人文科学とコンピュータ　18－2
（１９９３．５．２１）

オブジェクト指向によるチベット活字文献からの文字パターン識別

小島　正美（東北工大）　布宮　千夏子（米沢高技専）　川村　隆庸（日本ＩＢＭ）　秋山　庸子（東　北　大）　川添　良幸（東　北　大）　木村　正行（北陸先端大）

チベット文献の自動認識においては、１音節単位で文字を切り出し、切り出された文字間の関係を調べる必要がある。本研究は、切り出された文字から文字列パターンを作成し、チベット活字文献からの文字パターン識別をオブジェクト指向によるシュミレーション実験により行った。その結果、本手法はチベット文字を構成する１音節７パターンの識別に有効であることがわかった。

Automatic recognition for character pattern of printed Tibetan scripts by using object oriented method

Masami Kojima[*1]　Chikako Nunomiya[*2]　Takanobu Kawamura[*3]
Youko Akiyama[*4]　Yoshiyuki Kawazoe[*4]　Masayuki Kimura[*5]

[*1] Tohoku Institute of Technology
35-1, Kasumi-Cho, Yagiyama, Taihaku-Ku, Sendai 982, Japan.
[*2] Yamagata Prefectural Yonezawa Adult Technical Education
2736, Dounomae, Kubota, Yonezawa 999-21, Japan.
[*3] IBM Japan Tohoku Systems Engineering Co., Ltd.
4-6-1, Ichiban-Cho, Aoba-Ku, Sendai 980, Japan.
[*4] Institute for Material Research, Tohoku University.
1-1, Katahira, Aoba-Ku, Sendai 980, Japan.
[*5] Japan Advanced Institute of Science and Technology, Hokuriku.
15, Asahidai, Tatunokuchi-Machi, Nomi-Gun, Ishikawa 923-12, Japan.

It is necessary to separate each syllable from printed Tibetan manuscripts to recognize automatically the Tibetan characters. The relations between the Tibetan characters are very important to recognize them. These Tibetan characters are categorized as objects. We explain the simulation results by using object oriented method. From this experiment, it is confirmed to recognize all of seven patterns for Tibetan scripts.

1　はじめに

チベットに伝わった仏教が１２００年近くもチベット文化の主流を形成し、この様な仏教文化の受容とチベット仏教文化の形成・伝承を記すチベット語で書かれた文献資料は膨大な量として今日我々に残されている（文献１）。我々は、これらの文献のコンピュータによる可読化を現在試みている（文献２）。

本研究においてこれまで認識対象としてきたチベット文字は１音節単位で構成され、その単位は図１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。チベット活字文字の基本３０子音および４母音を図２に示す。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３）。

全てのチベット文字が図１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、図３に示す１音節７種類の構造に分類される。すなわちチベット文字の１音節構造は子音が１から４と母音１との組み合わせとなる。子音の上部か下部に母音がついている時は、その子音が基字となる。しかし、上部および下部に母音が付かない基字の場合、母音記号”ａ”を付けて読む。その場合、文字数が２文字と３文字の時にどの文字が基字に相当するかを判断しなければならない。そのためには文字パターン識別をする必要がある。文献２ではコンピュータによる可読化に、認識対象文字として木版刷チベット文献を使用した。

