オブジェクト指向設計法によるチベット活字文字認識について

○小島正美[1]　布宮千夏子[2]
川村隆庸[3]　秋山庸子[4]
川添良幸n[5]　木村正行[6]

Recognition of Printed Tibetan Characters by Object Oriented Designing

Masami Kojima[1], Chikako Nunomiya[2]
Takanobu Kawamura[3], Youko Akiyama[4]
Yoshiyuki Kawazoe[5], Masayuki Kimura[6]

Abstract

The set of Tibetan characters consist of basic 30 consents, 76 combination characters, and 4 vowels. Despite the limited number, there are many similar characters and thus, special treatment is necessary to recognize them automatically.

In this paper, each Tibetan character is assigned to have an Object Oriented dictionary which is created by combining the categorization and the character identification procedures. When the Euclidean distance between the candidate characters becomes shorter than the assumed threshold values, these characters are categorized as similar characters.

１．はじめに

インド仏教は、１２００年近くチベット文化の主流を形成し、チベット人固有の文化に大きな影響を及ぼしてきた。この間に蓄積されたチベット文献資料は、膨大な量の遺産として今日我々に残されている。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も大変意義がある［１，２］。また、これまで認識対象とするチベット文献は木版刷りだけが重要な文献であるとされてきた。しかし、最近中国において重要な文献が活字版で既に相当数出版されている。これらの文献をコンピュータによりローマナイズすることにより文献学を研究する人に有用な資料を提供することができる。そのため、活字版のコンピュータによる自動認識の重要性が今日再認識されてきた。今回認識対象としたチベット活字文献の一部［３］を図１に示す。

一般に文字認識を行なう場合、大きく分けて文字認識を行なう前までと後とに分けられる。前者は前処理部と言われ、行切り出し、傾き補正、ノイズ除去、正規化、文字切り出しが行われる。これまでの認識実験においては、文字を上部、基部、下部と分割し、各部において認識した結果を最後に統合して認識していた。基部における、誤認識の大きな原因は類似文字間で起きていることが分かっている。類似文字をその文字の特徴によってグループ化し、類似文字群毎の特徴に合わせた認識メソッドとその類似文字をカプセル化し、それを新たなオブジェクト類似辞書文字とした。この様にして作成された類似辞書文字を用いることにより、認識率の改善が実現できた［４］。さらに実用に向けた９９．９％以上の認識率を達成するために、本論文ではこれまでの様に文字を上部、基部、下部と分割

しないで縦方向に１文字とし、類似文字であるかどうかの判定はユークリッド距離による重ね合わせ法により第１位候補文字と第２位候補文字以降との距離が接近している場合とする。そこで、辞書文字同士がお互いに類似文字であると判別し、個々の文字が独自に有する文字特徴とメソッドにより、認識対象文字を判別する手法を試み、その有効性を検討する。

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２に示す音節構成であり、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＋付足字は図３－１，２に示す７６通りある［５］。基字＋付頭字＋付足字は、今回の認識実験で使用したチベット文献２頁から３０頁中に、図４に示す８個が存在した。これらの文字は重層字と呼ばれている。その他に図５に示すサンスクリット文字からの転写文字が５個存在した。チベット文字の総ての子音文字は母音記号「ａ」が内在しており、子音文字の上部に母音記号「ｉ」、「ｅ」、「ｏ」に相当する記号が付いた場合、または子音文字の下部に母音記号「ｕ」が付いた場合は、内在している母音記号「ａ」の効力がなくなり、付属した母音記号が優先する。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る。子音の数が２個の場合で、母音記号「ｉ」、「ｅ」、「ｏ」、「ｕ」が存在しない場合は、初めの文字に内在している母音記号「ａ」を付ける。子音の数が３個の場合で、母音記号「ｉ」、「ｅ」、「ｏ」、「ｕ」が存在しない場合は、９つの例外を除いて基本文字の子音に内在している母音記号「ａ」を付ける。子音の数が４個の場合で、母音記号「ｉ」、「ｅ」、「ｏ」、「ｕ」が存在しない場合は、基本文字の子音に内在している母音記号「ａ」を付ける。母音記号が付属した子音文字は基本文字となり、それ以外の子音文字からは内在している「ａ」を取り除く。今回認識実験に使用したチベット文字６８２５文字中、図６に示す基本３０子音［６］の出現頻度は全体のおよそ８８％を占めている。

３．クラス設計［７］

従来のプログラム開発においては、処理機能を表すプログラムとデータを蓄えるデータベースの２つを別々に設計しており、この事がプログラム開発を複雑で判りにくいものにしてきた。オブジェクト指向開発は、この機能とデータをオブジェクトとして１つにまとめ、設計を一元化することにより、プログラムの品質・生産生・拡張生を高めようとするものである。

