Vol.2021-CH-127 No.1
2021/8/28

デーヴァナーガリー文字 OCR の開発

加藤隆宏 [1,a)]　友成有紀 [2]　谷口力光 [3]　大澤留次郎 [4]
藤巻聡 [5]　岡田崇 [6]　橋本江美 [7]

概要：本発表は，多くのインド諸語表記に用いられる文字であるデーヴァナーガリー文字を読み取るための光学文字認識（OCR）ソフトウェアを開発するために，サンスクリット文献学の専門家とくずし字 AI-OCR 開発などを手がける凸版印刷株式会社との間で行った共同研究に関する報告である．

デーヴァナーガリー文字はヒンディー語，マラーティー語，ネパール語などの現代語のみならず，インド圏の文化や歴史などについて多くの史資料を残すサンスクリット語の表記のための主要な文字として使用されてきた．

サンスクリット文献学の分野において，サンスクリット語文献のデジタルアーカイブ化・テキストデータベース化は最重要課題であり，これまでドイツ，日本，インドを中心として様々なプロジェクトが展開されてきた．しかしながら，これらのプロジェクトはいずれも手作業（タイピング）によるデータ化が中心であり，個々の研究者の多大な時間と労力と引き換えに築かれてきたものである．今回の研究は，これまで手作業で行われてきたテキストデータ採取の方法を自動化するための OCR を開発し，それによりサンスクリット文献のテキストデータベース化を加速させることを目的とする．

重要なサンスクリット文献群を収めるアーナンダ・アーシュラマ・サンスクリット・シリーズ（Ānandāśrama Sanskrit Series）に収録された文献群を資料として用い，文字システムや文法構造についての専門知識を有する研究者と OCR 技術の開発者が共同して，矩形（データ採取の際に四角形で囲む文字の最小単位）の範囲設定，翻刻・データ化の方法などを検討した．こうして準備された「字形データベース」をもとにした AI-OCR を生成し，その読み取り精度を再検討した．

直近の課題としては活版文字に対応した AI-OCR を新たに開発することによって，将来に予想される手書き文字の OCR 開発事業の足掛かりとするとともに，この分野での着実な成果を目指した．

キーワード：デーヴァナーガリー，サンスクリット，光学文字認識，パターン認識，AI-OCR

Development of a *Devanāgarī* Optical Character Recognition (OCR) System

TAKAHIRO KATO[1,a)]　YŪKI TOMONARI [2]　CHIKAMITSU TANIGUCHI [3]
TOMEJIRO OSAWA[4]　SATOSHI FUJIMAKI[5]　TAKASHI OKADA[6]
EMI HASHIMOTO[7]

Abstract: This paper outlines some specific objectives of the research project cooperatively run by Sanskrit language experts and AI-OCR developers and discusses the process of designing “training data” through which an AI-OCR is generated. We also review some data obtained from the AI-OCR and clarify some problems found there.

Keywords: devanāgarī, Sanskrit, Optical Character Recognition, pattern recognition, AI-OCR

Vol.2021-CH-127 No.1
2021/8/28

1. 研究の背景と目的

現代の文献学研究において，検索可能なテキストデータを用いた研究は欠かせない方法の一つとなっている．近年，サンスクリット文献学の分野で盛んに取り組まれている写本校訂研究においても，テクスト批評の方法論として，本文以外の関連文書の文体・文法・用例などを検討することの有用性は早くから認められ，検索可能なテキストデータベースの利用によって本文批評の手法は飛躍的に向上した．

しかしながら，既存のデータベースプロジェクトiが提供するような，研究者それぞれの手入力によって作成されたデータベースには量的な限界があるのも事実であり，ある程度まとまった形のテキストデータベースを自動で構築するための文字認識技術の必要性がこれまで度々指摘されてきた．

このような状況をふまえ，本研究プロジェクトでは，サンスクリット語文献群を資料として用い，サンスクリット文献学を専門とする研究者とくずし字 AI-OCR 開発などを手がける凸版印刷株式会社の技術者との間で共同研究を行い，読み取り精度の高い OCR ソフトを開発することを第一の目的とした．以下，本稿では本プロジェクトで行った共同研究のうち，AI エンジンによるデータ分析の材料となる「字形データセット（教師データ）」を作成する過程で得られた諸課題とそれに対する考察と検討の結果，および生成された AI-OCR の読み取り結果についてまとめる．

2. 研究方法

2.1 ターゲット資料の選定

今回，OCR 生成のためのターゲット資料として，アーナンダ・アーシュラマ・サンスクリット・シリーズ（Ānandāśrama Sanskrit Series，現在 139 巻まで出ている）に収録された諸文献を用いた．同シリーズは，インド・プネー市にある出版元アーナンダ・アーシュラマが 1890 年代から 1930 年代頃までに出版した多くの重要サンスクリット文献を含んでいる．ここには神話・説話・文学・歴史・法典・科学・哲学等，あらゆるサンスクリット語著作がジャンルに偏らずに収録されており，多様な単語（＝多様な文字種）の採取が可能となると予想したからである．また，ジャンルの多様性は本研究プロジェクトのもう一つの柱である文献のデータベース化を行うのにも適していると言える．

