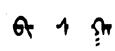
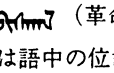
モンゴル語入出力インタフェースの実現と書誌データ検索への応用

満都拉[†]　藤井敦[†,††]　石川徹也[†]
[†]図書館情報大学
[††]科学技術振興事業団 CREST
{mandula,fujii,ishikawa}@ulis.ac.jp

1．はじめに

インターネットの普及に伴って、オンラインでの情報交換が日常化、国際化している。この高度情報化時代においてもモンゴル語の電子化されたデータベースは存在しない。モンゴル文字の特殊な構造や文字コードの問題のためにコンピュータ上で扱うことが困難だからである。本研究はモンゴル語文字コードの問題を解決し、モンゴル語入出力インタフェースを実現した。さらに、その入出力インタフェースを応用してモンゴル語書誌データ検索システムを構築した。

2．モンゴル文字の特徴

モンゴル文字は日本語の平仮名や片仮名と同じ表音文字である。原則として単語ごとに連ねて縦書きで綴り、行は左から右に進む。1 つの単語中では、各文字を英語の筆記体のように前後をつなげて表記する[1][5]。例えば  の 5 文字が連結して  （革命）という語になる。また、文字は語中の位置によって「語頭形」「語中形」「語尾形」（三形）のいずれかに変形する。さらに、母音には独立形がある。例えば、a（ア）と言う母音は、独立形が  、語頭形は 、語中形は  、語尾形は  に変形する。この「三形」は前後の文字と接続するため、2 つ以上の形に変形することがある。例えば、母音 （ア）は、語頭形  、語中形  、語尾形  など 2 つ以上の形に変形して前置の子音に適応する。

モンゴル文字は、母音を陽性、陰性、中性に分ける。また、語も陽性と陰性に分けられる。陽性語には陰性母音は入らず、陽性母音と中性母音から構成される。同じように、陰性語は陰性母音と中性母音から構成される。中性母音だけで構成された語を陰性語とする「母音調和」の規則がある。この規則によって、ある文字は同じ子音でも陰性語と陽性語で形が異なる。例えば、子音「g」は陽性語では  、陰性語では  などの形に変形する。また、同じ形で複数の音を表す同形異音文字があるため、字形だけでは発音を特定することができない場合がある。例えば には「a , e , n」の 3 通りの発音が対応する。しかも、同じ形の語でも発音が違うと意味も異なる。例えば、  という語は、「計算する（bodo）」「染める（bvdv）」という 2 通りの意味（発音）がある。この特徴はモンゴル文字を発音に基づいてコンピュータ上で扱う場合に大きな問題となる。

3．既存のモンゴル語文字コードと ISO

モンゴル文字をコンピュータ上で扱う研究は 1980 年代初め頃から行われ、中国、モンゴル国、日本、ドイツなどの様々な国や地域に固有の文字コードが存在する[2]。ここでは、代表的な 5 つの文字コードと ISO について説明する。

3.1　GB805487 コードと知能コード

GB805487 コードと知能コードでは字素、字母、音節の三つが混在してコード化されている。

字素は文字を構成する基本単位であり、子音か母音を表す。しかし、モンゴル語の母音がさらに字素に分割されることもある。例えば、母音 は と に分割され、それぞれ字素として扱われる。

字母は子音と母音の組み合わせであり、日本語の仮名に似ている。

音節は単語より小さい（または等しい）単位であり、複数の音韻によって発音上一つのまとまりをなす単位である。

当該コードの問題点を以下にまとめる。

- 文字の形に基づいてコード化されているため、発音に対応していない。
- 文字形と発音の相互変換ができない。
- 語を構成する文字の基本形を自動的に特定することができないため、計算機による形態素解析は困難である。

