E-046

伝統的モンゴル語テキストの音節分割アルゴリズム規則の設計
Design of Syllable Segmentation Algorithm Rule of traditional Mongolian Text

バトモンク† ウメルジアン ウスマン† 中平 勝子† 三上 喜貴†
Batumengke Osman Omarjan Katsuko T Nakahira Yoshiki Mikami

1 はじめに

モンゴル語の入力方式を構築するにあたり，多くのシステムにおいて音節入力方式が提案されている。音節を扱うメリットとして，モンゴル語には伝統的モンゴル文字とキリルモンゴル文字が存在するため，相互変換を容易に実現できることにある。

伝統的モンゴル文字―キリルモンゴル文字間の翻字手法は，これまで満都拉ら[1]，中里ら[2]によって研究されている。

中里ら[2]は、伝統的モンゴル文字から現代モンゴル文字へ翻字するために最適な変換単位と正字法の規則化について検討した。最適な変換単位の判断基準として、同じ単語をそれぞれ伝統的モンゴル語とキリルモンゴル文字で表記し、文字や音節などの単位で分割したときに、両モンゴル文字の分割数が等しくなることを条件とした。その結果、母音に挟まれる「g」を考慮した分割方法が比較的よかったものの、例外が多くて実用的ではない。また、彼らは変換単位や正字法の規則化について検討しただけで翻字手法を実現するに至ってない。伝統的モンゴル文字からキリルモンゴル文字への変換について検討した。しかし、キリルモンゴル文字から伝統的モンゴル文字への変換については検討していない。また、変換対象は単語であり、テキストではない。伝統的モンゴル文字とキリルモンゴル文字は助詞の表記法が異なるため、文章を対象とする場合は文脈に応じた助詞や活用語尾の処理が必要である。

満 都拉らの研究では、伝統的モンゴル文字とキリルモンゴル文字の表記規則に基づいた文字単位で相互変換する翻字手法を提案している。まず、原文字テキストから、単語抽出、助詞処理、長母音処理を行い、文字変換を正字法に適用し目的言語テキストを出力するような複雑な処理を行っている。この研究では、ローマ字転写による電子化方式を採用している。

しかし，いずれもローマ字転写を基本とした変換方式が採用されており，同形異音文字問題や文字と数字が混在するデータへの対応など，特に伝統的モンゴル語固有の問題について解決できない問題も多い。

以上のことを踏まえ，本稿では特に音節分割が難しいと考えられる伝統的モンゴル文字に着目し，文字コードを直接用いた音節分割アルゴリズムを提案する。本提案を応用することで，キリルモンゴル文字の音節分割についても同形異音文字問題は解決されると考えている。

2 モンゴル語音節の特徴

2.1 モンゴル文字の構成

モンゴル文字のアルファベットは 35 文字から構成されている[6]。

† 長岡技術科学大学
新潟県長岡市上富岡町 1603-1

表 1 伝統的モンゴル文字とキリルモンゴル文字の違い

	伝統的モンゴル語	キリルモンゴル語
母音数	8	13
子音数	27	20
円形文字	10	10
音符	0	2

表 1、伝統的モンゴル文字とキリルモンゴル文字[4]の母音/子音の数など、違いを示す。単語表記における違いは助詞の分かち書き、長母音の表記法、正字法の 3 点である。正字法とは、語を構成する子音と母音の接続に関する規制である。伝統的モンゴル語では助詞を分かち書きする場合がある。分かち書きされる助詞の先頭文字は語頭形ではなく語中形をとり、語の性によって字形は変わらない。しかし、直前にある自立語の末尾が子音か母音によって字形が異なる。

2.2 モンゴル語音節の構成

次に，モンゴル語音節の構成について述べる．本稿では，母音文字を V (Vowel)、子音文字を C (Consonant)と表記する[5]。

第一音が子音から始まり母音で終わる(CV)、または母音のみで構成した音節(VV)は開音節 “Open Syllable”と呼ばれる。単独母音を V、長母音を(VV)、2 重母音をV̈、単独子音を C、長子音を CC、2 重子音をC̈,3 重子音をC̈C̈で表現している。伝統的モンゴル語には一つの母音が一つの開音節になる。子音のみで構成された音節も単語も存在しない。子音が母音と結合して開音節になる場合は、音節頭は子音文字を使え、音節後は母音文字を使える。

