

現代モンゴル語の異種表記法の 相互変換システムの構築に向けて

中里致元

生出恭治

{chigen,k-oide}@human.is.tohoku.ac.jp

東北大学大学院 情報科学研究科
仙台市青葉区荒巻字青葉 09

概要: 現代モンゴル語には、歴史的な発音を反映していると考えられる表記法と、現代の発音を反映している表記法の二種類がある。二つの表記法は異なる国家で採用され、交流のないまま数十年が経過した結果、同じ言語であるにも関わらず、互いに読解不能な状況に陥っている。本稿では、モンゴル語の音韻の歴史的変化の法則を探ると共に、この二つの表記法が形式的な法則を用いた変換システムによって、どれだけ変換が可能なのかを測定する。

我々はまず伝統的表記と発音的表記の対応する単位を抽出し、その単位に基づいた変換規則を吟味、その結果を単位の切り方に反映させるという作業を繰り返した。また、従来モンゴル語学で通説となっている音韻変化の法則がどれだけ妥当なものであるかも計測した。

Towards Constructing of Inter-transcription System on Different Modern Mongolian Writing Methods

Chigen NAKAZATO

Kyoji OIDE

{chigen,k-oide}@human.is.tohoku.ac.jp

Graduate School of Information Sciences,

Tohoku University

09, Aoba, Aramaki, Aoba-ku, Sendai 980-8579, Japan.

Abstract: Modern Mongolian Language has two writing methods; one is a traditional one which reflects ancient pronunciation and the other is a new one which reflects the modern pronunciation. These two methods have been adopted in different countries so that mongolian people living in two separate countries cannot communicate in writing one another with the lapse of time. We try to search for diachronical changing rules of the Mongolian language, apply them to the transcription system and rate it. First, we try to separate each word into the equivalent units. Second, we investigate the transcription rules by using these units and feed back to the first step. In this process we examine the accuracy of common theories in existence on Mongolian diachronical changing rules.

1 はじめに

1.1 二つの表記法の歴史的な背景

現代モンゴル語の表記法は主に二種類ある。一つはモンゴル文字を用いるもので、これは歴史的な発音を反映していると考えられる。中華人民共和国の内モンゴル自治区を中心に、中国領内のモンゴル民族の間で通用している。もう一つはキリル文字を用いるもので、これは現代の発音を反映している。独立国であるモンゴル国で通用している。

中国とモンゴル国は政治的な理由により近年まで交流が余り盛んではなかった¹。その為、モンゴル国が1941年に公布したしたキリル文字によるモンゴル語表記法はモンゴル国や旧ソビエト連邦領内のモンゴル民族の間に広まるにとどまった。中国領内では従来のモンゴル文字²を使い続けて現在に至っている。

こうして半世紀が経過し、モンゴル国ではモンゴル文字を読み書きできる者が殆んどいなくなってしまった。1990年の民主化の後、民族主義の高まりと共にモンゴル文字復活の気運もあったのだが、教師がいらない・活字がない・紙がないなどの障害の中で、それも立ち消えになった。

一方の内モンゴルでは、民族文字であるモンゴル文字への誇りと愛着が強く、キリル文字への関心は高くない。キリル文字は多少の学習で読めるようになるが、複雑な正書法が災いして、書けるようになるまで学習を続ける者は少ない。

更に、会話での意思疎通には問題がないので、相手の表記法を身につけようとする動機は一層弱いものとなっている。モンゴル語にも方言差はあるが、モンゴル国で標準語とされているハルハ方言と、中国領内モンゴルで標準語とされているチャハル方言には大きな差はない。書写上では余りに異なる姿のせいで別の言語のように見えるが、あくまで一つの「現代モンゴル語」と見なして支障のない範囲なのである³。このような事情により現代モンゴル語は表記の上で分断され、同じ言語であるにも関わらず、互いに理解不能に陥ってしまっている。

