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　 ６ 月 ２８ 日現在

機関番号：３４３１５
研究種目：挑戦的研究（萌芽）
研究期間：2021～2023
課題番号：２１Ｋ１８３７２

研究課題名（和文）古文書解読熟練者の研究行為から抽出する音声認識電子テキストアーカイブ

研究課題名（英文）Digital text archiving for cursive writing documents by using the reading voice of which the experts read aloud.

研究代表者
赤間　亮（Akama, Ryo）

立命館大学・文学部・教授

研究者番号：７０２１２４１２

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：歴史的典籍・古文書の翻刻デジタルアーカイブを進める上で、ＡＩ画像認識を使った自動翻刻技術が進展している一方、古文書解読の能力を持つ人材は、高齢者が多く、キーボード入力が苦手な場合が多い。本研究では、解読能力を持つ人材に文書を読み上げてもらい、その音声により翻刻本文の「原稿」を記録するシステムの開発を行った。立命館大学ARCの古典籍ポータルをシステム基盤として、システムを組み込み、読み上げデータの記録保存が可能にすることができた。一方、画像認識と同様に、漢字仮名交じり本文としての正解率は、高くなく、現状では「粗稿」としての役割となっており、校正用原稿としての実用性にとどまる。

研究成果の学術的意義や社会的意義
本来日本語で書かれている古典籍や古文書を解読しようというモチベーションは、多くの人が持っており、そうしたモチベーションを活用したSNS型翻刻サイトとして「みんなで翻刻」は、著名なシステムとなっている。一方、オンラインでの翻刻作業を苦手とするが、長年の経験により解読能力が極めて高い人材を、古い文献の解読プロジェクトの協力メンバーとして参加してもらうことで、翻刻効率は大幅に向上できる。コンピュータシステムを使って翻刻本文アーカイブを試みる若手研究者と高い能力を有する高齢者とのコラボレーションを有効に機能させるシステムができたことは大きな意義があろう。

研究成果の概要（英文）：While automated transcription technology using AI image recognition is advancing in the digital archiving of transcribed historical books and documents, many of the people who have the ability to decipher ancient documents are elderly and often have difficulty with keyboard input. In this study, we developed a system to record the "manuscript" of the transcribed text by having a person with deciphering ability read the document aloud. We were able to incorporate the system using the ARC's old book Portal database as the system infrastructure, enabling the system to record and store the read-out data. As with image recognition, however, the rate of correct answers as a text mixed with kanji and kana is not high, and at present the system serves only as a "rough draft" and is only practical as a manuscript for proofreading.

研究分野：日本文学

キーワード：形態素解析　UNIDIC　古典籍　古文書　翻刻　翻刻支援

様　式　C－１９、F－１９－１（共通）

１．研究開始当初の背景

古典籍・文献のデジタルアーカイブが進展し、オンライン上でも大量に閲覧できるようになった。次の課題は、大量の古文書を「正確に解読」してビックデータとして分析できるデジタルテキストアーカイブを作ることである。

古典籍・古文書の翻刻デジタルアーカイブを進める上で、自動電子テキスト化の研究が行われてきたが、現在は、いわゆるくずし字で記録された古文書のデジタル画像から、ＡＩ画像認識を使った自動翻刻技術が進展している。

「正確な」テキストアーカイブのためには、文脈や表記のユレを考慮した上での確定作業が必要となる。しかし、AI のその正答率はいまだ十分ではなく、100%正確なテキストアーカイブをどのような手法で構築していくのかが大きな課題となっている。

一方、人の翻刻意欲に注目し、翻刻作業量を飛躍的に伸ばすことを目標に、SNS（クラウドソーシング）型翻刻サイトを運営し、翻刻に挑戦する人材を、発掘して進展させている「みんなで翻刻」システムもある。また、立命館大学アート・リサーチセンターでは、大規模な古典籍データベースをポータル型で運用しており、そのシステムにも AI による翻刻支援システムを組み入れてある。

２．研究の目的

本研究では、そうした状況を承け、翻刻本文アーカイブの一つの方策として、熟練者の解読能力を活用し、かつ熟練者の負担を出来る限り軽減して翻刻テキストを作成する方法を考案する。熟練者の読上げ音声を、読み上げ時の本文テキスト情報以外のノイズを極力軽減して、電子テキスト化し、立命館 ARC の初心者の育成・教育ができる「翻刻支援システム」と連携させ、ツールの一つとしてシステムに加え、従来の翻刻量を圧倒的に凌駕するシステムを開発するのが目的である。

