E-045

中期ウイグル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 Middle Uygur Text

ウメルジアン　ウスマン† 木村　哲也† 山田　耕一† 猿渡　隼斗† 三上　喜貴†
Osman Omarjan Tetsuya Kimura Koichi Yamada Hayato Saruwatari Yoshiki Mikami

概要

中期ウイグル語は、中央アジアに住んでいるウイグル人の一部が紀元 7 世紀から 19 世紀の終わりまでに使っていた言語である。この時代の間使われた中期ウイグル文字は、音を表現する文字で、表音文字(英：Phonogram)のカテゴリーに所属する。中期ウイグル文字列は、アルファベットで定義されて、母音、子音、分音符号、グリフと複雑なリガチャーの組み合わせにより構成されている。中期ウイグル語テキストには、独立系グリフが多く使われており、単語中に空白区(スペース)が使われている単語があり、単語を自動推定することが難しい。音節分割は単語を自動推定するために重要なプロセスであり、中期ウイグル語テキストの独立形文字、独立形グリフと独立形リガチャーを結合した単語の音節分割が必要である。一方で、中期ウイグル語の開音節の一部が現代ウイグル語の閉音節に変わるなど音韻上の変化があり、中期ウイグル語の音節構造解析などへの量的分析が行いづらい状況にある。中期ウイグル語と現代ウイグル語は上記の問題を除けは、同じ言語であり相互互換可能だが、そのためのテキスト処理が必要である。

本研究では、中期ウイグル語テキストの音節分割ソフトウェア(Syllable Segmentation of Uygur Text 1.0 professional Edition)と音節分割アルゴリズム基礎の設計を行った。その現状を報告する。

キーワード：中期ウイグル語音節、アルゴリズム規則

1　はじめに

中期ウイグル語はアルタイ(Altaic)語系のトルコ(Turkic)語系に属した言語である。中期ウイグル文字は、ソグド文字から得られて、モンゴル文字、満州文字、ブリヤート文字、カルムイク文字の前身である中央アジアの文字である。ウイグル文字は上から下の縦書きと右から左の横書きと二つ種類の書き方があり、アルファベットは 32 文字から構成されている。中期ウイグル語はスペースが単語の前後だけでなく、単語中の独立系グリフの間にも存在する。一方、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。ウイグル語には母音が存在しない音節はないので、母音を核として、母音と子音とを掛け合わせた音の連続的な表現でウイグル語の音節を生成する。音節はウイグル語構造の最小単位であり、ウイグル語の単語は、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。また、古代、中期、現代ウイグル語の音節を開音節、閉音節と二つに分ける事もできる。中期ウイグル文字の電子化にすると、コンピュータでアプリケーションを利用するときに障害がある。ウイグル文字コード表は、10 世紀の中期ウイグル語の名著であるクダトクビリク(英：Kutadgu Bilig)[1]の文字を基準に作っており、現在 ISO に提案中である。私が提案した中期ウイグル文字は以下の通りである。

表 1 中期ウイグル文字の分類

Acronym	Name	Glyph	Code Point
V	Vowels		X000-X007
C	Consonants		X008-X01F
D	Diacritics Marks		X020-X044
P	Punctuation Marks		X050-X05F
L	Lattice Digit		X060-X069
R	Round Digit		X070-X079
S	Signature Symbol		X080-X08E
B	Blank Space		X08F
N	Division Sign	|	FF5C

母音グループ内に 8 母音、子音グループ内に 24 子音、分音符号グループ内に 37 分音符号、句読点シンボルグループ内に 16 句読点シンボル、格子形数、円形数字数字グループ内に 20 数字、サインシンボルのグループ内に 15 サインシンボルである。

中期ウイグル文字は 128 の文字から構成されている。電子データの音節の分析と計算モデルの設計するために、文字の定義と略称を以下のように示す。

母音文字を V(Vowel)、子音文字を C(Consonant)、分音符号を D(Diacritics Marks)、上部分音符号を D̄、下部分音符号を D̲、上部分音符号の二重の組み合わせを D̿、下部分音符号の二重の組み合わせを D̳、母音と分音符号（二重分音符号）の組み合わせをそれぞれV̄、V̲、V̿、V̳子音と分音符号の組み合わせをC̄、C̲、C̿、C̳、句読点シンボルを P(Punctuation Marks)、格子形数字を L(Lattice Digit)、円形数字を R(Round Digit)、サインシンボルを S(Signature Symbol)、空白を B(Blank Space)、分節記号 N(Division Sign)で表現する。

