キリルモンゴル語 Web ページの縦書きモンゴル語への自動変換システムの開発

高　紅霞 ，　阪口　哲男
筑波大学図書館情報メディア研究科
〒305-8550 茨城県つくば市春日 1-2
E-mail: laura012@slis.tsukuba.ac.jp, saka@slis.tsukuba.ac.jp

抄録

モンゴル語には主に縦書きモンゴル語とキリルモンゴル語の２種類がある。縦書きモンゴル語は主に中国の内モンゴル自治区で使われ、キリルモンゴル語は主にモンゴル国で使われている。縦書きモンゴル語の特殊性のため電子化が遅れ、縦書きモンゴル語で提供される Web ページが限られている。一方、キリルモンゴル語は電子化が進んでおり、大量の情報がキリルモンゴル語によって発信されている。そこで、内モンゴル自治区など縦書きモンゴル語圏の人々がより多くの情報を得られるように、本論文ではキリルモンゴル語の Web ページを自動的に縦書きモンゴル語に変換する方式と Web ページ提示の問題について検討し、その自動変換システムの開発状況について報告する。

キーワード

縦書きモンゴル語、キリルモンゴル語、Web ページ、文字変換、レイアウト変換

Development of a System for Automatic Conversion of Cyrillic Mongolian Web Pages into Vertical Mongolian Script

Hongxia Gao, Tetsuo Sakaguchi
Graduate School of Library, Information and media Studies, University of Tsukuba
1-2, Kasuga, Tsukuba, Ibaraki, 305-8550, Japan
Email: laura012@slis.tsukuba.ac.jp, saka@slis.tsukuba.ac.jp

Abstract

The Mongolian language is mainly classified into two kinds, which are vertical Mongolian and Cyrillic Mongolian. Vertical Mongolian has mainly been used in Inner Mongolian Autonomous Region in China. Cyrillic Mongolian has mainly been used in Mongolia. However, the computerization of Vertical Mongolian is delayed because of its distinctiveness, and

Web pages in Vertical Mongolian are only a little. On the other hand, the computerization of Cyrillic Mongolian is advanced, and many Web pages are written in Cyrillic Mongolian. In order to make it easier and possible for people who use Vertical Mongolian script to obtain more information, this paper discusses the method of converting Cyrillic Mongolian Web pages automatically into Vertical Mongolian script and the problems of presentation of Web pages. This paper also describes automatic conversion system based on the method and its development.

Keywords
Vertical Mongolian, Cyrillic Mongolian, Web page, character conversion, layout conversion

1. はじめに

インターネットと Web の普及に伴って、大量の情報が Web を通じて発信されている。しかし、Web では縦書きモンゴル語の情報が少なく、縦書きモンゴル語を母語としている人々にとっては、Web 上の情報を利用するのは難しい。そこで、本研究ではキリルモンゴル語 Web ページを縦書きモンゴル語へ変換し、さらに変換後の Web ページを読みやすいレイアウトで提示するシステムの開発を行う。本稿ではそのシステムと開発状況について述べる。

2. モンゴル語 Web ページとその閲覧

モンゴル語には主に二つの種類がある。一つは縦書きモンゴル語で、もう一つはキリルモンゴル語である。現在は、縦書きモンゴル語の書き方の特殊性のために、モンゴル語の Web ページはキリルモンゴル語で書かれているものの方が多い。