初めに、我々は本研究のチベット文献の文字自動認識を実験の流れに従って記述し、対象とする文章中から名詞をオブジェクトとして洗い出した。次に、洗いだしたオブジェクトの機能をメソッドとして選択し、図７に示すようなメッセージ・インタラクト・ダイアグラム［８］を作成し，それを基に、図８に示す「文字クラス」および「辞書クラス」を考えた。図８に示す「文字クラス」は文字切り出し時に得る文字の属性である主要水平線（ＭＨＬ：Main Horizontal Line）の位置情報から属性として上部文字有無フラグを生成する。文字データとして２値イメージデータからアナログデータを生成する。文字の大きさ情報は文字データ取り込みの時の大きさである。２値イメージデータからホール数およ

び４方向のヒストグラム情報を得る。「辞書クラス」は属性として、上部有無フラグ、アナログデータ、ユークリッド距離を持っている。メソッドとして「ユークリッド距離を求める」がある。「辞書クラス」から「辞書文字クラス」へ各属性を継承し、「辞書文字クラス」では属性として新たにホールの個数、ヒストグラム情報、大きさ情報を持っている。「辞書クラス」から「候補文字群クラス」への関係として、候補文字を生成せよというメッセージにより、「候補文字群クラス」において候補文字を生成する。この場合、第１位候補文字と第２位候補文字以降との距離が実験で定めたしきい値以内の候補文字を生成する。候補文字が１個の場合はその文字が認識した文字となる。候補文字が複数個存在する場合は、候補文字同士で文字間の違いを見つける。ここで重要なことは、これまでの様に類似文字であると予め設定しないで、ユークリッド距離による重ね合わせ法により、実験で定めたしきい値以内になった文字をすべて候補文字として、それらの候補文字自身が各文字の持っている固有の認識メソッドにより、新たに認識文字となる文字を見つけることである。

これまで我々が行ってきたクラス設計においては、個々のチベット文字をインスタンスとして扱ってきた。本論文においては、個々のチベット文字をクラスと考える。すなわち、個々のチベット文字は個々に有している文字特徴およびメソッドを持っている。この様に考えることにより、これまでよりもより現実に則したオブジェクト指向設計法によるチベット文字認識システムの構築が実現可能となる。現在、候補文字同士の違いを候補文字自身が見つけ、学習して行くメソッドについて実験中である。

謝辞

本研究を進めるにあたり、大変貴重なアドバイスを頂いております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に感謝致します。また、大谷大学真宗総合研究所西蔵文献班のご好意によりチベット活字文字フォントを使用させて頂きましたことを感謝致します。なお、本研究は文部省科学研究費一般研究（Ｃ）の補助を得て行っている。

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研究」、東北大学特定領域研究組織ＴＵＲＮＳ017-報告書(1992. 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）The seminar on Tibet : TEXTS OF TIBETAN FOLK-TALKS, Ⅳ, The Toyo Bunko, 1984, pp. 2-30.

４）小島、布宮、川村、秋山、川添、木村：オブジェクト指向設計によるチベット文字認識について、情報処理学会第23回人文科学とコンピュータ研究会、23-2、(1994. 9)、pp. 9-15.

５）稲葉：チベット語古典文法学、法蔵館、(1966).

６）大谷大学真宗総合研究所編：大谷大学図書館所蔵西蔵文献目録索引、(1985).

７）Ｊ．ランボー、Ｍ．ブラハ、Ｗ．プレメラニ、Ｆ．エディ、Ｗ．ローレンセン、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、(1992. 7).
８）I. Jacobson : Object Oriented Software Engineering, Addison Wesley Publishin
g Company, (1992).