† 長岡技術科学大学
新潟県長岡市上富岡町 1603-1

1.1 中期ウイグル語のアルファベット

中期ウイグル語のアルファベットは 32 個の文字(8 個の母音と 24 個の子音)から成っている。

表2 中期ウイグル語のアルファベット

I /i/ (4)	E /e/ (3)	Ah /æ/ (2)	A / a / (1)
Uv /y/ (8)	U /u/ (7)	Ov /ø/ (6)	O /o/ (5)
Zhe/ ʤ / (12)	Te /t/ (11)	Pe /p/ (10)	Be /b/ (9)
			
Re /r / (16)	De /d/ (15)	He /x/ (14)	Che / ʧ / (13)
Shi / ʃ /(20)	Si /s / (19)	Zhee / ʒ /(18)	Ze /z/ (17)
			
Ke /k/(24)	Khi /q/ (23)	Fi /f/ (22)	Ghe / ˠ /(21)
			
Me /m/ (28)	Le /l/ (27)	Ngi /ŋ/ (26)	Ge /g/(25)
			
Ye /j/ (32)	Ve /v/ (31)	Hhe /h/ (30)	Ne /n/ (29)

1.2 開音節

第一音が子音から始まり母音で終わる、または母音のみで構成した音節は開音節 (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 で表現する。ウイグル語には、三重母音、三重子音は存在しない。

例： (Su,  :水)

中期ウイグル語には一つの母音が一つの開音節になる。しかし、中期ウイグル語には、子音のみで構成された音節も単語も存在しない。以下の単独母音、長母音、2 重母音は一つの開音節になる。

→ V、V̈、VV

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。

→ CCV、CCV̈、CCVV

→ CV、CV̈、CVV

中期ウイグル語の開音節を以下のよう 2 つのグループに分ける事ができる。

第一グループ：母音から構成された開音節

母音から構成された開音節は以下のように 3 つある。

表 3 第一グループ開音節の構成

音節後		開音節	音節頭		No
V	C	←	V	C	
		V	○		1
○		VV	○		2
		V̈	○		3

(1) 単独母音の開音節
(2) 単独長母音の開音節
(3) 単独 2 重母音の開音節

母音から構成された開音節は、音節頭が 2 重母音を連接する事が出来る。

→ VV̈

第二グループ：音の発音順番には、第一音が子音の結合で構成された開音節

子音と母音の結合で構成された開音節は以下のように 6 つある。

(1) 単独子音と単独母音から結合した開音節
(2) 長子音と単独母音から結合した開音節
(3) 単独子音と長母音から結合した開音節
(4) 単独子音と 2 重母音から結合した開音節
(5) 長子音と長母音から結合した開音節
(6) 長子音と 2 重母音から結合した開音節

表 4 第二グループ開音節の構成

音節後		開音節	音節頭		No
V	C	←	V	C	
○		VC		○	1
○		VCC		○	2
○		VVC		○	3
○		V̈C		○	4
○		VVCC		○	5
○		V̈CC		○	6
○		VC̈		○	7
○		VVC̈		○	8

表 5 中期ウイグル語の開音節のグループの例

	Open Syllable ←	Example ←	Example →	Code Point ←
1	V		u	E150
	VV		uu	E150 E152
	V̈		ui	E150 E13B
2	VC		tu	E152 E168
	VVC		too	E152 E143 E144
	V̈C		tou	E152 E143 E168
	VCC		ttu	E152 E169 E168
	VVCC		ttoo	E152 E143 E169 E144
	V̈CC		ttou	E152 E143 E169 E168
	CV		ot	E16B E150
	CCV		ott	E16B E169 E150
	CVV		uut	E16B E151 E150
	VVCC		uutt	E16B E151 E143 E150
	CV̈		uott	E16B E143 E150
	V̈CC		uott	E16B E169 E143 E150

