印度學佛教學研究第37卷第2號　平成元年3月

コンピューターによる仏教混淆梵語の研究（2）

——仏教混淆梵語のテキスト編集とデーヴァナーガリー文献の解読——

川　添　良　幸

1. 仏教混淆梵語研究に対するコンピューターの導入

コンピューターはその名の示す通り，数値計算を高速に行なう目的のために研究・開発された。第二次世界大戦中，米国において膨大な研究費が注ぎ込まれた後，弾道軌跡のシミュレーション用として，世界最初の電子式コンピューター，ENIACが登場したのは，終戦間もない1947年であった。この装置は，稼働するとフィラデルフィアの市街全体の電灯が暗くなると言われた程，電力を消費したが，大量の真空管を用いていたため故障が多く，しかも処理内容を変更するためには非常に複雑な配線をし直す必要があり，実用性は低かった。しかし，その後のコンピューター技術の進歩には目を瞠るものがあり，特に処理速度および記憶容量に関しては，実に5年に10倍という飛躍的な改良が行なわれた。さらに，ソフトウェアの充実に伴い，適用業務の範囲も急速に広がって行った。

所謂文科系の研究者にとって，コンピューターは長い間統計処理を高速に実行する装置に過ぎなかった。ところが最近，大容量の記憶装置の導入と高速なデータ・アクセスが可能となった上，データベース管理プログラムも充実したことにより，事情は一変した。知識は書物として蓄積するもの，という過去の常識は新しい技術により徐々に変更され，実際現在数多くの文献がコンピューター・データベース化されつつある。その長所は何といっても高速な検索機能にあり，一旦データベース化されたデータは，自在に編集することが可能である。例えば，本文を全て入力しておけば，索引はプログラムによって容易に作成される。もちろん，本文に対する追加や修正は何の問題もなく実行でき，それに対応して索引も容易に更新される。

我々のグループは，仏教混淆梵語で書かれた多くの写本を比較・対照し，文献学的研究を行なうため，数年前そのコンピューター・データベース化を企画した。まず，既に出版されている梵字の原典を集めた『法華経写本集成—全12巻—』[1]のローマ字化に着手した。現在までにその第1・2巻を完了し，索引および

逆引索引を付して出版している[2]。全巻は10年計画で完了する予定である。

現代のコンピューターは，処理が高速なことを第一の取柄としているが，最近では，その出力にも大きな進歩が見られ，図形や音声による処理結果の表示も可能となって来た。一方，入力に関しては，残念ながら未だに人間の情報処理能力には遥かに及ばないというのが，現状である。本研究のもう一つの目的は，仏教混淆梵語の記述文字であるデーヴァナーガリーの自動解読である。膨大な写本を全て人間が目で見ながらローマ字化して行くことは，写本の研究上非常な障害である。この作業を自動化できれば，研究者は本来の文献学に専念できる。我々は，ケルン－南條による法華経校訂本の自動解読を試み，現在ほぼ完全に処理できるまでに至っている[3]。今後，さらに進んだ情報処理技術を導入し，写本の解読をも試みる予定である。

2.　コンピューターによる仏教混淆梵語のテキスト編集

仏教混淆梵語で書かれた法華経写本は，現在までに30種類以上発見されている。それらは，大きくネパール—チベット，カシミール，中央アジアの3つのグループに分けられる。これらを集大成した前述の『法華経写本集成』は1977年から5年をかけて刊行されている。そこで集められた資料は文字数にして1億に達する膨大なものである。全ての資料の客観的な比較・検討を行なうには，そのコンピューター・データベース化が望ましい。現在までに，12巻中2巻分のローマ字化，およびそのコンピューター入力を完了している。

入力用にはパーソナル・コンピューター用エディター WORDMASTER(Micro Pro 社製) を用いている。キーパンチャーにより入力され，頁単位でチェックされたデータは，東北大学情報処理教育センターの IBM 3081 汎用大型コンピューター上に転送・蓄積され，常に端末からのアクセスが可能な状態にある。デーヴァナーガリー文字は，標準サンスクリットの語順にコード化されている。仏教混淆梵語では，標準サンスクリットの55文字中，43文字のみを使用するが，本研究ではより一般的な応用も考慮して55文字全てを取り扱える編集システムを開発した。キー入力は，英文のアルファベットはそのまま，それ以外の文字に対しては2または3ストロークを用いている。例えば，ṃは.mとキー入力する。この方法は覚えやすく，しかも時間的にも特殊なキー配列を設定するのと大差がない。出力には，特注のディジーホイールを用いているため，印字の質は活字並の最良のものである。その鮮明さは，最新のレーザービーム・プリンターをもって

しても追随困難である。

IBM 3081 上に蓄積されたデータを用いて，第1・2巻に出現する全ての単語，約1万1千語の索引および逆引索引が作成された。出力は標準サンスクリットの語順に従い，その単語の出現巻，頁，写本が併記されている。これにより，写本間の単語レベルでの相違は一目瞭然となった。また，逆引き索引は文法の研究に重要な役割を果たすものと期待されている。こうして作成された索引をキーとして，本文はランダム・ファイルとして保存されており，常に高速なアクセスが可能となっている。