¹今でも民族主義を警戒する中国政府によって、交流には様々な制限が伴っている。

²モンゴル文字による表記法は17世紀にほぼ固まったとされる。

³但し借用語は、モンゴル国ではロシア語から、内モンゴルでは中国語からのものが多い。

モンゴル語に馴染みのない読者の為に、モンゴル文字とキリル文字のサンプルを掲げておく。内容は全く同じものである。モンゴル文字は縦書きで、左から右に行が進む。

図1: モンゴル文字

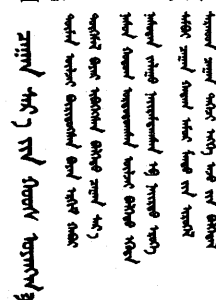


図2: キリル文字

ЦАГААН САРЫН ХОС УЯНГА

Түмэн өлзий бүрдсэн баян эрхэм хавар
Түүгээмэл баяр цэвэсэн бэлэгт цагаан сар

Насан хутаг уртадсан өлзий бэлэгт үүд
Настан залуу найрамдсан өв найрт өрх

Самбай цагаан хадаг амар мэндийн ерөөл
Сархад цагаан дэвж энх азын бэлэгдэл

1.2 文字コードと転写方法について

1.2.1 モンゴル文字

モンゴル文字は、系統としてはソグド系に属する。この系統の文字の特徴として、一文字の音価が一つとは限らないこと・語中の位置によって字形が変わることが挙げられる。その結果、単純に文字素を拾い出して翻字 (transliteration) するだけでは、十分な転写 (transcription) ができない。これはモンゴル文字を計算機で扱えるようにコード化する時にも問題になることである。

モンゴル文字のコードは、文字素の取り方からして異なった互換性のない方式が乱立している状態にある。中国やモンゴル国にも公的な規格が存在するが、寧ろアプリケーションやフォントに依存した民間コードの方が優勢である。ただ、2000年に更新されたISO/IEC10646-1 second edition (Unicode 3.0)にモンゴル文字が加わったので、いずれはコー

ドの標準化になるかも知れない。しかし現在の混乱状況は今暫く続くであろう。

本研究ではモンゴル文字のコードを直接には扱わず、転写文字の形でモンゴル文字を扱うことにした。正しく作られた転写方式なら、これを現行の様々なコードに機械的に容易に変換することができるし、他の転写方式に変換することも可能なので、特定のコードに束縛されずに済む。

モンゴル文字の転写方法にも様々なものがあるが、ここではベルリン自由大学の Infosystem Mongolei [1] で提唱されている方式を採用する。これはいわゆる ASCII コードの範囲内でモンゴル文字を転写するので、処理がし易いという利点を備えている。

1.2.2 キリル文字

キリル文字もコードが乱立しているという点で似たような状況にある。そこで、本研究ではキリル文字も転写文字の形で扱う。転写方式は我々独自のものを採用した [2]。英数字のみで構成されていることと、母音調和の判別が数字によって明らかになるという利点がある。

1.2.3 転写文字による表記

Infosystem Mongolei に従い、モンゴル文字の転写の頭には \$ 記号を、キリル文字の転写の頭には # 記号をつけて区別している。例えば図 1.2 で示した文の最初の一行は、

```
$caGan sar-a-yin qous uyangG-a  
#c7agaan s6ari5n xols6 o3y5nga
```

と転写する。

2 本研究の目的と意義

本研究の最終的な目標は、モンゴル文字表記とキリル文字表記の相互変換を実現することである。これを達成する方法として、巨大な対照テーブルを用意するということが考えられる。この方法なら音韻変化の法則を探る必要もなく、単純な手順によって変換が可能である。しかし、この方法には幾つかの問題点がある。一つは、モンゴル語の場合電子化さ

れたテキストが余り存在せず、テーブル作成に莫大な労力と時間がかかるということ。そもそも印刷媒体でもモンゴル語の辞書は数万語程度の規模のものしかなく、十分な大きさのテーブルを作るのは至難の業である。未知の単語には全く対応できないし、派生語や格曲用形をいちいち登録して行けば、テーブルは徒に巨大化するだけで実用的でもない。

