CA2015

動向レビュー

くずし字資料の解読を支援するデジタル技術

橋本雄太（はしもと ゆうた）*

1. はじめに

古文書や古記録、古典籍など江戸時代以前に刊行ないし筆記された文字資料の大多数は、現代人にとって解読困難な「くずし字」で書かれている。このため人文情報学分野では、くずし字で書かれた資料の解読を支援あるいは自動化するための取り組みが続けられてきた。本稿では、こうしたくずし字資料を扱う研究開発の現状と課題をまとめる。

そもそも「くずし字」とは、文字資料のうち、楷書の文字の点画を省略した手書き文字、あるいは手書き文字を基にした版本の文字を総称して指す用語である。書道史においては草書や行書など、点画の省略段階で名称を区分し、「くずし字」という総称は用いられないが、古文書や古典籍を扱う日本史学や日本文学分野では「くずし字」という用語が使用される傾向にある。

現代人がくずし字で書かれた文字資料を解読するにあたり、まず課題となるのが、現代では使用機会が極めて限定される「変体仮名」を解読することである。たとえば現在の「す」という文字は、漢字の「寸」から派生した文字である（派生元の漢字を「字母」という）が、江戸時代には、同じ「す」という音を表現するために、「寿」「須」「数」などの漢字を字母とする複数の字体が用いられていた。こうした平仮名の異体字を「変体仮名」という。現在では、看板などの限定された場所を除けば、変体仮名を目にする機会はほとんどない。

変体仮名に加えて、前近代史料の解読を困難にしている要素が、漢字の草書表記である。たとえば「前」という漢字を草書体で筆記する場合には、筆で速記するために字画を大きく省略した形で書かれる。さらに文字は連綿体で書かれることが多いため、初学者には一字一字の区分の判読も難しい。草書の形態は時代や地域によってさまざまに変化があるが、江戸時代には御家流（青蓮院流）と呼ばれる書体が普及し、寺小屋における教育を通じて武士層から農民層まで同一の書体で文字を書くことが可能になった。しかし、明治時代に活版印刷が導入されると、楷書体による漢字表記が普及し、書道などの一部の場を別にすれば、次第に草書体漢字も利用の場を失っていった。

伝来するくずし字資料の点数については、いくつかの推計が発表されている。たとえば近世日本文学研究者の中野三敏氏は、江戸時代以前に刊行された書籍の点数は100万点を優に超すと見積もっている(1)。また同時に、このうち現在までに活字化された書籍は1万点にも満たないとの推測を明らかにしている。したがって、中野氏の推測によると、活字で読むことのできる江戸時代以前の刊行資料は、全体の1%にも満たないことになる。また伝来する文字資料には刊行資料のみならず、古文書・古記録など、多数の手書き資料も含まれる。日本史学では、刊行資料よりもこれらの手書き資料が研究資料として頻繁に利用される。古文書・古記録はその多くが図書館や博物館、文書館に収蔵されており、近年はデジタル化も活発に進められているが、旧家の蔵など文化学術機関の管理の及ばない場所にも大量に保存されている。地域歴史資料の保存活動に関わる歴史学者の奥村弘氏によれば、このような地域に私蔵されている歴史資料の数は「控えめに考えても二〇億点以上」にのぼるという(2)。

このようにくずし字資料は膨大な点数が残されており、少数の専門家の手によってそのすべてを翻刻（テキスト化）することはまず不可能である。そこで人文情報学分野では、画像認識やクラウドソーシングなどの情報技術を駆使して、くずし字資料の大規模テキスト化を実現するための取り組みがなされてきた。

2. くずし字資料大規模テキスト化がひらく可能性

実際にくずし字資料の大規模かつ高精度なテキスト化が実現した場合、それがもたらす恩恵として考えられるものには次のようなものがある。

①資料からの情報発見・情報抽出の効率化

資料の大規模なテキスト化が実現すれば、膨大な点数の江戸時代以前の刊行資料・筆記資料を対象に、全文検索をかけることが可能になる。たとえば特定の人名や地名に関連する情報をまとめて取得することが可能になり、資料から情報を抽出する効率は飛躍的に向上するだろう。これは歴史学や古典文学など資料ベースで研究を行う分野の研究者にとって福音となるはずである。

