社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－ＤＤ－62（７）
2007／7／27

レガシ符号チベット文字TrueTypeフォントにおけるマルチバイト文字符号とその自動識別可能性

鈴木 俊哉† 佐藤 大††

† 〒739-8511 東広島市鏡山 1-4-2 広島大学情報メディア教育研究センター
†† 〒980-8574 仙台市青葉区星陵町 1-1 東北大学病院メディカル IT センター
E-mail: †mpsuzuki@hiroshima-u.ac.jp, ††satodai@sic.med.tohoku.ac.jp

あらまし 情報交換を目的とする文字集合の規格として，現在は ISO 10646 文字集合，文字符号化方式としては Unicode を用いることが国際的な標準である．しかし，インド系文字は音素分解にもとづき文字符号を定義しているため，表示・印刷には複雑なレンダリング処理が必要となる．広く普及しているローマ字専用の処理系ではこれが不可能であるため，ローマ字用の処理系でもインド系文字が扱えるよう，様々な図形分解に基く符号化方式が応急処置的に作成されてきた．これらにより符号化されたデータが広く配布されているが，符号化方式は標準化されていないため，文書のアーカイブやデータ抽出に問題がある．本稿は，北方ブラフミ文字の例としてチベット文字のレガシ方式をとり，符号化方式の分析と自動識別方法を検討する．チベット文字レガシ符号の中には漢字集合に匹敵する大きさを持つものがあり，クメール文字とは異なり PDF などではフォントがサブセット化されている可能性がある．このような場合の識別方式についても検討する．

キーワード チベット文字，Unicode，TrueType フォント，情報交換，アーカイブ

Legacy Multi-Byte Encodings for Tibetan Scripts and Auto-Detection.

suzuki toshiya† and Dai SATO††

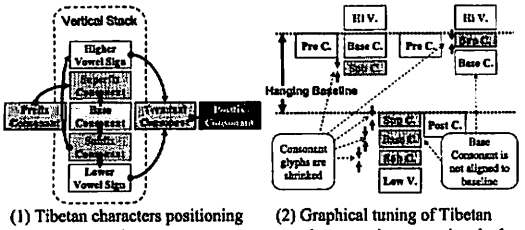
† Information Media Centrer, Hiroshima Univ., Kagamiyama 1-4-2, Higashi-Hiroshima-shi, 739–8511 Japan
†† Strategic Informatics Center, Tohoku Univ. Hosp. Seiryo-cho 1-1, Aoba-ku, Sendai-shi, 980-8574 Japan
E-mail: †mpsuzuki@hiroshima-u.ac.jp, ††satodai@sic.med.tohoku.ac.jp

Abstract For Brahmic scripts used in India, Central and South East Asia, ISO/IEC 10646 defines the character sets by phonetic decomposition of Brahmic scripts with refering Indic national standard IS 13194. These character sets expect the complicated text layout system to render the coded text for display or printing purpose. To reduce the implementation cost by reusing Roman text layout system, there are various non-standardized legacy encoding schemes for Brahmic scripts. In this report, we investigate the legacy encoding schemes for Tibetan script as an example of northern Brahmic script which uses many ligatures. We found legacy encoding system for precomposed Tibetan glyphs. The glyphset of these encodings are so large that the subsetted font in the documents are expected to be quite smaller than whole glyph set. We discuss the encoding detection algorithm by subsetted font.

Key words Tibet, Unicode, TrueType Font, Information Interchange, Archive

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，複雑な合字規則を持つ．国際標準である ISO 10646 文字集合では，ブラフミ系文字の殆どは音素単位での文字を定義し，これを用いたテキスト符号化方式の Unicode では合字生成や詳細な位置調整については表示系が処理するよう規格化されている．これはインドの国家規格である IS 13194:1991 に習った符号化である．しかし，このような符号化方式では，符号化テキスト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]．これらの用字系について，インターネット上で流通しているデジタル文書の多くは，図形単位で符号化した図形文字フォント（レガシフォント）と，この図形文字を表示順に符号化（レガシ符号化）したテキストで構成されている．しかし，このようなブラフミ系文字のレガシ符号については標準がほとんどないため，仕様の不明確な符号化方式が乱立しており，ローマ字レンダラによる符号化テキストの表示は容易となるが，そ



(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．(1) 母音記号 (Vowel Sign)，前加字 (Prefix Consonant)，上加字 (Superfix Consonant)，基字 (Base Consonant)，下加字 (Subfix Consonant)，後加字 (Terminal Consonant)，重後加字 (Postfix Consonant) の位置関係．(2) 実際のチベット語テキストレイアウトにおける上加字，基字，下加字の位置調整．

