様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

令和　６ 年　 ５ 月 ２０ 日現在

機関番号：５３７０１
研究種目：基盤研究(C)（一般）
研究期間：2020～2023
課題番号：２０Ｋ００６５４

研究課題名（和文）書き下し文生成を目的とする訓点資料の高精度電子化

研究課題名（英文）A High-accuracy digitisation of kunten material to generate transcriptions

研究代表者
田島　孝治（Tajima, Koji）

岐阜工業高等専門学校・その他部局等・准教授

研究者番号：９０６１１６４０

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究では、訓点資料の書き下し文の自動生成を目指し、訓点資料の高精度な電子化及びその解析を試みた。具体的な資料としては、国立国語研究所蔵　尚書（古活字版第三種本）の巻１～巻９を対称としたデータベースを構築し、そのデータを用いた画像からの文字抽出、訓点抽出を行った。この研究の成果は、国際会議である「第33回日本資料専門家欧州協会年次大会」で発表した。訓点については朱色のヲコト点のみに注目して抽出したが、位置まで正確に抽出できる割合は60%程度であり、文字の形状、使い方の特徴の考慮など、より文献に考慮した分析を行っていく必要がある。

研究成果の学術的意義や社会的意義
本研究の成果は、訓点資料という、解読に必要な知識が多いため限られた研究者しか解析できない資料を、計算機を用いて自動分析する仕組みを構築したことに社会的な意味がある。国語辞典や漢和辞典には、特定の単語の実例として漢籍（漢文による訓点資料）を提示していることが多いが、訓点研究を専門としない研究者が実際の漢文資料を使って、実例を理解することは、必要となる知識が不足するため極めて難しい。本研究では、誰もが平易な形で資料の訓点を詳細に把握できる、資料の訓点情報を詳細に記録したデータベースを構築した。
また、データ使った訓点資料の自動認識も行い、文字の位置に関しては自動で抽出できる成果が得られている。

研究成果の概要（英文）：In this study, we tried to create a high-accuracy digitisation and analysis of the kunten material with the aim of automatically generating transcriptions of the kunten material. We developed a database of the National Institute for Japanese Language and Linguistics (NINJAL) collection of the Shosho (old printed type editions, Type 3), volumes 1 to 9. The results of this research were presented at the international conference, the 33rd EAJRS Conference.
As for the kunten, we focused only on the vermilion Wokoto-ten and extracted them. However, our method was only successful for about 60% in extracting the correct position. We found that more analysis using information from the text, such as the shape of the characters and the characteristics of how they are used, is needed to be more precise.

研究分野：情報工学

キーワード：訓点資料　データベース　自動解析　ヲコト点　書き下し文

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

東アジア諸国においては、漢文が共通書記言語であった。そのため、日本を含む各国において漢文で記述された古典籍が多く残されている。日本語史の研究においては、漢文訓点資料と呼ばれる古典籍が研究対象となっている。これには、漢文の本文に加え訓点と呼ばれる注釈が付与されている。訓点は漢文を自言語として理解するための注釈であり、漢籍、仏典、国書など奈良・平安・鎌倉時代を中心にそれ以降も広く用いられてきた。訓点は、時代や利用していた人々の流派によって様々な種類が存在する。そのため漢文訓点資料は、文学、歴史学、仏教学、文化学など広く史的研究を行う他分野の研究者に活用されている。しかし、訓点研究を専門としない研究者にとってその内容を理解することは容易ではなく、扱いたい漢文訓点資料の書き下し文が手に入らない場合、研究資料として活用することは難しかった。

２．研究の目的

本研究における最大の目的は、コンピュータを用いて漢文訓点資料の可用性を高める方法を検討し、またそれを実装することである。この目的のために、本研究では、書き下し文を自動的に生成して資料と共に提供する仕組みを作ることを目指す。

３．研究の方法

漢文訓点資料の可用性を高め、研究利用を促進し、専門家だけでなく一般の人々にわかりやすい状態で漢文訓点資料を電子化したデータを提供することが重要である。漢文訓点資料の研究成果は、一般的に書き下し文として示されることが多かった。本研究では、書き下し文生成の過程を自動化することを目指す。書き下し文を自動的に生成、提供するためには、訓点の解読方法をどのように集約し、計算機で取り扱えるようにするかを考えなければならない。本研究では、解読書の一つであるヲコト点図の電子化データを整備し[1]、訓点解読の自動化を試みる。また、書き下しの結果だけでなく、どの訓点をどう解釈してその書き下し文が作られたかを、明示的に確認できるようにする方法も検討し、わかりやすい状態で資料を提供できるようにする。