1.3 閉音節

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

→ VC　　→ CVC

例： (Yaz, :春)

中期ウイグル語の閉音節を以下のよう 2 つのグループに分ける事が出来る。

第一グループ：音の発音順番には、第一音が母音の結合で構成された閉音節

音節頭が母音から構成された閉音節は以下のように 3 つある。

(1) 単独母音と単独子音から結合した閉音節
(2) 長母音と単独子音から結合した閉音節
(3) 長母音と長子音から結合した閉音節
(4) 2 重母音と単独子音から結合した閉音節
(5) 2 重母音と長子音から結合した閉音節

表 6 第一グループ閉音節の構成

音節後		閉音節	音節頭		No
V	C	←	V	C	
	○	CV	○		1
	○	CVV	○		2
	○	CCVV	○		3
	○	CV̈	○		4
	○	CCV̈	○		5

第二グループ：音節頭と音節後には子音文字の結合で構成された閉音節

音節頭と音節後には子音文字の結合で構成された閉音節は以下のように 8 つある。

表 7 第二グループ閉音節の構成

音節後		閉音節	音節頭		No
V	C	←	V	C	
	○	CVC		○	1
	○	CVCC		○	2
	○	CVVC		○	3
	○	CVVCC		○	4
	○	CCVVC		○	5
	○	CV̈C		○	6
	○	CV̈CC		○	7
	○	CCV̈C		○	8

(1) 単独子音と単独母音から結合した閉音節
(2) 長子音、単独母音と単独子音から結合した閉音節
(3) 単独子音、長母音、単独子音から結合した閉音節
(4) 長子音、長母音と単独子音から結合した閉音節
(5) 単独子音、長母音、長子音から結合した閉音節
(6) 単独子音、2 重母音、単独子音から結合した閉音節
(7) 長子音、2 重母音、単独子音から結合した閉音節
(8) 単独子音、2 重母音、長子音から結合した閉音節

表 8 ウイグル語の閉音節のグループ例

	Closed Syllable ←	Example ←	Example →	Code Point ←
1	CV	مەت	ot	E16B E142
	CVV	مەەت	oot	E16B E143 E142
	CCV	مەتت	ott	E16B E169 E142
	CCVV	مەەتت	oott	E16B E169 E143 E142
	CV̈	مەۇت	out	E16B E151 E142
	CCV̈	مەۇتت	outt	E16B E151 E151 E142
2	CVC	تۇت	tut	E16B E151 E168
	CCVC	تۇتت	tutt	E16B E169 E151 E168
	CVCC	تۇتت	ttut	E16B E151 E169 E168
	CVVC	تۇتت	tutt	E16B E143 E151 E168
	CVVCC	تۇتت	ttuut	E16B E151 E151 E169 E168
	CCVVC	تۇتت	tuutt	E16B E151 E151 E168
	CV̈C	تۇت	tuot	E16B E143 E151 E168
	CV̈CC	تۇتت	ttuot	E16B E143 E151 E169 E168
	CCV̈C	تۇتت	toutt	E16B E169 E143 E151 E168

1.4 長母音

長母音VVは同じ音の二つと二つ以上の繰り返しで形成する。長母音の前側と後側が接した音と一つの音節になる。ウイグル語の 8 母音の長母音は以下のように示す。表 8 の空白区にはウイグル語の長母音は存在しない。

表 9 ウイグル語の長母音の連接規則

	a	ə	e	i	o	ŏ	u	ü
a	aa		aae aee	aai aii	aao aoo		aau auu	
ə		əə		əəi əii				
e			ee	eei eii				
i	iia iaa	iiə iəə	iie iee	ii	iio ioo		iiu uui	
o	ooa oaa			ooi oii	oo		oou ouu	
ŏ		ŏŏə ŏəə		oŏi oii		ŏŏ		ŏŏü ŏüü
u	uua uaa		uue uee	uui uii	uuo uoo		uu	
ü		üüə üəə	üüe üee	üüi üii				üü