の情報交換性は低い．我々は，レガシフォントによるデジタル文書の流通例として，南方ブラフミ系文字に属するクメール文字のレガシ符号を調査し，その自動識別方法を提案しているが [2], [3]，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，チベット文字のレガシ符号の調査結果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特徴

2.1 北方系・南方系ブラフミ文字の相違点

ブラフミ文字はその歴史的発展の違いから，大きく北方系と南方系の 2 つに分けられる．南方ブラフミ系文字に比べ，北方ブラフミ系の文字では複合子音の表記に様々な結合文字を用いる．このため，ISO 10646 で定義される音素文字数に対し，表示用の図形文字数の比率は非常に大きくなる傾向が知られる．クメール文字のレガシ符号では，全てが 8 ビット符号単一フォント用に設計されていた [2]．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 倍の図形文字，さらにチベット文字の場合には約 20 倍の図形文字が提供されている．このように北方ブラフミ文字では図形文字の数が 8 ビット符号に収まり切らず，レガシ符号の設計方針もクメール文字などとは異なる可能性がある．

2.2 チベット文字の合成規則

チベット文字における結合文字の合成規則は，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は，母音記号は子音文字の上下にしか配置されないが，複合子音を表記する際，追加の子音字を基底となる子音字の上下左右に配置する．その規則を図 1 に示す．本稿では，中国の藏文学の用語に習い，各位置に配置される子音文字について，基底となる子音字との位置関係をもとに，

- 前加字 (基底子音字の左側)
- 上加字 (基底子音字の上側)
- 基字 (基底子音字)
- 下加字 (基底子音字の下側)
- 後加字 (基底子音字の右側)
- 重後加字 (後加字の右側)

と分類する．母音記号は上加字の上，または，下加字の下にしか配置されない．

チベット語を表記するために必要な文字集合は子音字 30 字，母音記号 4 字から成る．さらに，サンスクリット語の転写のために導入された子音字が 5 種類ある．全ての子音字は基字になりうるが，上加字になりうる子音字は 3 種類，下加字になりうる子音字は 4 種類に限られている．

単純に計算すると，縦方向の子音字の組み合わせ数は

	符号化方式	N_f	N_{glyph}	備考
8bit single	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	MacOS
8bit multi	Sirlin TeX (gtib)	1	158	7bit font × 2
	THF	1	182	7bit font × 2
	SUZTIB	1	312	7bit font × 16
	RABTEN	1	894	8bit font × 6
	TibetanMachine Web	1	915	7bit font × 10
	TibetanMachine	1	1010	8bit font × 5
16bit	Yagpo!_Wylie	1	1045	個体字環境
	RABTEN Web	1	2059	
	TibetBT	1	4601	個体字環境

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

上加字＋基字	3×30	$= 90$
基字＋下加字	30×4	$= 120$
上加字＋基字＋下加字	$3 \times 30 \times 4$	$= 360$
計		570

と見つもることができる．これに，4 種類の母音記号を組み合わせた数を考えると，その組み合わせ総数は 3000 個を越える．実際にはチベット語に存在しない発音の結合文字は作られないので、必要な結合文字数はこれよりも少ないが，全て合成済みグリフによって結合文字を提供しようとすれば，8 ビット符号では不可能なことは明らかである．現代チベット語の表記においては，縦方向の子音字の結合は上加字，基字，下加字の最大 3 文字の結合であるが，サンスクリット語表記や古典文献では 4 つ以上の子音字が結合する場合もあり，これらも含めると結合文字の総数はさらに増える．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をベースラインにぶらさげるように揃えるが，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，上加字・下加字の他に，字形自体は変化しない基字について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

チベット語およびチベット文字の電子化は仏典研究と密接に関連するため，ブラフミ文字の中でも最も長い電子化の歴史がある文字の一つである．従って，商用のレガシ符号フォントも多数存在する [4] が，本稿では，インターネット上で配布されている 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．クメール文字レガシフォントの調査 [2] と同様に基本的には Windows 用 16 ビット Unicode 符号表をもとに符号表を作成するが，MacOS 専用のファイル形式で配布されているものについては MacOS 用 8 ビット符号表を用いた．

