サンスクリット文献の索引作成に関する大型計算機の応用

小野基・小田淳一

1. はじめに

テキストの用語索引は当該文献の言語学的研究や原典解釈そのものを補助する手段として有用であるばかりでなく，そのテキストの含まれる言語文化一般の研究に寄与するところが少なくない。しかし，従来その手作業による作成は多大な労力と時間を要するのが常であった。ところが近年になってコンピュータが登場し，その文字処理能力の向上や特殊な文字法を図形パターンとして出力する技術の発達により，様々な言語のテキストの用語索引がコンピュータによって作成可能となった。本報告はそれらの技術をサンスクリット文献に応用し，他の言語にはない幾多の特殊性を有するサンスクリットの場合に生じる諸問題を考慮しつつ，大型計算機による用語索引の作成を試みたものである。なお対象テキストとして Dharmakīrti の *Pramāṇavārttikasvavṛtti* (ed. by R. Gnoli, Roma, 1960) を取り上げた。

2. 処理過程と使用機器

文献の入力から作成された索引の印刷に至る過程は大別して以下の四段階に分けられる。

Ⅰ 印刷用字体の作成
Ⅱ テキストの入力
Ⅲ テキストの処理
Ⅳ プリンタによる印刷

Ⅰ，Ⅱ及びⅢの過程は大型汎用コンピュータ（FACOM M 380）上で行われ，印刷用字体と処理されたテキストは共に MT（磁気テープ）に出力される。一方，Ⅳはレーザービームプリンタ（CANON LBP 3500）によるオフラインの出力（単体としての使用）となる。機器はいずれも筑波大学学術情報処理センターのものを用いた。

3. 処理内容

各段階の処理内容を以下に示す。

Ⅰ　印刷用字体の作成

文字の印刷は，文字コードと実際の文字パターンに割り付けられたコードをプリンタが照合することによって行われる。従って通常のアルファベットに存在しない特殊文字を印刷する場合は，まずその文字パターンを作成してコードを割り付ける必要がある。特殊文字パターンの作成及びコード割り付けの手順は次の通りである。

① 「印刷用字体ファイル」（18×24 ドットの文字パターンを格納したもの）を複写する。

② 特殊文字のパターンを別のファイル上で作成する。

③ 作成したパターンにコードを割りつける。

④ 作成したパターンを「印刷用字体ファイル」内の当該コードのパターンと差し替える。

⑤ 修正した「印刷用字体ファイル」を MT に出力する。

Ⅱ　テキストの入力

テキストの入力は，(a)入力した文字コードから印刷のための特殊文字を表すコードへの変換，(b)索引化のための処理（単語への分割，語の位置の同定，文字順の並べ替え等），(c)入力データから原テキストへの還元，の三点が容易に行なわれるよう考慮して，次のような方法をとる。なお入力は端末から直接行なわれる。

① 入力データは 1 レコード 255 バイト（文字）とする。

② 各レコードの先頭 6 桁はそのレコードの ID（識別子；テキスト ID が 1 桁，ページ ID が 3 桁，行 ID が 2 桁）とする。

③ 印刷時に特殊文字となる文字は 3 バイト（特殊文字であることを示す 1 バイトのマーク@，英字 1 バイト，記号 1 バイト）で入力する。（例：ā→@a_）

Ⅲ　テキストの処理

入力されたテキストは索引化のための様々な処理を経て印刷用の特殊文字コードに変換され，MT に出力される。その手順は次の通りである。

① 入力データを単語に分割し，それぞれの語の後にテキスト名とテキスト中の位置（ページ番号，行番号）を付し，「単語ファイル」とする。

② 「単語ファイル」を並べ替える。但し，サンスクリットのアルファベットはシステムによって提供されているソートプログラムが扱える英字アルファベットと文字順が異なるので，次のような処理を加える。

1) 各文字が特殊文字に変換された場合のサンスクリットにおける文字順を識別し，それぞれの文字を対応する順の英字に変換してデータの末尾につける。

2) データ末尾の英字の文字列をソートキーとしてレコードを並べ替え，「サンスクリットアルファベット順ファイル」とする。

③「サンスクリットアルファベット順ファイル」の文字の中で特殊文字で印刷する指定のあるものを印刷用の文字コードに変換して「印刷編集用ファイル」とする。

④「印刷用ファイル」をプリンタが要求する形式に編集して「編集ファイル」とする。

⑤「編集ファイル」を MT に出力する。

Ⅳ　プリンタによる印刷

LBP 3500 による印刷は次のように行われる。

①　Ⅰ—⑤の「印刷用字体ファイル」をプリンタが要求するパターン部として読み込ませる。

②　Ⅲ—⑤の「編集ファイル」をプリンタが要求するテキスト部として読み込ませる。

③　印刷を行う。

4. 問題点

本報告は試験的に行ったものであるので作業の過程で様々な問題が生じ，その対策に費した時間が全過程の殆どを占めたが，因みに文字パターンの作成とテキストの入力・修正を除いた実際のテキスト（単語数約 36000）処理時間は CPU タイムにして約 4 分である。

なお，次の事項が今後特に検討されるべき問題点であると思われる。

A　入力方法の問題

i）特殊文字の入力方法は現在の 3 バイト方式で良いか。あるいは特殊文字 1 文字に対して特殊記号 1 バイトを充てる方が入力並びに処理に有利ではないか。

ii）連声（Sandhi）の問題，特に単語の語頭に音変化が生じる場合をどう取り扱うか（本報告では本来の形に還元して入力している）。

iii）索引では語の変化（曲用や活用）形が必ずしもその原形に隣接して位置するとは限らないが，入力時にそれらの屈折語尾をどのように取り扱うべきか。

iv）サンスクリット特有の，頻出する合成語表現を入力時にどう分割すべきか（語頭の否定辞の取り扱いを含む）。

B　アルゴリズムの効率

本報告は大型汎用コンピュータを用いたために，作業領域や処理時間についてはあまり考慮せず，使用したプログラムのアルゴリズムはごく単純なものである。しかし，より小さなシステム（パソコン等）を利用して同様の作業を行なおうとする時にはアルゴリズムを大幅に修正して効率の良い処理をせねばならない。

5. 今後の課題

A　KWIC (Key Word In Context) の作成

単語の文脈上の位置を明示できる索引 KWIC の作成を準備中である。この索引による

ならばテキストの統辞上の特性の把握が可能となり，また 4-A に挙げた問題点の幾つかは解消されよう。

B　チベット語索引，梵蔵・蔵梵対照索引の作成

Ⅲ—②のソートプログラムをチベット語用に書き替えることによりチベット語のテキストの索引作成が，またⅢ—①のサンスクリットの「単語ファイル」に対応するチベット語とその ID とを書き加えることにより梵蔵・蔵梵対照索引の作成が可能である。

6. おわりに

本報告は，富山県立高岡商業高校教諭高木哲也氏（本学会会員）が筑波大学大学院在学中に発案した計画を著者らが引き継いだものである。本報告に何らかの価値ありとすれば，それはまず同氏に帰せられるべきものであろう。また国立教育研究所の及川昭文先生は印刷用字体ファイル及び入出力処理に必要な多くのプログラムを快く提供して下さった。更に東京大学大型計算機センターの三宅郷久先生にはテキスト処理用のプログラムの一部を作成して頂いた。著者らはこれらの方々の御協力により本報告をまとめることができた。ここに謝意を表する次第である。

小野（筑波大学大学院），小田（尚美学園短大専任講師）