具体的な手順として次の順に研究を進める計画を立てた。

①『尚書（古活字版）』に対する語順点、仮名点を反映させた書き下し文の生成
②漢文訓点資料を機械学習させ、訓点情報を自動認識する方式の検討
③他の漢文訓点資料の電子化方式の検討と書き下し文生成

４．研究成果

①のステップは、応募者らがこれまでに行ってきた『尚書（古活字版）』を対象としている。この資料については、ヲコト点のみを用いた簡易的な書き下し文の自動生成を行ったことがあり、ここにヲコト点以外の訓点を反映させる方法を検討した。訓点研究者が作成する書き下し文と生成した書き下し文を比較するために、巻１に絞り、訓点研究者による手動の書き下し文を生成し、これにより近い文章の自動生成を行った。書き下し文について検索できるデータベースを構築した[2]。データベースの構築においては、N-gram により漢文を分割し、部分一致により検索できるように工夫した。また、ヲコト点については書き下し順を一意に定めるための手法を作った。

次に、これらの漢文訓点資料の電子化結果を機械学習させ、その結果を用いた漢文訓点資料の自動的な電子化とそこからの書き下し文生成に取り組んだ。第一ステップとして、現在の訓点資料から、文字、および訓点を自動的に抽出する方法を検討し、これを実装した[3][4]。具体的には、尚書の資料画像を用いて、本文を抽出する方法を検討した。カメラ画像を用いた判断を目指し、既存の OCR を用いて抽出してみたが、この方法で抽出できる文字は 50%以下であり、また正しい文字としても認識できていなかった。OCR に文字を学習させることも検討したが、バリエーションが高いことや、他の平仮名、片仮名、漢字により訓点が重ねて書き込まれているため、OCR だけでは認識が困難であると考え、資料の翻刻されたテキストデータを使う方法を検討した。翻刻データを使えば、文字数と物理的な行数が把握できる。これと資料画像の二値化画像から抽出した画素の固まりを用いて文字位置判定を行った。二値化により得られた黒い画素をつなげてグループ化し、小さな固まりはノイズとして除去することで、文字と思われる固まりを判定させ、文字数、行数を合わせることで適切に抽出することができた。ただし、資料画像が曲がっていたり、翻刻データに間違いがある場合にはうまく抽出できない。これらについてはデータの補正方法も用意して対策した。この結果、『尚書（古活字版）』については、すべての文字の抽出を行うことができた。

次のステップとして、文字の周辺の訓点抽出を行った。解釈方法が点図データベースにより一

位に定めることが容易な朱色のヲコト点について、まずは抽出を行った。色フィルタにより画像から朱色を抽出し、文字と同様に画素の固まりをグループ化することで訓点を抽出できた。ヲコト点については、付与されている位置が大切であるため、文字を５×５のグリッドに収め、この一回り大きな７×７のグリッドを生成し、どのグリッド内に収まっているかを計算して、ヲコト点の位置情報とした。しかしながら、この方法では適切な位置として抽出されたヲコト点の割合は半数程度であり、まだ改善が必要であることが分かった。文章とヲコト点の関係性、文字との共起関係などを使う方法について現在は検討を進めている。また、漢字、平仮名、片仮名で書かれた点の抽出については、現在データベース新たに設計し、手動により抽出したデータを作るとともに自動抽出に方法を検討している。

③の他の資料については、画像の確認および対象の検討を行ったが、この研究期間においては、書き下し文の生成には至らなかった。『尚書（古活字版)』については、活字資料であり、文字及び行の配置が固定的である。一方で手書き資料においては、文字数や行数が固定的ではなく、これまでに検討してきた方法だけでは適切な抽出ができないことが推測される。国語研究所で画像公開をしている、金剛頂一切如来真実摂大乗現証大教王経、悉曇蔵、和漢朗詠集を候補として検討している。