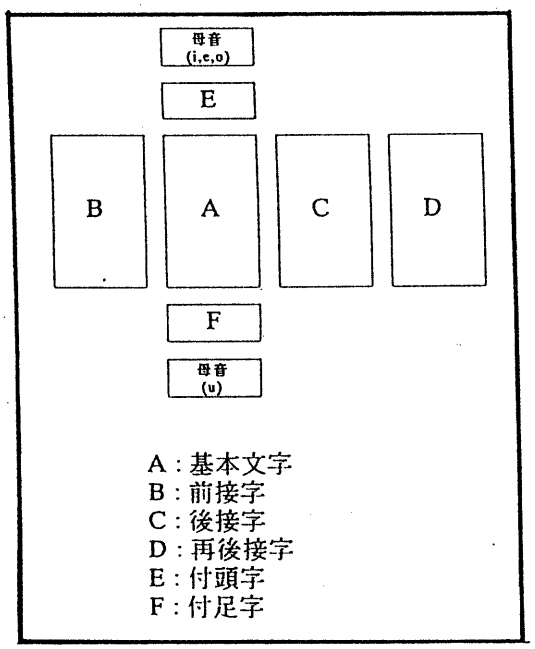

図１　チベット文字の１音節構成

図２　チベット活字文字の基本３０子音および４母音

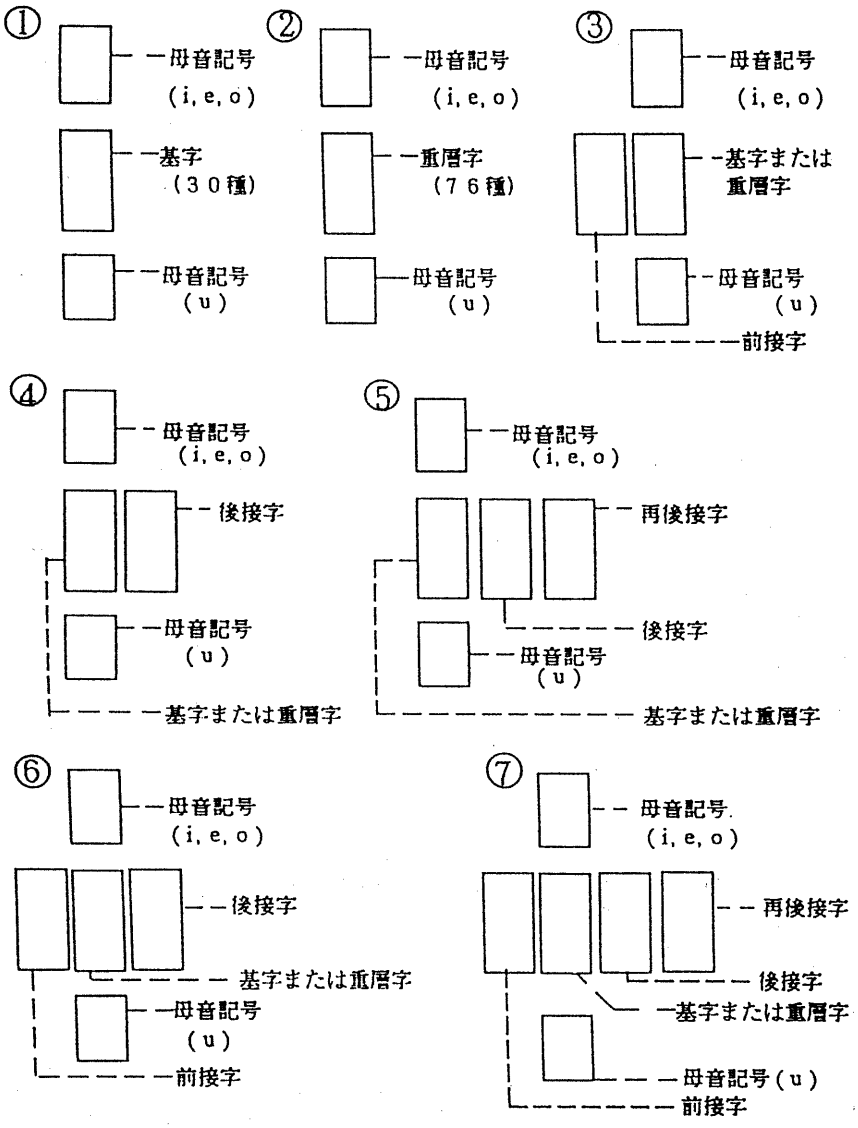

図3　チベット文字の音節構造
（1音節が7通りの構造をとる）

この場合、チベット文字１音節の区切り点が識別困難であった。そこで、今回はチベット活字文字により、チベット文字１音節単位で認識を試みるため、１音節７パターン識別のシュミレーション実験を行った。

