4D-6

チベット文字 TrueType フォントのレガシ符号化方式とその自動識別可能性

鈴木 俊哉 †　　佐藤 大 ‡
† 広島大学情報メディア教育研究センター　　‡ 東北大学病院メディカル IT センター

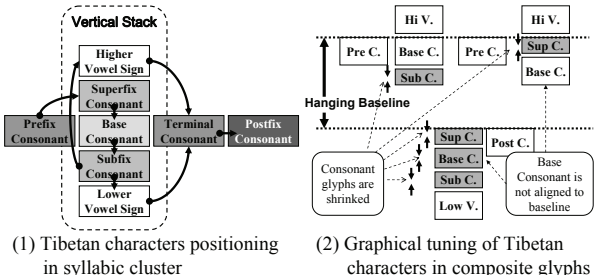
1　背景

南アジア・東南アジア・中央アジアで用いられる，ブラフミ文字を起源とする文字は，基底文字となる子音文字に対し上下左右に母音記号や声調記号を配置し，複雑な合字規則を持ち，さらに配置の順序は発音の順序と一致しないという特徴を持つ．従って，ISO 10646 文字集合を使った Unicode があるが，符号化文字列から字形を決定し描画する処理 (レンダリング) が複雑となり，広く利用されるには到っていない [1]．インターネット上で流通しているブラフミ系文字のデジタルドキュメントの多くは，音素ではなく，図形単位で分解した文字 (図形文字) を表示順序に符号化したテキストと，図形文字を表示するためのフォント (レガシフォント) によって構成されている．我々は，レガシフォントによるデジタルドキュメントの流通例として，南方ブラフミ系文字に属するクメール文字のレガシフォントについて調査し，レガシ符号の自動識別方法を提案しているが [2]，これと対照的な性格を持つ北方ブラフミ系文字の例として，チベット文字についての調査と自動識別可能性について報告する．

2　チベット文字レンダリングの問題

南方ブラフミ系であるクメール文字の Unicode フォントでは音素文字 114 個に対しの約 2 倍の字形が提供されており，レガシ符号もほぼ全てが 8 ビット符号単一フォントであった．しかし，北方ブラフミ系文字の代表であるデヴァナガリ文字の Unicode フォントでは音素文字 106 個に対し約 6 倍の字形，チベット文字の場合には約 20 倍の図形文字数が提供されている．チベット文字では要求される図形文字の数が多いため，レガシ符号フォントの設計方針も異なる可能性がある．

チベット文字におけるリガチャ合成は，他のブラフミ系文字と同様に，複合子音の表記のために導入されたものである．チベット文字において複合子音を表記する際，追加の子音字を基底子音字の上下左右 (右側には 2 つの子音文字を付加する場合もある) に配置する．



(1) Tibetan characters positioning in syllabic cluster　　(2) Graphical tuning of Tibetan characters in composite glyphs

図 1: チベット文字の合成規則

Encoding of Non-Unicode Tibetan Fonts and Auto Detection
†suzuki toshiya　‡Sato Dai
†Information Media Center, Hiroshima Univ.
‡Strategic Informatics Center, Tohoku Univ. Hospital

Encoding		N_f	N_{glyph}	備考
7/8bit	Ü-chan	1	70	Wylie 転写
	Tsampa	1	92	Wylie 転写
	GPLTibetan	4	126 - 135	
	TCRC	3	187	
	LTibetan	7	151 - 221	
	Sambohta Web	1	221	
	Tibetan Modern A	1	222	
	cTibTeX (ctib)	1	243	合成済 gtib
	TLK	3	250	内部符号
multi 7/8bit	Sirlin TeX (gtib)	1	158	2 個 1 組
	THF	1	182	2 個 1 組
	SUZTIB	1	312	16 個 1 組
	RABTEN	1	894	6 個 1 組
	TibetanMachine Web	1	915	10 個 1 組
	TibetanMachine	1	1010	5 個 1 組
16bit	Yagpo!_Wylie	1	1045	簡体字環境
	RABTEN Web	1	2059	
	TibetBT	1	4601	簡体字環境

表 1: 調査したフォントの符号化方式．フォント数 (N_f)，グリフ数 (N_{glyph}).