②テキストマイニングの活発化

自然言語処理技術を駆使して大量の文書データを分析し、そこから有益な情報を取り出す手法を「テキストマイニング」という。人文情報学分野では、歴史資料や古典文学作品を対象にしたテキストマイニングは最も活発な研究テーマのひとつであるが、もっぱらその

*国立歴史民俗博物館

対象はテキストデータベースの整備が進んだ欧米資料に限定されてきた。日本語資料の大規模テキスト化が実現すれば、この領域でもテキストマイニング研究が活発化することは間違いないだろう。

③異分野における歴史資料の利用活性化

大量のテキストデータが利用可能になることで、自然科学など、歴史学や文学とは異なる分野において歴史資料の活用が進むかもしれない。たとえば地震学では、古くから歴史資料の研究資料としての活用が進んでおり、江戸時代以前に発生した地震についての記録を編纂した長大な資料集が刊行されている。また歴史資料とは一見無縁に見える天文学においても、過去の天文イベントについての記述を古文書や古記録に求めるタイプの研究[3]が近年登場し、話題を呼んだ。大規模テキスト化の実現を通じて異分野の研究者が膨大な歴史資料にアクセス可能になれば、こうした新しいタイプの研究利用はいっそう活発化するだろう。

④一般層における利用活性化

歴史資料を利用するのは研究者だけとは限らない。郷土史やファミリーヒストリーに関わる調査のため、あるいは創作活動のための資料として利用するため、歴史資料の解読に関心を寄せる人々は少なくないが、多くの場合くずし字の解読が障害となっている。大規模テキスト化はこの構造を解消し、一般の人々がより容易に一次資料に触れることを可能にするかもしれない。

3. くずし字解読に関わる人文情報学研究

日本におけるくずし字資料のテキスト化に関わる研究の歴史は長く、すでに1999年には古文書を対象とした光学文字認識（OCR）開発を目標とする科研費研究が始まっている[4]。しかしながら、当時利用可能だった機械学習アルゴリズムやコンピューターの計算能力、また機械学習のための訓練用データは今日と比べて非常に限定されており、古文書OCRと呼びうるシステムの開発までには至らなかった。ただしこの一連の研究は、画像検索機能付き電子くずし字辞典や、n-gramを利用した翻刻支援システム、古文書のレイアウト認識アルゴリズムなど、古文書解読を支援する多数の成果を生み出した。

2010年代に入ると深層学習が画像認識を含む多数の分野において従来手法を上回る性能を発揮することが分かり、くずし字の自動認識への応用に期待が高まった。さらにこの分野で画期をなしたのは、2016年の「日本古典籍字形データセット」（後に「日本古典籍くずし字データセット」に改称）の公開である[5]。国立情報学研究所と国文学研究資料館が共同で公開したこのデータセットは、江戸時代の古典籍に書かれたくずし字の1文字ずつの字形画像データや文字座標データを含むもので、2019年時点で4,328文字種にわたる約100万字の字形を収録している。容易に利用可能なオープンなデータセットが公開されたことで、多数の研究者やエンジニアがくずし字の自動認識研究に参入した。さらに2019年には電子情報通信学会パターン認識･メディア理解（PRMU）研究会による「PRMUくずし字認識チャレンジ[6]」やKaggleくずし字認識コンペティション[7]など、くずし字認識アルゴリズムの精度を競う複数のコンペティションが開催された。こうしてくずし字の自動認識は、国際的にも認知度の高い研究テーマとなっていった。

以下では、この過程で公開された、くずし字解読に関する主要なアプリケーションを紹介しよう。

3.1 木簡・くずし字解読システムー MOJIZO ー

「MOJIZO[8]」は、奈良文化財研究所（以下「奈文研」）と東京大学史料編纂所が共同開発し2016年に公開した文字画像の検索システムである。MOJIZOはクエリとして与えられた文字画像を解析し、奈文研および史料編纂所のデータベースから類似した文字画像を検索することで文字の識別を行う。2017年にはスマートフォンおよびタブレット対応版も公開された。