そこで我々は二つの表記法の間の違いの法則性を探り、それに基づいた変換システムを構築することにした。この方法だと、上記の問題は回避できる。当然ながら、総てが法則通りには行かないが、どこまで法則化できるかが判明することだけでも大きな意味があると言えよう。

我々には、モンゴル語がどのように変化して来たかという言語学的な興味もある。勿論、モンゴル文字の綴りが過去の発音そのものというわけではないし、キリル文字の綴りが現代の発音そのままというわけでもない。だが、深い関係があることは確かであり、本研究はモンゴル語学にも貢献し得ると期待できる。つまり、モンゴル語の音韻の歴史的変化については先行研究が幾つかあるが、本研究はこれらを検証・補強するという意味でも役立つ。

更に、政治的にだけでなく、文字という形で文化的にも分断されてしまっている両国のモンゴル民族の橋渡しになれば、学問上の意義を越えた成果となるだろう。

3 対照辞書の作成

本研究に先立ち、約 23000 語に及ぶモンゴル文字・キリル文字対照電子辞書を作成した。底本には、現代モンゴル語の国語辞典として定評のある「mongGul kelen-u tobci tayilburi toli」(ulaGanbaGatur, 1966) を用い、誤植や綴りの複数あるものの確認などの為に「sin-e _usug-un _durim-un qadamal toli」(.kokeqota, 1987)、「mongGul bicig-un qadamal toli」(ulaGanbaGatur, 1992)、「mongGul kitad toli」(.kokeqota, 1976)「mongGul kelen-u barimziy-a abiyen-u toli」(.kokeqota, 1984)などを参考にして

この電子辞書はネットワーク経由で既に一般に公開している [2]。

4 モンゴル文字からキリル文字への変換

発音は経済性の原則に従って同化・弱化・省略など、単純化して行く傾向があると言われている。モンゴル語の場合も、古い発音を反映しているモンゴル文字表記の方が複雑で情報量に富む場合が多い。そこでまず、モンゴル文字からキリル文字への変換法則を見て行くことにした。

また、キリル文字の表記では正書法により、接尾辞などの付加の際に語幹の形が変わる場合がある。実際に文書を変換する場合には、この事例に往々にしてぶつかるわけだが、変換法則を取り出すにあたっては基本となる辞書形を対象とし、こうした異形態には後で対応することにする。

4.1 変換単位の導出と検証

変換を考える時には、何が何に変わるのかという変換の単位が判っていないなければならない。単位の導出は変換法則の導出と表裏一体のものであり、並行して行なう必要がある。

その際、取り出した単位が単位とするのに相応しいものかどうかを検証しなければならない。その方法として、同一単語を単位に切り分けた時に、モンゴル文字とキリル文字で単位の個数が等しいかどうかを一つの目安とした。また、母音に比べて子音は変化しにくいので、取り出した単位の子音を対照することでも検証が可能である。ただ、それだけでは判断し切れないので、目視による検証を併用せざるを得ない。

まず、扱える最小の単位として「音素のようなもの⁴」による切り分けを行なった。但し、モンゴル文字の形動詞形については語末の qu,ku を一つにまとめて単位としている。また、正書法の規則により、母音の直後の yi,Yi についても一つにまとめている。

次に、長母音や二重母音をまとめて、「母音と子音」による切り分けを行なった。

次に「音節のようなもの」による切り分けを行なった。子音の直前で切っているのので、母音を含まない

単位も生じる。

更に、“母音に挟まれた g”を単位に加えてみた。古代モンゴル語には長母音・短母音の区別はなかったとされており、現代モンゴル語にある長母音は、母音に挟まれた g が消失して生じたとするのが通説である [3]。モンゴル文字の綴りには古代モンゴル語の音が反映されているので、g は前後の母音を含めて一つの単位とすることができる。

但し、長母音の成立過程としては通説は正しいのだろうが、母音に挟まれた g が常に消失して長母音になるわけではない。g が消失しない例も多く存在する。語義によって消失する・しないが分かれるという説もあり [4]、そうだとすると形式的な処理だけで変換を行なうのは困難になる。