1.5 2 重母音

音を発音の際に前の音を長く後の音を短く、前の音を短く後の音を長く、前と後の音が同じで長い発音をしている二つの母音から構成した二つの音の組み合わせは 2 重母音(Double Vowel(V̈))と呼ばれる。ウイグル語の 2 重母音は一つの音節である。

ae “ەې”, ai “ەي”, ao “ەو”, au “ەۇ”
əi “ېي”
ei “ېي”
ia “ىا”, iə “ىې”, io “ىو”, iu “ىۇ”
oa “ئا”, oi “ئى”, ou “ئۇ”
ŏi “ۆى”, ŏü “ۆۈ”
ua “ۇا”, ue “ۇې”, ui “ۇى”, uo “ۇو”
üi “ۈى”, üə “ۈې”, üe “ۈې”

1.6 長子音

長子音CCは、持続時間の長い子音である。同じ文字を二つ書き、音は促音となる。ウイグル語では開音節と閉音節の前側には長子音を連接する事はできない。しかし、開音節と閉音節の後側には長子音を連接する事ができる。

例：(att, ئەتت), (tatt, تاتت)

1.7 2 重子音

音を発音の際に前の音が長く後の音が短く、前の音が短く後の音が長く、前と後の音が同じ長い発音している二つの子音から構成した二つの音の組み合わせは 2 重子音

(Double Consonant(C̈))と呼ばれる。ウイグル語の 2 重子音は単独では音節に成らないが、母音と連接して一つの音節になる。ウイグル語には開音節と閉音節の前側には 2 重子音は存在しない。開音節と閉音節の後側には 2 重子音C̈が存在している。

例：(Halk, ئۆلكە: 人民)

2 音節分割規則の分析

ウイグル語の音節を“Syllable(S), 独立形、前側、中側、後側の文字を(X)で表現する。ウイグル語の音節の規則は S_1 から S_{22} までの 22 部分に分類する事が出来る。$S_1 = X_1$ は一つの文字、一つの母音から構成された単独開音節である。ウイグル語には $S_1 = X_1$ しか一つの文字から構成された単独音節は存在しない。

表 10 ウイグル語の音節分割の規則

後側	中側	前側	独立形	音節
X_4	X_3	X_2	X_1	
			V	S_1
C		V		S_2
CC		V		S_3
$\ddot{C}$		V		S_4
C		VV		S_5
CC		VV		S_6
$\ddot{C}$		VV		S_7
C		$\ddot{V}$		S_8
CC		$\ddot{V}$		S_9
$\ddot{C}$		$\ddot{V}$		S_{10}
V		C		S_{11}
VV		C		S_{12}
$\ddot{V}$		C		S_{13}
C	V	C		S_{14}
C	VV	C		S_{15}
C	$\ddot{V}$	C		S_{16}
CC	V	C		S_{17}
CC	VV	C		S_{18}
CC	$\ddot{V}$	C		S_{19}
$\ddot{C}$	V	C		S_{20}
$\ddot{C}$	VV	C		S_{21}
$\ddot{C}$	$\ddot{V}$	C		S_{22}

$S_2 = (X_2 + X_4) \sim S_{13} = (X_2 + X_4)$ 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。$S_2 = (X_2 + X_4) \sim S_{10} = (X_2 + X_4)$ は前側が開音節である。$S_{11} = (X_2 + X_4) \sim S_{13} = (X_2 + X_4)$ は後側が開音節である。(表 10 参照) $S_{14} = (X_2 + X_3 + X_4) \sim S_{22} = (X_2 + X_3 + X_4)$ は閉音節である。表 10 のウイグル語の音節分割の規則の X_1, X_2, X_3, X_4 の空白区にはウイグル語の母音も子音も存在しない。

図 1. ウイグル語の音節分割の規則

例：図 1. の(V)は独立形 $S_1 = X_1$ であり、前側と後側が開音節のカテゴリーに所属する。(CVC)は閉音節 S_{14} である図 1 に示す。

3 ウイグルテキスト処理の問題と解決手法

3.1 問題

中期ウイグルテキストを現代ウイグル語テキストに翻訳する際には以下のように二つの問題がある。

3.1.1 音節の問題

ウイグル中期ウイグル語の音節規則である後側が開音節 $S_{11} = (X_2 + X_4) \sim S_{13} = (X_2 + X_4)$ の規則は、現代ウイグル語の閉音節 $S_{14} = (X_2 + X_3 + X_4)$ に変化した。