2.1 キリルモンゴル語と縦書きモンゴル語

モンゴル語はアルタイ語族に属して、現在主にモンゴル国と中国の内モンゴル自治区で使われている。モンゴル国では 1940 年代初期に、それまで使われていた縦書きのモンゴル文字に代わって、キリル文字で書かれるようになった。1990 年代初期から縦書きのモンゴル文字が再び利用されつつある。現在モンゴル国ではキリル文字と縦書きのモンゴル文字が共に使われている。一方中国の内モンゴル自治区では、現在も縦書きのモンゴル文字が使われていて、キリル文字はほとんど使われていない。本稿では縦書きのモンゴル文字で記述するモンゴル語を縦書きモンゴル語と呼び、キリル文字を用いるものをキリルモンゴル語と呼ぶ。モンゴル国で使われているキリルモンゴル語は、ロシア語で用いられるキリル文字に二つの母音字を追加したものを用いる。キリルモンゴル語は左から右への横書きで、行は上から下へ進む。一方、縦書きモンゴル語は上から下への縦書きで、行は左から右へ進む。縦書きモンゴル文字用の縦書きレイアウトをサポートしないシステムではテキストを横書きにする必要がある。この際、モンゴル文字を反時計回り

に 90 度回転しなければならない。

2.2　モンゴル語 Web ページ閲覧の支援

中国の内モンゴル自治区では、キリルモンゴル語の読み書きができる人は少ない。また、縦書きモンゴル語テキストの電子化は遅れ、Windows Vista 以外の OS とブラウザは縦書きモンゴル文字に対応していない。しかし、Windows Vista であっても、縦書きモンゴル語の行は上から下へ進む横書きとなっている。それに対し、キリルモンゴル語の方は、テキストの電子化が進んでいる。インターネットの普及に伴い、大量の情報がキリルモンゴル語によって発信されている。中国の内モンゴル自治区の人々にとってはモンゴル国との情報交換のためにも、Web 上のキリルモンゴル語の情報を利用することができるとよいと思われる。そこで、本研究では、モンゴル語の Web ページの閲覧を支援するためにキリルモンゴル語 Web ページの縦書きモンゴル語への自動変換システムの開発を進めている。

3．キリルモンゴル語から縦書きモンゴル語への変換

本研究におけるキリルモンゴル語から縦書きモンゴル語への変換は三つのステップからなる。まずは、助詞処理と複数形処理をする。次に、モンゴル語の電子辞書[2]を利用して単語単位にキリルモンゴル語から縦書きモンゴル語へ変換する。最後に、辞書に含まれていない語に対しては、[1]で提案されたキリルモンゴル文字から縦書きモンゴル文字への翻字手法を利用してキリルモンゴル語から縦書きモンゴル語へ変換する。

本研究におけるキリルモンゴル語から縦書きモンゴル語への変換の流れは図 1 のようになっている。変換機能は Ruby を用いて実現し、辞書は MySQL を用いたデータベースとして格納している。