4.1.1 音素のようなものによる切り分け

例えば \$zirGuG-a-n, #zo3rgaa-n という語なら、

一致	○		○	○		×	×	×
\$	z	i	r	G	u	G	-a	-n
#	z	o3	r	g	a	a	-n	

と区切る。この場合、単位の数は一致しないし、同じ位置にある子音も一致しないもの含まれる。

実験結果

辞書の総数:22943

単位の数が一致しない単語の数:12832

半分以上の単語で単位の数が一致せず、切り分け方として適切でない。そこで数や位置などが変わりが易い母音を一まとめにしてみた。

4.1.2 母音と子音による切り分け

これは同じ例で示せば、

一致	○		○	○		×		×
\$	z	i	r	G	u	G	-a	-n
#	z	o3	r	g	aa	-n		

と区切る方法である。この場合も単位の数は一致しないし、子音も一致しないものが含まれる。

実験結果

単位の数が一致しない単語の数:12454

音素のようなものによる切り分けと余り差がない。

⁴あくまで文字の話なので、音素と一致するわけではない。音節の場合も同様。

4.1.3 音節のようなものによる切り分け

一致	○	○	○	×	×
\$	zi	r	Gu	G-a	-n
#	zo3	r	gaa	-n	

のように区切る。これでも単位の数は一貫しないし、子音も一致しないものも含まれる。

実験結果

単位の数が一貫しない単語の数:4750

例に挙げた語では差が余り見られないが、実験結果を見るとある程度意味のある切り分けになって来たようなので、これを基にして切り分け方の吟味をすることにした。

子音の一貫についても、前方からの一貫と後方からの一貫の両方を組み合わせて、どのような文字列が一貫しなくなるのかを調べた。つまり、

一致	○	○	△		○
\$	zi	r	Gu	G-a	-n
#	zo3	r	gaa		-n

上の表で△で示されるような部分を抽出して、問題点を洗い出した。この例では、“母音に挟まれたg”の処理が適切でないことが分かる。

4.1.4 母音に挟まれたgを考慮した切り分け

一致	○	○	○	○
\$	zi	r	GuG-a	-n
#	zo3	r	gaa	-n

のように区切る。今度は単位の数が一貫。子音も一貫している。また、変換法則を減らす為には、

一致	○		○	○		○
\$	z	i	r	G	uG-a	-n
#	z	o3	r	g	aa	-n

のように区切るとより良い。実験では、このように母音に挟まれたg以外にも更に細かく調整をした。

実験結果

単位の数が一貫しない単語の数:2115

総単語数の一割以下にまで、問題のあると思われる切り方をしている語が減少した。単位の切り分けについては、まずまずの結果が出せたと言えよう。

但し、解決しなければならない問題はまだまだ多

い。例えば子音が一致していると言っても、これは切り分けの便宜の為に暫定的に決めた条件に照らして一致しているにすぎない。一例を挙げるなら、4.3で後述するように、\$bは#bか#wになった時に一致していると見なしているが、実際に変換規則を作る場合には、どんな場合に#bになり、どんな場合に#wになるのかを記述しなければならない。

また、母音については殆んど手をつけていない状態である。我々は残念ながらまだ山の麓に辿り着いたにすぎない。以下に、これから越えなければならない峠を展望しておく。

4.2 連続していない単位

変換の単位が連続した文字列であるとは限らない。例えば、第一音節のiは後続する音節の母音の影響を受けて同化現象を起こすことが知られている[5]。

\$miq-a → #max

\$nidu-n → #nu4d-en

最初の例では、iはqの後ろのaに同化してaに変わり、二番目の例ではdの後ろのuに同化してu4に変化している。間に挟まれたqやdを含めて変換法則を作るとすれば、法則が徒に増えることになって好ましくない。