3.1.2 グリフの問題

中期ウイグル語と現代ウイグル語のグリフは語頭形、語中形、語末形、独立形と 4 種類に分ける。中期ウイグル語のグリフの独立形式の一部が現代のウイグル語では語中形と語末形に変化した。

(A) 現代ウイグル語　　(B) 中期ウイグル語

発音(Bol sa)

発音(Bo sa)
トルキスタンのカラハン朝時代の紀元 10 世紀に書かれた文献作品
右から左の横書きの中期ウイグル語テキスト
Kutadgu Bilig
(180 page, Line6)
[2]

図 2. 中期ウイグル語と現代ウイグル語の音節規則の比較的分析

例：図 2 の(B)中期ウイグル語“Bo sa, ᠊ ᠉ ᠉”の“s, ᠉”, “a, ᠊”は独立形グリフである。中期ウイグル語で使っていたこの独立形グリフ“s, ᠉”と“a, ᠊”は現代のウイグル語では、語中形 “s, ᠉”と語末形“a, ᠊”に遷移した。中期ウイグル語の開音節を表現している文字は、グリフの独立形で表現

して、二つの独立形グリフの間には空白区が存在している。空白区から音節認識する事が出来ないので、中期ウイグル語の独立形グリフ使った開音節と閉音節を分割する事が必要である。

例：図 2 の中期ウイグル語を空白区から音節分割して、機械的認識する際には“Bo si a, ー ﾍ ﾍ”と三つの音節に分ける事になり、違う単語になる。

3.2 解決手法

中期ウイグル語と現代ウイグルを相互互換する際に発生した開音節から閉音節に変化した部分の問題と二つ以上の独立形グリフの間の空白区の問題を解決するために、中期開音節規則と現代閉音節規則の交換の規則を設計した。中期ウイグル語の開音節を(a)で表現して、(b)の値は、a= S_{11} ~ S_{13} である。 現代ウイグル語の閉音節を(B) で表現して、(B) の値は、B= S_{14} ~ S_{22} である(表 10 参照) 。中期の開音節 $S_{11}=CV$ ， $S_{12}=CVV$ ， $S_{13}=C\ddot{V}$ を現代の音節 CVC に変換するために、閉音節規則 S_{14} ~ S_{22} と基礎とし以下のような音節交換規則を設計した。

a=b;

a= $C+V+C+V=CVCV$;

b= $C+VC+C+V=CVCCV$;

空白区は“Blank Space (Bs)” で表現する。中期ウイグル語のグリフ G_l , G_m , G_r , G_n の中に G_l 間の空白区があり、現代ウイグル語では単語の間に空白区はない。(図 2 の(A) 参照) 。

Bs =0020;

a= $C+V+Bs+C+Bs+V$;

a= $CV \quad C \quad V$ (図 2 の(B)参照) 。

中期ウイグル語(a= $CV \quad C \quad V$)を現代ウイグル語に交換すると、別単語になる。

a= $C+V+Bs+C+Bs+V$ を現代のテキストに変換すると以下のように二つの形になる。

$$b=\begin{cases} C+VC+C+V+Bs=CVCCV \\ C+VC+Bs+C+V+Bs=CVCCV \end{cases}$$

0020＋0020＋0020=1;

$C+V+Bs+Bs+Bs+C+Bs+V$;

3.3 技術モデルの設計

ウイグル文字の開音節を x 、閉音節を y で割り当てを行った。

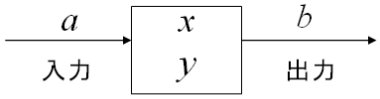
$x.V=x_1$ 、 $y.C=y_1$ 、 $V.x=x_2$ 、 $C.y=y_2$

$x.V.x=x_3$ 、 $y.C.y=y_3$ 、 $x.VV=x_4$

$y.CC=y_4$ 、 $VV.x=x_5$ 、 $CC.y=y_5$

$x.VV.x=x_6$ 、 $y.CC.y=y_6$

ウイグル文書の多くのデータを統計的に処理することによって、統計的モデリング “Statistical Modeling” を得ることができる。図 3 に、その過程を示す。