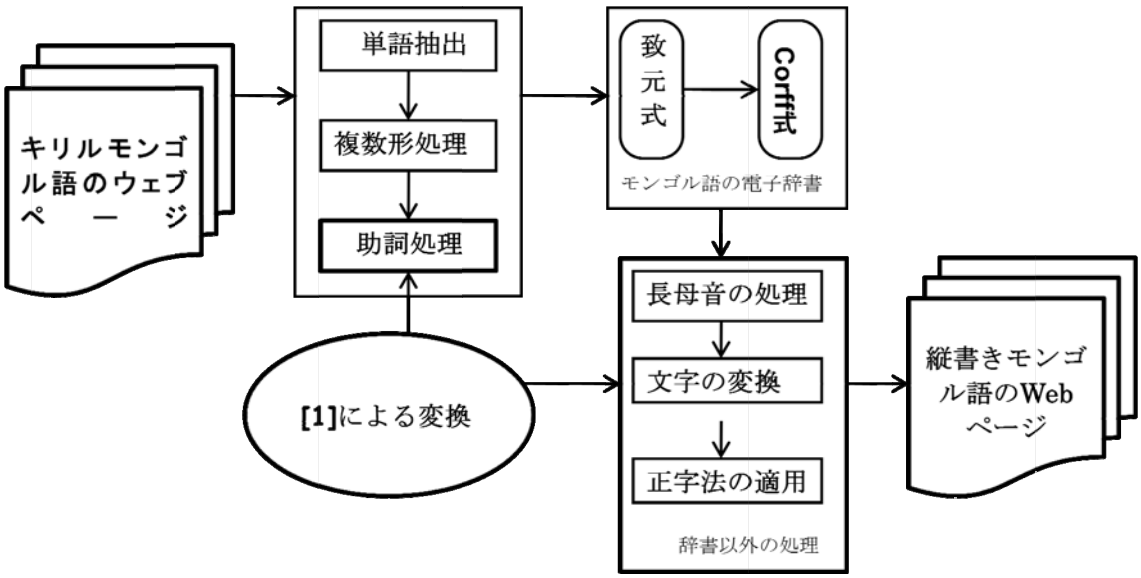

図 1 キリルモンゴル語から縦書きモンゴル語への変換の流れ

3.1 先行研究におけるキリルモンゴル語から縦書きモンゴル語への変換

キリルモンゴル語から縦書きモンゴル語へ変換をした研究としては「伝統的モンゴル語と現代モンゴル語を対象とした双方向的な翻字手法」[1]がある。この研究の手法によって、キリルモンゴル語または縦書きモンゴル語の一方のモンゴル語で書かれたテキストを文字単位で他方のモンゴル語に変換することができる。ここで、キリルモンゴル文字から縦書きモンゴル文字への変換手法について簡単に説明する。

- 助詞処理　キリルモンゴル語では、助詞は前の自立語に語尾として接続されている。しかし、縦書きモンゴル語の助詞は分かち書きされている。そのため、キリルモンゴル語の語尾を分割して、縦書きモンゴル語の助詞に変換する。
- 長母音の処理　キリルモンゴル語では長母音が１３種類ある。しかし、縦書きモンゴル語では長母音を記述する文字がないため、まずは長母音が存在しているかを調べて、[1]で作成された長母音処理規則に従って、長母音を処理する。
- 文字変換　キリルモンゴル文字を縦書きモンゴル文字との対応表に基づいて変換する。
- 正字法の処理　縦書きモンゴル語では弱化母音を記述しているが、キリルモンゴル語では弱化母音を記述していない。また、縦書きモンゴル語では語を構成する時に子音が３つ以上連続することがないが、キリルモンゴル語では子音が３つ連続することがある。以上によりキリルモンゴル語から縦書きモンゴルへ変換するときには母音が欠落することがある。そこで、縦書きモンゴル語の正字法に従って欠落した母音を[1]の手法によって補完する。

[1]によるキリルモンゴル語から縦書きモンゴル語への翻字精度は 80.6%である。[1]の翻字手法は不規則な語や固有名詞に対処していない。

本研究では、[1]で提案しているキリルモンゴル語から縦書きモンゴルへ変換する手法を少し手を加えて利用している。

3.2 モンゴル語の電子辞書による変換手法

[1]で提案しているキリルモンゴル文字から縦書きモンゴル文字への翻字手法は不規則な語や固有名詞に対していない。そこで、翻字精度を上げるために中里致元の「モンゴル語　電子化計画」[2]のモンゴル語電子辞書も利用する。

この辞書はキリルモンゴル語が致元式ローマ字で表記され、縦書きモンゴル語が Corff 式ローマ字[3]で表記されている、キリルモンゴル語と縦書きモンゴル語の対応を記述した電子辞書である。致元式は中里致元の独自の表記方法であって、英数字のみを使用し、キリルモンゴル文字と対応している。Corff 式はドイツのベルリン自由大学の Corff Oliver 博士が提案した MLS(Mongolian Language Support)に使われた縦書きモンゴル語の英数字による入力方式である。

電子辞書を利用して変換するには、まずは、キリルモンゴル文字を致元式に変換し、次に辞書を参照して Corff 式表記を得、最後に Corff 式を縦書きモンゴル文字に変換する。致元式、Corff 式のどちらも英数字を用いるので、元々英数字の部分と区別するためのマークを付けることにより、モンゴル語と英数字が混乱するのを避ける。

