人文科学とコンピュータ　17－2
（１９９３．　３．　５）

木版刷チベット文字辞書作成のための文字特徴抽出

秋山　庸子　（東北大）　小島　正美　（東北工大）　川添　良幸　（東北大）　木村　正行　（北陸先端大学院大）

本研究では、木版刷チベット仏典の自動認識を行うための基礎実験の一つとして、チベット文字を構成する音節ごとの切り出しを検討する。原理的には各音節は区切り点により識別されるが、本研究では対象としている文献は古く不鮮明で、その自動認識は困難である。今回は基本子音３５１７文字の１文字切り出し法は統計的手法により行い、１文字切り出しは８０％の精度で可能となった。また、切り出された各基本子音についてその頻度を調べた結果、出現頻度の多い文字は全て異なる３種類の類似文字群に分類出来る事が分かった。

EXTRACTION OF CHARACTERISTIC FEATURES IN TIBETAN WOODEN-BLOCK MANUSCRIPTS TO CREATE A CHARACTER DICTIONARY

Youko AKiyama*　Masami Kojima**　Yoshiyuki Kawazoe*　Masayuki Kimura***

* Institute for Material Research, Tohoku University
1-1 Katahira, Aoba-Ku, Sendai 980, Japan.
** Tohoku Institute of Technology
35-1, Kasumi-Cho, Yagiyama, Taihaku-Ku, Sendai 982, Japan.
*** Japan Advanced Institute of Science and Technology, Hokuriku.
15, Asahidai, Tatunokuchi-Machi, Nomi-Gun, Ishikawa 923-12, Japan.

It is important to separate each syllable from wooden-block printed Tibetan manuscripts to recognize automatically the Tibetan characters. However, the small triangle symboles, which indicate the separation of syllables, are very difficult to be recognized.

The purpose of the present study is to improve the rate of character segmentation by using the knowledge of specific structures of Tibetan syllables, and to extract the characters efficiently with characteristic features.

1　はじめに

チベットに伝わった仏教は、その後1200年近くチベット文化の主流を形成し、その間に蓄積された膨大な量のチベット仏教文献資料が今日我々に残されている（文献１）。これらの文献のコンピュータによる可読化の研究は、インド原典、チベット訳文献、漢訳文献などの研究者から強く望まれている（文献２）。

本研究においてこれまで認識対象としてきた北京版チベット大蔵経（文献３）の中の正法白蓮華経の冒頭部分を図１に示す。認識実験は大きく分けて文字認識を行う前までと後とに分けられ、前者は前処理と言われる。

前処理においては、行切り出し、切り出した行の傾き補正、文字切り出し、ノイズ除去および正規化が行われる。前処理の善し悪しが全体の認識結果を決定するので、認識対象の文字により最適の処理を行う必要がある。ここで取り扱っているチベット文字は１音節単位で構成され、その単位は図２に示す様に基字、付頭字、付足字、前接字、後接字、再後接字、母音記号の７種の要素から構成される。なお、基字＋付頭字あるいは基字＋付足字の形をなす場合、その部分を重層字ともいう。母音記号のうちｉ、ｅ、ｏに相当する記号は基字あるいは重層字の上部に付き、母音記号ｕは下部に付く。上部および下部に母音記号がついていない基字は、通常のローマ字風の記述に従えば母音記号ａを含む（文献４、５）。

全てのチベット文字が図２に示す構成要素を持つわけではなく、前接字、後接字あるいは再後接字を持たない字もある。実際のチベット文字はそれらを組み合わせる事により、１音節７種類の構造に分類される。チベット文字基本30子音を図３－ａに、4母音を図３－ｂに示す。これらの基本子音が前接字、後接字あるいは再後接字に存在する場合には母音を付けず、基字に存在する時のみ母音を付加する。（それ以外の母音の識別は言語の知識によって行われる。）そのため、チベット文献の文字に表音記号を付ける場合、１文節の切り出しからさらに１音節の切り出しが必要となる。この場合、図４に示す本来の１音節切り出しのための区切り点を検出して、その後１音節の切り出しを行えば良い。しかし、図１に示す木版刷りチベット文献においては、区切り点が文字に付着したり、墨によるノイズとの識別が困難なため、区切り点による１音節切り出しは困難である。