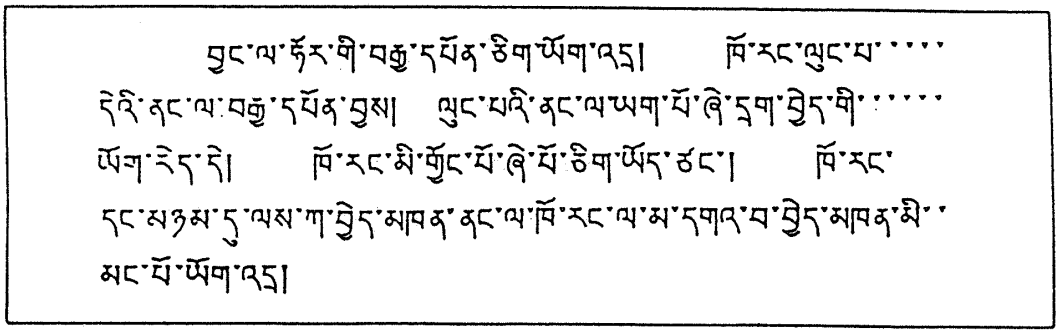


図１　認識実験を行なったチベット活字文献の一部

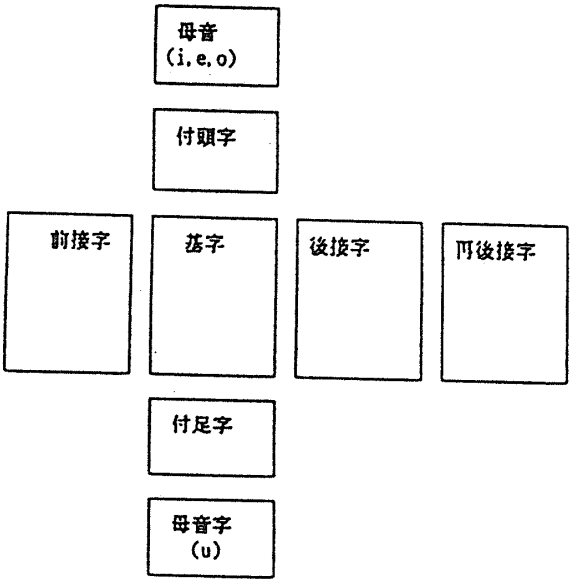
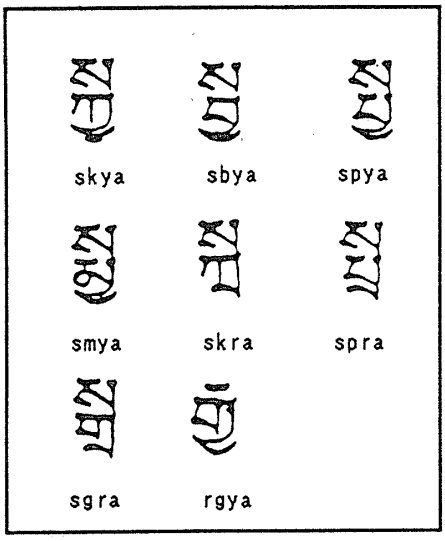