3.2 SUDAR コードと NEC コード

SUDAR コードと NEC コードでは、各文字は字母でコード化されており、同形異音文字を区別している。さらに、一文字の異なる形を別々にコードしており、ある音節も文字コード単位になっている。しかし、このコードでは「アリガリ文字」が含まれていないため、アリガリ文字を用いて表記する外来語の入出力ができない。

3.3 MLS コード

MLS (Mongolian Language Support) コードは字素でコード化されており、文字を基本形に分割している。しかし、例えば母音  を  に分割しているため、元の  に復元することが出来ない。

3.4 モンゴル語文字 ISO

モンゴル文字コードの ISO 案は、中国の内モンゴル大学、北京大学、モンゴル国立大学、モンゴル国立師範大学、モンゴル国家規格度量衡統一センター、国連大学などが参加して提案され、2000 年に ISO や Unicode に登録、2001 年 9 月に実用フォントを公開する予定だった[4]。しかし、未だに文字フォントが実装されていない。

ISO によるモンゴル文字コードの大きな特徴は、1 バイトのコードページ中にモンゴル文字とそこから派生した文字体系、トド文字、マンジュ（満州）文字、シベ（錫伯）文字の 4 種類を含めた点にある。この 4 種類の文字それぞれに、一つの音に対して独立形あるいは語頭形に対応する唯一のコードが割り当てられている。すなわち、文字が示す音と文字形の間には一対一の対応がある。これを「Basic Character Set」として情報伝達の規格とする。言い換えれば、Unicode は文字の音価を伝達するコード体系である。

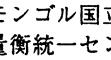
しかし、モンゴル文字は一つの音を示す文字の形が語頭、語中、語尾などの位置によって異なるため、表示と印刷には語中形・語尾形を含む「Presentation Form」（フォントに相当するもの）が必要である。この Presentation Form は、Basic Character Set のコードからアルゴリズムを通じて機械的に生成される 1 バイトのコードに対応している。4 種類の文字は大部分が共通した形を持つので 1 バイトに十分収まる。音価のデータは Presentation Form の生成アルゴリズムによって文字形のデータに変換される。

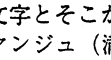
また、このアルゴリズムで生成できない不規則な変化形などは、Basic Character Set のコードに制御コードを付加することによって生成させる。この制御コードには、語頭形・語中形・語尾形を示す「Position Selecter」、変異形を示す「Varient Selecter」、語中のモンゴル語スペースを示す「Monglian Space」の 3 つがある。

ISO 規格は、モンゴル文字（トド・マンジュ・シベ文字を含む）の Basic Character Set、アルゴリズム、制御コードの 3 つの規格から成り立つ。

4．既存のモンゴル語入力方式

モンゴル文字に関する既存の代表的な入力方式として、次の 2 種類がある。

まず、ドイツのベルリン自由大学の Corff Oliver が提案した入力方式（Corff 方式）がある。モンゴル語では 7 つの母音  において 4 と 5 番目、6 と 7 番目の形が同じでもそれぞれ発音が異なる。しかし、Corff 方式はどちらも同じ文字 o と u に揃えて入力するため結局は a, e, i, o, u の 5 つの母音で入力することに相当する。また、この方式で入力したデータは同形異音文字を区別できないという問題がある。

次に、中国の内モンゴル大学で作られた入力方式（内大方式）がある[3]。内大方式ではモンゴル語の 7 つの母音  を区別して A, E, I, 0, O, V, U で入力する。また、子音をそれぞれ区別して入力するので同形異音文字を区別することができる。しかし、この方式では

同形異音文字を区別するために数字のゼロを用いて4番目の母音  を入力しているため文字と数字が混在するデータを扱うことが困難である。

5．本研究の入出力インタフェース

既存のモンゴル語文字コードを利用すると同形異音文字を区別できない、アリガリ文字が表現できないなどの問題がある(3章参照)。本研究では同形異音文字を区別し、アリガリ文字も扱うため、モンゴル語の発音情報をローマ字で入力しASCII コードで保存する。すると、モンゴル語の読み情報が失われず、同形異音文字を区別でき、アリガリ文字の情報も保存できる。また、音韻に基づいているため、いずれ公開されるUnicodeとの相互変換も比較的容易である。