２．１音節文字切り出し

本研究において実験に使用したチベット活字文献を図４に示す。図４のチベット活字文献から行切り出しを行い傾き補正を行った後、チベット文字の１音節を示す逆三角形の小さな区切り点を自動的に検出して１音節単位に文字切り出しを行った例を図５に示す。サンプル数２０２０音節文字中切り出し成功は１９７８音節文字で音節文字切り出し率は９８％となる。切り出し失敗のほとんどは行切り出し時に混入するノイズなどである。図５に示すように、１音節間には１文字分のスペースを入れている。切り出し成功した１音節文字について、１文字毎の文字列パターンを作成し、その文字列パターンを入力パターンとした。

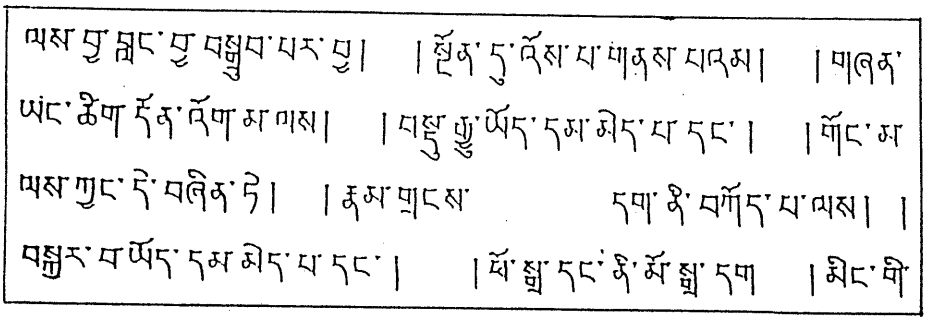

図４　チベット活字文献

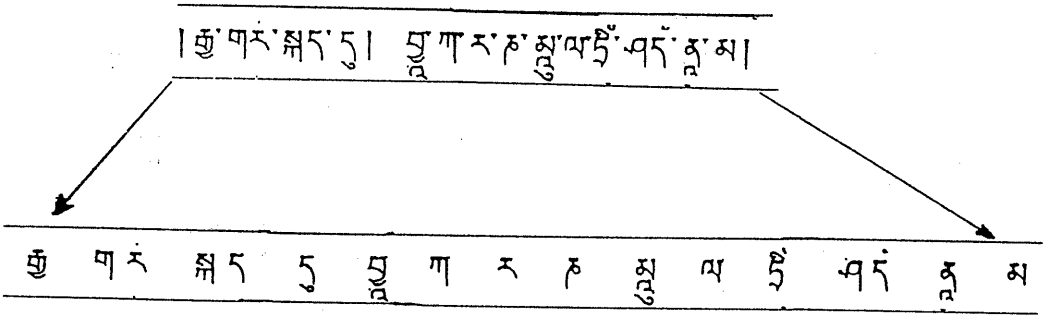

図５　チベット文字１音節単位の文字切り出し例

3　データ構造

1音節文字パターン1～7クラスのデータ構造を図6に示す。本来はこれらのデータは、イメージスキャナでチベット文字の1文字を取り込み、予め作成しているテキストデータと参照して認識し、その結果得られるものである（文献4）。本研究では、チベット文字1文字単位では正しく認識されたものとして、1音節1文字分のデータ作成を行った。図6において、番号1～3までは基本子音または重層字の表音記号、番号4は母音”ｉ”、”ｅ”、”ｏ”,”ｕ”、番号5は重層字を示す記号”ｑ”が入る。もし、母音および重層字が存在しない場合は、番号4および5の所に”－”記号を入れる。また、1音節内での文字間には”．”記号を入れ、1音節の区切りには”／”記号を挿入する。これらのチベット1音節文字が文字クラスというオブジェクトに相当する。