2.2 転写方式

従来の研究ではデーヴァナーガリー文字を ISO 方式や KH（京都ハーヴァード）方式に基づきローマ字転写して記録したデータが広く用いられてきた．本研究ではデジタルデータ化の方法として，デーヴァナーガリー文字 Unicode に対応付ける方法を採った．デーヴァナーガリー文字をデータ化する際には比較的長く続く文節や単語をどこで区切るのかという困難があるが，デーヴァナーガリー文字 Unicode に一対一で対応づけることによって，この問題を解消することができると考えたからである．

また，ローマ字転写による表記はインドにおいて必ずしも一般的とは言えない．そのため，この方式は主にインド人研究者やインド諸語の使用者などの利便性にも考慮したものでもある．

2.3 結合文字ii

デーヴァナーガリー文字は子音文字 33 字（特殊な結合文字 kṣa, tra, jña の 3 字を加え 36 文字とすることもある）と，母音文字 14 字と，その他いくつかの記号により構成される，左書きの音素音節文字（アブギダ）である．これらの文字のうち，一つの文字の上下左右に別の文字やそれに由来する要素・記号が様々に結合することで多様な結合文字が形成される．

OCR を行うにあたり問題となるのは，(1) 数百種類以上存在する結合文字をどのように取り扱うか（ある程度の構成要素に分解するか，音節をまとめて 1 字とみなすか），(2) 特殊な形で結合する短母音 i の記号をどのように処理するか，という二点である．

デーヴァナーガリーの結合文字は，理論上は無限に作りうるが，現実には最多で 6 つの要素の結合に限定され，活版印刷用の活字ともなればパターンがある程度限定できる．そのため，完成した一音節を一矩形として音節単位の翻刻作業を行うのが妥当であると判断した．また，この矩形の設定方法は母音記号や子音 r の記号（後述）の正確な機械認識のためにも有効であると予測した．

本研究は科研費 20K20692「デーヴァナーガリー文字 OCR の開発とデータベースの構築」の成果の一部である．本稿では，本研究プロジェクトの二つの柱（OCR 開発とデーベース構築）のうち，特にデーヴァーナーガリー文字 OCR の開発に関する研究成果を報告し，データベース構築部分の成果については別稿を期す．

1 東京大学人文社会系研究科
2 公益財団法人中村元東方研究所
3 東京大学人文社会系研究科
4 凸版印刷株式会社情報コミュニケーション事業本部
5 凸版印刷株式会社情報コミュニケーション事業本部
6 凸版印刷株式会社総合研究所
7 凸版印刷株式会社総合研究所
a) tkhrkt@l.u-tokyo.ac.jp

i 例えば，GRETIL（Göttingen Register of Electronic Texts in Indian Languages），SARIT（SEARCH AND RETRIEVAL OF INDIC TEXTS），Cologne Digital Sanskrit Dictionaries（University of Cologne），TITUS（Thesaurus Indogermanischer Text- und Sprachmaterialien）等．

ii ここでいう「結合文字」とは，ブラーフミー系文字でいうところの「結合文字（saṃyukta-akṣara）」ではなく，コンピューター製版でいうところの「結合文字（combining character）」である．前者が子音のみの連続（saṃyoga）を表す文字である一方，後者は出力される際に合成されうる一切の文字を指し，母音記号・アヌスヴァーラ・アヌナーシカ・ヴィサルガなどの音節を形成する記号文字も含む．

2.4 字形データベースの開発

字形データベース（教師データ）の作成については，既存の文字認識ツールを複数テストしてこれらの弱点を分析することから始めた．

Google のデーヴァナーガリー文字 OCR の精度は数年前と比べて大きく上昇しサンスクリットも対象言語に挙げている．しかし，サンスクリットのテキストを読み取らせてみたところ，以下の 4 つのケースにおいて文字が正しく認識されないことがわかった．

（1）結合文字が長大である場合，（2）頭線（シローレーカー）上に付された鼻音記号（アヌスヴァーラ）の位置にヴァリアントがある場合，（3）頭線の上下に分かれる母音記号で，頭線の上下間に一定のずれがある場合，（4）結合文字が改行により次行冒頭に飛ばされた場合，（5）子音 r が頭線上に鉤状の子音記号によって示される場合，である．

（2），（3）は近年の印刷物では稀有ではあるが，活版印刷による版本では頻繁に起こるものである．サンスクリットで書かれた重要著作の版本で質の良いものの多くがこのような活版印刷によるものであることを考慮すると，Google によるデーヴァナーガリー文字 OCR への改善点として，これら 2 点が特に重要な位置を占めることになるだろう．

次に，ind.senz による OCR ソフトウェアを検証する．ind.senz はデーヴァナーガリー文字用 OCR を，ヒンディー語，マラーティー語，サンスクリット語の言語別に提供している．今回のターゲット資料をサンスクリット語文献としたことに鑑みて，このうちのサンスクリット語用 OCRiii を検討対象とする．