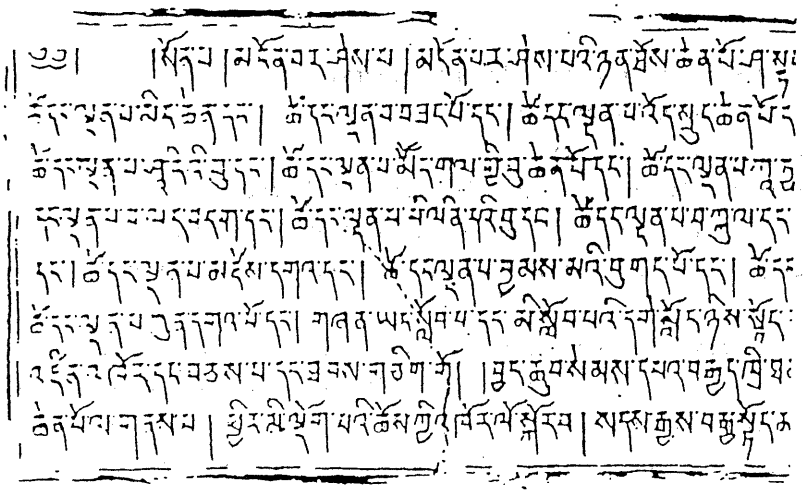

図１．北京版チベット大蔵経の中の正法白蓮華経の冒頭部分

そこで、区切り点が検出されない状態においてもチベット文字の１音節切り出しを行うために、チベット文字の１字をそれぞれオブジェクトと考え、オブジェクト間の文法および構造関係から１音節構造を推定し切り出しを行う方法を検討したい。

本研究はオブジェクトのしてのチベット文字の辞書作成を行うために、その基礎実験としてチベット文字基本30子音について、１文字および繋がり文字の切り出し率、出現頻度および文字幅などの特徴量を北京版チベット大蔵経の中の任意の4248文字について調べた。

2　文字特徴抽出

入力パターンとして木版刷りチベット文献をイメージリーダIBM6392により１行ずつパーソナルコンピュータ(PC)IBM5550へ取り込む。次にPCから汎用大型コンピュータIBM3090-SJ2へ行切り出し後のデータを送る。読み取った文字行は傾きを持っているので、ＬＰＰ（Local Projection Profile）法（文献６）により傾き補正を行う。ここで扱う北京版チベット大蔵経においては、図１に示す様に文字同志が複雑に重なり合っているために、縦方向射影による単純な文字切り出し法は適用できない。そこで、従来の方法では切り出せない文字に対する処理として、チベット文字の特徴ある横棒（主要水平線：ＭＨＬ）より上部に存在する母音を分割して、下部に存在する文字は単純に縦方向射影による切り出しを行った（文献７）。本研究で切り出し対象とした基本子音は全体の文字数の88％を占める。切り出し対象とした基本子音3517文字中2852文字が１文字切り出しが可能で、その割合は81％となった。切り出しが出来なかった繋がり文字のほとんどが２文字繋がり文字である。１文字群と２文字繋がり文字群の文字幅に対する頻度分布を図５に示す。図５において、横軸は文