図２　チベット文字の音節構成

図３－１　２文字の組み合わせからなる７６種類の重層字
（その１）

図３－２　２文字の組み合わせからなる７６種類の重層字
（その２）



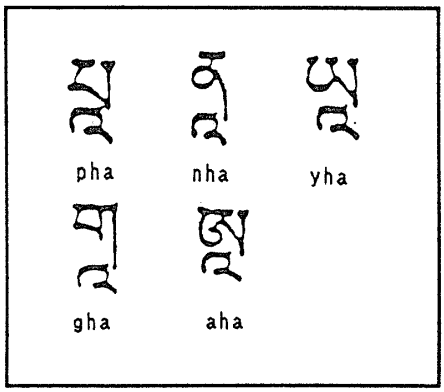


図５　サンスクリット文字からの転写文字

図４　３文字の組み合わせで出現した重層字

図6　チベット文字基本３０子音と４母音

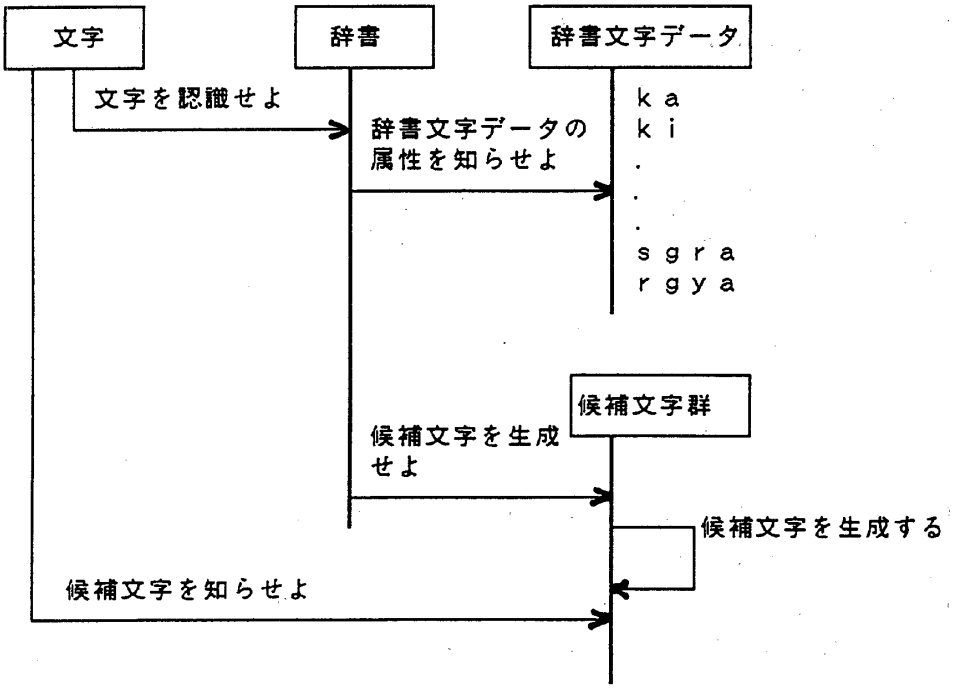


図7　メッセージ・インタラクト・ダイアグラム

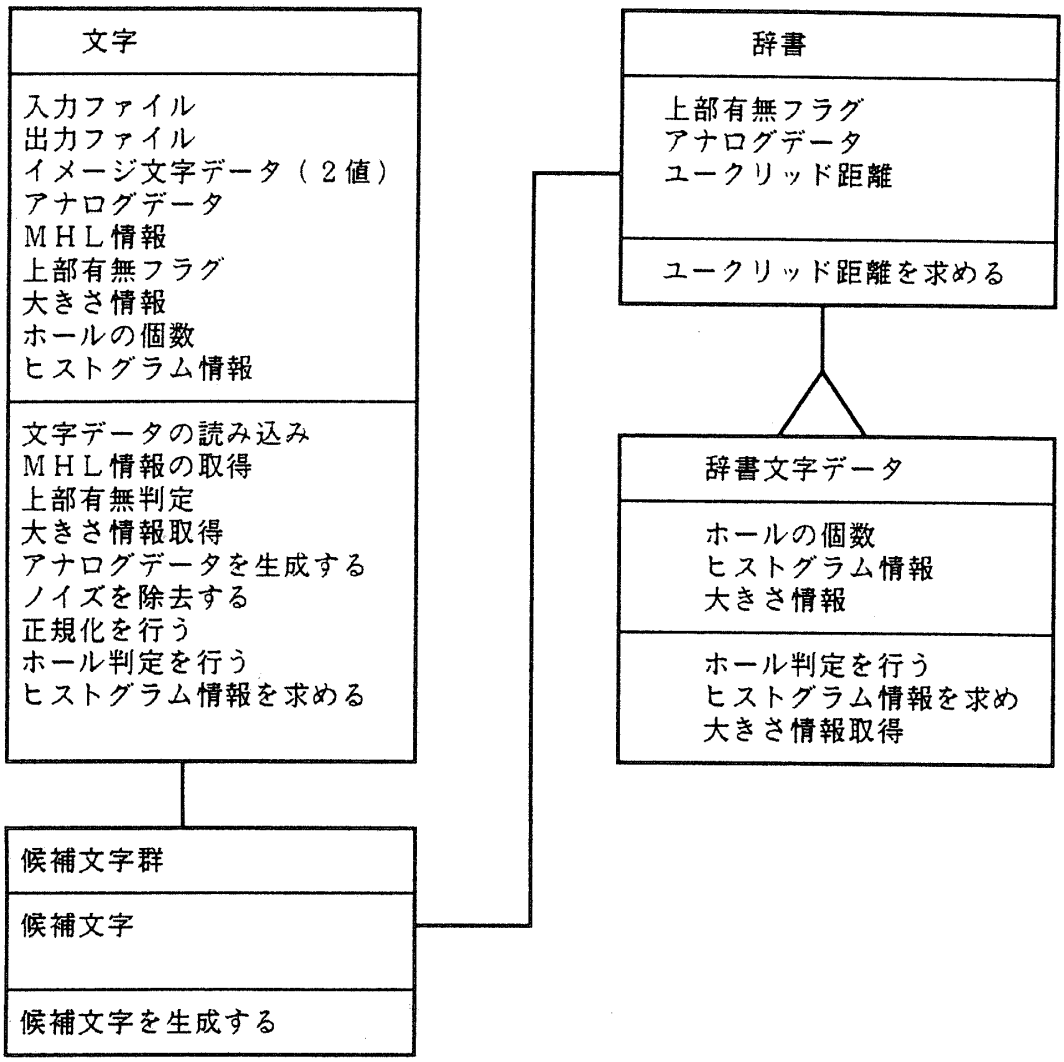


図８　文字クラスおよび辞書文字クラス

1東北工業大学・助教授　Associate Professor, Tohoku Institute of Technology.
2山形県・技師　Techinical Staff, Yamagata Prefecture.
3日本ＩＢＭ・次長　Senior Advisery SE, ＩＢＭ Japan, Ltd.
4東北大学・技官 Techinical Staff, Tohoku University.
5東北大学・教授　Professor, Tohoku University.
6北陸先端大・教授 Professor, ＪＡＩＳＴ, hokuriku