逆に前の母音によって後続する母音が影響を受ける現象もある。これは母音調和とよばれ、いわゆるアルタイ系の言語の特徴である⁵。

\$baGatur → #baatar

\$olan → #ollo1n

4.3 キリル文字の正書法に由来する規則

音韻論的に/b/であるものを表記する時に、キリル文字ではbかwを用いる。語頭では必ずbになり、語中ではl,m,w,nの直後を除きwになる。語末では必ずwになる。モンゴル文字でbと表記されているものがbになるかwになるかは⁶、モンゴル文字からの変換法則ではなく、純粋にキリル文字の正書法の問題であり、区別すべきと言える。

⁵この場合、モンゴル文字が母音調和していないわけではなく、別の母音調和規則に従っている。

⁶モンゴル文字のbが常にbかwになるということではない。

このような煩雑な正書法規則がキリル文字には幾つもある。中にはロシア語の正書法をそのまま踏襲したものもあって、必ずしも発音に忠実とは言えない理由の一つとなっている。これに対処する為には、一旦変換した後でキリル文字の正書法に基づいた変換をもう一度かける必要があるだろう。特に、実際の文章では語尾・接尾辞を加える際にキリル文字の正書法に従って語形が変わることがあるので、変換システムの実用化にあたっては避けて通れない道である。

正書法で語幹の母音が増減（または移動）する例
\$ekile ku → #exl ex
\$ekile n-e → #exel ne

更に、キリル文字にある大文字・小文字の区別を、モンゴル文字側から形式的な方法で判定することは極めて困難である⁷。例えば固有名詞は大文字で始めるという規則がキリル文字にはあるが、それが固有名詞であるかどうかはモンゴル文字の形からだけでは判断できない。仮に辞書にその情報を載せて変換の際に参照させるとしても、完全ではない。その固有名詞が形容詞的に用いられている場合には小文字で書くという規則がキリル文字にはあって、これは文脈から判断するしかないのである。

4.4 変換法則の検証

単位への切り分けと変換法則の抽出の繰り返しも、同じ辞書データを基にしていたら、やがては限界に達する。変換システムの精度が向上していることを実証するには、未知の文章を変換させてみなければならない。一般には同一の文章をモンゴル文字とキリル文字の両方で記述するという事は稀なのだが、内モンゴルで発行されている「solongG-a」という雑誌はロゼッタストーンのように二種類の表記法で同一の内容を掲載している。変換法則の検証には、これを用いることにした。しかし、検証に十分なだけの量の文章を入力するのには、膨大な労力と時間を要するのは言うまでもないことである。

5 おわりに

我々は当初、変換法則の導出はそれほど困難ではないと考えていた。しかしその目算に反して、変換単位の切り分けとの試行錯誤の段階で足踏みを続けることになっている。その最大の原因は、4.1でも触れた「母音に挟まれたg」にある。どうした場合にgが消えて長母音化するのか、文字の前後の環境からだけでは判断するのが難しいことが分かって来た。

母音に挟まれたgはモンゴル文字では頻出する形なので、形式的な法則でこれが処理できないとすると、他の方法を検討しなければならない。例えばvladimirCovは、親族関係や社会的な関係を表わす語・動植物や鉱物などの名称・身体部位の名称・家具や衣服などの名称などの場合には長母音化しないとの説を述べている[4]。

これには例外が多く、変換法則としては余り有用ではないのだが、それなりの示唆に富むものである。より精度の高い変換システムを構築する為には、今取り組んでいるような形式的処理だけではなく、語義などの周辺情報を踏まえた複雑な処理が必要になるだろう。

参考文献

- [1] Oliver Corff, URL: <<http://userpage.fu-berlin.de/~corff/infomong.html>>.
- [2] 中里致元, モンゴル語電子化計画. URL: <http://texa.human.is.tohoku.ac.jp/~chigen/md_cnt-j.htm>.
- [3] cinggeltey, *odu _uy-e-yin mongGul kelen-u _zuy.* _kokeqota, 1979.
- [4] vladimirCov, *mongGul bicig-un kele ba qalq-a ayalGun-u qaricaGuluGsan kelen-u _zuy.* lEninggraD, 1929.
- [5] N.Poppe, *Khalkha Mongolische Grammatik.* Wiesbaden, 1951.

⁷モンゴル文字には大文字・小文字というセットは存在しない。