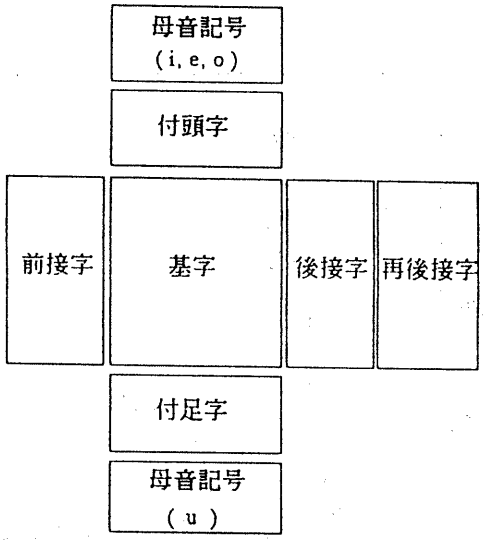

図２．チベット文字の１音節構成

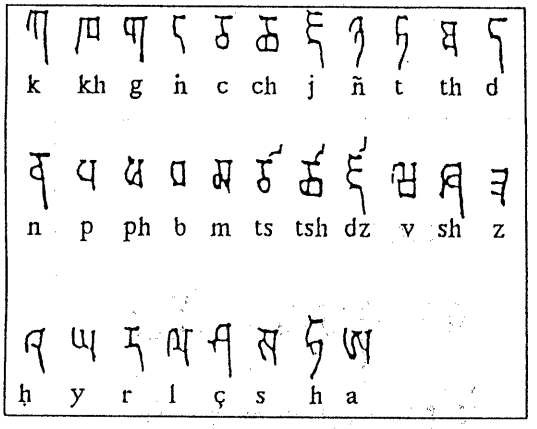

図３－ａ　基本子音３０文字

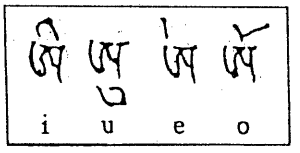

図３－ｂ　４母音（ｉ，ｕ，ｅ，ｏ）

図３．チベット文字基本３０子音と４母音

字幅で、その値は相対値で表し縦軸は各文字幅の頻度を表している。図の左側のピークは１文字群(SUM1)、右側のピークは２文字繋がり文字群(SUM2)をそれぞれ表し、SUM1の出現頻度は図の出現頻度のスケール値の４倍と等価である。図５において、重なり合っている部分が１文字群と２文字繋がり文字群との識別が困難なところであるが、統計的にみてこの部分は１文字群の0.5%以内と大変少ない（文献８）。そのため１文字群の切り出し位置において切り出しを行った。２文字繋がり文字が１文字と判定された数をεとする。１文字群の数をSUM1、１文字群の切り出し精度をηとする。$\eta=(SUM1-\varepsilon)/SUM1$より、切り出し率99.5%以内で１文字群切り出しが可能となる。

次に、基本30子音の出現頻度を図６に示す。図６の横軸は文字の出現頻度の多い字種を順に並べ、縦軸はその頻度を表す。字種の違いにより極端に出現頻度に差がある（文献９）。任意の基本子音3517文字中出現頻度が100を越える文字は”ｇａ”、”ｎａ！”、”ｄａ”、”ｎａ”、”ｐａ”、”ｂａ”、”ｍａ”、”ｓａ”、”ｈａ！”、”ｒａ”の10文字である。ここで、”！”は表音の鼻音を表す記号とする。これら10文字はそれぞれ3種の類似文字群に分類出来る。すなわち、類似文字Ⅰ群は”ｂａ”、”ｐａ”、”ｈａ！”、類似文字Ⅱ群は”ｍａ”、”ｓａ”、類似文字Ⅲ群は”ｄａ”、”ｎａ”、”ｎａ！”、”ｒａ”となる。この他にも数個の類似文字群があるが、類似文字Ⅰ、Ⅱ、Ⅲ群とは出現頻度が少ないので、以下ではこれら3種の類似文字群にだけ注目する事にする。

我々はこれまで木版刷りチベット文献中の文字自動認識を行ってきている。木版刷りチベット文献の自動認識は、まず文字の形を表すランレングス法と重ね合わせ法とを併用して行い、この認識手法だけでは特定出来ない場合には類似文字に対しては第１位～５位までの候補文字から第１位の文字の推論を行う２段階法により行ってきた。その結果、北京版大蔵経2～82頁中辞書文字として使用した297文字に対するクローズ実験において90%

