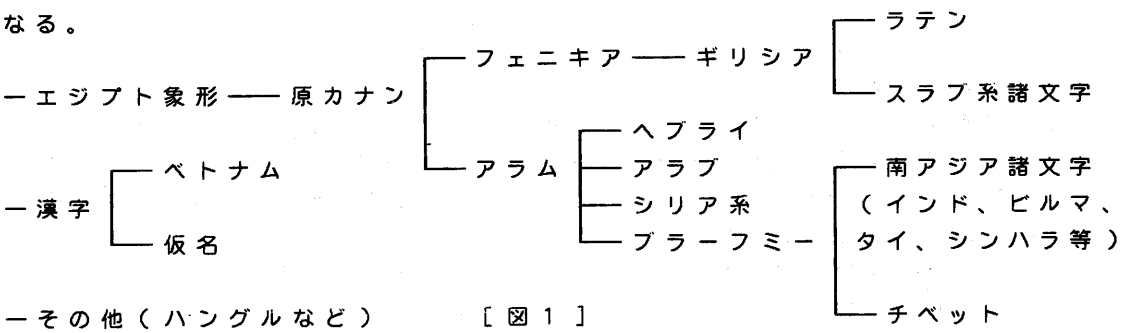
民族学の分野では世界の文化を対象としており、従がって取扱う言語も現在使われているもの、すでに使われなくなっているものを含めて全世界にまたがっている。それらをラテン文字化することなく、そのままの形で自由にコンピュータによって処理出来るならば、それぞれの言語に関して、あるいはそれを通じて行なわれる種々の学術活動を飛躍的に向上させるであろうことは想像に難くない。

現在漢字に関しては我国でひじょうな進歩がみられるが、それ以外の言語に対しては実用を目指した研究がほとんどなされていないように思える。それは、単に技術だけで出来る問題ではなく、言語の知識と具体的に処理すべき情報（ニーズ）が無ければならないからであろう。しかし逆にニーズがあってもそれを実現する技術がうまく得られなければ解決できない。

この小論の目的は、各種の言語にたいする具体的処理法の提案ではなく、英語や日本語を扱っている場合とは一寸違った問題のあることを指摘し、エンド・ユーザー側からの要求を出してその解決法を求めようとするものである。

２．英語等と異なる点

現在使用されている主要な文字を、系統的に大きく分類すると図１のようになる。



［図１］

一般に英語などに比べてアジア・アフリカの言語を扱う場合に異なる事は次の様な事柄である。ａ）文字要素の配置　ｂ）文字の入力　ｃ）文字の出力　ｄ）語句の切れめ　ｅ）改行の位置　ｆ）語順　ｇ）編修

２．１　文字要素の配置

ローマ字は基本的には一次元的に並べられている。ドイツ語やフランス語、スペイン語などでは母音の上部に記号が付られることもあるが、その数はそれほど多くなく、それらを付けた形が一つの文字としてあらかじめ用意されている場合が多い。しかしタイ文字やアラブ文字の場合には上下に要素を配置して２次元的に表現することが多い。

２．２　文字の入力

特定の熟練者が入力するのではなく、誰もがあまり訓練しなくても良くするには、出来るだけキーの数を少なくすることが望ましい。また端末を多目的に利用するために既存の英文タイプライター程度の鍵盤にしておくことも必要である。

図１の中の主要な言語を入力するのに必要な要素の数を表１に示す。ここでは数字や句読点・記号類は数えていない。また語中や語末での変化形、結合文字なども数えていない。

言語	基本要素数	言語	基本要素数
サンスクリット	６４	ヘブライ	４８
タ　イ	７４	ギリシア	４８
チベット	３４	ラテン	５４
アラブ	３７	ハングル	２４

［表１］

２．３　文字の出力

文字毎の大きさが異なり、また語中、語末の変化や上下の配置など一定のパターンではなく柔軟な制御を必要とする場合が多い。

２．４　語句の切れめ

ヨーロッパ語は単語の切れ目が明確であるが、アジア・アフリカの言語は必ずしもそうではない。したがって後の処理のことを考えて、入力時に適当な単位に切ることが必要である。辞書を持つことにより自動的に分割をする試みもタイ語などでやられているが、むつかしい問題が存在する言語もある。