3.2 くずし字学習支援アプリケーション

くずし字の自動認識を目指すのではなく、人間の学習支援を目的とした教育用アプリケーションも複数公開されている。代表的なものには、いずれもモバイルアプリケーションである、早稲田大学文学学術院と米･カリフォルニア大学ロサンゼルス校（UCLA）が共同開発した「変体仮名あぷり[9]」や、大阪大学文学研究科が中心となって開発した「くずし字学習支援アプリKuLA[10]」がある。

3.3 みんなで翻刻

京都大学古地震研究会が2017年1月に公開した「みんなで翻刻[11]」は、市民参加によりくずし字資料の大規模翻刻を実現するクラウドソーシング型のシステムである（E2353参照）。もともと「みんなで翻刻」は江戸時代以前の災害史料の翻刻のために開発されたシステムであったが、2019年3月までに当初の目標であった東京大学地震研究所図書室の所蔵する和古書約500点の翻刻が完了したため、同年7月に翻刻の対象を歴史資料一般に拡張した新バージョンが公開された。このバージョンの「みんなで翻刻」は、デジタルアーカイブの画像共有にかかわる国際標準IIIF（CA1989

参照）に対応しており、IIIF 準拠の任意のデジタルアーカイブと連携することが可能である。また、後述する凸版印刷および人文学オープンデータ共同利用センター（CODH）の開発したくずし字認識システムと連携しており、1 文字単位でくずし字の読みの候補を、字形の類似度に基づくスコアとともに提示する機能を参加者に提供している。2022 年 1 月時点で合計 18 件の翻刻プロジェクトが進行しており、入力文字数の合計は新旧バージョン合わせて 2,000 万字を超えている。

3.4 くずし字解読支援システム「ふみのは」

くずし字の自動認識は学術機関のみならず産業界でも研究されてきた。そうした企業の代表が印刷会社大手の凸版印刷である。同社は 2015 年に、江戸期以前のくずし字を 80% 以上の精度で認識するシステムを開発したとの報道発表を公開している[12]。さらに同社は深層学習技術を取り入れたくずし字認識システムの開発に取り組み、2021 年にくずし字資料の解読支援システム「ふみのは[13]」としてサービス公開を行った。

「ふみのは」システムでは複数文字を一括で認識する機能が組み込まれており、資料の翻刻を効率的に行うことができる。また資料のテキスト化を支援するだけでなく、解読結果を博物館展示やオンライン公開するためのソリューションや、教育機関向けに教員や学生が共同で解読を行うためのプラットフォームをも提供している点に特徴がある。たとえば立命館大学アートリサーチセンターの「くずし字翻刻学習・指導システム」が「ふみのは」の機能を一部取り入れている（E2179 参照）。

3.5 KuroNet

KuroNet[14] は、2019 年に人文学オープンデータ共同利用センター（CODH）が発表したくずし字資料の自動認識システムである。KuroNet の最大の特徴は、資料画像中の文字位置の特定から文字種の識別まで、文字認識に関わる一連の処理を人間の介在なしに行える点にある。この機能のために KuroNet は U-Net と呼ばれる医療分野の画像認識のために開発された人工知能（AI）を組み込んでおり、複雑なレイアウトの資料でも高い精度で文字位置を特定することができる[15]。また同年、KuroNet を利用した一般向けサービスとして「KuroNet くずし字認識サービス」も公開された。これにより IIIF 形式で公開されている画像であれば、誰でも KuroNet による文字認識を試すことが可能になった。