3.3 複数形の変換

モンゴル語の複数形は主に二つの形式で表されている。一つは前の単語の末尾に複数を表すスペルをつなぐ形式であって、もう一つは前の単語の後に複数を表す単語を置く形式である。

キリルモンゴル語では前の単語の末尾につながっている複数を表すスペルが、縦書きモンゴル語では分けられていることがある。また、縦書きモンゴル語では対象が人間の場合だけに使われる複数形がキリルモンゴル語では物にも使われていることがある。

これに対しては、キリルモンゴル語の単語の末尾が表１の左の覧のいずれかであれば右の覧の縦書きモンゴル語の複数形を表すスペルに置き換えることによって解決する。

表１ キリルモンゴル語の複数形と縦書きモンゴル語の複数形の対照表

キリルモンゴル語の複数形	縦書きモンゴル語の複数形
～иууд	～[illegible]、～[illegible]
～иуул	～[illegible]
～{й,н,л,р} д	～ [illegible]、 ～ [illegible]
～йс	～ [illegible] 、～[illegible]
～нууд	～△[illegible]、 ～△[illegible]、 ～△ [illegible]

（注:△は空白）

4. Webページのレイアウト変換

前述したようにWindows Vistaは縦書きモンゴル語の文字表示に対応しているが、特に指定しなければ、文字が横書きで行は上から下へ進む。しかし、縦書きモンゴル語では文字が縦書きで、行は左から右へ進むべきである。そこで、本研究では、WebページのHTML記述を解析し、Style Sheet指定を付与することによって変換後のWebページのレイアウトを見やすくする。ただし、現在のWebブラウザでは縦書の際に行が右から左に進む場合にしか対応していないので、現時点では行の方向が逆になっている。

4.1 レイアウト変換の問題点

縦書きモンゴル文字が横倒しでは読みにくく、ユーザにとって不便である。また、キリルモンゴル語Webページの全体のレイアウトが横書きを主体としているので、縦書きに変換するとレイアウトが崩れて、かえって読みにくくなる可能性がある。キリルモンゴル語と縦書きモンゴル語の書字方向の違いを考慮して、文字変換後のWebページのレイアウトを読みやすく変更することに工夫しなければならない。

4.2 レイアウトの変換方式

キリルモンゴル語の Web ページとその変換後の縦書きモンゴル語のページの内容を対応づけできるようにして読みやすくすることを目標とする。

現在、文字変換後のWebページのレイアウト変換は、Ruby用のHTMLパーサであるHpricot[4]

ライブラリを用いて実現している。現時点で、縦書きを指定する Style Sheet(writing-mode)を付与している HTML 要素を図 2 に示す。図 2 に示している HTML 要素に Style Sheet を付与した場合は、その子要素が図 2 にあるものであっても Style Sheet は付与しない。例えば、「div」要素が「td」要素の子要素である場合は「td」要素に Style Sheet を付与し、「div」には付与しない。

a address blockquote caption center div dl fieldset form h1 h2 h3 h4 h5 h6 hr li p pre td th

図 2　縦書指定の Style Sheet を付ける HTML 要素

図 3 はキリルモンゴル語の Web ページ例である。このページを開発中のシステムで変換した結果を図 4 に示す。図 4 では変換後のモンゴル語が縦書きに表示され、レイアウトも元の Web ページ（図 3）とおおよそ対応できるようになっている。しかし、本来は左から右に進むべき縦書きモンゴル語の行は右から左に進んでいる。また、画像が右回りに 90 度回転している。本来縦書きを指定する writing-mode はテキストに対するものであるが Windows Vista の Internet Explorer では画像にも作用している。