第一音が子音から始まり子音で終わる(CVC)、または第一音が母音から始まり子音で終わる(VC)音節は閉音節“Closed Syllable”と呼ばれる。

表 2 に，伝統的モンゴル語における開音節／閉音節の音節構成を示す．音節を構成する音素数は 1 ～ 5 個まで存在する。

表 2．伝統的モンゴル語の開音節、閉音節の構成

開音節		閉音節	
1	V	1	VC
	VV		VVC
	V̈		VC̈
2	CV		V̈C
	CVV	2	CVC
	CV̈		CVC̈
			C̈VC
			CVVC
			CV̈C
			CVC̈
			C̈VC̈

2.3 モンゴル語音節分割規則の分析

表 3 に伝統的モンゴル語の音節(Syllable)分割則を示す。

音節 S に含まれる各文字の並びについて，独立形を X_1、語頭を X_2，語中を X_3，語尾を X_4 で表す。

このうち，X_1 はそれ単体で音節を作る単独開音節である．独立形以外については，X_2～X_4 はいくつかの文字が並ぶことによって音節を形成する．基本的にその組み合わせは語頭―語尾，および語頭―語中―語尾の２つに分類され，それぞれの要素に母音（長母音，二重母音を含む）および子音（多重子音を含む）が含まれる．これらの組み合わせをすべてとると，モンゴル語の音節の規則は S_1 から S_{16} まで 16 分類する事が出来る。独立形の後ろにほかの文字を入力すると S_2~S_{16} に移る．S_2～S_4 は母音と子音文字の組み合わせから構成した開音節である。S_5～S_{16} は閉音節である。表 3 のモンゴル語の音節分割の規則の X_1,X_2,X_3,X_4 の空白区にはモンゴル文字の母音も子音も存在しない。

表 3. 伝統的モンゴル語の音節分割の規則

独立形	前側	中側	後側	音節
X_1	X_2	X_3	X_4	
V				S_1
	C		V	S_2
	C		VV	S_3
	C		V̈	S_4
	V		C	S_5
	V		C̈	S_6
	V		C̈̈	S_7
	VV		C	S_8
	V̈		C	S_9
	C	V	C	S_{10}
	C	VV	C	S_{11}
	C	V̈	C	S_{12}
	C	V	C̈	S_{13}
	C	V	C̈̈	S_{14}
	C̈	V	C	S_{15}
	C̈	V	C̈	S_{16}

2.4 同系異音問題

モンゴル語には同形異音問題が存在する．同形異音問題を解決する手法として，モンゴル文字自由選択記号(FVS1, FVS2, FVS3)および母音間隔記号(MVS)を用いるため，見かけ上音節が FVS(1-3)や MVS によって分断されることがある。

同形異音問題を引き起こす FVS1-3 および MVS は，それ自体は単なる同形異音を指定する記号に過ぎないため，その両側にある文字を接続して 1 音節とする。

すなわち、FVS1、FVS2、FVS3 は同じ条件の下で、同一文字の不同な自由変形を区別する。これが単語の中で関係する文字の後に入る。同じ条件の下で、モンゴル文字の変化は以下である [3]。

例：独立形 ᠠ、その後に FVS1 を入力する独立形が変わって「ᠬ」 ようになる。ᠠ の語頭形が ᠠ と ᠬ の 2 種類ある。FVS2 と FVS3 も同様のふるまいをする [6]。

ᠳᠠᠭᠠᠨ daγan U1833+U1820+U182D+U1820+U1828

ᠳ[FVS1]ᠠᠭᠠᠨ d[FVS1]aγan U1833+U180B+1820+U182D+ U1820+U1828

頭形の文字「ᠳ(d)」の語頭形の後ろに FVS1 を入力することでそれのグリフを選択している。その FVS1 を直前と直後の文字と一音節を判断する。MVS は語末形の ᠠ(a)、ᠨ(e) とその前に入る子音の中に入る。

例：

ᠰᠠᠷᠠ sara U1830 +U1820+U1837+U1820

ᠰᠠᠷ᠋ᠠ sar[MVS]a U1830 +U1820 +U1837+U180E+ U1820

この二つの文字は発音同じであるが、意味は違う。この例で示しているとおり、語末形「ᠠ(a)」の前に[MVS]がきたら、それのコードを認識することで直前と直後に入る文字と一音節になると判断する。

