

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

ا	ب	پ	ق
I /i/ (4)	E /e/ (3)	Ah /æ/ (2)	A /a/ (1)
ك	خ	ع	ه
Uv /y/ (8)	U /u/ (7)	Ov /ø/ (6)	O /o/ (5)
ز	ص	س	و
Zhe /ʒ/ (12)	Te /t/ (11)	Pe /p/ (10)	Be /b/ (9)
ر	د	ذ	ف
Re /r/ (16)	De /d/ (15)	He /x/ (14)	Che /ʃ/ (13)
ش	ن	چ	ج
Shi / / (20)	Si /s/ (19)	Zhee / / (18)	Ze /z/ (17)
ك	ق	ف	غ
Ke /k/ (24)	Khi /q/ (23)	Fi /f/ (22)	Ghe /g/ (21)
م	ل	ن	ج
Me /m/ (28)	Le /l/ (27)	Ngi /ŋ/ (26)	Ge /g/ (25)
ي	و	ه	ن
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V̇、単独子音をC、長子音をCC、2重子音をĊ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, سۇ: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、V̇、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、CCV̇、CCVV

CV、CV̇、CVV

中期ウイグル語の開音節を以下のような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
		V̇			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。

V̇V̇

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
		V̇C			4
		VVCC			5
		V̇CC			6
		V̇Ċ			7
		VV̇Ċ			8

表5 中期ウイグル語の開音節のグループの例

	Open Syllable	Example		Code Point
1	V	ئ	u	E150
	VV	ئئ	uu	E150 E152
	V̇	ئ̇	ui	E150 E13B
2	VC	ئ	tu	E152 E168
	VVC	ئئ	too	E152 E143 E144
	V̇C	ئ̇	tou	E152 E143 E168
	VCC	ئئئ	ttu	E152 E169 E168
	VVCC	ئئئئ	ttoo	E152 E143 E169 E144
	V̇CC	ئ̇ئئ	ttou	E152 E143 E169 E168
	CV	ئ	ot	E16B E150
	CCV	ئئئ	ott	E16B E169 E150
	CVV	ئئئ	uut	E16B E151 E150
	VVCC	ئئئئ	uutt	E16B E151 E143 E150
	ĊV	ئ̇	uott	E16B E143 E150
	V̇CC	ئ̇ئئ	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, ياز:春)

中期ウイグル語の閉音節を以下のような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
		CV̇			4
		CCV̇			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
		CV̇C			6
		CV̇CC			7
		CCV̇C			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		Code Point
1	CV	توت	ot	E16B E142
	CVV	توتت	oot	E16B E143 E142
	CCV	توتت	ott	E16B E169 E142
	CCVV	توتتت	oott	E16B E169 E143 E142
	CV̇	توت	out	E16B E151 E142
	CCV̇	توتت	outt	E16B E151 E151 E142
2	CVC	توت	tut	E16B E151 E168
	CCVC	توتت	tutt	E16B E169 E151 E168

CVCC	توتت	ttut	E16B E151 E169 E168
CVVC	توتت	tutt	E16B E143 E151 E168
CVVCC	توتتت	ttuut	E16B E151 E151 E169 E168
CCVVC	توتت	tuutt	E16B E151 E151 E168
CV̇C	توت	tuot	E16B E143 E151 E168
CV̇CC	توتت	ttuot	E16B E143 E151 E169 E168
CCV̇C	توت	touutt	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て形成する。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接続規則

	a	ə	e	i	o	ö	u	ü
a	aa		aae aee	aaī aaii	aaō aoo		aau auu	
ə		əə		əəī əaii				
e			ee	eeī eii				
i	iia iaa	iīə iəə	iie iee	iī iio	iio ioo		iīu uii	
o	ooa oaa			ooī oii	oo		ooū ouu	
ö		ööə öəə		ööī öii		öö		ööü öüü
u	uua uaa		uue uee	uui uii	uuo uoo		uu	
ü		üüə üəə	üüe üee	üüī 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(V̇))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 “سۆ”, ai “سې”, ao “سو”, au “سۇ”
 əi “سېي”,
 ei “سېي”
 ia “سېي”, iə “سېي”, io “سېي”, iu “سېي”
 oa “سېي”, oi “سېي”, ou “سېي”
 öi “سېي”, öü “سېي”
 ua “سېي”, ue “سېي”, ui “سېي”, uo “سېي”
 üi “سېي”, üə “سېي”, üe “سېي”

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接続する事はできない。しかし、開音節と閉音節の後側には長子音を接続することができる。

例：(att, تاتت), (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(Ć))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、母音と接続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Ćが存在している。

例：(Halk, ھەلک: 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S₁からS₂₂までの22部分に分類する事が出来る。

S₁=X₁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S₁=X₁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
Ć		V		S ₄
C		VV		S ₅
CC		VV		S ₆
Ć		VV		S ₇
C		Ṽ		S ₈
CC		Ṽ		S ₉
Ć		Ṽ		S ₁₀
V		C		S ₁₁
VV		C		S ₁₂
Ṽ		C		S ₁₃
C	V	C		S ₁₄
C	VV	C		S ₁₅
C	Ṽ	C		S ₁₆
CC	V	C		S ₁₇
CC	VV	C		S ₁₈
CC	Ṽ	C		S ₁₉
Ć	V	C		S ₂₀
Ć	VV	C		S ₂₁
Ć	Ṽ	C		S ₂₂

