木版刷チベット文献中の文字特徴抽出

小島正美・秋山庸子・川添良幸・木村正行

1. はじめに

我々は既に影印北京版西蔵大蔵経［文献 1］を例とした木版刷チベット文献の自動認識について報告している［文献 2］。この研究により，誤認識の多くは出現頻度の高い類似文字であることが分かっており，それらに推論のアルゴリズムを適用する事により，認識率の向上を図ってきた［文献 3］。本論文では，文字認識の実用レベルである認識率99%の達成を目指し，木版刷チベット文字の特徴を抽出すること，およびその結果に基ずくオブジェクト指向プログラムの検討について述べる。

2. 文字切り出し

本研究においてこれまで認識対象としてきた影印北京版西蔵大蔵経の冒頭部分を図1に示す。一般的に，文字認識を行うためには行切り出しを行い，切り出した行から文字切り出しを行う。図1に示す様に，文字同士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。この方法により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は2文字繋がり文字である。

また，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。そのため，チベット文献中に表れる文字を表音記号化するためには，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3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に示す様に本研究の対象となる木版刷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

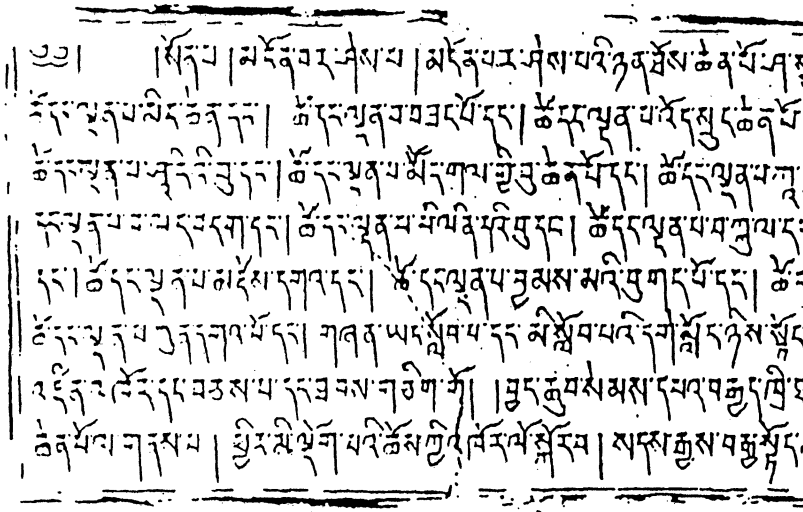

図1　北京版チベット大蔵経の中の正法白蓮華経の冒頭部分

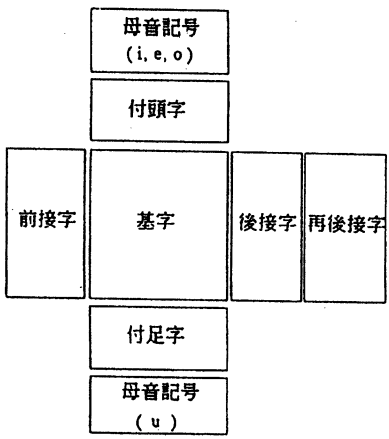

図2　チベット文字の1音節構成

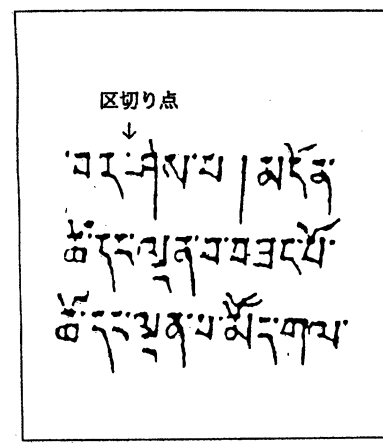

図3　チベット文字1音節表示の区切り点

の識別が困難なため，区切り点だけによる1音節切り出しが困難である。

この様に，木版刷チベット文字の自動認識を行う場合，チベット文字の構造上の特徴を考慮する必要がある。1音節文字の切り出しのためには，前節で述べた様にチベット文字の1音節を表示している区切り点だけでは，1音節文字の切り出しを行うことは出来ない。そこで，チベット文字それぞれをオブジェクトと考