クメール文字レガシ符号では，非互換な符号化方式でありながら，多くのレガシ符号の G0 面に類似した文字配列が見られフォント間の差異が見えにくい問題があった．本稿で調査したチベット文字レガシ符号には，そのような類似した文字配列は見られなかった．この一つの理由としてチベット語の結合文字の指定が複雑なために入力支援プログラムが必要であって，単にフォント内の文字配列によってキーマップを設定しても実用性に乏しいという問題が考えられる．

調査したフォントの符号化方式は，表 1 に示すように，大別して

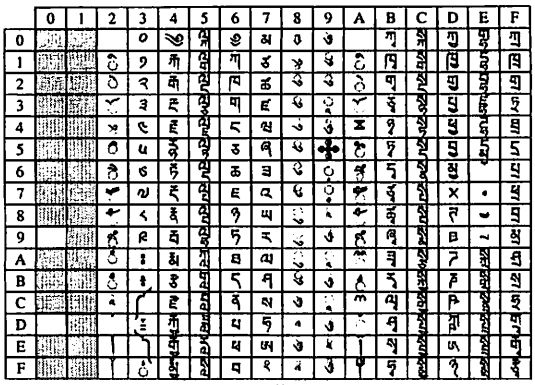


図 2 Tibetan Modern A 符号 (使用フォント: TIBMDA.TTF)

- 単一の 7 または 8 ビットフォント (合成用グリフ集合)
- 複数の 7 または 8 ビットフォント (合成用グリフ集合または合成済みグリフ集合)
- 単一の 16 ビットフォント (合成済みグリフ集合)

の 3 つであった．以下にそれぞれの詳細を整理する．この中で、図形文字をフォント中の文字をメトリクスによって以下の 2 つに分類する．

- Spacing Characer (SC): コードポイントに対して文字がわりあてられており，文字の幅が有限であって通常のローマ字のように印字されるもの
- Non Spacing Character (NSC): コードポイントに対して文字がわりあてられており，文字の幅がゼロか，または非常に小さく設定してあり，アクセント記号のように重ね打ちするもの

3.1 8 ビット単一フォント

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のは，レンダラを含めて無償配布されている大谷大学の TLK である．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．これにより，35 文字の子音字に対して合計で 140 個のコードポイントを消費し，8 ビット空間の大半は図案文字の重複符号化で占められている．また，下加字として異なる字形を持つ 4 つの文字については，下加字形を個別に 4~10 個のコードポイントで符号化されている．上加字形については個別には符号化せず，上加字と基字の合成済みのグリフを 25 個符号化している．また，複数の下加字の合成グリフ，下加字と母音記号の合成済みグリフも符号化されている．母音記号は，基字または下加字の下に配置される場合のみ複数のコードポイントで符号化されているが，上に配置される母音記号はメトリクス調整の必要が小さいため，多くとも 2 つのグリフしか符号化されていない．LTibetan 符号も TLK と同様に重ね打ちを多用するが，単体の子音文字は 109 個と TLK の 7 割程度に抑え，合成済みグリフにわりあてている．

TLK, LTibetan とも，コードポイントの半分以上を重ね打ち用の図案文字にわりあてているが，このような符号化方式では 1 つの結合文字を合成する際の重ね打ちの回数が増えるので，符号表が与えられたとしても互換性のある書体を設計するのは難しい．この問題を回避するため，より多くの合成済みグリフを収録した符号化方式がありうる。最も特徴的な例として Tibetan Modern A 符号を図 2 に示す．これと類似した方針で設計されたレガシ符号としては TLK との相異点は以下のように整理される．

- TCRC 符号では 211 文字中合成済みグリフが 45 字含まれているが，通常の子音文字の他に重ね打ち用の子音文字が NSC として符号化されている．
- Sambhota Web 符号では 211 字中 97 字が合成済みグリフであり，子音文字の一部が複数のコードポイントで符号化されているものの，メトリクスは SC であって重ね打ちはできない．
- Tibetan Modern A では 222 字中 99 字が合成済みグリフであり，子音文字は 1 個のコードポイントでのみ符号化されている．

これらの符号化方式では，重ね打ちが容易なものであってもできる限り合成済みグリフで符号化している．合成済みグリフを多数符号化したことによって，図案文字の数は減っており，重ね打ちにおけるメトリクスの調整ができないリガチャが増えていると考えられる．また，これらの合成済みグリフを主とするフォントの傾向として，合成済み文字の配列は発音順ではなく，表示順 (上加字がある場合は上加字，上加字がない場合は基字) の順番で並べられていることも特徴である [5, p. 312].