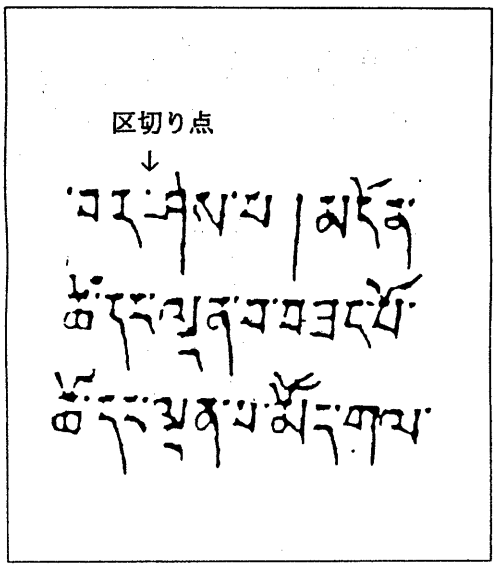

図４．チベット文字１音節表示の区切り点

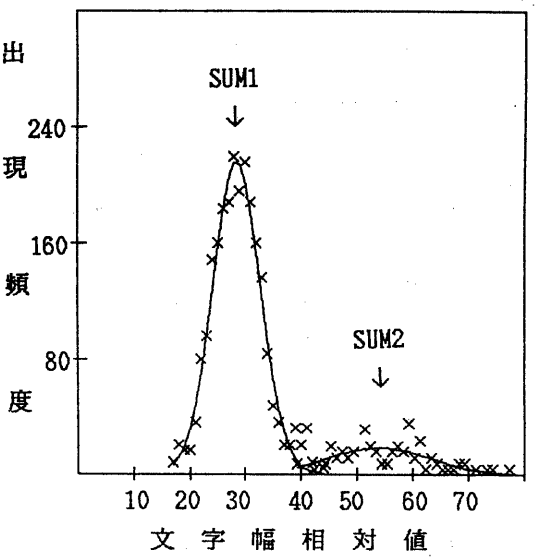

図５．１文字群及び２文字繋がり文字群の文字幅相対値に対する出現頻度

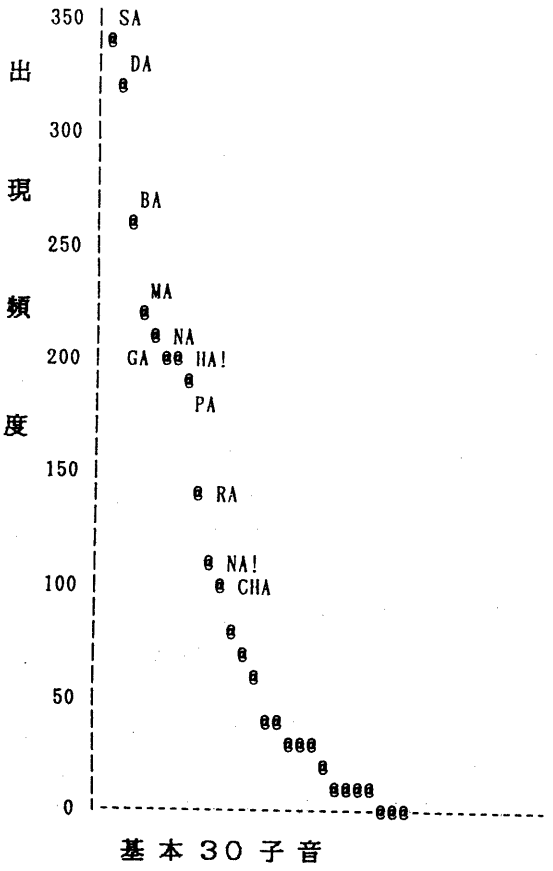

図6．基本30子音に対する出現頻度

表1．2文字繋がり文字パターンに対する出現頻度

繋がり文字	出現頻度
DANA!	33
DAGA	23
DABA	17
DASA	16
HA!DA	15
NA!SA	14
DANA	10
BABA	7
RABA	7
DADA	7
DAPA	7
GASA	6
DAKA	6
YADA	6
NA!HA!	5
RANA	5
MADA	5
RAGA	5

