ローマ字によるビルマ文字入力方式

丁 塵辰　　内山 将夫　　隅田 英一郎

国立研究開発法人 情報通信研究機構
先進的音声翻訳研究開発推進センター 先進的翻訳技術研究室

{chenchen.ding, mutiyama, eiichiro.sumita}@nict.go.jp

1　はじめに

ビルマ（ミャンマー）文字はアブギダ文字の一種であり、主にビルマ語の表記に用いられる文字である。 現在、Unicode 方式によるビルマ文字の符号化はまだ十分に普及しておらず、旧式タイプライターから流用してきた字形による組立方式が広く使われている。文字の符号化方式に生じる混乱を避け、また利用しやすい入力インターフェースを提供するため、本研究はローマ字によるビルマ文字の入力手法を考案した。具体的、Unicode における各ビルマ文字の名称、及び「ミャンマー言語委員会」の制定した転写方式 [1] に参照し、一般的に受容されるラテン文字化方式を採用した。更に、広く使われる QWERTY 配列キーボードを考慮に入れ、頻繁に使われるビルマ文字（又は複数文字の組合）を打ちやすい位置に配置した。提案方式は、全体的に自然でありながら、効率を図るため若干記憶負担を要することになる。

2　入力方式

図 1 は提案方式を示している。 基本的に「h」と「g」キー、又は一部のキーの連打により文字変換を実現する。「a」、「o」、「x」の三つのキーには、出現頻度高い文字組合を配置した。「q」キーは綴り上に生じうる曖昧性を解消するため、区切り記号として用いられる。提案方式において、26 個のローマ字小文字により、63 個の Unicode ビルマ文字で構成する任意の文字列を入力できるようになる。ただし、正しい綴りの文字列が間違い綴りよりもっと入力しやすいである。

大文字おローマ字はショートカットとして配置された。利用者の好みにより、打鍵数・シフトキー利用回数のトレードオフが可能になる。シフトキー以外の修飾キーは一切利用しないことで、他のソフトウェア間に潜在的衝突を回避できるようになる。

Unicode 方式によるビルマ文字の符号化に、いくつかの文字に関する正規化処理が求められる。

- Unicode 文字「102C」と「102B」は同じ文字の二つ異字形である。「102B」が「1001」、「1002」、「1004」、「1012」、「1015」、「101D」の後に使われる。 この 2 字形の使い分けを自動的に行う。
- Unicode 文字「1037」と「103A」が同時に出現する場合は順番が決まっていない。 Unicode の推奨する「1037 + 103A」が必ずしも遵守されないため、この 2 文字の順番の調整を自動的に行う。
- Unicode 文字「1025」と「102E」の組合は「1026」に自動的に変換する。
- Unicode 文字「102A」は「101E」と「103C」の組合と同じ字形になり、すでに提案方式に取り入れられる。「1029」は「102A」、「1031」、「102C」、「103A」の組合まで分解可能なので、各文字ばらばら入力される場合「1029」に自動的に変換する。

3　実データによる考察

本節で従来のビルマキーボードレイアウトと提案方式における打鍵頻度の分布を調べる。

ビルマ語は口語と文語の二つ文体に分けられる。これに応じて、2 種類のコーパス上で考察を行った。一つは文語表現の *Asian Language Treebank* (ALT) [2] に含まれるビルマ語データであり、もう一つは口語表現の *Basic Travel Expression Corpus* (BTEC) [3] データである。ALT データは 2 万文の新聞記事で構成され、BTEC データは 40 万文の旅行対話文で構成される。