それは、全自動化とは反対のアプローチ、熟練者による「翻刻」作業の究極的な効率化をも併用されるべきであるというアイディアによる。

３．研究の方法

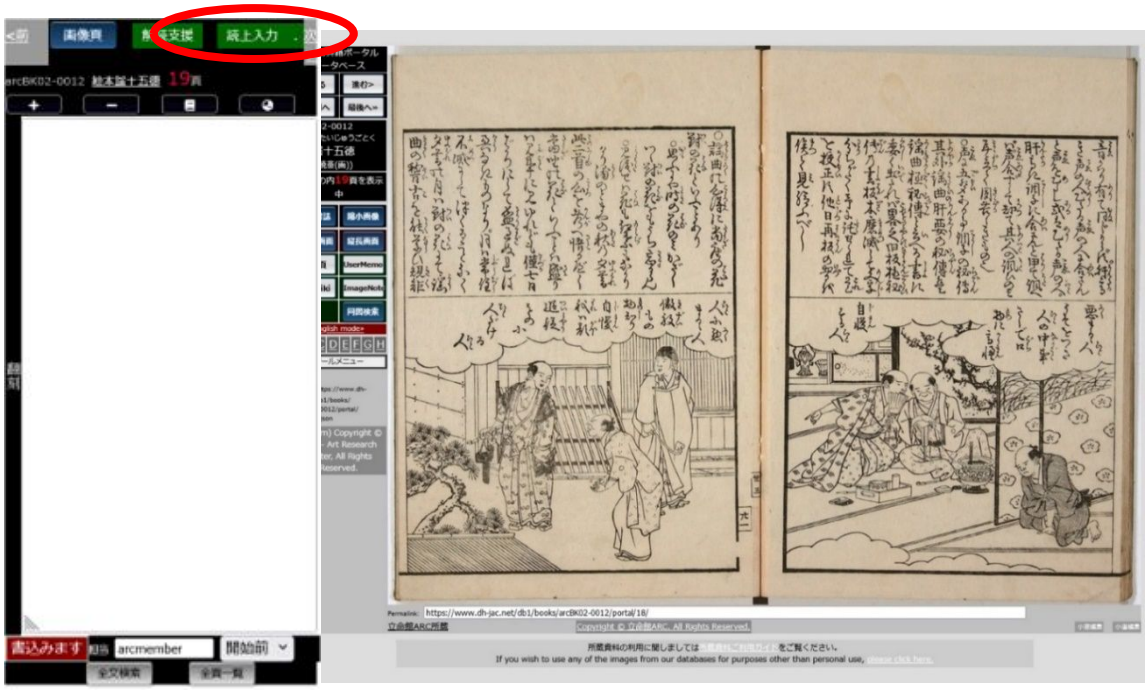
研究代表者らが開発した「AI くずし字解読支援システム」が稼働しているのを踏まえて、本研究では、

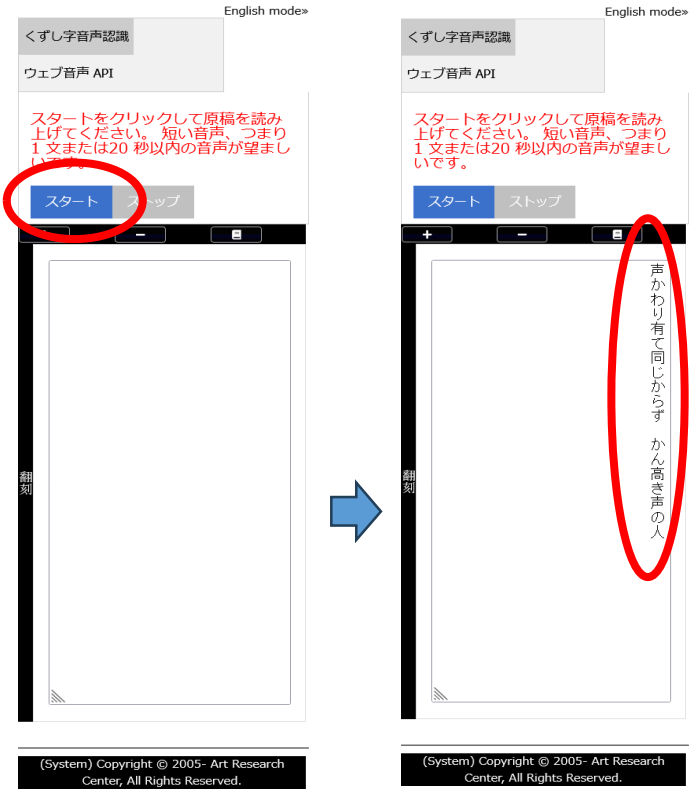
（1）熟練者の参画によって得られる読上げ音声のテキスト化

（2）解読初心者の育成・教育とを連携させたシステムの開発

を行い、従来の翻刻量を圧倒的に凌駕するスピードで翻刻テキストを生成できるシステムを提案・開発する。

①「くずし字翻刻システム」にくずし字翻刻の音声入力システムを実装





まず、二つの音声認識システムを設置し、
1) 翻刻する作品本文が現代日本語と類似している場合、「汎用日本語音声認識システム」を使用して音声入力を行う。
2) 翻刻する作品本文が現代日本語と類似していない場合は、「古文用音声認識システム」を使用することにした。