4．オブジェクト指向による文字パターン認識

データ作成をオブジェクト指向の考え方（文献5）により作成した場合、そのデータを活用するプログラムもオブジェクト指向にしなければならない。

本研究において、チベット活字文献からの文字パターン識別をオブジェクト指向によるプログラミングにより作成した。そのクラス設計を図7に示す。文字クラスとしてチベット1音節文字が相応する。これらの文字クラスは弁別子文字構造を使って弁別子文字タイプの7種類に分類できる。まず弁別子文字構造において、スーパクラス文字は母音クラス、基本子音クラス、重層字クラス、前接字クラス、後接字クラス、再後接字クラスの6個のサブクラスから構成されている。6個のサブクラスの内、基本子音クラスだけは文字パターン識別に直接には関与しないので、今回このクラスは作成していない。基本子音クラスを除いた各サブクラスにおいて、メソッド”判定する”がそれぞれ実行される。次に弁別子文字タイプでは1音節の文字数により、文字数1に文字タイプ1クラス、文字タイプ2クラス、文字数2に、文字タイプ3クラス、文字タイプ4クラス、文字数3に文字タイプ5クラス、文字タイプ6クラス、文字数4に文字タイプ7クラスがそれぞれ対応している。1音節の文字数によってパターンの分類が確定するのは、文字タイプ7クラスだけである。

文字数1の場合、入力文字が基本子音の時は文字タイプ1クラスで重層字の時は文字タイプ2クラスと識別する。

```
文字パターン1・2クラス　　／1 2 3 4 5／
文字パターン3・4クラス　　／1 2 3 4 5・1 2 3 4 5／
文字パターン5・6クラス　　／1 2 3 4 5・1 2 3 4 5・1 2 3 4 5／
文字パターン7クラス　　／1 2 3 4 5・1 2 3 4 5・1 2 3 4 5・1 2 3 4 5／


／　　－－　1音節の区切り記号
・　　－－　1音節内での1文字区切り記号
1～3　－－　基本子音または重層字の表音記号が入る場所
4　　－－　母音記号が入る場所
5　　－－　重層字表示記号”ｑ”が入る場所
```

