第１０回石田（實）記念財団研究助成金

チベット仏教原典の自動認識に関する研究

Research on Automatic Character Recognition of Tibetan Buddhist Texts

研究成果報告書

１９９９年１１月１９日

研究代表者　小島正美
東北工業大学・通信工学科

共同研究者　川添良幸
東北大学・金属材料研究所

共同研究者　木村正行
北陸先端科学技術大学院大学

１．はじめに

本論文で認識対象としている文献は、図１に示すデルゲ版チベット文献である。この文献の大蔵経と蔵外だけでも表裏木版刷紙で18万枚に及ぶ膨大なもので、現在、東北大学中央図書館に最重要資料として保管されている。これらの文献をコンピュータで自動認識することができれば、インド原典、チベット訳文献、漢訳文献などの研究者が本来の文献学に専念できる点において大変意義がある[文献１、２]。

著者らは東北大学文学部印度哲学研究室の協力を得ながら1989年から木版刷チベット文献の自動認識の研究に取組み、これまで正しく切り出しされた基本30子音については90％程度の認識率を得ている[文献３]。しかし、チベット文字の1音節構造は子音１ないし４個と母音の組み合わせからなり、１音節単位での表音認識を行わなければならず、そのためには図１に示すような大変小さな塊のような１音節区切り記号（ツェック)の識別が必要である。また、個々の文字が複雑に重なり合い、繁がっているため１音節単位での文字切り出しが大変困難である。