２．５　改行の位置

行末をそろえる問題も言語毎にそれぞれの事情がある。英語の様に音節で切るのではなく、意味ある単位で切らねばならぬものもある。

２．６　語順

単語を辞書順に並べること、即ちソーティングが簡単でないものが多い。　たとえばタイ語では、ある母音は次の子音の後にまわしてからソートしなければならない。

２．７　編修

２次元的な構造を持っているのでエデイターに工夫がいる。

３．アラビア語とサンスクリット語の例

目下、国立民族学博物館では、アジア・アフリカ地域に行なわれる幾つかの文字のコンピュータ入出力システムを開発中であるが、ここでは、アジア・アフリカ地域での三大文字系統のうちアラブ文字系（北アフリカ、アラブ諸国、イスラエル、イラン、蒙古など）とブラーフミー文字系（インド、パキスタン、スリランカ、ビルマ、タイなど）の二つの系統のそれぞれ代表的な文字、すなわちアラブ文字とデーヴァナーガリー文字について、その主要な問題点を列挙し、あわせてこれまでに行ったデータ操作結果を報告したい。

上記の二大文字系に第三の漢字系の文字を合わせれば、ラテン文字（ローマ字）を使用する地域（中・南アフリカ、オーストラリア、トルコ、ベトナムなど）、あるいはハングルなどの特殊な文字を使用する極く少数の地域（南北朝鮮など）を除いてアジア・アフリカの全域をカバーすることになるが、漢字系統についてはここではとりあつかわない。（図１参照）

３．１　アラブ文字

ａ．入力

アラブ文字は表音文字であって、基本となる文字数は２９である。これ以外に、普通のアラブ語の書物では記されないのであるが、短母音ａ，ｉ，ｕおよび重子音（ｄｄ，ｌｌ，ｓｓなど）を示す記号が、弁別記号として基本文字の上下に付加されることがある。これらの弁別記号は、もちろん基本文字と合体した形でなく、別々に（基本文字＋弁別記号という２操作で）入力するのが得策である。

また、１単位となる綴り（１語と、それに接頭、あるいは接尾される冠詞、前置詞、代名詞など）は連書され、しかも同一文字の形がこの単位の先頭、中央、末尾、独立に出るかによって変化するので、綴りの切れ目を入力時に指示しておく必要がある。

さらにインデックス作成の際などに必要な辞書配列を得るためには「語根」とよばれる子音（ふつう３個）をその他の文字から区別しておかなければならない。たとえば「書」を意味する語根‘ｋｔｂ’は次のように関連した種々の言葉をつくる。

ｋａｔａｂａ	彼は書いた
ｙａｋｔｕｂｕ	彼は書く
ｍａｋｔｕｂｕ	机
ｍａｋｔａｂａｈ	図書館
ｋａｔｉｂ	書記

辞書配列ではこれらはひとまとめに置かれるので、標識として例えばｋ、ｔ、ｂの３子音を大文字シフトで入れればよい。

アラブ語の特徴は、上のような語の派生システムが今も生きており、必要が生じた時にはいつでもこの手続きによって新語を作り出すことができるという点であるから、この語根－派生語関係は重要である。これはコンピュータに語の意味を理解させる際にも見逃されてはならない点であろう。

またアラブ文字は右から左へと書いてゆくので、端末画面の動きも現行の逆であったほうが使い易いことは言うまでもない。

ｂ．出力

入力を上に記したように簡約化し得るのはコンピュータの出力機能がすぐれているからであって、現行のタイプライターでは60～70のキーがどうしても必要のところを30～40ですませることができる。従って出力に際しては次のような種々の工夫を要する。

まずアラブ文字には種々の書体がある。また先に触れたように基本文字の形は、綴りの１単位のどの部分に来るかによって通常変化する。書体によっては１文字が20以上に変化するものがある。