図6　1音節文字パターン1～7クラスのデータ構造

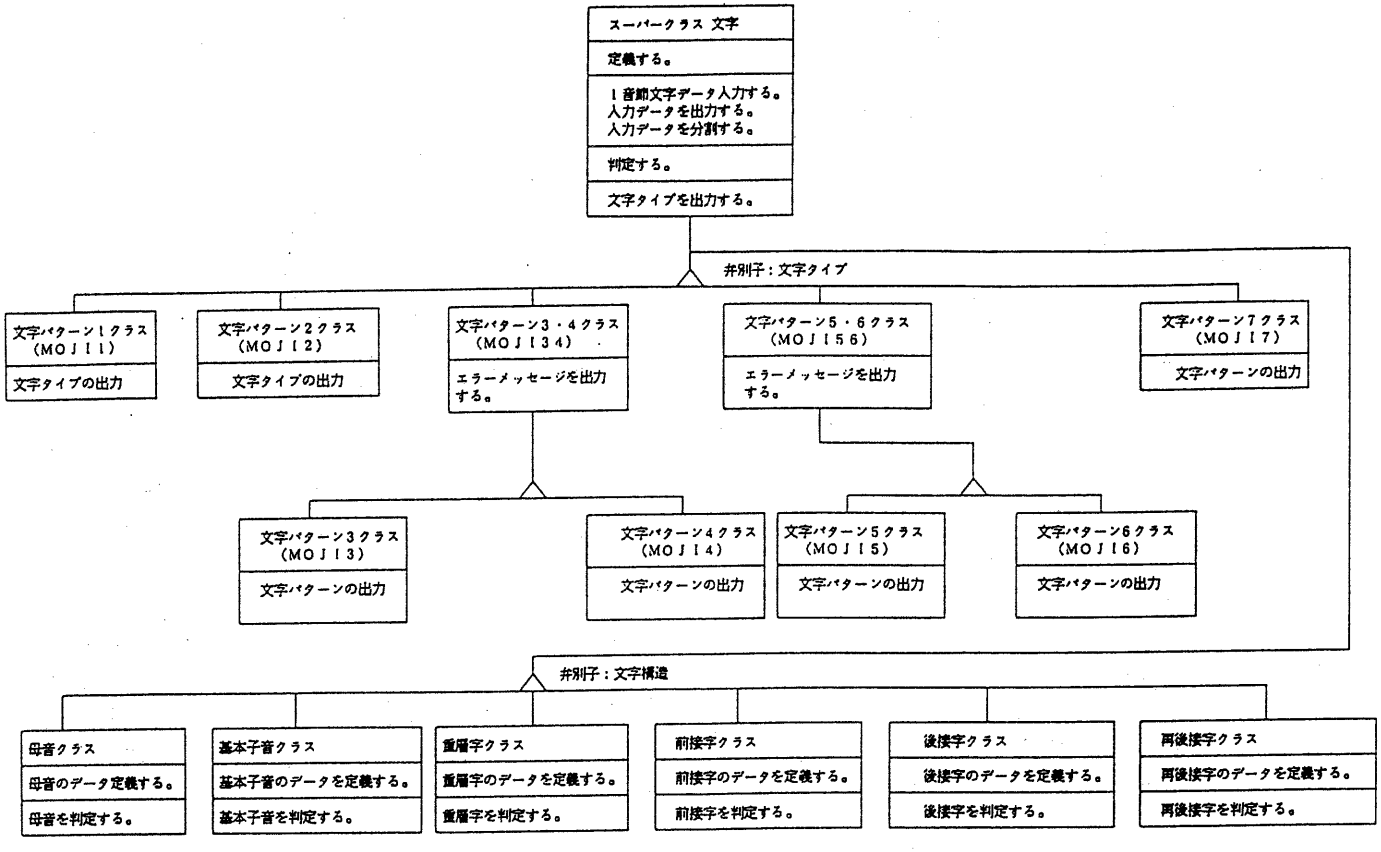

図7　オブジェクト指向によるチベット活字文献からの文字パターン識別クラス設計図

文字数２における文字タイプ３クラスと文字タイプ４クラスの識別は、入力文字の第２番目の文字が重層字かまたは第２番目の文字に母音が付く（識別条件１）または入力文字の第１番目の文字は前接字で２番目の文字は後接字でない（識別条件２）が成立した時は文字タイプ３クラスと識別する。入力文字の第１番目の文字が重層字かまたは第１番目の文字に母音が付く（識別条件１）または入力文字の第１番目の文字は前接字でなく、第２番目の文字は後接字である（識別条件２）かまたは入力文字の第１番目の文字は前接字で第２番目の文字は後接字である（識別条件３）時は文字タイプ４クラスと識別する。それ以外は文字タイプ３４クラスのエラーメッセージを出力する。

文字数３における文字タイプ５クラスと文字タイプ６クラスの識別は、入力文字の第１番目の文字が重層字かまたは第１番目の文字に母音が付く（識別条件１）または入力文字の第１番目の文字は前接字でなく第２番目の文字は後接字で第３番目の文字が再後接字（識別条件２）の時は文字タイプ５クラスと識別する。入力文字の第２番目の文字が重層字かまたは第２番目の文字に母音が付く（識別条件１）または入力文字の第１番目の文字は前接字で、第３番目の文字は後接字（識別条件２）の時は文字タイプ６クラスと識別する。それ以外は文字タイプ５６クラスのエラーメッセージを出力する。

図７において、スーパクラスとして文字クラスがあり、そのサブクラスとして文字タイプ１クラス、文字タイプ２クラス、文字タイプ３・４クラス、文字タイプ５・６クラス、文字タイプ７クラスがそれぞれある。さらに文字タイプ３クラス、文字タイプ４クラスは文字タイプ３・４クラスに汎化される。文字タイプ５クラス、文字タイプ６クラスは文字タイプ５・６クラスに汎化される。