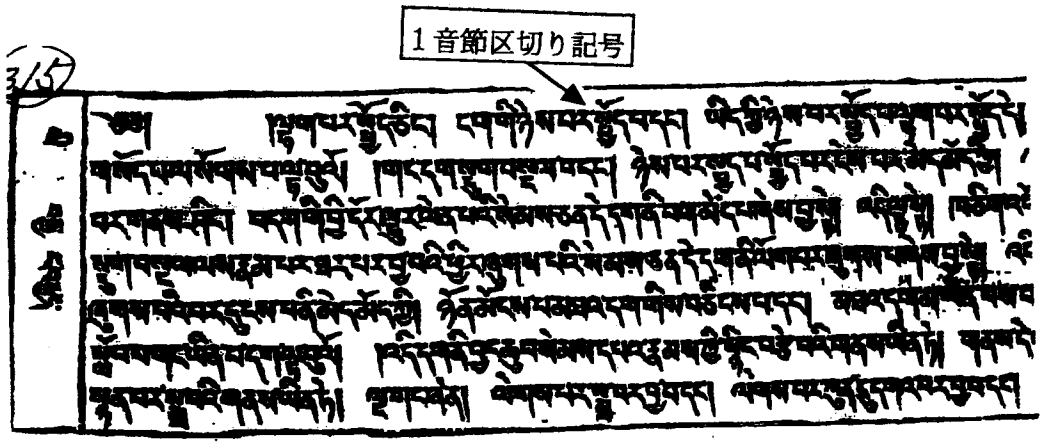

図１　実験に使用したデルゲ版チベット文献の一部

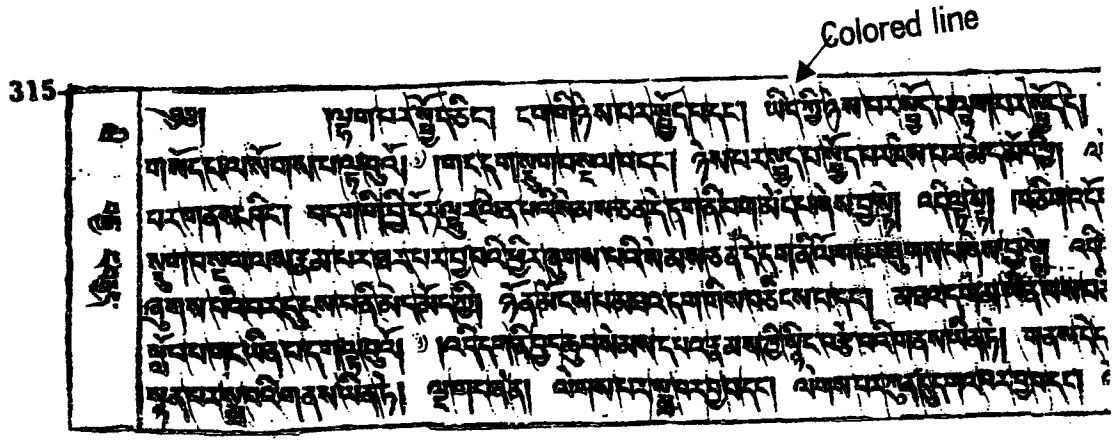

図２　カラー線を挿入したデルゲ版チベット文献の一部

そこで、著者らはチベット文献を自動認識する過程において、作業工程を分析し、どの箇所にコンピュータを積極的に導入すべきかを検討し、チベット学者らが現時点で真に使用可能なシステムを検討した。その結果、チベット１音節文字切り出しの部分に、チベット学者の知見を利用して図２に示すようにカラー線を挿入してもらい、そのカラー線をコンピュータで自動識別することが、現段階ではもっとも効率の良いシステム設計となることが分かった。その場合、切り出しされた１音節文字をクラス１音節文字として、１音節文字個々について文字特有の属性を持たせて、その属性に合った認識メソッドを適用することにより文字認識率の向上を目指すシステム構築[文献４～６]を行っている。

２．チベット文字

チベット文字の１音節構成の最大要素は図３に示すように、基字、付頭字、付足字、前接字、後接字、再後接字、母音の７種から構成される。なお、基字+ 付頭字、基字+ 付足字は重層字と呼ばれる。さらにサンスクリット文字からの転写文字などがある。母音記号のうち”ｉ”、”ｅ”、”ｏ”に相当する記号は上部に付き、母音記号”ｕ”に相当する記号は下部に付く。チベット文字は単体で母音”ａ”を内在している。そのため、例えばチベット文字単体に母音記号”ｉ”が付加された場合は、チベット文字に内在されている母音記号”ａ”が無視されることになる。母音記号を付加したこれらの文字を総計すると615字種程度となる。チベット文字の１音節構造は子音１乃至４個と母音の組み合わせからなる。子音の数が２個以上の場合、どの子音が基字となるか判定しなければならず、分節パターンの識別が必要となる[文献７]。

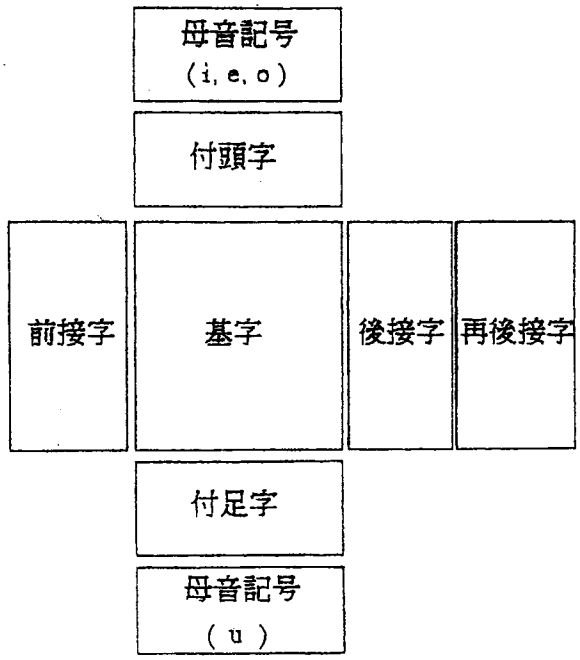

図３ チベット文字の１音節構成の最大要素

３．実験

本論文で用いた木版刷チベット文献の大きさは幅がおよそ１０cm、長さが４０cmである。予め、Ａ３サイズのカラースキャナ上に載せることが可能なサイズに縮小を行い、１音節単位にチベット学者の知見によりカラー線を挿入した文献をスキャナ上に載せて、イメージデータの取り込みを行った。使用したスキャナはSHARP-jx-610である。初めに、水平方向の斜影をとり、そのピークとなるMHL（