作成された索引の一番簡単な利用例として，法華経第1巻の中の全ての単語の語長の出現頻度を求めた。その結果，この巻では7文字の単語の出現頻度が一番大きく，最大長の単語は43文字の『kumārabhūtasyānabhiniṣkrāntagṛhāvāsasyāṣṭau』であることが分った。この程度の処理は瞬時に終了する。

次に開発したプログラムは，単語検索システムである。第1・2巻に出現する全単語に対して，端末から単語を入力することにより，その単語の出現頁・写本の1行分のデータを出力する。複数箇所出現に対しては，その全てを出力する。これにより，写本の系統関係等を客観的に検討するための基礎資料が得られる。現在，出力をコンコーダンスに変更しているところであり，完成後は文献学的研究に大いに寄与するものと思われる。

以上のように，我々のグループはコンピューターを用いて，テキストを編集し，完成したデータベースを利用して，法華経写本の文献学的研究を進めている。この方法は，人間による索引作成に対して，高速，完全，再編集可能，等の優れた点を多く持っている。特に，今後とも新しい写本が発見される可能性は高く，再編集の容易さの重要性は明らかである。

3.　デーヴァナーガリー文献の自動解読

我国に於ては，仏典と言えば漢訳を指すのが普通であるが，もちろん，その原典はインドに於てサンスクリット語で書かれたものである。サンスクリット語の記述文字はデーヴァナーガリーと呼ばれ，現在でもヒンズー語等で使用されている。しかし，残念ながら現存する多くの写本は古く汚れており，その読解は困難である。そのため，多大な時間を費やして専門家によるローマ字転写が行なわれ，その資料に基づいて写本の比較・検討がなされている。前節で述べたように，我々のグループは，既に法華経写本のローマ字化と，そのコンピューター・

データベース化を試み，それに基き，高速な索引作成，語彙検索等を可能とした。しかし，そこにおいても，写本解読およびコンピューターに対するデータ入力は，旧来の方法に依っている。今後，データベース作成の高速化を図るためには，コンピューター導入のメリットの他の側面，即ちデータの自動入力に研究の重点を置かなければならない。

コンピューターによる文字自動認識技法には，パターンマッチング法および構造解析法が知られている。前者は，あらかじめ全ての文字に対して標準的なパターンを作成・登録しておき，読み込んだ文字との対応を調べ，認識するものである。一方後者では，まず全文字の図形的な特徴をヒューリスティックに選びだして纏めておき，新たに読み込んだ文字の特徴とそれらとの比較によって認識を行なう。我々は，約3年前，世界に先駆けて，デーヴァナーガリー文字の自動解読の研究を開始した。まず，既存の認識技術を調査し，標準的なパターンマッチング法および構造解析法の両者をデーヴァナーガリー文字認識に適用した。その後，デーヴァナーガリー文字に特徴的な太い横棒および縦棒を考慮した認識方法の改良を始めとして，様々な技法開発に取り組んできた。

一般に，文字認識の始めの手順は，画像としての文字データのコンピューターへの取り込みである。これには，イメージスキャナーと呼ばれる光学的に対象となるパターンをドット単位でコンピューター・データ化する装置を使用する。我々の使用している装置は，1mm に8ドットの精度である。この精度は高いほど良いと単純には言えるものではなく，読取対象文字の複雑さや，最終的なシステムの認識速度の要求等から決定される。次は，画像として取り込まれたデータから，1文字ずつに切り出す作業である。人間はいとも容易に文字の認識を行なっているように見えるが，現状のコンピューターで同様のことを実現するのは極めて困難である。すなわち，コンピューターは非常に高速ではあるが，逐次的にページ内の全てのイメージを調べ，データの多い個所を行とする。次に各行内で文字単位の認識を行なう。活字の日本語は字幅が一定であるのでこの操作は容易であるが，デーヴァナーガリー文字の字幅には相当のばらつきがあり，文字の切目の情報などで切り出さねばならない。印字の質にも依存するが，完全な切り出しは極めて困難である。

通常の文字認識技法では，各文字が切り出された後，その自動認識が行なわれる。まず，構造解析法では，切り出された文字の骨組みだけが利用される。この方法は細線化と呼ばれ，その結果から文字の特徴量が抽出される。今回採用して

いるデーヴァナーガリー文字の特徴としては，例えば太い横棒の有り無し，太い縦棒の有り無し，その位置および数，終端数，交差数等が挙げられる。現在，イメージスキャナーによるイメージの取り込み，および切り出しの済んだ1行分のデータは，IBM3081 により約10秒でほぼ完全に解読される。ケルン－南條本は1頁に約12行ずつ，全体で490頁あるが，全巻の自動認識は約20時間で済んでしまう計算になる。

今回，試験的に始めから10頁を自動認識対象とした。結果の一部を図1に示す。ケルン－南條本は印刷本ではあるが，80年以上も以前のものであり，印字の品質は良くない。現在は，イメージスキャナーによるデータ取り込み，ファイル転送，文字切り出し，等に多くの時間が必要であり，本格的な利用には，全体のシステム化が必要である。この段階になれば，専門家は不要となり，装置数を増やすことにより，原理的にはいくらでも作業時間の短縮が図れる。