3 音節分割アルゴリズムの構築

通常の文章を入力する際，音節の切れ目を気にして入力することはありえない．そのため，文書中に含まれる総音節数がわからない．この状態から音節を抽出する方法の構築が必要となる。

本稿では音節分割アルゴリズムを次の様に考える。

音節分割とは単語を音節ごとに区切り、単語を構成する要素を抽出することができる。モンゴルには母音が存在しない音節はないので、母音を核として、母音と子音を掛け合わせた音の連続的な表現でモンゴル語の音節を生成する。音節はモンゴル語構造の最小単位であり、モンゴル語の単語を、一音節の単語、二音節の単語、三音節の単語とそれ以上の音節の単語に分類することができる。2 文字の連接アルゴリズムを基礎とし、モンゴル語テキストを用いて、音節構造の量的分析を行った。

本アルゴリズムを構築するにあたり，表５にモンゴル文字の２～５音素の連接アルゴリズムにおける音節数計算表を示す。行要素には n-1 番目までの音素分のパターンを，列要素には n 番目の音素パターンを示し，これらの交点が音節の切断状態（＝音節数）となっている。表中の音節数には，表４の様な切断状態が定義されている。

本アルゴリズムにおいて，開音節と閉音節の音節境界状態と定義を以下のように行う。句読点シンボルを P (Punctuation Marks)、数字を D（ Digit)、空白を S (Space)、N (Format Controls)，分節記号｜(Division Sign, U+FF5C) で表現する。

この中で句読点シンボル P、数字 D、空白 S がテキストの中に含まれるとそれを文章と文章の区切り、単語と単語の区切りと考える。したがって、これらの文字は音節切断条件とみなす。

モンゴル語テキストの１音節の時、母音の独立形だけと判断する。それはテキストの音節分割を行う過程で一単語の最後の音素に区切りをつける。これにより、音節分割が、テキストから単語を対象に移る。

2 文字以降については次のように判定する。

表 5 に 2-4 文字までの連接アルゴリズムテンプレート抜粋を示す。

テンプレートにおける開音節と閉音節の音節境界状態と定義を表４に行う。

2 文字の連接は次の様に解釈する．たとえば，表５では、A から B の流れで、母音 V の次に母音 V が入力された場合、表２に基づいて、VV が一つの音節になると判断する。C が一つ入力された後に V が一つ入力された場合、表２に基づいて、CV が一つの音節になる。V の後に P が入力された場合、P の後で一区切りとして、一つの音節になる。

３文字の連接では同様に V の次に VV が入力された場合、母音が３文字繋がって音節になることないので、不法順序と判断する。C の次に VV が入力されたときに表２

に基づいて、CVV は一つの音節になると判断する。V+CV =2 の連接は 2 音節になるため、後ろの音素を切断し、V+CV とする。これと同様に４・５文字の連接もこの方法で判断する。

表４.音節境界状態と定義

切断状態	定義
≠	不法順序
0	音節分割がない
1	1 つの文字の分割
2	2 つの文字の分割
$n_{(1,2,3,\ldots\ldots n)}$	3 つの以上の文字の分割

表 5. モンゴル文字の 2-4 文字の連接アルゴリズムテンプレート(抜粋)

		B					
		V	C	P	D	S	N
A	*V*	1	1	1	1	1	1
	C	1	≠	0	0	0	0
	N	1	0	0	0	0	0
	V V	≠	1	1	1	1	1
	C V	1	1	1	1	1	1
	N *V*	1	1	1	1	1	1
	V C	1	1	1	1	1	1
	C C	≠	≠	≠	≠	≠	≠
	N C	1	≠	≠	≠	≠	≠
	V V C	2	≠	1	1	1	1
	C V C	2	1	1	1	1	1
	N V C	2	1	1	1	1	1
	V C C	2	≠	1	1	1	1
	C C C	≠	≠	≠	≠	≠	≠
	N C C	1	≠	≠	≠	≠	≠
	V V V	≠	≠	≠	≠	≠	≠
	C V V	≠	1	1	1	1	1
	N V V	≠	1	1	1	1	1
	V C V	2	2	2	2	2	2
	C C V	≠	1	≠	≠	≠	≠
	N C V	1	1	1	1	1	1