Main Horizontal Line)を検出し、図４に示すように（P１+P2）／２（上部母音の領域を十分に取り込んだ位置）から次のP３までを切り出す。このように行切り出しを行うために、十分な行切り出し範囲の指定を行っているのは、木版刷チベット

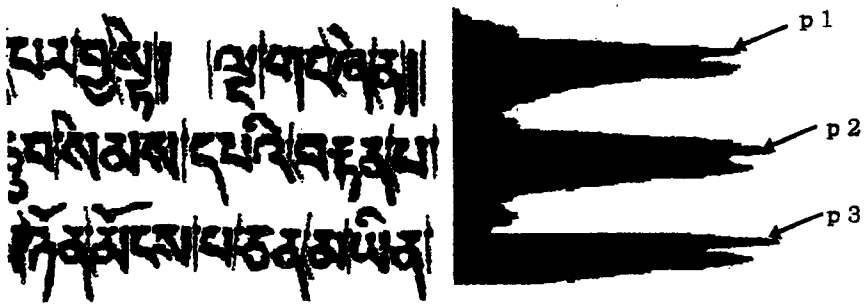

図４　水平方向斜影図

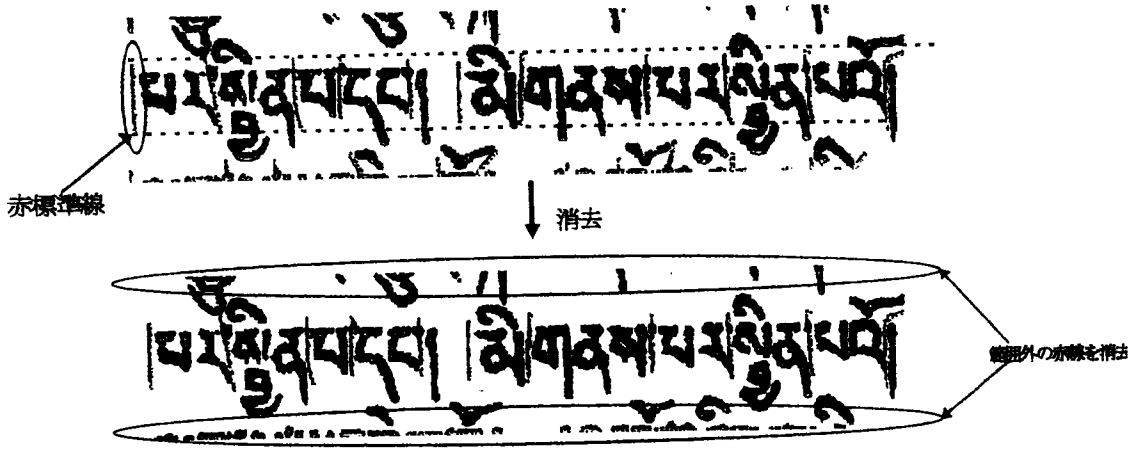

図５　行切り出し例

文献は活字版と異なり行が大きな波線状となり、行間の最大ピークが必ずしもMHLとはならないためである。このようにして図５に示すように行切り出しを行う。次に図６、７に示すように、カラー線で囲まれた部分を識別して、１音節文字を切り出し、切り出し対象文字以外はノイズとみなして除去する。このようにして、4000音節文字に対して95％の割合で１音節文字切り出しが