3.2 8 ビット複数フォント

重ね打ち用の図案設計の困難を回避するため結合済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．このうち，Sirlin, THF は合成済みグリフを含まない図案文字集合で，本来は 8 ビット単一フォントに収まる文字集合であるが，7 ビット符号に収めるために 2 つのフォントに分割したものである．また，SuzTib 符号もフォントファイル数は多いが，これはグリフの分類のために増えており，実際のグリフ数はファイル数 (16 個) に比べると少なく 312 個である．SuzTib 符号は合成済みグリフも含むが，312 個のグリフ中合成用の NSC が 87 個あり，8 ビット単一フォントの設計と同様と言える．

これらに対し，多数の合成済みグリフを含めたフォントとしては RABTEN, Tibetan Machine, Tibetan Machine for Web がある．これらのフォントにおいて，重ね打ち用の NSC グリフは，

- RABTEN 符号では全グリフ 894 個に対し NSC は 2 個
- Tibetan Machine for Web 符号においては全グリフ 915 個に対し NSC は 115 個，
- Tibetan Machine 符号においては全グリフ 1010 個に対し NSC は 131 個，

である．これらの符号では，少なくとも子音字合成に関しては合成済みグリフで表示するが，Tibetan Machine および Tibetan Machine for Web は母音記号については重ね打ちで表示する．ただし，重ね打ちに際してフォント切替が発生しないよう，各構成フォント内の合成済みグリフに重ね打ちし得る母音記号を全て収録している．このため，各フォント内で重複して収録されている NSC グリフがあり，NSC グリフの数は実際の描画データ数の 4~5 倍の数になっている．例えば，Tibetan Machine 符号を構成する 5 つのフォントで，`0xD2-0xD9` のコードポイントは同一の母音記号 (-u) に，`0xE0-0xE6` のコードポイントは同一の組み合わせ母音記号 (-uu) に，割り当てられている．

3.3 16 ビット単一フォント

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 ビット単一フォント用のレガシ符号である．

RABTEN Web 符号は，RABTEN 符号で表示可能な全ての

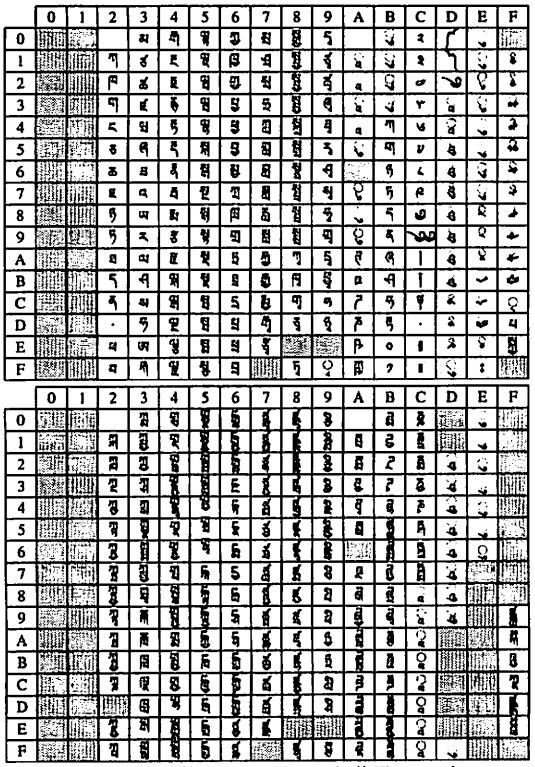


図 3 Tibetan Machine 符号の 0 面と 1 面 (使用フォント: Timn.ttf, Tims1.ttf)

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，使用するコードポイントは 0x1400-0x1CFF, 0xE000-0xE0FF, 0xE400-0xE4FF である．ほとんどの合成済みグリフは 0x1400-0x1CFF) に割り当てられている．

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 統合漢字のブロック (0x4E00-0x9F00) で，GB2312 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. フォントによる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の符号化方式を識別するアルゴリズムについて検討する．フォント中のあるコードポイントに注目した場合，先の文字メトリクスの種別 (SC と NSC) の他，明示的には文字がわりあてられていない状態 (未使用状態) がありうる．TrueType フォントの場合，この 3 つの状態を識別する情報 (コードポイントわりあて，および，文字のメトリクス) は，グリフ描画データとは別個に格納されており，ラスタライズせずに高速に抽出することができる．従って，各コードポイントに対する 3 つの状態の分布のみでフォントを識別することができれば，符号化テキストに対するビットパターンマッチングや，文書をラスタ化した後の文字認識といった方法よりも，高速かつ信頼性の高い符号化方式識別が可能となる．