3.6 くずし字認識アプリ「みを（miwo）」

CODH が 2021 年 8 月に公開した「みを[16]」は、iOS および Android で利用可能なモバイルデバイス向けのくずし字認識アプリである。開発者のカラーヌワット・タリン氏によると、上述の KuroNet をベースとして、Kaggle くずし字コンペティションで上位を獲得したモデルのアルゴリズムを一部取り入れているという。Twitter では「みを」のユーザーが「#miwoapp」というハッシュタグを付けて利用レポートを投稿しており、日本画家や家系研究者、筆跡研究家など、様々な背景の人々が「みを」を利用してくずし字資料の解読に取り組んでいることがうかがえる。カラーヌワット・タリン氏はアプリの機能改善を継続することを明言しており、今後の発展が期待される。

4. 今後の課題と展望

最後に、くずし字解読に関わる人文情報学研究の今後の課題と展望について述べよう。

「みを」に代表される、これまで開発されたくずし字認識システムには、いまだ大きな制約が複数存在している。第一に、AI の訓練に利用されている「日本古典籍くずし字データセット」は、江戸時代に出版された木版本から文字画像を採取しているため、古文書や古記録などの手書きの資料に対しては認識精度が著しく低下してしまう。第二に、現在のくずし字認識 AI は、字形のみを手がかりに文字の識別を行っているため、字形が類似する異なる文字を判別できないという問題がある。そこで、くずし字認識 AI が発展する可能性として、次の二つの方向が考えられる。

第一は言語モデルの導入である。文章中の文字や語句の出現は一定の確率に従うことが知られており、与えられたテキストデータから文字や語句の出現確率を学習させた確率的モデルを自然言語処理分野では「言語モデル」と呼ぶ。この言語モデルを用いて、字形ベースの文字認識によって得られた複数の認識候補から、文脈的により確からしい候補を選択することが可能である。この仕組みはすでに現代語を対象とした OCR システムや手書き文字認識システムの実装に広く利用されており、くずし字の自動認識システムにおいても認識精度の向上に寄与することが期待される。しかしながら、この手法を試すにはまず日本語の歴史資料を対象とした言語モデルを構築せねばならず、そのためには AI に学習させるためのテキストコーパスを準備する必要がある。日本語の歴史資料を対象としたコーパスには国立国語研究所が公開する「日本語歴史コーパス[17]」があるが、こうしたコーパスを利用して構築した言語モデルが、どの程度くずし字認識の精度向上に貢献するかは未知数であり、今後の研究が待たれる。

第二の方向は、クラウドソーシングシステムとの連

携である。例えば、すでに「みんなで翻刻」には一文字単位でくずし字認識を行うAIが組み込まれているが、KuroNetのように文書画像全体の認識が可能なより高度なAIを導入することで、人間が文脈や背景知識をもとにAIの誤認識を修正するという協業関係が生まれるかもしれない。また、「みんなで翻刻」で翻刻された2,000万字の翻刻テキストを、くずし字認識AIの訓練に用いることができれば、認識精度の向上に大きく寄与する可能性がある。ただし、「みんなで翻刻」の翻刻文は文字座標についての情報を持たないため、これを字形データセットとして加工するには様々な工夫が必要である。また、ボランティアによる翻刻は必ずしも正確ではないので、データセット中の誤りに対して頑健なモデルを構築する必要もあるだろう。また、「みんなで翻刻」にとっては、AIとの連携が進むことで、初学者が翻刻に参加する敷居が下がるというメリットがある。2022年1月時点で「みんなで翻刻」の参加登録者は2,000人に達しているが、さらなる参加者増が実現するかもしれない。

以上のように、くずし字解読にかかわる人文情報学研究には、様々な発展の余地が残されている。その探求を続けることは、情報技術と文化・歴史との関係、そしてAIと専門家、市民との関係について考える契機をも与えてくれる。今後とも多くの人々を惹き付ける魅力的なテーマであり続けるだろう。