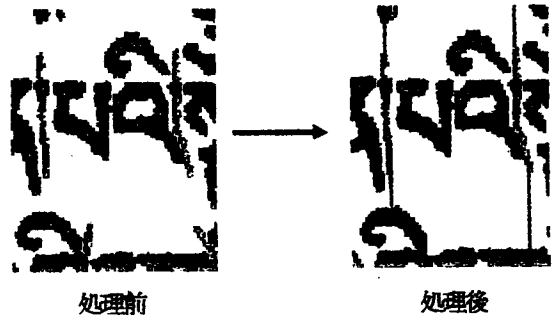

図６　１音節文字切り出し領域を決定

成功している。失敗した１音節文字切り出し率５％のうち、３％は上下行からの繋がり文字で、２％はノイズとみなして上部母音を削除したために起きている[文献８]。

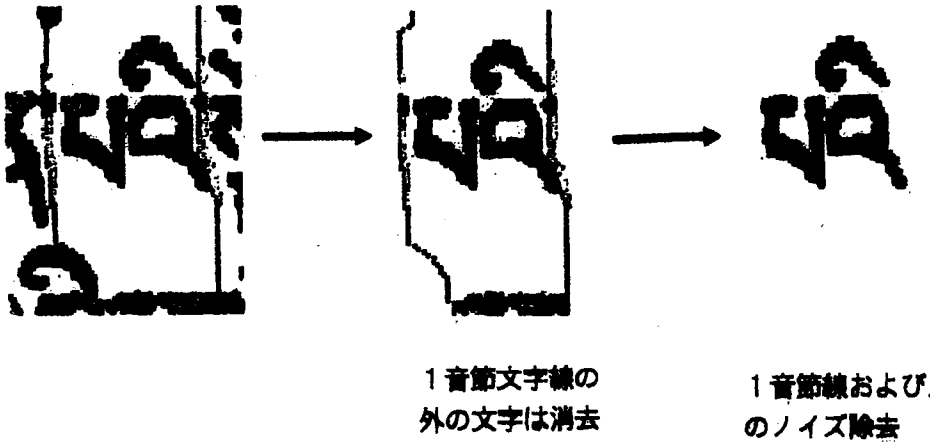

図７　１音節文字切り出し例

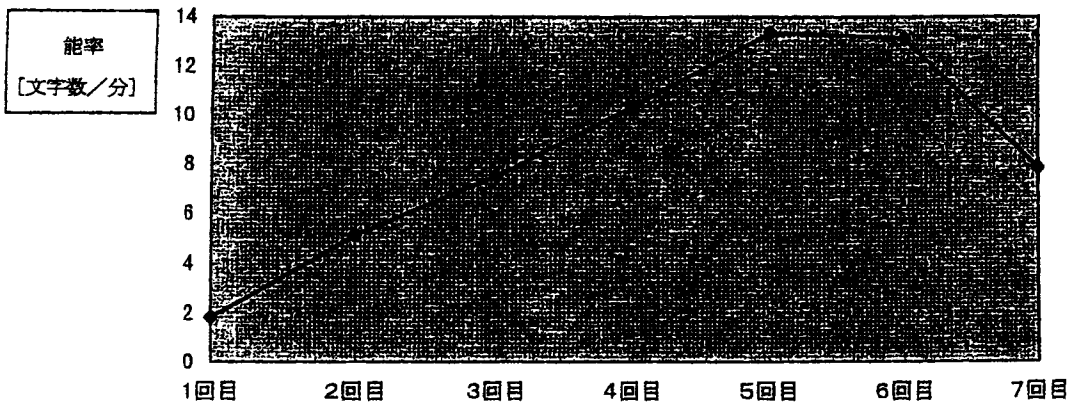

図８　経験による１分あたりの入力文字数の変化

ここで、チベット文献の認識工程の作業時間を調べてみる。そのなかで、もっとも作業時間を要するところは１音節文字毎にカラー線を挿入する部分ではなく、１音節文字をキーボードから打ち込む箇所である。実際に、手動によりキーボードからチベット表音文字を打ち込んでもらい、表音１文字打ち込みの１分あたりの文字数を打ち込む回数ごとに示したのが図８である。打ち込み者はあらかじめチベット

表音文字の学習を行って、それから打ち込みを行った。打ち込み作業を始めてから、5日目までは学習効果のため1分あたりの打ち込みの文字数が上昇したが、6日で上昇が留まった。1分あたりの打ち込み文字数は個人差はあるが、ほぼ同様な傾向となった。この作業は認識工程のなかでもっとも労力を必要とする部分である。そこで、この部分にコンピュータを積極的に導入することにより、チベット学者が必要とするシステムが可能であると思われる。