このデータを用いて、次の連接アルゴリズムの実装を行う。

4　連接アルゴリズムの実装

伝統的モンゴル語テキストを現代のモンゴル語テキスト変換するために、まず、伝統的モンゴル語の音節構造の規則の設計を行った。開音節 S_1~S_4、閉音節 S_5~S_{16} の設計を行った。そのあと、S_1~S_4 、S_5~S_{16} と音節分割に用いた文字記号“｜”の設計を行った。

図 1 に，以上を踏まえた音節分割アルゴリズムを示す.

```
read code [1];(文字コードの読み込み)
P[1]={V or C};  (code[1][1]の文字コードより判定)
read glyph[1] = {独、頭、中、尾}  (グリフの読み込み);

if(P[1] = V && glyph  = {独}){
          音節= 1;
          break;
     }else{
     T = 1; # T は音節連結テンプレートの値
     while(T!=2){
        X = X × P[i]        (×： 直積)
        i ++;
     read code[i];
      P[i] = { V or C};        (code[i]より判定)
     read  T = 音節連結テンプレート(X,  P[i]);
      }
      for(k=1; k<=I; k++){
      Y = Y × (translate code[i] to char);
      }
print Y;  (音節文字)
```

5　アルゴリズムの検証

伝統的モンゴル語の音節の分析と分割するために Visual C#2008[7] を使った。設計したソフトは[9]をもとに改良した。

Visual C#を使って開発した音節分割システムは以下のとおりである。ここでは入力されたテキストが自動的に音節分割する。

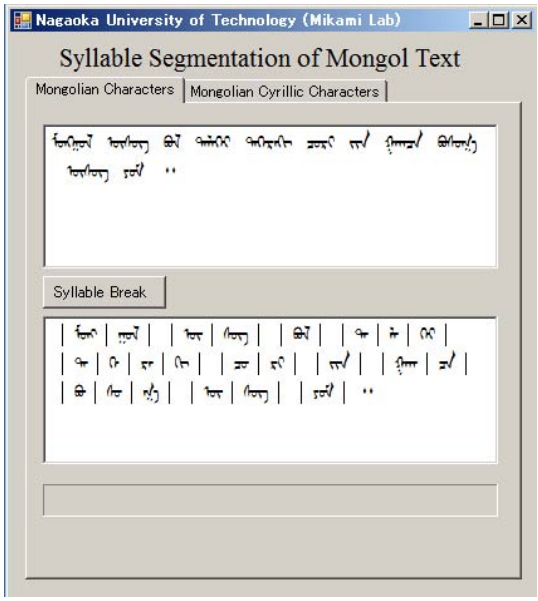

図 2. 音節分割分析システム

ソフトは二つの部分に分類される。

①電子データを入力部分

このソフトを使うにあたって,電子ファイルを図４の上の画面に貼り付けて、「Syllable Break」ボタンを押して、実行する。

②：音節分割部分

電子データを入力した後、実行ボタンが押されたら、下の画面に分割された内容が表示される。その音節分割

図１　モンゴル文字の音節分割アルゴリズム

を文字記号“｜”で表示している。

6　まとめ

本研究では伝統的モンゴル文字の音節分割の規則の設計を行い、文字の連接アルゴリズムを基礎とし、モンゴル語テキストを用いて、音節構造の量的分析を行う環境を構築した。

今後の課題として、実テキストに対して本アルゴリズムを適用し、その精度を測定する。また、これを入力システムに適用させることも課題の一つである。

このアルゴリズムを用いて、より多くのモンゴル文字テキストを音節分割し、単語中の音節データ群を抽出し、統計を取る。また、そのデータを用いて音節入力による予測変換入力方式に活用させることである。

7　参考文献

[1] 満都拉，藤井敦，石川徹也，“電津尾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．6，pp．2733-2745，2006

[2] 中里致元，生出恭治，“現代モンゴル語の異種表記法の相互交換システムの構築に向けて”，情報処理学会研究報告，2022-CH-53，pp．41－46，2003.

[3] The Unicode Standard, Version 6.0.

[4] Yoshiki Mikami. A History of Character Codes in Asia. 2002-03-20.

[5]The Worlds Writing Systems, edited by P. Daniels, W. Bright, New York, Oxford University Press, 1996.

[6]Basis of Cyrillic letters. edited by Burin Tokkuto. 2005.

[7]Visual Studio 2008