モンゴル語は表音文字であり、また文字が語頭、語中、語尾の位置によって形が変わるため、入力の時に語中の位置を指定する必要がある。本研究では、モンゴル語の発音をアルファベットで入力し、スペースキーによって単語を区切る。すると、スペースの後ろにある文字は語頭形で、スペースの前に入る文字は語尾形を表して、前後に文字がつながっている文字は語中形と判断することができる。入力を終了してリターンキーによって確定されると、アルファベットの綴りをモンゴル文字に変換して表示する。

本研究の入力方式は、内大方式（4章参照）に改良を加えて利用する。内大方式は、母音と子音の同形異音文字を区別してモンゴル語の特徴を表現することができる。しかし、この方式では、四番目の母音 を数字のゼロで入力している。これは、数字が混在しているデータを扱うときに混乱を招く。そこで、本研究では をアルファベット大文字の O で入力する。また、内大方式では、ほとんどがアルファベットの大文字で入力しているため利便性が悪い。本研究では全て小文字で入力する。ただし、大文字はモンゴル文字の特殊な形を表すために用いる。例えば、ｄとｔの語中形 や語尾形の をそれぞれＤとＴで入力する。また、文字の表記をなるべくアルファベットに限定して他の記号を用いないようにするために、子音 はｘで入力する（内大方式では$で入力する）。

6．モンゴル語書誌データ検索システム

本研究で実現したモンゴル語書誌データ検索システムの構成を図1に示す。

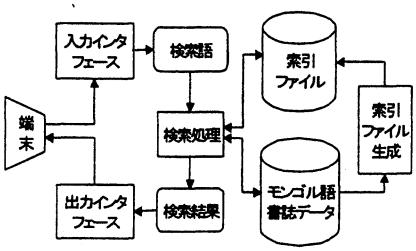


図1 モンゴル語書誌データ検索システムの構成

内モンゴルでは、図書館を公共図書館、大学図書館、研究所資料室の3種類に分類している。公共図書館は一般市民向けのため蔵書が特定の分野に偏らないように幅広い分野を含む図書を収集している。しかし、大学図書館や研究所の資料室は、大学や研究所の専門によって分野が偏ることがある。本研究では幅広い分野を扱うために、内モンゴル自治区で最大の公共図書館である「内モンゴル図書館」の蔵書リストを利用した。内モンゴル図書館の蔵書リストの書誌データは書名、著者名、出版社、出版地、出版年、サイズ、ページ数、ISBN、分類番号などの項目から構成されている。この図書館では、現在約120万件の書誌を所蔵している。当該蔵書リストの一部（1183件）を入手し、書誌項目が欠落してない508件を本研究で構築した入力インタフェースによって人手で入力して電子化した。本検索システムでは、検索質問を入力画面の質問入力テキストボックスにモンゴル語の発音に基づいてアルファベットで入力する。「学校」という意味のモンゴル語キーワード (sorgagoli) を入力した例を図2に示す。

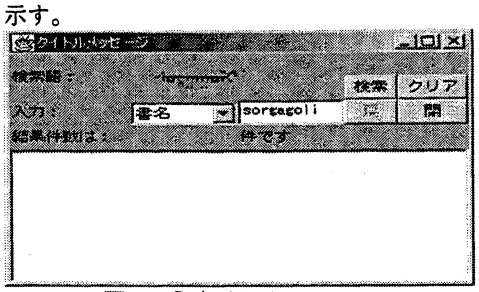


図2 入力インタフェース

ユーザが入力する時にミスがないとは限らないため、入力した検索語が正しいかどうかを確認する必要がある。ユーザがアルファベットを入力しリターンキーによって確定するごとに、システムは入力されたアルファベットをモンゴル文字に変換する。同時に、アルファベットとモンゴル文字がテキストボックス上に併記して表示される。そこで、ユーザは画面上のモンゴル文字とアルファベットの綴りが合っているかどうかを確認することができる。