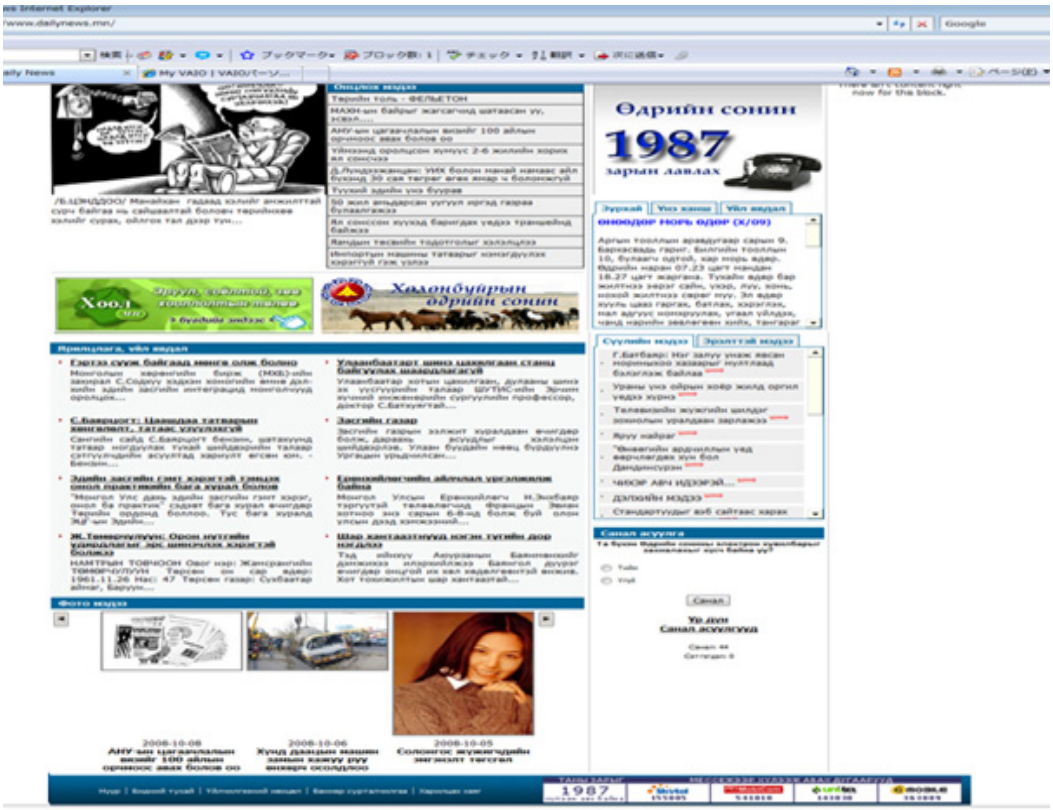

図 3　キリルモンゴル語の Web ページ例

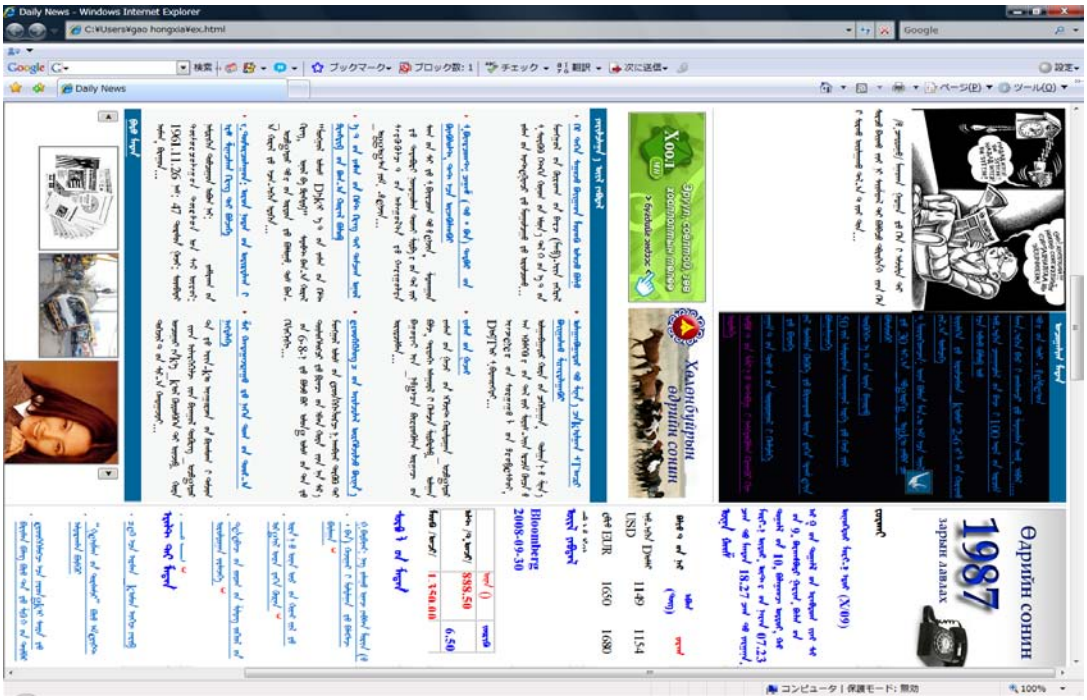

図 4　ページ変換例

5. 関連研究

同じ言語における文字変換という点では「漢字かな自動変換機能等を備えたインターネット学習システムの開発」[5]という研究がある。[5]では、インターネット上の情報を子供たちでも読めるようにするために、子供たちの学年に応じた漢字をかなに自動的に変換する機能を構築している。しかし、漢字をかな文字に置き換えてしまうため、文字数が増えてレイアウトが変わり、見づらくなると論文中で述べているがその解決法は提示されていない。また、富士通ラーニングメディアは Web 上の漢字をひらがな変換して表示できる子供向け Web ブラウザ「ひらがな・なびぃ」[6]を公開している。そこでも、漢字をひらがなにすることでレイアウトが崩れたり、大きくなったりしている。

本研究で行っている変換は同じ言語における文字変換という点で共通しているが、キリルモンゴル語から縦書きモンゴル語に変換するときに文字の量が変わる上に、文字の書き方と進む方向が違っており、レイアウトの問題解決がより重要となる。

6. おわりに

本研究では Web におけるキリルモンゴル語から縦書きモンゴル語への自動変換とその変換後の Web ページのレイアウトを提示するシステムの開発を行っている。