4.1 完全フォントによる識別

まず，フォントが完全な形で取得できた場合の識別方式について整理する．ブラフミ系文字の 8 ビットレガシ符号の場合，G0, G1 面に本来存在しない管の重ね打ち用の NSC が多数配置されるため，文字のメトリクスにより少なくとも標準的な符号

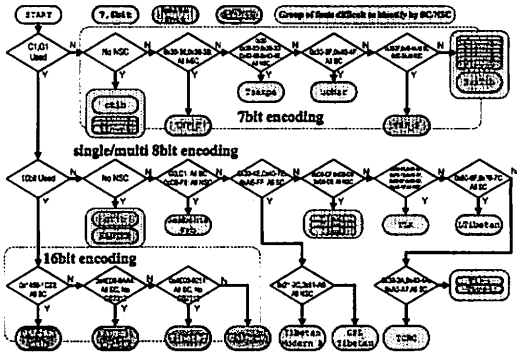


図 4 チベット文字レガシフォントの符号識別手順

化ではないことは容易に検知できる．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する．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する．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い．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 となり，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．以上を踏まえ，図 4 に識別アルゴリズムの例を示す．

チベット文字の合成済みグリフ集合を 16 ビット符号化した場合，全ての文字が SC となり，一般の CJK 漢字フォントと識別することが困難となる．また，簡体字とチベット文字の混植の需要が少なくないこと [6] を考慮すると，CJK 漢字フォントとチベット文字 16 ビットレガシ符号の識別は重要な課題である．本稿では，符号化テキストではなく，フォントのみの符号化方式識別を考えるので，チベット文字 16 ビットレガシ符号において使用されるコードポイント分布と，標準的な CJK 漢字フォントのコードポイント分布の比較からの識別方法を検討する．CJK 統合漢字ブロックにおいて，各国の国家規格に定義された漢字集合と，各用字系のフォントがグリフをわりあてているコードポイントが，どの程度重なっているか (交差数) を表 2 に示した．まず，チベット文字 16 ビットレガシ符号は GB2312 との交差数は 0 であることから，少なくとも簡体字フォントではないことは容易に決定できる．が，さらに台湾・日本・韓国の一般の漢字フォントでも GB2312 との交差数は 30%以上であり，これがほとんど 0 であるチベット文字 16 ビットレガシ符号とは明らかに分布が異なることがわかる．また，Yagpo 符号 (1045 字) と TibetBT 符号 (4601 字) は提供する文字数に大きな差があるが，各漢字集合との交差数の分布の割合は大きな違いがないことがわかる．例えば，GB7589, GB16500 との交差数はどちらの符号化方式でも，それぞれ 35%と 30%程度である．従って，交差数の分布では，完全なフォントが入手できている場合，Yagpo 符号および TibetBT 符号を CJK 漢字フォントから識別することは容易であるが，漢字集合との比較によってこの 2 つの符号化方式を識別することはできない．TibetBT 符号は

地域	漢字集合	台湾		中国		日本		韓国		チベット			
		bkai00mp.ttf		dwfszk.ttf		ipam.ttf		batang.ttf		Yagpo_2-4.ttf		Tibetbt.TTF	
台湾	CNS11643:1992-1	5410	41.4%	5413	19.7%	4473	66.8%	3873	79.2%	127	13.1%	641	14.0%
	CNS11643:1992-2	7650	58.5%	7650	27.9%	1056	15.8%	475	9.7%	452	46.7%	2088	45.5%
	CNS11643:1992-15	0	0.0%	246	0.9%	69	1.0%	12	0.2%	10	1.0%	41	0.9%
中国	GB2312:1980	4384	33.5%	6763	24.6%	3378	50.4%	2684	54.9%	0	0.0%	44	1.0%
	GB7589:1987	3598	27.5%	7226	26.3%	409	6.1%	196	4.0%	339	35.0%	1552	33.8%
	GB7590:1987	1639	12.5%	4068	14.8%	217	3.2%	77	1.6%	230	23.8%	1018	22.2%
	GB8565:1989	47	0.4%	332	1.2%	55	0.8%	32	0.7%	12	1.2%	61	1.3%
	GB12345:1990	2093	16.0%	2352	8.6%	1583	23.6%	1308	26.8%	122	12.6%	652	14.2%
	GB16500:1995	1307	10.0%	3778	13.8%	1040	15.5%	323	6.6%	265	27.4%	1265	27.5%
	HKscs	7	0.1%	2804	10.2%	501	7.5%	194	4.0%	109	11.3%	614	13.4%
日本	JISX0208:1990	5326	40.7%	6356	23.1%	6356	94.9%	4085	83.6%	223	23.0%	940	20.5%
	JISX0212:1990	3883	29.7%	5801	21.1%	279	4.2%	386	7.9%	335	34.6%	1553	33.8%
韓国	KSC5601:1987	4354	33.3%	4888	17.8%	4180	62.4%	4888	100.0%	136	14.0%	624	13.6%
	KSC5657:1991	2400	18.4%	2856	10.4%	1233	18.4%	0	0.0%	139	14.4%	619	13.5%
	PKSC5700-1:1994	5711	43.7%	7911	28.8%	902	13.5%	0	0.0%	506	52.3%	2243	48.8%
ベトナム	TCVN5773:1993	431	3.3%	732	2.7%	158	2.4%	93	1.9%	66	6.8%	210	4.6%
	VHN01:1998	3127	23.9%	3311	12.1%	2771	41.4%	2483	50.8%	92	9.5%	462	10.1%
	VHN02:1998	610	4.7%	914	3.3%	479	7.2%	391	8.0%	41	4.2%	151	3.3%