上図のように翻刻システムを開き、赤丸の「読上入力」をクリックする。すると左図のようなウィンドウが現れ、「スタート」ボタンを押すと、早速録音が開始される。

その録音ファイルは、早速、音声認識用サーバーに転送され、音声認識プログラムを用いて、認識結果が生成される。この認識結果は、直接、翻刻ウィンドウに反映され、表示される。

しかし、原本の漢字仮名交じり文は、江戸時代や明治前期の作品が中心となるため、現在の漢字宛や送り仮名の規則とはことなり、法則が一定していない。

そのため、読み上げ結果から生成されるテキストの結果と原本の本文との相違が数多く生じており、校正者が修正をしていく必要がある。

しかし、読み上げ本文が正しければ、間違った漢字宛や区切りであっても、本文と見比べることでかなりの精度まで結果を向上させることができる。

このあと、書き込みボタンをクリックすることで、翻刻システムの翻刻ウィンドウに読み上げ原稿がコピーされると、AI くずし字翻刻支援システムが使えるため、再校を進めることができ、これを以て、ほぼ完ぺきな本文を完成させることがでる。

②一方、読み上げ結果は、古文用音声認識システムの精度に依存する。この精度を上げることにより、校正者の負担をできるだけ軽減することで、翻刻スピードを飛躍的に上げることが可能である。そのためには、形態素解析データの構築が必要であり、そのための実証実験と独自の学修データ作成の試みを行った。

既存のデータセットとしては、国立国語研究所に登載されている「日本語歴史コーパス　江戸時代編　洒落本」が最も解析精度が高いことを確認した。この正誤判定に使った買った江戸中期の役者評判記「役者自虎通　京都・江戸」の本文 4000 語を使い、役者評判記の形態素解析データとして活用した。Web 茶まめによる近世江戸口語 UniDic 解析結果を元に修正を加えるもので、修正の負荷軽減のため、音声認識システム構築に必要な最小限の項目に絞り、修正した。これをもとに、MeCab のユーザー辞書を作成する。

これによって精度が少しずつ向上した環境で、音声認識によって正確なテキストがアーカイブされ、それをもとにまた、役者評判記形態素解析データを増強する。こうした循環的な作業により、古文のテキストマイニングにも活用できる形態素解析辞書が成長していく。

4．研究成果

研究成果としては、ARC 古典籍ポータルデータベースの翻刻システムに、従来の翻刻システムのインターフェイスを崩すことなく、くずし字翻刻向け音声入力システムを設置することができた。支援システムを不要とする上級翻刻者や、AI 翻刻支援システムを活用する中級者、初級者は、このシステムとバッティングすることなく、音声入力システムを立ち上げることができ、録音ボタンをクリックするだけで、次々と読み上げ原稿が翻刻ウィンドウに書き込まれる。原本の画像は、その横に表示されており、その粗稿にあたる翻刻本文を参考にしながら、翻刻が苦手な初級者が AI 翻刻支援システムも併せて使いながら、解読していくという仕組として実効性が確認できた。

一方、本研究では、読み上げ音声から、正確な漢字仮名交じりの本文の翻刻結果を獲得するワンランク上の目標も設定し直している。それには、古文用 DNN-HMM ベースの音声認識を活用し、認識システムの精度を改善していく必要がある。そのために、古文学習データ量を増強し、近世語の形態素解析データを増強していく必要があることが確認された。

また、商用音声認識 API の custom 言語モデル機能の使用を試みたり、自己教師あり学習モデ

ルを用いて End-to-End 古文音声認識システムを構築したりしている。また、入力ディバイスについても、音声入力にノイズが混入する可能性を考慮し、外部騒音の影響を受けにくい皮膚密着型マイクを使用する翻刻システムの実験も実施し、それぞれの実験結果を得ることができた。今後、この実験データをもとに、より精度の高い「音声読上式翻刻システム」に繋げていきたい。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 張宇涛，戸塚史織，耿毓庭，岩居健太，西浦敬信，赤間亮
２．発表標題 自己教師あり学習モデルを用いたくずし字翻刻のための古文音声認識システムの構築
３．学会等名 情報処理学会第86回全国大会
４．発表年 2024年

１．発表者名 Zhang Yutao、西浦敬信、戸塚史織、赤間亮
２．発表標題 古文用音声認識システムの構築とくずし字翻刻の音声入力システム
３．学会等名 2022年度「日本文化デジタル・ヒューマニティーズ拠点」プロジェクト成果発表会
４．発表年 2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

日本古典籍デジタル研究所：古典籍翻刻プロジェクト https://www.arc.ritsumei.ac.jp/lib/vm/J-book/A/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	西浦　敬信 (Nishiura Nobutaka) (70343275)	立命館大学・情報理工学部・教授 (34315)	

６．研究組織（つづき）

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	山路　正憲 (Yamaji Masanori) (00899049)	立命館大学・衣笠総合研究機構・研究員 (34315)	削除：2022年4月20日

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