次に、文字認識を行う場合、個々の文字を文字クラスと考え、その文字が切り出された時点で保持している特徴を、その文字固有の属性として認識時まで継承させる必要がある。すなわち、文字切り出し部と認識部とに分けた場合、文字切り出し時に各文字が有している文字特有の特徴（属性）が、文字認識部では失われることになる。すなわち、本論文では、1音節文字単位での文字認識を考えているので、重ね合わせ法による文字認識を行う場合、正規化処理を行った時点で、図9に示すように文字切り出し時の1音節文字の大きさ情報が失われてしまう。このような事のないように、1音節文字の大きさ情報および1音節文字の複雑さ情報[文献9]を1音節文字切出し時に取得し、認識対象としている1音節文字に属性として持たせる必要がある。また1音節文字は文字数および文字の種類によりパターン1から7まで分節できる。これらのパターンの出現頻度に関して、文字数で分類すれば1文字と2文字を合わせた出現頻度はおよそ82%となり、4文字出現頻度は1%程度と極端に異なっている。また、類似文字群としての属性は、活字文字と木版刷り文字とでは大きく異なってくる。例えば、活字文字の誤認識の多くは類似文字"pa"と"ba"に集中しているが、木版刷り文字の誤認識はその他に"da"　と"nga"に集中している点である。そのため、活字文字で有効な類似文字認識メソッド[文献10～12]は

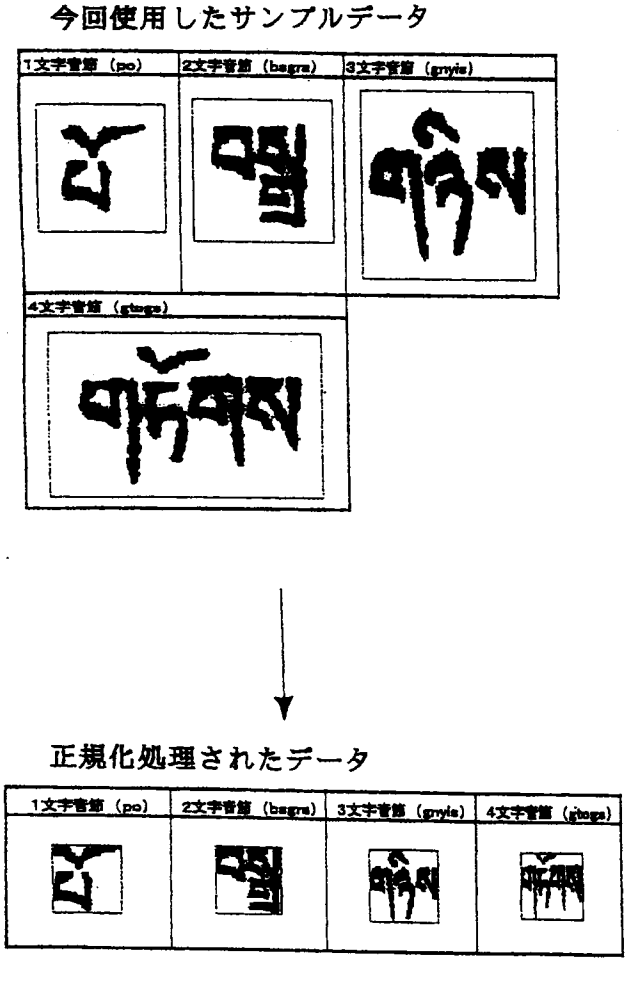

図9　1音節文字の正規化処理例

そのまま木版刷り類似文字認識に効果的に作用しないことが分かった。すなわち、木版刷チベット文献では文字の出現頻度は活字版文献と共通しているが、文字構造では、活字版と異なる特徴を有するので、適切な特徴を１音節文字の属性として持たせる必要がある。また、図１０に示すように、１音節文字切り出し文字クラスから１音節文字認識対象文字クラスへの継承を考える必要がある。