S₂=(X₂+X₄)~S₁₃=(X₂+X₄)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。S₂=(X₂+X₄)

~S₁₀=(X₂+X₄)は前側が開音節である。S₁₁=(X₂+X₄)~S₁₃=(X₂+X₄)は後側が開音節である。(表10参照)S₁₄=(X₂+X₃+X₄)~S₂₂=(X₂+X₃+X₄)は閉音節である。表10のウイグル語の音節分割の規則のX₁, X₂, X₃, X₄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S₁=X₁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S₁₄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

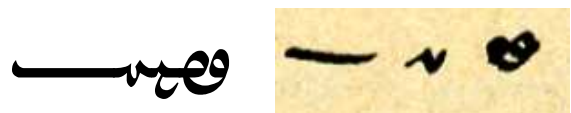
ウイグル中期ウイグル語の音節規則である後側が開音節S₁₁=(X₂+X₄)~S₁₃=(X₂+X₄)の規則は、現代ウイグル語の閉音節S₁₄=(X₂+X₃+X₄)に変化した。

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で、中期ウイグル語の独立形グリフを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, — ۱ ۰”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、 $a = S_{11} \sim S_{13}$ 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、 $B = S_{14} \sim S_{22}$ 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 $S_{14} \sim S_{22}$ 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)。

$Bs = 0020$;

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

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

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。

$$b = \left\{ \begin{array}{l} C + VC + C + V + Bs = CVCCV \\ C + VC + Bs + C + V + Bs = CVCCV \end{array} \right\}$$

$$0020 + 0020 + 0020 = 1;$$

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

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。

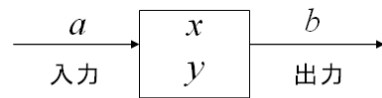
$$x.V = x_1, \quad y.C = y_1, \quad V.x = x_2, \quad C.y = y_2$$

$$x.V.x = x_3, \quad y.C.y = y_3, \quad x.VV = x_4$$

$$y.CC = y_4, \quad VV.x = x_5, \quad CC.y = y_5$$

$$x.VV.x = x_6, \quad y.CC.y = y_6$$

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種類のもの考えられる。

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

$$x.y = x_7, \quad x.xy = x_8$$

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

$$y.x = y_7, \quad y.yx = y_8, \quad y.xy = y_9$$

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 $S_{11} \sim S_{13}$ 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

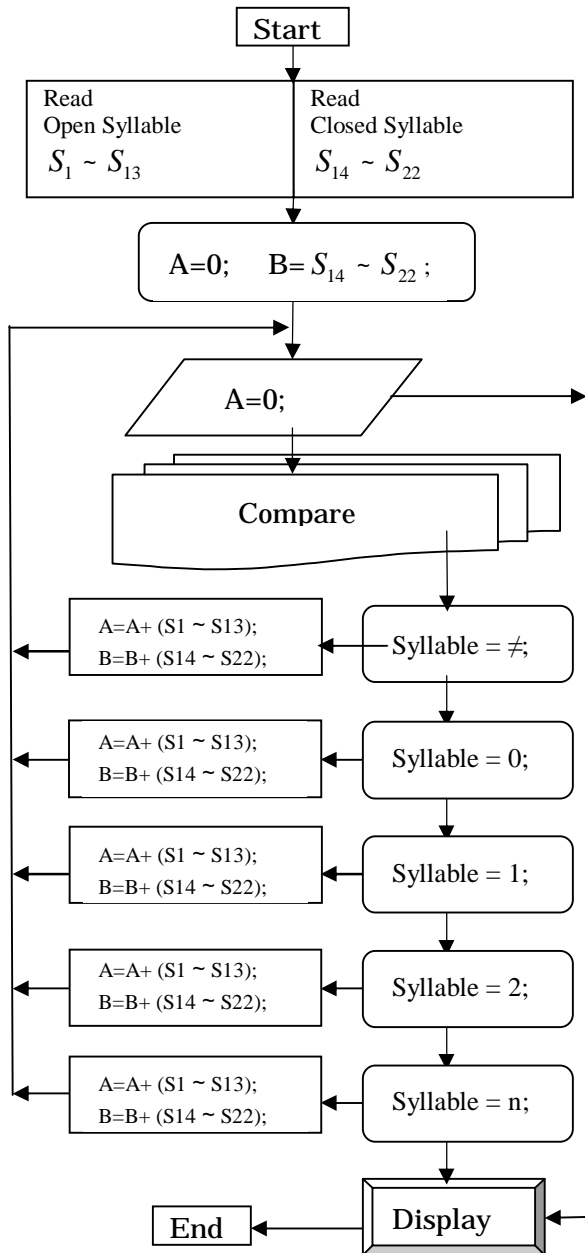
ソフトは三つ部分に分類される。

- (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 $S_1 \sim S_{13}$ 、閉音節 $S_{14} \sim S_{22}$ 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文字の接続アルゴリズムの規則

2文字の接続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, …, n)	3つの以上の文字の分割

ウイグル語テキストの2文字の接続では、 $V+V=1$, $V+C=1$, $V+D=1$, $V+P=1$, $V+L=1$, $V+R=1$, $V+Ss=1$, $V+Bs=1$,

$C+V=1$ の接続は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文字の接続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文字の接続アルゴリズムの規則

ウイグル語テキストの3文字の接続では、 $V+CV=2$ の接続は2音節になる。

表 13 ウイグル文字の3文字の接続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ピ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.