音節分割

図 3. ウイグル語の統計的モデリング法

音節分割入出力関係の入力を a ，出力を b とするとき、モデル構造として次の４種類のものが考えられる。

$x_{(1,2,3,4,5,6)}$ (開音節 Open Syllable) モデル

$x.y=x_7$ 、 $x.xy=x_8$

$y_{(1,2,3,4,5,6)}$ (閉音節 Closed Syllable) モデル

$y.x=y_7$ 、 $y.yx=y_8$ 、 $y.xy=y_9$

3.4 音節分割システムの開発

C#を使って音節分割システムの開発を行った。中期ウイグル語の音節規則である後側が開音節 S_{11} ～ S_{13} の現代ウイグル語の閉音節に変わった部分を交換すると中期ウイグル語テキストグリフの空白区“ Bs ”の問題を解決するために、中期ウイグル語テキストの音節分割のソフトを開発した。

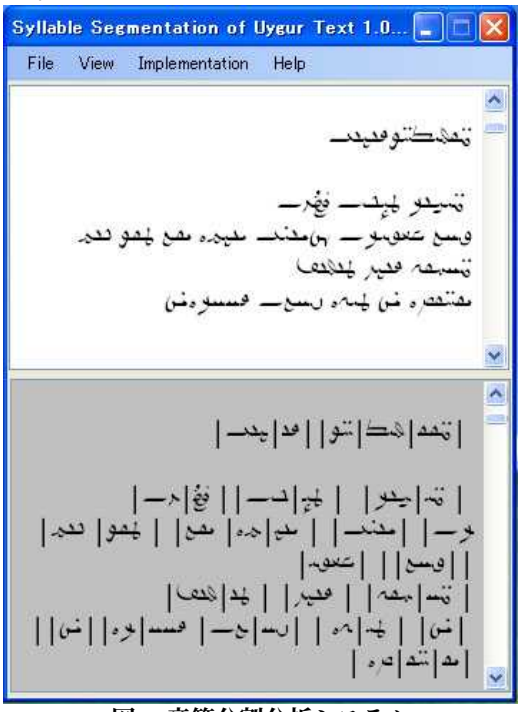


図 4. 音節分割分析システム

ソフトは三つ部分に分類される。

(1) 電子データの入力部分

(2) 音節分割部分

(3) 中期ウイグルテキストを現代ウイグル語に変換する部分

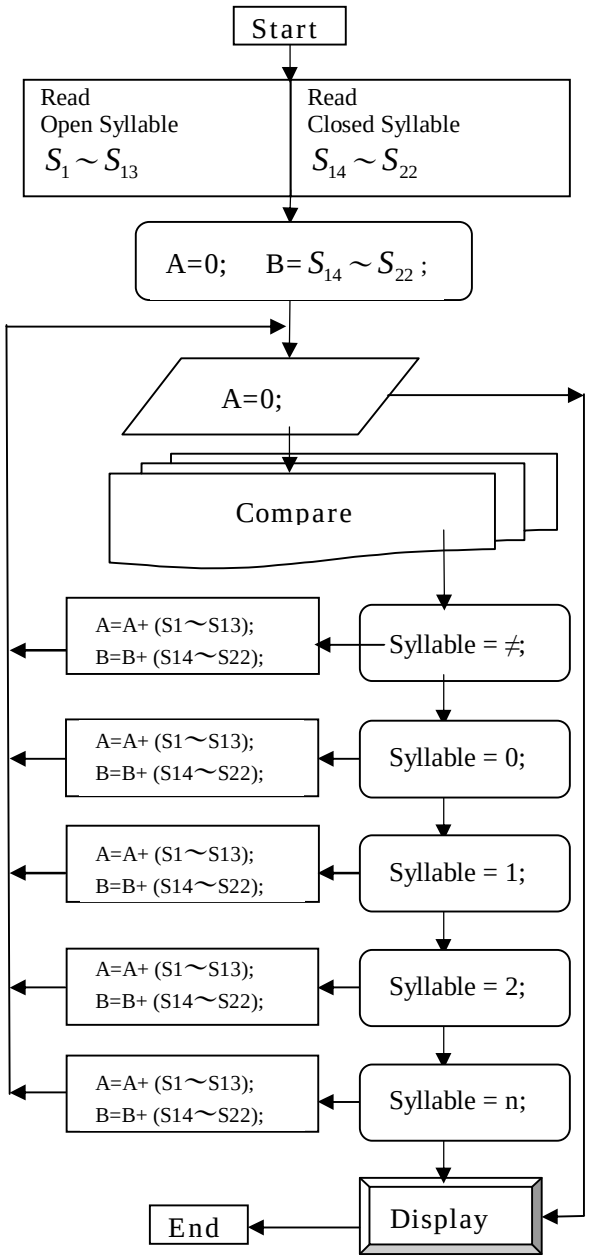
(4) 問題を報告する部分

(5) データを記録する部分

4 実装

中期ウイグル語の音節構造を量的分析するために、開音節 S_1 ~ S_{13} , 閉音節 S_{14} ~ S_{22} とウイグル文字 V, C, D, P, L,

R, Ss, Bs の音節分割アルゴリズムの規則を設計した。以下に示す。



4.1 2 文字の連接アルゴリズムの規則

2 文字の連接アルゴリズムの規則を基礎とし中期ウイグル語のテキストを実装して、音節構造の量的分析を行った。

表 11 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

ウイグル語テキストの 2 文字の連接では、V+V=1, V+C=1, V+D=1, V+P=1, V+L=1, V+R=1, V+Ss=1, V+Bs=1, C+V=1 の連接は 1 音節になる。これ以外の文字は音節に成らない。

表 12 ウイグル文字の 2 文字の連接アルゴリズムの規則

	V	C	D	P	L	R	Ss	Bs
V	1	1	1	1	1	1	1	1