参考文献

[1] 堤 智昭, 田島 孝治, 小助川 貞次, 高田 智和：訓点データベースと点図の自動判別, 情報処理学会論文誌, Vol.63, No.2, pp.283-292 (2022).
[2] 中村 海翔, 田島 孝治, 堤 智昭, 高田 智和, 小助川 貞次：書き下し文での訓点情報検索を可能とする訓点資料データベースの試作, じんもんこん 2022 論文集, Vol.2022, pp. 283-288 (2022).
[3] 苫米地 康太, 田島 孝治：訓点資料画像の文字位置検出と訓点資料データベースへの追加, 研究報告人文科学とコンピュータ (CH), Vol.2023-CH-132, No.14, pp.1-4 (2023).
[4] Tajima, Kōji: The Improvements of the Search-ability for Shōsho Kunten Database The 33rd EAJRS Conference (2023). （国際学会における口頭発表）

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　0件）

| | |
|---|---|
| １．著者名<br>堤 智昭、田島 孝治、小助川 貞次、高田 智和 | ４．巻<br>63 |
| ２．論文標題<br>訓点データベースと点図の自動判別 | ５．発行年<br>2022年 |
| ３．雑誌名<br>情報処理学会論文誌 | ６．最初と最後の頁<br>283～292 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.20729/00216234 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

| | |
|---|---|
| １．著者名<br>堤 智昭 , 田島 孝治 , 高田 智和 , 小助川 貞次 | ４．巻<br>2021 |
| ２．論文標題<br>訓点データベースを用いたヲコト点図の機械的分類手法の検討 | ５．発行年<br>2021年 |
| ３．雑誌名<br>じんもんこん2021論文集 | ６．最初と最後の頁<br>182～187 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>なし | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

| | |
|---|---|
| １．著者名<br>堤 智昭，田島 孝治，高田 智和，小助川 貞次 | ４．巻<br>2020 |
| ２．論文標題<br>訓点研究支援のための基盤システムの設計・実装 | ５．発行年<br>2020年 |
| ３．雑誌名<br>じんもんこん2020論文集 | ６．最初と最後の頁<br>89 - 94 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>なし | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

〔学会発表〕　計5件（うち招待講演　0件／うち国際学会　1件）

| |
|---|
| １．発表者名<br>中村 海翔，田島 孝治，堤 智昭，高田 智和，小助川 貞次 |
| ２．発表標題<br>書き下し文での訓点情報検索を可能とする訓点資料データベースの試作 |
| ３．学会等名<br>情報処理学会　人文科学とコンピュータシンポジウム2022 |
| ４．発表年<br>2022年 |

１．発表者名
堤 智昭 , 田島 孝治 , 高田 智和 , 小助川 貞次

２．発表標題
訓点データベースを用いたヲコト点図の機械的分類手法の検討

３．学会等名
じんもんこん2021

４．発表年
2021年

１．発表者名
堤 智昭, 田島 孝治, 高田 智和, 小助川 貞次

２．発表標題
訓点研究支援のための基盤システムの設計・実装

３．学会等名
じんもんこん2020

４．発表年
2020年

１．発表者名
苫米地 康太, 田島 孝治

２．発表標題
訓点資料画像の文字位置検出と訓点資料データベースへの追加

３．学会等名
研究報告人文科学とコンピュータ研究会

４．発表年
2023年

１．発表者名
Koji Tajima

２．発表標題
The Improvements of the Search-ability for Shosho Kunten Database

３．学会等名
The 33rd EAJRS Conference（国際学会）

４．発表年
2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

**6．研究組織**

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（機関番号） | 備考 |
|---|---|---|---|
| 研究分担者 | 堤　智昭<br>(Tsutsumi Tomoaki)<br>(80759035) | 筑波大学・人文社会系・助教<br>(12102) | |
| 研究分担者 | 小助川　貞次<br>(Kosukegawa Teiji)<br>(20201486) | 富山大学・学術研究部人文科学系・教授<br>(13201) | |
| 研究分担者 | 高田　智和<br>(Takada Tomokazu)<br>(90415612) | 大学共同利用機関法人人間文化研究機構国立国語研究所・言語変化研究領域・教授<br>(62618) | |

**7．科研費を使用して開催した国際研究集会**

〔国際研究集会〕　計0件

**8．本研究に関連して実施した国際共同研究の実施状況**

| 共同研究相手国 | 相手方研究機関 |
|---|---|