図 2 は ALT データ上に従来のビルマ文字キーボードにおける打鍵頻度の分布を示す。この配置はキーボード上のローマ字と関係がない。図 3 は小文字のみを

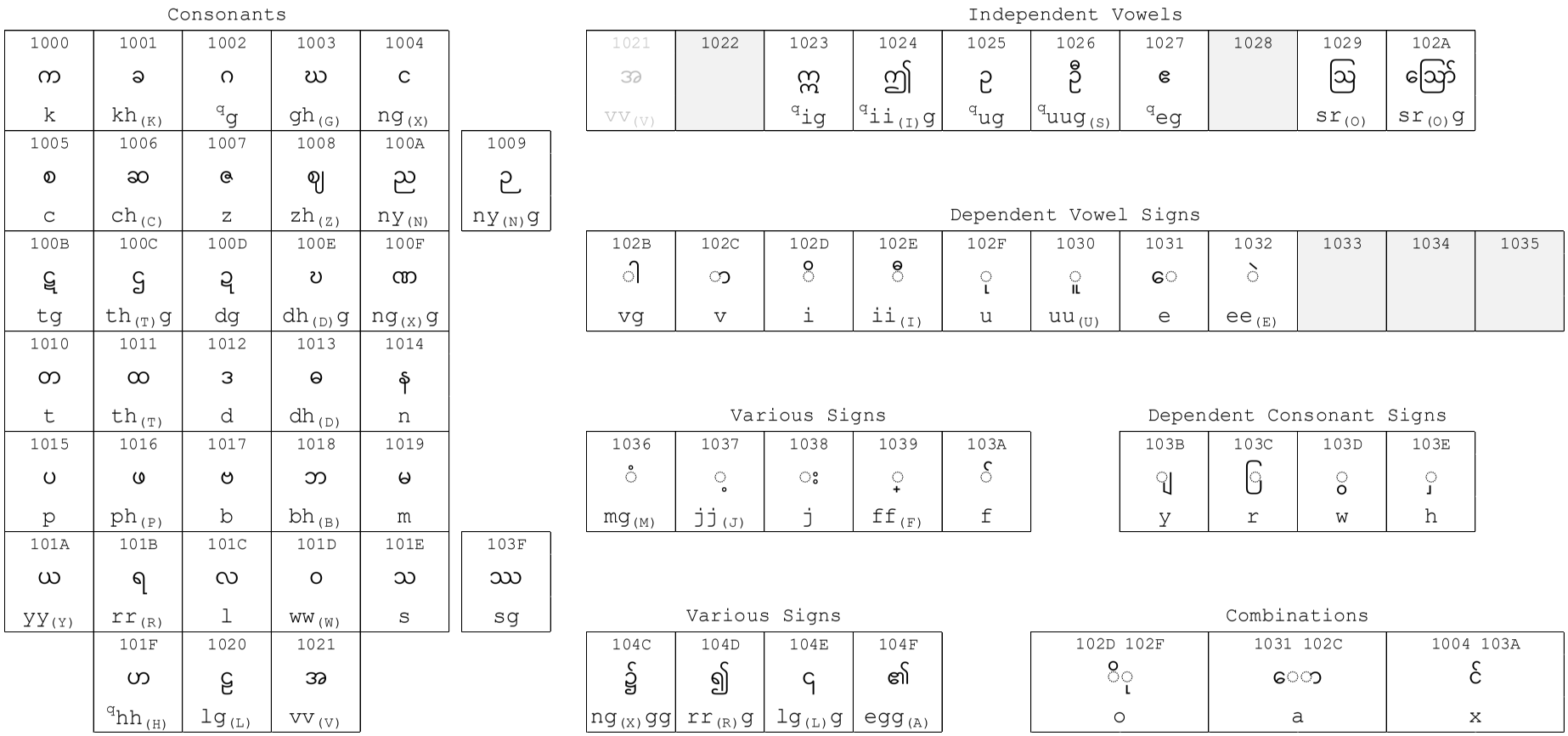

図 1: 提案の入力方式。各セルに上から下に「Unicode コード」・「ビルマ文字」・「ローマ字による打鍵」が示されている。括弧内部の大文字はショートカットである。上付き文字の「q」は曖昧性を解消するため利用しうることを示す。