その規則を図 1 に示す．母音記号は上加字 (Superfix) の上，または，下加字 (Subfix) の下にしか配置されない．また，文字を横方向に並べる際，基本的には多くのブラフミ文字と同様に基字 (Base Consonant) の位置をぶらさがりベースラインに揃えるが，上加字が結合している場合は，基字ではなく上加字を揃える．その結果，欧文レンダラを前提として音素文字を図形符号化しようとすれば，基字も様々なメトリクスのものを用意しなければならない．

3　チベット文字レガシフォントの調査結果

チベット仏典の研究と関連するため，チベット文字の情報処理はブラフミ系文字の情報処理の中でも最も長い歴史を持つものの 1 つである．早期にデジタル化されたチベット語データは，基本的に 特定の専用アプリケーションで作成および利用されていたため，多様な字形を利用するため，8 ビット符号化された複数フォントを切替えるなど，汎用アプリケーションの流用という目的には合致しない方式も存在する．本稿では，インターネット上で無償配布されている TrueType フォントについて調査した．対象フォントの一覧を表 1 に示す．

8 ビット単一フォントのうち，結合文字が最も少なく，重ね打ちを多用する TLK の符号表を図 2 に示す．このフォントでは，1 つの子音文字の基字形に対して 4 つの異なるメトリクスが想定され，4 つのコードポイントで符号化されている．全てのブロックで，文字の配列は辞書順である．このように 8 ビット単一フォントのレガシ符号は，音素に対応する図素文字を様々なメトリクスで符号化し，重ね打ちによって結合文字を表示する手法で設計されている．チベット文字の場合，縦方向に 3-5 文字の重ね打ちが生じるので，結合文字が十分な品質に達するように図素文字を整理することに困難がある．

重ね打ち用の図素設計の困難を回避するため結合済

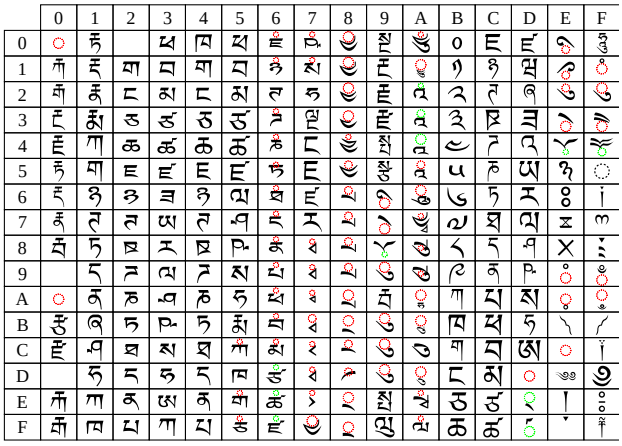


図 2: TLK フォント符号 (使用フォント:Kailasa)

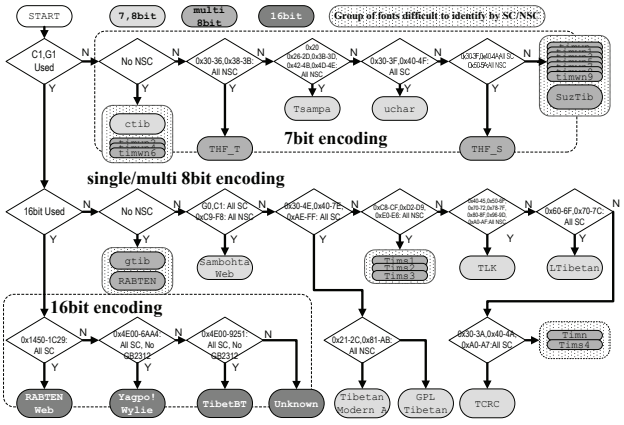


図 3: チベット文字レガシフォントの符号識別手順

みの字形を符号化するには，8 ビット複数フォントか 16 ビット単一フォントの方式が考えられる．今回調査したフォントでは，複数フォントによるレガシ符号は Sirlin，THF，SuzTib，RABTEN，Tibetan Machine，Tibetan Machine for Web，である．これらのほとんどは特殊なレンダラは用いずに，8 ビット単位でしかコードポイントを扱えない処理系でチベット文字入力方式を実装するために設計されたものである．フォント数が多く情報交換性には問題があり，Web 上で使用例が確認されたのは THF と Tibetan Machine for Web のみである．