文字タイプがどのクラスに所属するかを出力するのは、それぞれどのサブクラスにも共通するメソッドである。ここではそのメソッドを”文字タイプを出力する”とする。

また、１音節の文字タイプを識別する時、文字タイプ３・４クラスおよび文字タイプ５・６クラスにおいては識別条件１、識別条件２、識別条件３だけでは識別困難な文字が存在する場合が考えられる。その時のために文字タイプ３・４クラスおよび文字タイプ５・６クラスには”エラーメッセージを出力する”というメソッドを付け加えた。

```
Please input data --->d--i-              Please input data --->d----.n----
****** 文字ﾊﾟﾀｰﾝ1です。 *******            ****** 文字ﾊﾟﾀｰﾝ4です。 *******

Please input data --->h!---              Please input data --->g----.n!---
****** 文字ﾊﾟﾀｰﾝ1です。 *******            ****** 文字ﾊﾟﾀｰﾝ4です。 *******

Please input data --->gy--q              Please input data --->l--i-.g----.s----
****** 文字ﾊﾟﾀｰﾝ2です。 *******            ****** 文字ﾊﾟﾀｰﾝ5です。 *******

Please input data --->phy-q              Please input data --->sh---.g----.d----
****** 文字ﾊﾟﾀｰﾝ2です。 *******            ****** 文字ﾊﾟﾀｰﾝ5です。 *******

Please input data --->b----.h!-u-        Please input data --->g----.g----.d----
****** 文字ﾊﾟﾀｰﾝ3です。 *******            ****** 文字ﾊﾟﾀｰﾝ5,6ｴﾗｰ  *******

Please input data --->b----.h!-i-        Please input data --->g----.n--u-.n!---.s----
****** 文字ﾊﾟﾀｰﾝ3です。 *******            ****** 文字ﾊﾟﾀｰﾝ7です。 *******
```

図８　実験結果の一部

ここで、エラーメッセージとなるチベット文字のほとんどは特殊文字である。そのために、認識しようとしている文献によりその文字の種類および出現頻度が異なってくる。エラーメッセージとなる文字の出現頻度を調べ、出現頻度別に文字を分類し、出現頻度の多い文字種から順に新たな識別条件を文字数が２の時は文字タイプ３・４クラスに文字数が３の時には文字タイプ５・６クラスに新たなメソッドを付加したり、クラスの追加を考えるだけで良い。

この様に、オブジェクト指向によるプログラムは、現在完成しているプログラムに影響を与えずプログラムの拡張が容易に実現できる利点がある。例えば、弁別子文字構造におけるクラス設計において、文字パターン識別だけでなく最終的な文字認識を行おうとした場合、基本子音のクラスの追加と、重層字クラスの拡張を行う必要がある。正しく切り出された１９７８音節文字中、チベット文字で表音出来ない９文字を除いた１９６９音節文字で実験を行い、その結果の一部を図８に示す。エラー表示されたのは、１音節文字中に母音の数が２個以上存在したパターンが９個と、パターン５と６の識別時ミスが５個の合わせて１４個である。

５．まとめ

チベット活字文献中から１音節切り出しはサンプル数２０２０音節文字中、切り出し成功したのは１９７８音節文字で切り出し成功率は９８％である。切り出し成功した音節文字からチベット文字として表音できる１９６９音節文字を１音節毎に１文字単位で正しく認識されたものとみなし、入力データを作成した。その入力データに対する文字パターン識別のシュミレーション実験を行った結果、パターン識別は９９％の精度で実現できた。エラーメッセージは１音節中に母音が２個存在するパターンとパターン５と６の識別ミスにより出力された。

今後の課題は、１音節切り出し後の文字パターン認識を、オブジェクト指向による辞書文字を使用して行いたい。

６．謝辞

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助教授、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、実験をスムースにできる様に心配りして頂いた東北大学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）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）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).

３）稲葉：チベット語古典文法学、法蔵館 (1966).

４）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).

５）Ｊ．ランボー、M. ブラハ、W. プレメラニ、Ｆ．エディ、W. ローレンス、羽生田栄一訳：オブジェクト指向方法論ＯＭＴ－モデル化と設計－、トッパン、(1992. 7).