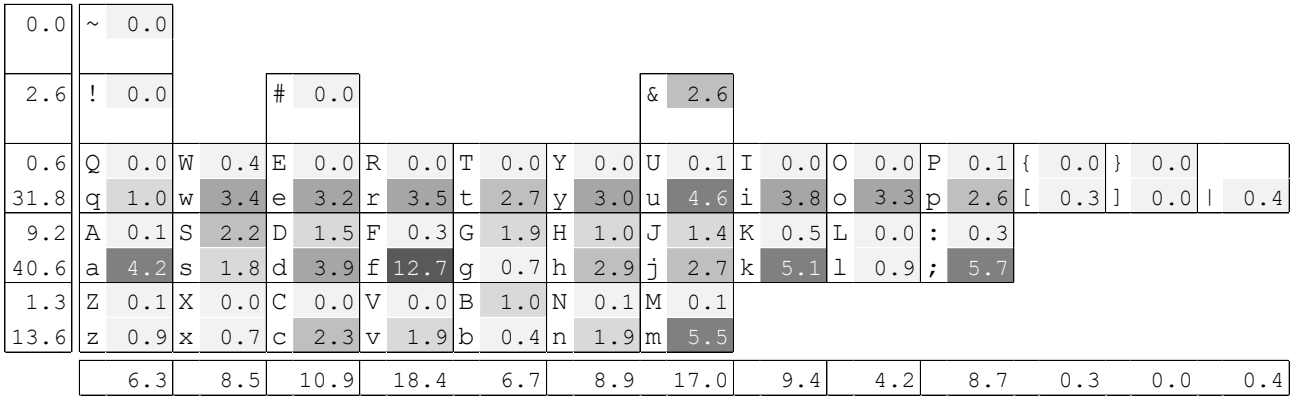

図 2: ALT データ上従来方式の打鍵頻度（%）。

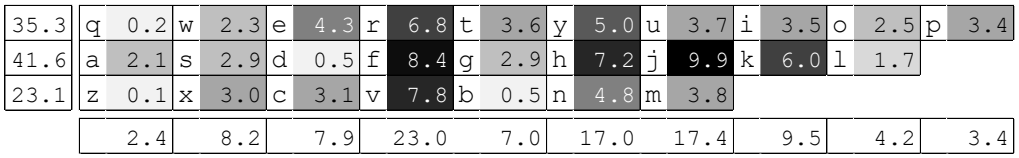

図 3: ALT データ上提案方式の打鍵頻度（%）。 大文字のショートカットが利用されない。

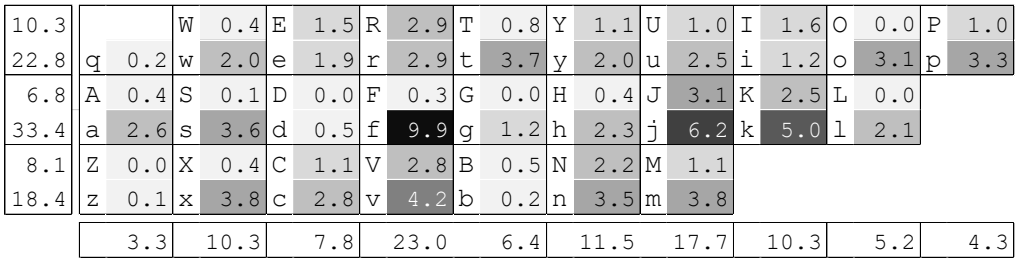

図 4: ALT データ上提案方式の打鍵頻度（%）。 すべて大文字のショートカットが利用される。

用いる場合において提案方式の打鍵頻度分布である。打鍵がキーボード中央に集中していることが見える。打鍵の分布がコンパクトになると共に、図 2 より図 3 の打鍵数が 16.1% に増える。これは一部文字の入力は複数回の打鍵を要することからである。図 4 は提案方式にすべて可能の大文字ショートカットを利用する場合の分布である。コンパクトでありながら、従来方式の図 2 より 7.8% の打鍵数が節約できる。これはシフトキーを従来方式より頻繁に使うからである。打鍵時の指の負担を直観的に示すため、シンプルな重み付き数値化を試算した。重みとして「fjdk」を 0 に、「asghl;ertuicvnm,」を 1 に、他のキーの重みを隣接するキーの中の一番小さい重みに 1 を足すことにし、シフトキーの重みを 2 にする。以上の重み付きにより、図 2 より図 3 は指の負担が 13.9% 低減し、図 4 は 4.6% 増加することが分かる。利用者の好みにより、重みの振り方が変わる一方、図 3 と図 4 は基本的に打鍵数・シフトキー利用回数のトレードオフを可能にしている。

図 5、6、7 はぞれぞれ図 2、 3、4 に対応する BTEC データ上の分布となる。全体的に同じ傾向が見られる。図 5 と比べ、図 6 には 13.0% の打鍵数が増え、図 7 には 8.0% の打鍵数が減る。ALT と BTEC データ上の分布と比べ、「s」と「k」キーの利用頻度には最も差異があることが見られる。「s」で入力 Unicode 文字「101E」は文語に頻繁に使われる助詞に現れるので、口語である BTEC データ上の利用頻度が著しく下がる。「k」で入力 Unicode 文字「1000」はビルマ語に「私」に該当する第一人称代名詞の各変体に出現するので、口語である BTEC データにより頻出する。前述の重み付き数値化から、図 5 より図 6 は指の負担が 16.0% 低減し、図 7 は 2.1% 増加する。