の認識率を得ている（文献１０）。認識ミスの主な原因は、極めて類似した文字の出現頻度が多いためである。出現頻度により重み付けを行った距離計算などを考慮する必要がある。

3　繋がり文字

２文字繋がり文字は切り出し対象としている基本子音3517文字中542文字と、その割合は15％と少ないが、チベット文字の１音節での文字認識を考えた場合、無視出来ない割合となっている。２文字繋がり文字パターンに対する出現頻度を表１に示す。表１において、初めの文字が”ｄａ”の場合が圧倒的に多いという極めて特徴的な傾向を示している。２文字繋がり文字については、１音節の区切り点が明確に識別出来ないため、チベット文字１音節の中での繋がり文字かあるいは２音節に渡った繋がり文字かは判別が困難である。チベット文字は１音節単位で認識しなければならないため、この結果は大変重要な解決しなければならない問題となる。

3　まとめ

認識用の辞書としてこれまでの様に単に文字のイメージデータとか構造情報を参照するだけでは精度の良い文字認識は出来ない。各辞書文字をオブジェクトと考え、オブジェクトにウェイトを置いた辞書作成を考える必要がある。本研究はそのための基礎実験として、木版刷りチベット文献中の基本子音3517文字について特徴抽出を行った。その結果チベット文字の出現頻度は文字種によって際だった特徴があり、その出現頻度の多い10文字は全て3種の類似文字群に分類できることがわかった。１文字群と２文字繋がり文字群との識別は、統計的手法で切り出し文字幅を決定し、１文字群の切り出しは99.5%の精度で実行できる事が分かった。また、２文字繋がり文字群となり易い文字列がある事が分かった。これについてはさらに詳しく調べていきたい。今後は、これらのチベット文字特徴およびチベット文字の文法を取り込んだオブジェクト指向チベット辞書文字の作成を行いたい。

謝辞

本研究を進めるにあたり、貴重なご意見を頂いた東北大学文学部塚本啓祥教授、磯田熙文助教授、仙台電波高専山崎守一助教授、および熱心にご討論して頂いた東北大学金属材料研究所川添研究室の皆様に深謝致します。また、実験をスムースに出来るように心配りして頂いた東北大学情報処理教育センター職員の皆様、金属材料研究所材料科学情報室の皆様に心から感謝致します。

参考文献

１）塚本：インド文学の形成と展開、「サンスクリット・チベット語のコンピュータによる総合的研究」、東北大学特定領域研究組織ＴＵＲＮＳ017－報告書(1989. 2)；磯田：チベット文字の特色とコンピュータ利用について、ibid.

２）川添：コンピュータによる仏教混淆梵語の研究(2)、印度学仏教学研究37巻第2号(1989.3).

３）大谷監修：影印北京版西蔵大蔵経、世界聖典刊行協会編、京都、p.1-279 (1955. 7).

４）稲葉：チベット語古典文法学、法蔵館 (1966).

５）小島、川添、木村：木版刷チベット文献の文字自動認識の試み、情報知識学会誌、vol. 2 No. 1 (1991. 12).

６）秋山、増田：書式指定情報によらない紙面構成要素抽出法、電子通信学会論文誌(D), J66-D No. 1 (1983. 1).

７）秋山、小島、川添、木村：上部母音別認識法によるチベット文献中の文字自動切り出しについて、平成４年度電気関係学会東北支部連、2C6(1992. 8).

８）舟久保：パターン認識、共立出版 (1991. 12).

９）奈良、川添：文科系のための計算機とプログラミング、昭晃堂 (1986. 5).

１０）小島、川添、木村：推論を用いたチベット文献中の文字自動認識、印度学仏教学研究第41巻第1号 (1992. 12).