今後の課題としては、より見やすいレイアウトの調整を進める必要がある。最終的には、インターネットから利用可能として、複数の人に使用してもらって、評価を行いたい。 今後の課題としては、現在のブラウザでは、変換後の Web ページで縦書きモンゴル語の行が右から左へ進ん

でいるので、行の進む方向を修正する必要がある。また、キリルモンゴル語の単語を縦書きモンゴル語に変換する場合、候補が複数ある場合があるので、その部分は読み手が正しい単語を選べる機能を構築する必要がある。

参考文献

[1] 満都拉，藤井敦，石川徹也．伝統的モンゴル語と現代モンゴル語を対象とした双方向的な翻字手法．情報処理学会論文誌，Vol.47，No.8，pp.2733-2744，2006.

[2] 中里致元，モンゴル語　電子化計画．
http://texa.human.is. to hoku.ac.jp/~chigen/md_cnt_j.htm

[3] Oliver Corff，The Transliteration Principles of MLS.
http://userpage.fu-berlin.de/~corff/im/MLS/translit.unx

[4] Hpricot pylori*style wiki
http://tam.qmix.org/wki/Hpricot.html

[5] 榎本聡，室田真男，清水康敬．漢字かな自動変換機能等を備えたインターネット学習システムの開発．電子情報通信学会論文誌，Vol．J83-D-I，No.3，pp.384-394，2000.

[6] ひらがな・なびぃ．　http://kids.knowledgewing.com/free35

[7] 嘎拉桑朋斯格，蒙古国基立尔蒙古文正字法，内蒙古人民出版社，呼和浩特,2001.