C	1	0	0	0	0	0	0	0
D	1	0	0	≠	≠	≠	≠	≠
P	1	0	≠	0	0	0	0	0
L	1	0	≠	0	0	0	0	0
R	1	0	≠	0	0	0	0	0
Ss	1	0	≠	0	0	0	0	0
Bs	1	0	≠	0	0	0	0	0

4.2 3 文字の連接アルゴリズムの規則

ウイグル語テキストの 3 文字の連接では、V+ CV =2 の連接は 2 音節になる。

表 13 ウイグル文字の 3 文字の連接アルゴリズムの規則

	V	C	D	P	L	R	Ss	Bs
VV	≠	1	1	1	1	1	1	1
CV	2	≠	0	0	0	0	0	0
DV	1	1	0	≠	≠	≠	≠	≠
PV	1	1	≠	0	0	0	0	0
LV	1	1	≠	0	0	0	0	0
RV	1	1	≠	0	0	0	0	0
SsV	1	1	≠	0	0	0	0	0
BsV	1	1	≠	0	0	0	0	0

5 まとめ

文化遺産である文献作品クダトクビリクの右から左の横書きのウイグル語テキストを用いて音節の実装を行った。中期ウイグル語と現代ウイグル語の総合互換する際の難しさは、中期ウイグル語の開音節の一部が現代ウイグル語の閉音節に変化した。また、中期ウイグル語のグリフの独立形式の一部が現代のウイグル語では語中形と語末形に変化した事に起因する。C#を使って二つ言語の音節構造を比較的分析して a=C+V+C+V=CVCV と b=C+VC+C+V=CVCCV の交換は音節分割構造の交換の中心点である事がわかった。

参考文献

[1] Kutadgu Bilig, Istanbul, Alaeddin Kiral Basimeve, 1942.
[2] Omarjan Osman. ISO/IEC JTC 1/SC 2/WG 2 N3102/2008-10-23.
[3] Yoshiki Mikami. A History of Character Codes in Asia. 2002-03-20.
[4] Turfanforschung of manuscripts from the Berlin Turfan-Collection, Digitales Turfan-Archiv.
[5] Omarjan Osman. Syllable Segmentation rule for Asian language 1.0 professional Edition. 2010-12-01.