コンピュータによる出力に際しては、むろん全ての書体及び変化形を備えることも可能であろうが、タイプライターのように、なるべく字幅、字高、下方への伸びのそろった書体を選び、しかも変化形の数を制限すれば文字数を70くらいに抑えることもできる。そうしても見たところ不自然な感じはない。

しかし字幅を自由にし、同時に文字合成の方法によれば、さらに文字数を減少させることも可能である。すなわち、「語頭形＋尾部形」によって「語尾形」と「独立形」を導くことができる。これによれば文字数38、尾部形３の計41文字でしかもタイプライターの字形より読み易い形で打出すことが可能である（打出し例は図２参照）。

[図２]

هَذا نَمُوذَج مِن الحُروف العَرَبِيَّة المِعْيَارِيَّة
المَشْكُولَة الَّتي تَحْلُّ جَمِيع مَشَاكِلِنَا الطِّبَاعِيَّة
وتَضْمَن لَنَا إدراج لُغَتِنَا فِي مَيَادِين العُلُوم
والتِّكْنُولُوجيا الحَدِيثَة عَلى أَسَاسِ المُحَافَظَة عَلى
كِتَابَتِنَا الأَصِيلَة وفَصَاحَة لِسانِنَا العَرَبِي المُبين.

弁別記号は、（１）a，i，uの母音記号、（２）重子音記号、（３）-an,-in, -un の語末（Tanwinと呼ばれる）、（４）語末子音（子音の後に母音が来ないこと）、を示す記号があり、それらを互に組みあわせると２２種類となり、図２はこれを用いて打出したものである。

行末揃えは、行末に来た綴りの中に適当な長さの横棒（連結線）を挿入することによって可能である。

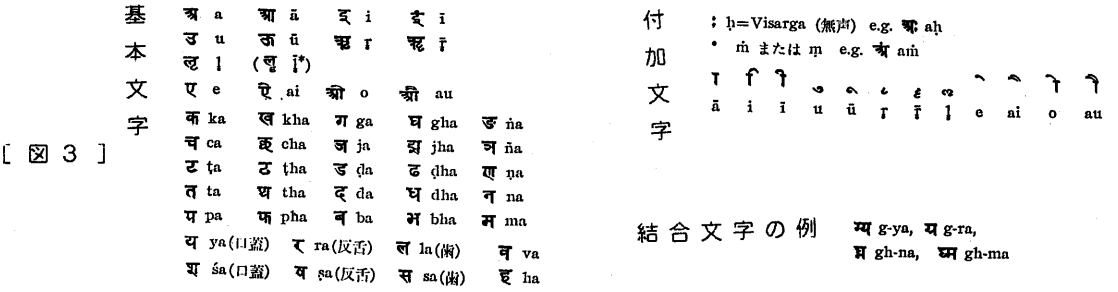
３．２　デーヴァナーガリー文字

現代インドで使用される諸文字、タイ、ビルマ、カンボジア、ラオスなど東南アジア諸国の文字、チベット文字、スリランカのシンハラ文字などの基となったブラーフミー文字は、紀元前７～８世紀の頃に当時のアラム文字を下敷にして作られたものと推定されているが、アラム文字と基本的に異なる点は、ブラーフミーには単独子音を表す独立の字形はなく、全ての子音記号は母音 'a' を伴っている「音節記号」である、という点である。この伝統はその末えいである上記の各文字に原則的には受けつがれている。それら現代諸文字の中で中心的位置を占めるのは、インドの公用語であるヒンディー語をはじめ、マラーティ語や古典サンスクリットの文献にも使用されるデーヴァナーガリー文字である。ここではこれによってサンスクリットを文字にする際の主要な問題点をあげるにとどめたい。その他の諸文字は、文字の形こそ違え、システムとしては大同小異であるから、わずかの修正によって応用可能と考える。

ａ．入力

基本となる文字は母音と子音あわせて47、それらの前後・上下に弁別記号として付加される要素が17、合計64文字である。（図３参照）

[図３]