え，オブジェクト間の文法および構造関係から，1音節構造を推定して切り出しを行う方法を現在検討している。

3. 文字特徴抽出

基本30子音の出現頻度を図4に示す。図4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，その出現頻度が極端に異なる。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”，“cha”の11文字である。ここで“！”は鼻音を表す記号とする。これらの中の9文字はそれぞれ3種の類似文字群に分類できる。すなわち，類字文字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。この他にも数種の類似文字群が存在するが，類似文字Ⅰ，Ⅱ，Ⅲ群と比較して出現頻度が極端に少ないので，まずこれらの類似文字Ⅰ，Ⅱ，Ⅲ群に注目する事にする。

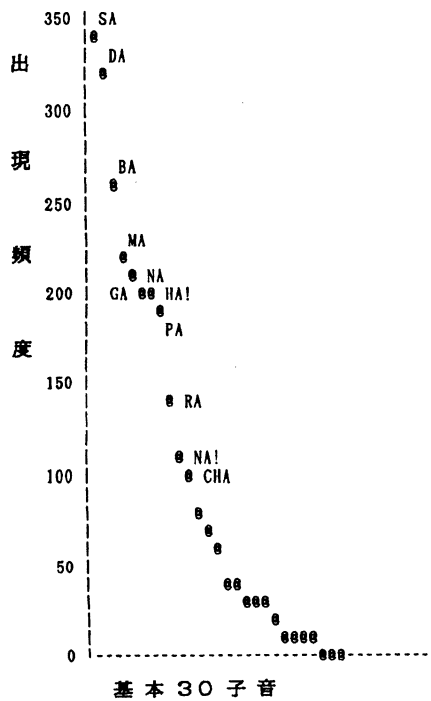

図4　基本30子音に対する出現頻度

我々はこれまで木版刷チベット文献中の文字認識を行ってきている。木版刷チベット文献の文字自動認識は，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，この認識手法では特定できない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%の認識率を得ている[文献 3]。誤認識の主たる原因は，極めて類似した文字の出現頻度が高いためである。

現在，出現頻度が高く誤認識し易い“ba”，“pa”，“ha!”についてオブジ

ェクト指向によるチベット活字辞書を作成し実験を行っている。本手法による認識方法はこれまでの認識方法とは異なり認識しようとする文字の特徴に依存して認識プログラム本体を変更する必要がない。すなわち，オブジェクト指向文字辞書は各辞書文字とその文字の特徴を具備した認識メソッドから構成され，認識候補文字により異なった認識メソッドが実行できる。

4. まとめ

木版刷チベット文献中の文字自動認識において，文字の切り出しから認識までを高精度化するためには文字の特徴情報を考慮する必要がある。そのため，文字そのものをオブジェクトと考え，オブジェクト指向による方法を取り入れた文字の切り出しおよび認識を行う必要がある。例えば認識用の辞書として，これまでの様に単に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。そこで，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ことにより，実用的な認識システムの実現が図れる。

今後は，誤認識し易い類似文字を取り上げて，類似文字特有の認識メソッドを検討し，文字認識率の向上を目指して行きたい。

謝辞

本研究にあたり，貴重なご意見を頂いた宝仙学園短大塚本啓祥学長，東北大学文学部磯田熙文教授，伊藤道哉助手，仙台電波高専山崎守一教授，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様に心配りして頂いた東北大学情報処理教育センター並びに金属材料研究所材料科学情報室の皆様に感謝致します。

参考文献

1) 大谷監修：影印北京版西蔵大蔵経，世界聖典刊行協会，京都，p.1-279 (1955.7).
2) 小島，川添，木村：チベット文献自動認識，印度学仏教学研究，第39巻第2号 (1991.3).
3) 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).

〈キーワード〉 木版刷チベット文献，北京版西蔵大蔵経，オブジェクト指向文字辞書，文字特徴抽出，類似文字

小島正美（東北工業大学助教授）・秋山庸子（東北大学技官）・川添良幸（東北大学教授，理博）・木村正行（北陸先端科学技術大学院大学教授，工博）