テストページで発見された誤認識全 43 件のうち，文字の掠れに起因すると思われるものを除外すると，大きく次の 3 つの場合に誤認識が生じやすいことが判明した．

（1）頭線（シローレーカー）上に付された鼻音記号（アヌスヴァーラ）の位置にヴァリアントがある場合，（2）頭線の上下に分かれる母音記号で，頭線の上下間に一定のずれがある場合，（3）特定の子音連続がある場合，である．

これらは上記で検討した Google による OCR にも共通するものであるが，SanskritOCR においては次のような特徴が見られた．まず（1）の場合には，存在する鼻音記号が認識されない場合のみならず，刊本には存在しない鼻音記号が認識される例も見られた．（2）については，母音 o，au，ī が ā に誤認識される（つまり，母音記号の頭線上部分が無視される）場合が殆どであった．最後に（3）については，子音連続の 2 文字目以降（例えば，sru という音節の子音 r）が認識されない場合や，dv などの特定の子音連続が他の単子音として誤認識される場合が多く見られた．

これら既存のツールの問題点をふまえ，字形データベースの作成時に，矩形（データ採取の際に四角形で囲む文字の最小単位）の範囲設定，翻刻・データ化の方法等について，文字システムや文法構造についての専門知識を活かしつつ検討した．また，採取した文字をもとに作成した出現頻度表を分析し，出現頻度の低い文字を選んで採取すると同時に，出現頻度が極端に低い文字種については対応する文字を完本の画像データをもとに合成してデータに追加することも試みた．これらの方法によって字形データベースの文字種・文字数がさらに充実した．なお，これまでに採取できた文字種は，翻刻による文字種（音節種）が 1477 音節，翻刻による収録文字数（ユニコード文字の数≠音節数）が 14596 文字となった．これにフォントから合成した 2829 音節（3504 文字）を追加した．

3. OCR の検証結果とその評価

3.1 1 文字 OCR の検証結果

以下，前節で述べた字形データベースをもとに生成した AI-OCR による 1 文字 OCR の検証結果を報告する．

・検証結果（予稿提出時点）

総文字数：2,433 文字

（1）認識結果（候補のうち確信度が最も高い 1 文字）が正解文字の場合のみを正解とする

正解文字数：2,234 文字

認識精度：91.82%

（2）認識結果の候補に正解文字が含まれている場合に正解とする

正解文字数：2,323 文字

認識精度：95.48 %

3.2 評価

上記で得られた AI-OCR の検証結果を精査したところ誤認識が起きているケースとして次のような事例が見られた．

（1）活字のサイズ・潰れ・印刷の不鮮明に由来するもの

（2）鮮明な活字であっても紛らわしい字形をしているもの

（3）（ヴァリアントの）字形を採集できていなかったもの

（4）同じ字形でも前後の文字の影響により異なる字形と認識されていると思しきもの

この中で，（1）については常に一定程度の混入を避けえないものであるが，このようなケースでも必ず認識失敗しているというわけでもなく，一部が欠けていても正しく認識するケースは散見された．他方，（2）と（3）については，さらなる字形の採取により教師データを増やしていく必要があり，今後の増補・調整が課題である．

また，総文字数 2,433 字中，656 文字含まれていた頭線上に記号が存在する文字については，（1）のケースが少なからず混入している中で 90.40%（593 文字）の認識成功率であった．また，事例数は限定されるが，上付きの母音記号と子音 r の鉤状の上付き記号が同時に付されるケース（**2.4**

iii SanskritOCR; http://www.indsenz.com/int/index.php?content=about

に示した既存 OCR ツールでも読み取りに失敗することの多い字形）である 11 例では，81.82%（9 例）の認識成功率であった．また，母音記号 i を含む音節については 283 例中の 93.29%（264 例）の認識成功率であった．4 文字以上の長大な子音連結については今回の評価用データには含まれなかったが，採集済の字形に合致するものであれば，3 文字程度の結合子音も適切に認識していた．

（2），（4）のケースについては，n-gram を用いた調整なども視野に入れる必要があり，これらは今後の課題となるだろう．

4. まとめと今後の課題

（1）生成された AI-OCR を使って，ターゲットとするテキスト群（今回はアーナンダ・アーシュラマ・シリーズの文献群）の文字認識テストを継続する．

（2）上述の AI-OCR のテスト結果の評価（3.2）に基づき，これをさらにチューニングして AI-OCR の精度を高める．

（3）AI-OCR の精度がある程度の水準まで達した時点で，ターゲットとするテキスト群をスキャンしてデジタル化．

（4）デジタル化されたテキスト群を検索可能な電子テキストデータベース化する．

（5）字形データベース作成に使用したアーナンダ・アーシュラマ・シリーズ以外のテキストに対して OCR 認識実験を実施し，活版の形状が異なるものに応用できるかどうかなどを検証する．この結果を分析することによって，今後の課題につなげる．

（6）今回作成した AI-OCR の手書き写本資料への応用を検討する．