現在，既存のパターンマッチング法に改良を加え，文字の切り出しと認識を同時に行なうという新しい文字認識手法を開発中である。これにより，切り出し率を実用レベルまで向上させる予定である。さらに，今後の課題として，写本の自動読取を計画しているが，その場合には，現在の1字ずつを単位とする読取ではほとんど解読できず，仏教混淆梵語の文法知識を持つ人工知能システム開発が必須である。その基礎研究として，現在サンスクリット辞書の電子化と，写本文字の蒐集を行なっている。

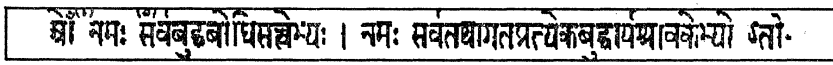
4. まとめ

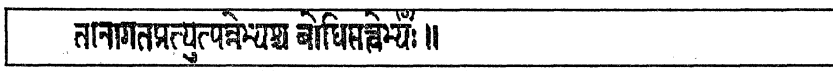
仏教混淆梵語のテキスト編集にコンピューターを導入し，30種以上の法華経写本に対してその全てのデータベース化を開始した。現在，第2巻までの入力を終了し，そこまでに出現する全ての単語の索引・検索等を可能とした。また，同時にデーヴァナーガリー文字によって記述された文献の自動認識を試みている。現在のところ，ケルン－南條による校訂本を，1頁約2分で認識できるまでに至っており，写本の自動認識の研究も予定している。

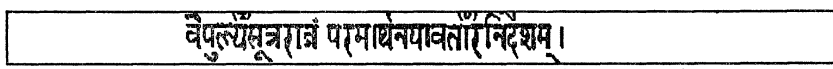
謝辞

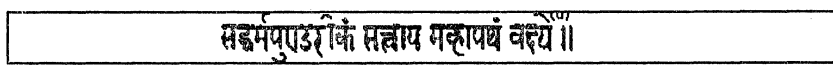
この研究は，東北大学情報処理教育センターの全ての職員の支援の下になされている。仏教混淆梵語のテキスト編集のプログラムは東北大学工学部大学院生の永島完司（現在，富士フィルム）によって作成された。デーヴァナーガリー文献の自動解読は，

センターの金井浩助手，工学部大学院生の K. ジャヤンティ（現在，グラフィカ）および鈴木昭浩の協力によってなされている。また，テキスト入力および文字データベース作成には，仙台電子計算機専門学校の御協力を得ている。これら全ての人々に深甚なる感謝の意を表する次第である。









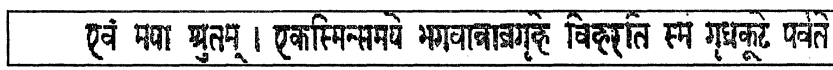


図1　ケルン－南條による法華経校訂本の自動解読結果。枠内がスキャナーで読み取られたオリジナルのデーヴァナーガリー活字文字による法華経の一部であり，各行の下に解読結果が対応するローマ字で示されている。

参考文献

1. 『梵文法華経写本集成』(Sanskrit Manuscripts of Saddharmapuṇḍarīka Collected from Nepal, Kashmir and Central Asia)［中村瑞隆・塚本啓祥・田賀龍彦・久留宮圓秀・伊東瑞叡・三友健容・三友量順共編］第1-12巻，東京，1977-1982年。［梵文法華経刊行会］
2. 『梵文法華経写本集成ローマ字本・索引』(Sanskrit Manuscripts of Saddharmapuṇḍarīka Collected from Nepal, Kashmir and Central Asia, Romanized Text and Index)［塚本啓祥・田賀龍彦・三友量順・山崎守一・川添良幸（第2巻）共編］第1巻，東京，1986年，第2巻，1988年。［梵文法華経刊行会］
3. 『数種の重要な持徴量を用いたデーヴァナーガリー文字認識の試み』(An Approach to Devanagari Character Recognition Using Outstanding Features)［ジャヤンティ クリシュナマチャリ・鈴木昭浩・金井　浩・牧野正三・川添良幸・木村正行・城戸健一］電子情報通信学会研究会報告 PRU87-103. 1987年。

＜キーワード＞　法華経，データベース，デーヴァナーガリー，自動解読

（東北大学助教授）

NEW PUBLICATION

Sanskrit Manuscripts of
Saddharmapuṇḍarika

Collected from Nepal, Kashmir and Central Asia

*

Romanized Text and Index

梵文法華経写本集成 — ローマ字本・索引

By K. Tsukamoto, R. Taga, R. Mitomo & M. Yamazaki;
Computer-programing directed by Y. Kawazoe

Vol. 1, 1986; Vol. 2, 1988 (in 14 vols.); each volume ¥25,000
Publisher:Society for the Study of Saddharmapundarika Manuscripts,
Yayoi-innsatsu Ltd., 1-14-15 Moto-Asakusa, Taito-ku, Tokyo 111 JAPAN