表2 CJK 漢字および 16 ビットレガシ符号チベット文字フォントの，各種漢字集合との交差点数

符号化方式	全文書数	全文字数	全文字種数	割合
Big5(*)	1232	18879141	7665	58.7%
GB2312(*)	4904	23380084	5918	87.5%
TibetBT(*)	2379	7233799	1581	34.3%
TCRC	1823	7208058	187	100%

表3 調査したレガシ符号チベット語文書と中国語文書の文字種数 (*) ASCII 文字は除外している

使用するが Yagpo 符号が使用しないコードポイントなどによって識別せざるを得ない．

4.2 サブセットフォントによる符号化方式識別

次に，PDF などの文書のように，チベット語レガシ符号フォントが文書中で使用されている文字のみにサブセット化された場合の識別可能性を検討する．まず，サブセット化が行なわれる環境では，フォントにおける文字集合がレガシ符号全体に対して，どの程度の部分集合となるかを調査する．本稿では，対象として以下の 3 つのサイトの HTML コンテンツを調査した．

- `www.tibet.cn`: GB2312 による中国語文書と TibetBT 符号によるチベット語文書
- `www.tibet.net`: TCRC 符号によるチベット語文書，英字文書など
- `www.xizang-zhiye.org`: `www.tibet.net` の中国語翻訳 (GB2312 または Big5 符号) 文書

表 3 に，調査した文書より得られた部分集合の割合を示す (ただし，HTML タグは除いた)．また，文書長 (文字数) と，それに含まれる文字種数の関係を図 5 に示す．

この結果から，8 ビットレガシ符号である TCRC 符号は，10000 文字程度の文書を取得すれば 187 個中 140 個程度のコードポイントをサンプリングできるが，16 ビットレガシ符号では，大量の文書を取得したとしても，サンプリングできる合成済みチベット文字は 2000 種未満，単一の文書の場合には 200~300 種程度になってしまうことがわかる．16 ビットのチベット文字レガシ符号は合成済みグリフであって，SC/NSC の情報は得られないため，コードポイントに対して重みをつけずに比較しても符号化方式の識別は困難である．

4.3 サブセット済フォントからの符号化方式識別

フォントからの符号化方式識別の利点はその確実性であるが，上で示したように文字集合に対しごく一部の分布情報しか得られない場合には，コードポイントの使用・未使用について重みをかけた上で評価しなければならない．今回の調査で見つかった 3 つの 16 ビットレガシ符号 (Yagpo, RABTEN Web, TibetBT)