４．まとめ

チベット仏教原典として、現在、大量の木版刷文献が存在し、それをコンピュータにより自動認識化することがチベット学者らから望まれている、その場合のもっとも困難なところは１音節文字切り出しの部分であった。チベット文献の自動認識を考えた場合、もっとも大変な作業は、キーボードから文字を打ち込む作業であり、その部分をコンピュータで行うことが可能であれば、作業効率を飛躍的に向上させることが可能である。本論文では、１音節文字切り出しにカラー線を挿入することにより、文字切り出しを行い、１音節文字切り出し時の文字属性を、切り出した文字に属性として持たせ、１音節文字認識時まで継承させることを提案し、その有効性を示した。従来までの文字認識と大きく異なるところは、認識プログラムとデータを分離しないで、１音節文字をクラスと考え、１音節文字ごとにそれぞれ固有の認識メソッドを持たせた点である。

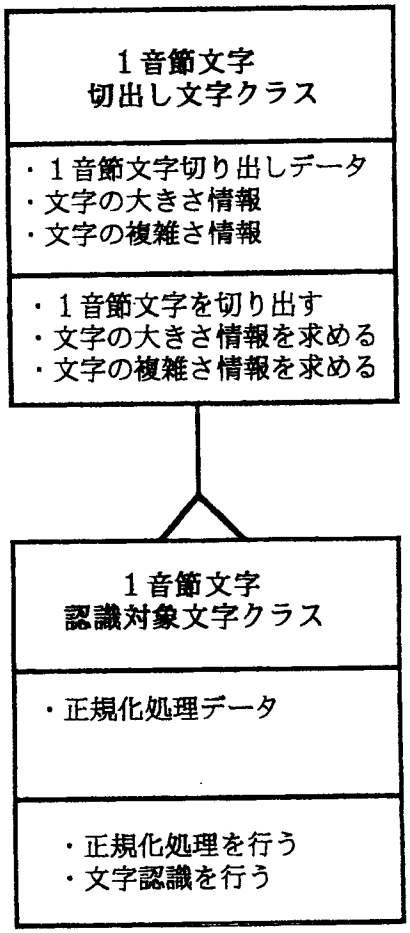

図１０　木版刷りチベット文献の文字認識クラス図

謝辞

実験を進めるにあたり、チベット文献の資料提供ならびにチベット文法情報等でアドバイスをいただきました宝仙学園学長塚本啓祥教授並びに東北大学文学部磯田熙文教授に心から感謝いたします。また、本研究に対してご援助していただきました 石田（實）記念財団に深く感謝いたします。

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研究」、東北大学特定領域研究組織ＴＵＲＮＳ017-報告書（Feb. 1989）；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、Vol. 37, No.2 (March 1989).

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、1号、pp. 158-161、（Decem. 1992).

4）Martin, J. : Principle of Object Oriented Analysis and Design, Englewood Cliffs (1993).

5）Jacobson, I. : Object Oriented Software Engineering, Addison Wesley Publishing Company, (1992).

6)J. ランボー, M. ブラハ, W. プレメラニ, F. エディ, W. ローレン、羽生田訳：オブジェクト指向方法論OMT - モデル化と設計 - 、トッパン、 (July 1992).

7）小島、布宮、川村、秋山、川添、木村：オブジェクト指向によるチベット活字パターン識別、情報処理学会人文科学とコンピュータ研究会、18-2、（May 1993).

8）小島、川添、木村：オブジェクト指向設計を考慮した木版刷チベット文献のイメージ文字認識、情報処理学会人文科学とコンピュータ研究会、38-3、（May 1998).

9）山下、小島、木村：音節構造解析による活字チベット文字認識の高速化、情報処理学会人文科学とコンピュータシンポジウム、pp. 53-60、(Sep. 1999).

10）小島、布宮、川村、秋山、川添、木村：オブジェクト指向設計によりチベット文字認識について、情報処理学会「人文科学とコンピュータ研究会」23-2、（Sep. 1994）.

11）Masami Kojima, Yoshiyuki Kawazoe and Masayuki Kimura : Automatic Tibetan Script Recognition by Computer; 7th Seminar of the International Association for Tibetan Studies, pp. 527-533, (April 1997).

12）Masami Kojima, Yoshiyuki Kawazoe and Masayuki Kimura : Automatic Recognition of Tibetan Buddhist Text by Computer ; 1999 EBTI, ECAI, SEER & PNC Joint Meeting, pp. 387-393, (January 1999).