出力インタフェースは、システムが動作状況をユーザに知らせ、処理結果を表示する。モンゴル語は縦書き表記なので画面上に縦書きで表示することが理想的である。しかし、本システムでは、モンゴル語を左に 90 度横に倒して表示する。これは本研究で利用した文字フォントが横倒しになっているからである。検索処理部からアルファベット表記で返された検索結果をモンゴル文字に変換してモンゴル文字で表示する。

検索処理部はブーリアン検索によって書誌データを検索する。本システムでは書名、著者名、分類番号による検索が可能である。最初、画面上に表示されるのは、検索ヒット件数と検索された項目との簡単な書誌情報である。例えば、検索された項目が書名の場合は図 3 のようになる。

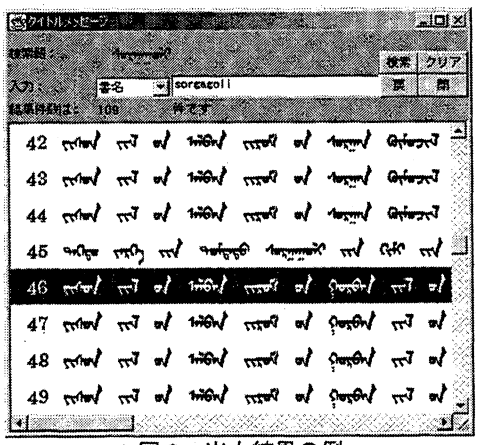


図 3　出力結果の例

この書誌情報一覧から必要とする書誌を選び、詳細情報を得ることができる。例を図 4 に示す。

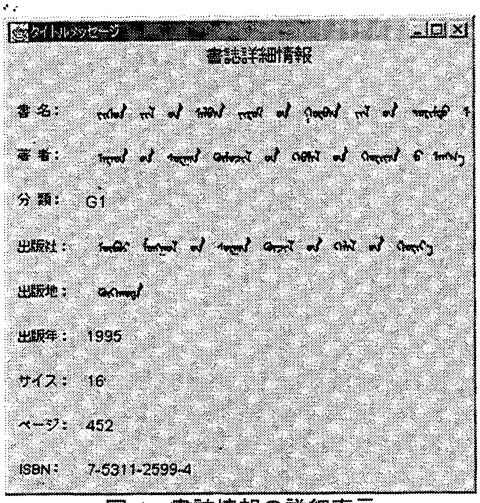


図 4　書誌情報の詳細表示

7．おわりに

本研究ではモンゴル語入出力インタフェースを実装し、モンゴル語書誌データを電子化した。また、この入出力インタフェースを応用してモンゴル語書誌データ検索システムを実現した。

今後の研究課題として、まずモンゴル語の縦書き表示を実現する必要がある。また、現在のモンゴル語書誌データ検索システムの研究は実験レベルであり、実用レベルに到達するまでには大規模な書誌データを用いた動作確認や評価実験が必要である。

参考文献

[1] 満都拉, モンゴル語専門用語の由来分析, 第 13 回専門用語研究シンポジウム, pp. 13—20, 2000.

[2] 確精扎布, 蒙古文編碼, 内蒙古大学出版社, 2000.

[3] Ochir ら, WINDOWS 環境でモンゴル文字入力方法の検討, 内蒙古大学学報, pp. 102—108, 2000.

[4] 上村明, モンゴル語処理の現状, bit. Vol. 30, No. 6, pp.70—71, 1998.

[5] 亀井孝、河野六郎、千野栄一. 言語学大辞典 第 4 巻 世界言語編 (下—2), 三省堂, 1992.