入力方式に 8 ビット制限がなければ，16 ビット単一フォントにより結合文字を提供する方式がもっとも単純である．RABTEN Web，Yagpo!_Wylie，TibetBT が 16 ビットレガシ符号である．RABTEN Web 符号は，RABTEN 符号で表示可能な全ての結合文字を結合済み字形の 16 ビットの文字集合として定義したものである．Yagpo!_Wylie および TibetBT は簡体中文環境用に設計された符号化方式であり，Unicode における CJK 包摂漢字のブロック (0x4E00-0x9F00) で，GB2312:1980 と衝突しないコードポイントに結合済み字形を配置したものである．

4 符号化方式の自動識別

今回調査したチベット文字レガシ符号について，形状認識的な機能を用いずに，各符号化位置のメトリクス種別のみで符号化方式を識別するアルゴリズムについて図 3 に一例を示す．各符号化位置について得られる情報は，未使用コードポイント，通常文字に使用 (Spacing Character)，重ね打ち文字に使用 (Non Spacing Character) の 3 種類である．識別手順として，7 ビット符号，8 ビット符号，16 ビット符号に分類した後，SC または NSC が集中的に並んでいるブロックを探すことで識別を行なうこととし，より大きなブロックに SC または NSC が集中しているものを先に候補から除外するよう設計した．また，未使用コードポイントが外字収録に利用される可能性を考え，比較の際には空白文字以外の使用済みコードポイントに限定してメトリクス情報を利用した．このアルゴリズムにより，調査された 8 ビット単一フォント用符号，16 ビット単一フォント用符号は完全に識別できる．しかし，8 ビット複数フォント用の符号について，たとえば Tibetan Machine 符号を構成する 5 個のフォントのうち，どのフォントであるかを特定することはできなかった．これは，1 つの符号化方式を選んだ場合，それを構成するフォント群において，使用されるコードポイントの分布が類似しており，また，結合文字を符号化しているためにメトリクスが全て SC であるため，フォント群のうちのどのフォントであるかを識別することが困難なためである．

5 まとめ

本稿では，レガシ符号チベット文字フォント 65 個 (31 書体) を調査し，18 種の符号化方式を得，図形認識機能を用いずに符号化方式を識別する手法を提案した．提案手法では，8 ビット単一フォント用符号，16 ビット単一フォント用符号を特定可能である．これにより，北方ブラフミ系文字でも，クメール文字と同様に図形認識を用いずに高速に符号化が識別できる可能性が示された．今後の課題としては，8 ビット複数フォント用符号の識別の問題が残る．8 ビット複数フォント用符号では，基本的には結合文字を符号化しているため，メトリクスが全て SC となり情報量が少ないため，今回の手法では完全な識別ができていない．より詳細なメトリクスの比較が課題と考えられる．また，今回の調査では 16 ビット単一フォントの符号は 3 種類しか得られなかったが，中文 DOS 上のチベット語環境では多くの 16 ビット符号が用いられていた報告があり [3]，さらに調査が必要である．

参考文献

[1] Yoshiki Mikami and Zavarsky Pavol. Writing systems and character codes in the world (2). *IPSJ Magazine*, Vol. 46, No. 9, pp. 1046–1052, 2005.

[2] suzuki toshiya and Sato Dai. インド系言語レガシデータの符号化方式自動識別. 情処研報, Vol. 2006-DD-56, No. 1, pp. 1–8, 2006.

[3] Chen Yu-Zhong and Yu Shi-Wen. Tibetan information processing: Past, present, and future. *Proceedings of China-Japan Joint Conference to Promote Cooperation in National Language Processing 2002*, pp. 336–345, 2002.