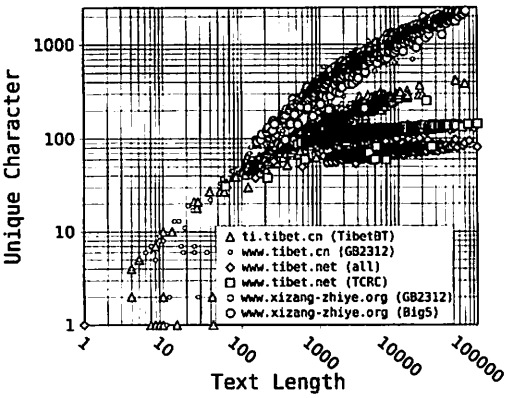


図5 レガシ符号チベット文字文書および中国語文書における文書長と文字種数の関係

をとると，RABTEN 符号は ASCII 文字と混在できない符号化であり，また，合成済みチベット文字をカナダ先住民族文字，オガム文字，ルーン文字などの領域にわりあてているため，標準的な符号化では短かい文書の中で混在する可能性は極めて低く，容易に除外できると考えられる．これに対し，Yagpo 符号と TibetBT 符号は，どちらも合成済みチベット文字を CJK 統合漢字領域における GB2312 の補集合を用いて符号化しているため，漢字符号とチベット文字レガシ符号の間での識別も問題となる．ここで，16 ビットのチベット文字レガシ符号の識別について以下の 2 つの課題が考えられる．

- GB2312 補集合であることが検出できれば少なくとも CJK 漢字フォントではなく，16 ビットのチベット文字レガシ符号である
- チベット文字特有のコードポイント偏差により，チベット文字レガシ符号を特定する

4.3.1 GB2312 補集合の検出

表 2 に示すように，台湾・日本・韓国のフォントは GB2312 漢字集合と 30~50%の交差確率を持っているので，漢字の出現頻度を無視した場合，7 個 (日本，韓国のフォントの場合) から 12 個 (台湾のフォントの場合) 以上の漢字をサンプリングし，全て GB2312 漢字集合に含まれていないものであれば，誤答率 1%未満で漢字文書ではないと推定できる．もし，高い頻度で使用される漢字の集合がこれらの用字系で共通であれば，さらに少ないサンプル数での識別が期待できる．図 5 で調査した文

頻度	Big5		GB2312		TibetBT		TCRC	
順位	出現率	文字符号	出現率	文字符号	出現率	文字符号	出現率	文字符号
1	3.39%	U+FF0C (，)	5.85%	U+3000 ( )	28.0%	0xFE3D [tsek]	22.7%	0x2D [tsek]
2	3.12%	U+7684 (的)	2.87%	U+FF0C (，)	5.52%	0x7E8D (Sa)	5.09%	0xC5 (Sa)
3	1.76%	U+3002 (。)	2.64%	U+85CF (藏)	4.40%	0x7E73 (Ga)	4.96%	0xDB (-i)
4	1.40%	U+85CF (藏)	2.16%	U+7684 (的)	4.17%	0x7E74 (Nga)	4.73%	0xF4 (-o)
5	1.13%	U+897F (西)	1.94%	U+897F (西)	3.84%	0x7E7B (Da)	4.70%	0x47 (Ga)
6	1.08%	U+4EBA (人)	1.40%	U+3002 (。)	3.50%	0x7E7C (Na)	4.41%	0x68 (rJa)
7	0.99%	U+4E00 (一)	1.25%	U+4E2D (中)	3.26%	0x7E7F (Ba)	3.72%	0x6D (Na)
8	0.94%	U+4E2D (中)	1.03%	U+3001 (、)	2.51%	0xFE40 [shey]	3.32%	0x7A (Ba)
9	0.90%	U+5728 (在)	0.77%	U+56FD (国)	2.35%	0x7E80 (Ma)	3.28%	0x50 (Nga)
10	0.85%	U+662F (是)	0.65%	U+4E00 (一)	2.18%	0x7E8A (Ra)	2.60%	0xBA (-a)

表 4　調査した中国語文書，レガシ符号チベット語文書において頻出した文字とその出現頻度

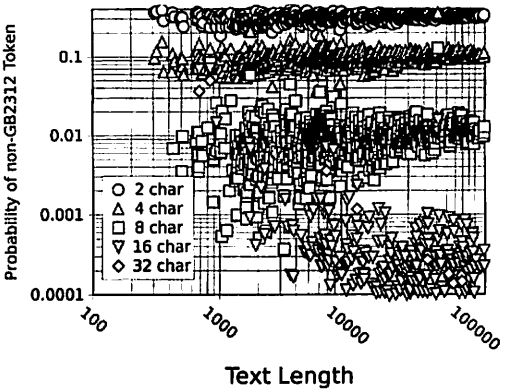


図 6　Big5 符号化文書における GB2312 漢字の出現確率

書のうち，Big5 で符号化された文書に対して，一定長の文字列を走査した際に，その中に GB2312 に属する漢字が 1 つも含まれていない確率を調べたものを図 6 に示す．この結果から，任意の Big5 符号化文書から 8 個の文字列を取った場合，GB2312 漢字が全く出現しない確率は期待値 1%以下となる．この数値は，出現頻度を無視した漢字集合の交差数から見つもられるサンプル数 (13 文字) よりも短い．以上の結果から，今回の調査で発見された GB2312 補集合を用いた 16 ビットのレガシ符号を識別するためには，CJK 統合漢字領域のコードポイントを 8 個以上サンプルすれば良いことがわかった．