但し、数個の子音が連続する時にはそれぞれの本来の形とは異なる一つの結合文字をつくる。その数は普通に用いられるものだけで180 に達する。従ってもし表記形をそのままの形でキーと対応させようとするならば、少なくとも180 個程度の結合文字をボードに用意しなければならない。このようにすればサンスクリットを読めないキーパンチャーにも入力は可能であるが、しかしかなりの非能率となることは否めない。

サンスクリット文において子音連続が極めて頻繁に起こることを考慮すれば、むしろ、入力時には子音記号を音節記号として扱わずａ母音を伴わない単独子音として扱っておき、子音連続は、たとえば 'tva' は 't(a)', 'v(a)' の二子音と母音 'a' に分かって入力することにした方が容易であろう。このようにすれば端末には（数字を除いて）上記の64個のキーを用意するだけですむことになるからである。

但し、サンスクリットに限って言えば、古く18世紀末よりローマ字化システムが確立し、多量の文献がローマ字で出版されて来たという特殊事情があるので、入力はローマ字で行うことが可能である。この場合アルファベットにない音価の文字は２文字で入力することとする。民族学博物館においては１をguttural，２をpalatal，３をcerebralおよびanusvara，visarga と定め、（例えば‘s’は‘s2’と入力）Udanavarga，Dharmasamuccaya の２種のサンスクリット仏典を入力し、種々の分析を遂行中である。

工夫を要するのは「連声」の入力である。連声とは二つの単語の連接時に、前の語末と後の語頭に生じる音変化であって、サンスクリット文はこの変化の結果をそのまま文字に表記する。例えば、‘na asti iha’（彼はここにいない）は‘nastiha’となり、‘devas asti’（神はある）は‘devo´sti’と表記される。また、語末子音の後に次の語の語頭母音が来るときには連書される。たとえば、‘tam iva’は‘tamiva’と表記される。従って、単語インデックスは連声、連書が行なわれた表記形からは作れず、各単語の原形を何らかの形で入力時に指示しておかなければならない。

あるいはコンピュータに十分な量の単語（数万語?)を先に記憶させておけば、自動的に連声・連書の法則を用いて文章を独立の単語に分解させることができるであろうか。

b．出力

出力に際しては、まず、あらかじめ先に述べた180 ほどの「結合文字」を用意しなければならない。縦・横に特に長いものが多いので工夫を要する。

次に、語中の母音記号、visarga (h)，anusvara (m)，avagraha (')，など17種の弁別記号が基本図形の上下・左右に附加される。これらはアラブ文字と違って数も多く、また比較的独立した形を持っているので現行の出版物においてもなされているとおり、別に作って合成できる。但し特殊の形を作るもの８種(ru,ru,su,su,sr,hu,hu,hr）については文字を用意しなければならない。

謝辞　アラブ語とその文字に関しては、大阪外国語大学の福原信義氏に御協力を頂いた。

参考資料

１）梅棹忠夫（編）『人文科学研究支援のためのコンピュータアプリケーションの開発』文部省科学研究費研究成果報告書　昭和５６年３月
２）杉田繁治『研究博物館と情報処理－国立民族学博物館での経験』情報処理学会誌　第２３巻第３号　昭和５７年３月
３）杉田繁治『人文科学者のためのコンピュータ・システムの条件』情報処理学会第２５回全国大会　昭和５７年１０月
４）日経コンピュータ『コンピュータを研究に駆使する国立民族学博物館』１９８２年１２月１３日号
５）坂本恭章『アジア・アフリカの諸言語の電算機処理』情報管理　Vol.22 No.7 1979
６）及川昭文、中山和彦『レーザービームプリンタを利用したタイ文字出力システム』情報処理学会第２０回全国大会　昭和５４年７月
　　杉田繁治
７）Sugita,shigeharu『Text Processing of Thai Language－The Three Seals Law』　ＣＯＬＩＮＧ　８０　：１９８０
８）中西亮『世界の文字』みずうみ書房、１９７５
９）西田龍雄編『世界の文字』大修館、１９８１