以上の 2 コーパス上の考察により、異なる文体にも拘わらず、従来方式より提案方式には指使いがもっと合理的・効率的であるが分かる。

0.0	~ 0.0												
2.1	! 0.0		# 0.0				& 2.1						
0.4	Q 0.0	W 0.3	E 0.0	R 0.0	T 0.0	Y 0.0	U 0.0	I 0.0	O 0.0	P 0.1	{ 0.0	} 0.0	
31.9	q 0.9	w 4.3	e 3.9	r 2.9	t 2.5	y 3.1	u 7.0	i 3.0	o 1.7	p 2.3	[0.3	] 0.0	| 0.0
10.2	A 0.2	S 1.6	D 1.9	F 0.3	G 2.2	H 0.8	J 1.7	K 1.2	L 0.0	: 0.3			
41.1	a 4.7	s 2.4	d 4.4	f 11.8	g 1.0	h 2.9	j 1.9	k 5.6	l 0.9	; 5.5			
1.7	Z 0.3	X 0.0	C 0.0	V 0.0	B 1.3	N 0.1	M 0.0						
12.7	z 0.6	x 0.6	c 1.7	v 2.0	b 0.9	n 0.7	m 6.2						
	6.7	9.2	11.9	17.0	7.9	7.6	18.9	9.8	2.6	8.2	0.3	0.0	0.0

図 5: BTEC データ上従来方式の打鍵頻度（%）。

36.2	q 0.1	w 2.4	e 4.9	r 5.5	t 4.3	y 5.2	u 3.4	i 4.1	o 3.1	p 3.2
40.9	a 2.2	s 1.5	d 1.1	f 8.6	g 1.8	h 6.2	j 10.0	k 7.7	l 1.8	
22.6	z 0.3	x 2.4	c 2.8	v 8.5	b 0.9	n 4.5	m 3.2			
	2.6	6.3	8.8	22.6	7.0	15.9	16.6	11.8	4.9	3.2

図 6: BTEC データ上提案方式の打鍵頻度（%）。 大文字のショートカットが利用されない。

10.2		W 0.3	E 1.8	R 2.4	T 0.7	Y 1.5	U 0.9	I 2.0	O 0.0	P 0.6
24.7	q 0.2	w 2.4	e 2.3	r 2.1	t 4.6	y 2.6	u 2.2	i 1.0	o 3.9	p 3.4
5.6	A 0.0	S 0.0	D 0.0	F 0.3	G 0.0	H 0.3	J 3.2	K 1.8	L 0.0	
34.3	a 2.8	s 1.8	d 1.3	f 9.9	g 0.9	h 1.8	j 6.0	k 7.6	l 2.2	
6.8	Z 0.0	X 0.4	C 1.0	V 2.6	B 1.0	N 0.9	M 0.9			
18.4	z 0.3	x 2.9	c 2.5	v 5.2	b 0.2	n 4.2	m 3.1			
	3.3	7.8	8.9	22.5	7.4	11.3	16.3	12.4	6.1	4.0

図 7: BTEC データ上提案方式の打鍵頻度（%）。 すべて大文字のショートカットが利用される。

4 おわりに

本研究は、ローマ字を用いるビルマ文字入力手法を提案した。ビルマ文字の出現頻度、及び QWERTY キーボードのレイアウトを考慮に入れ、効率化を図った。提案方式において、打鍵数・シフトキー利用回数間のトレードオフはカスタマイズできる。

現在、提案方式の実用化に取り掛かっている。利用者のフィードバックによる調整、又は辞書などの言語資源の導入はこれからの課題となる。

5 謝辞

ミャンマーの University of Computer Studies, Yangon の自然言語処理研究室の協力に感謝する。ソフトウェア開発に協力をいただく株式会社シルク・ラボラトリの森田裕樹氏に感謝する。

参考文献

[1] Department of the Myanmar Language Commission, *Myanmar-English dictionary (Myanma-anggalip abidan)*. Ministry of Education, the Republic of the Union of Myanmar, 12 ed., 2014.

[2] H. Riza, M. Purwoadi, T. Uliniansyah, A. A. Ti, S. M. Aljunied, L. C. Mai, V. T. Thang, N. P. Thai, V. Chea, R. Sun, S. Sam, S. Seng, Khin Mar Soe, Khin Thandar Nwet, M. Utiyama, and C. Ding, “Introduction of the Asian language treebank,” in *Proc. of O-COCOSDA*, pp. 1–6, 2016.

[3] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, “Creating corpora for speech-to-speech translation,” in *Proc. of EUROSPEECH*, pp. 381–384, 2003.