4.3.2　チベット文字特有パターンの検出

次に，合成済みチベット文字のレガシ符号について考えた場合，もっとも特徴的なパターンは音節区切りに使われる Tsek 記号 (Unicode における U+0F0B) 挿入と予想される．レガシ符号化されたチベット語文書における文字の出現頻度を，TibetBT 符号と TCRC 符号について示すと，表 4 のようになる．両符号化方式とも Tsek 記号は 20%以上の出現率であるのに対し，その他の文字は 6%以下の出現率であるから，明確な偏りがあり，Tsek 記号の符号位置を用いた符号化方式の識別が期待できる．ただし，8 ビット単一・複数フォント符号の Tsampa, LTibetan, TCRC, Sambhota Web, Tibetan Modern A, THF, Tibetan Machine, RABTEN の符号化方式では全て Tsek 記号を 0x2D で符号化しているため，この方法を用いることはできない．Yagpo 符号では Tsek 記号は ASCII 領域 (0x002E) で符号化しているが，TibetBT 符号では Tsek 記号は CJK Compatibility Forms(縦書き字形) 領域で符号化している．たとえば，TibetBT 符号においては Tsek 記号の出現頻度が 28%であるから，任意の 4 文字をサンプリングし 0xFE3D が含まれていなければ，誤答率 1%以下で TibetBT 符号ではないと推定できる．

以上の 2 つの結果から，サブセット化されたフォントでも，8 文字以上のグリフが含まれていれば，誤答率 1%以下で符号化方式を推定できることがわかった．

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，標準化されていないレガシ符号化方式を 18 種類得た．これらの符号化方式について，フォントを入手することで，図形認識機能を用いずに符号化方式を識別できることを示した．また，web 上で公開されている 16 ビットのレガシ符号化文書を調査し，フォントがサブセット化されうる環境では長い文書を用いても文字集合全体を尽くすことは困難なことを示した．

サブセット化されたフォントからの符号化方式の識別可能性について，16 ビットのチベット文字レガシ符号が CJK 統合漢字の漢字領域を用いていることに注目し，16 ビットレガシ符号であることがわかっている場合には 4 文字以上，中国語文書との混同の可能性がある場合には 8 文字以上が含まれるフォントであれば，誤答率を 1%未満で符号化方式を推定できることを示した．この 16 ビットレガシ符号の推定方法は，フォントを用いずに符号化テキストに対して直接適用することもできるが，8 ビットレガシ符号は類似性が高く適用が困難である．

今回の調査では合成済みチベット文字の 16 ビット符号化方式は 3 種類しか発見できなかったが，中文 DOS 上のチベット文字環境では多くの 16 ビットレガシ符号が用いられていた報告があり，また，先頃 GB/T 20542:2006 として公布された国家規格 [7], [8]) に対する適用性も検討が必要である．

文　献

[1] Y. Mikami and Z. Pavol: "Writing systems and character codes in the world (2)", IPSJ Magazine, **46**, 9, pp. 1046–1052 (2005).

[2] suzuki toshiya and S. 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, **2006-DD-56**, 1, pp. 1–8 (2006).

[3] suzuki toshiya, M. Yamato, Y. Nagato and Y. Mikami: "Encodings in legacy khmer truetype fonts", Document Numérique, **10** *in printing*, (2007).

[4] Tibetan and H. D. Library: "List of tibetan fonts" (2006). http://www.thdl.org/tools/fonts/tibfonts.xml.

[5] 三上："文字符号の歴史 -アジア編-", 共立出版 (2002). ISBN 4-320-12040-X.

[6] C. Yu-Zhong and Y. 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).

[7] China Electronic Standardization Institute: "GB/T 20542:2006 Information technology - Tibetan coded character set - Extension A" (2006). http://www.nits.gov.cn/scl002/tibetCodeExADescription.asp.

[8] A. West: "Mapping Tables for Precomposed Tibetan" (2006). http://www.babelstone.co.uk/Tibetan/PrecomposedTibetan.html.