(1) 中野三敏. 和本のすすめ―江戸を読み解くために. 岩波書店, 2011, 260p.
(2) 奥村弘. 大震災と地域歴史遺産：災害に強い地域文化形成における大学の役割. 名古屋大学大学文書資料室紀要. 2013, No. 21, p. 133-164. https://doi.org/10.18999/bulnua.21.133, (参照 2022-01-04).
(3) 早川尚志, 岩橋清美. 特集, 歴史書から探る太陽活動：東アジアの歴史書に記録されたキャリントン・イベント. 天文月報. 2017, 110(7), p. 455-463. https://www.asj.or.jp/geppou/archive_open/2017_110_07/110_455.pdf, (参照 2022-01-04).
(4) "手書き文字OCR技術を援用した古文書翻刻支援システムの開発". 科学研究費助成事業データベース. https://kaken.nii.ac.jp/ja/grant/KAKENHI-PROJECT-11558045/, (参照 2022-01-04).
(5) "日本古典籍くずし字データセット". 人文学オープンデータ共同利用センター. http://codh.rois.ac.jp/char-shape/, (参照 2022-01-04).
(6) 第23回PRMUアルゴリズムコンテスト くずし字認識チャレンジ2019. https://sites.google.com/view/alcon2019, (参照 2022-01-04).
(7) "Kaggleコンペティション：くずし字認識". 人文学オープンデータ共同利用センター. http://codh.rois.ac.jp/competition/kaggle/, (参照 2022-01-04). Kaggleは国際的な機械学習コンペプラットフォームである。
(8) MOJIZO　木簡・くずし字解読システム. https://mojizo.nabunken.go.jp/, (参照 2022-01-04).
(9) "源氏物語から蕎麦屋の看板までマスター 変体仮名あぷり・The Hentaigana App 早大・UCLAで共同開発". 早稲田大学. 2012-11-02. https://www.waseda.jp/top/news/34162, (参照 2022-01-04).
(10) くずし字学習支援アプリKuLA. https://kula.honkoku.org/, (参照 2022-01-04).
(11) みんなで翻刻. https://honkoku.org/, (参照 2022-01-04).
(12) 山本純子, 大澤留次郎. 古典籍翻刻の省力化：くずし字を含む新方式OCR技術の開発. 情報管理. 2016, 58(11), p. 819-827. https://doi.org/10.1241/johokanri.58.819, (参照 2022-01-04).
"凸版印刷、江戸期以前のくずし字を高精度でテキストデータ化する新方式OCR技術を開発～江戸期以前のくずし字が80%以上の精度でOCR処理可能に～". 凸版印刷. 2015-07-03. https://www.toppan.co.jp/news/2015/07/newsrelease150703_2.html, (参照 2022-01-21).
(13) 古文書解読とくずし字資料の利活用サービス「ふみのは」. https://www.toppan.co.jp/biz/fuminoha/, (参照 2022-01-04).
(14) "KuroNetくずし字認識サービス(AI OCR)". 人文学オープンデータ共同利用センター. http://codh.rois.ac.jp/kuronet/, (参照 2022-01-04).
(15) カラーヌワット・タリン, 北本朝展. "くずし字認識の進化とサービス化の展開". 人文学とコンピュータシンポジウム じんもんこん2020論文集. 2020-12-12/13, 情報処理学会人文科学とコンピュータ研究会. 2020, p. 3-10. http://id.nii.ac.jp/1001/00208568/, (参照 2022-01-04).
Tarin, Clanuwat; Lamb, Alex; Kitamoto, Asanobu. "Kuronet: Pre-modern Japanese kuzushiji character recognition with deep learning". 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019. https://doi.org/10.1109/ICDAR.2019.00103, (accessed 2022-01-04).
(16) "みを(miwo) - AIくずし字認識アプリ". 人文学オープンデータ共同利用センター. http://codh.rois.ac.jp/miwo/, (参照 2022-01-04).
(17) "日本語歴史コーパス". 国立国語研究所コーパス開発センター. https://ccd.ninjal.ac.jp/chj/, (参照 2022-01-04).

Ref.
耒代誠仁ほか. "木簡およびくずし字のデジタルアーカイブを文字画像で検索するサービスの実装". 人文学とコンピュータシンポジウム じんもんこん2016論文集. 東京, 2016-12-09/11, 情報処理学会人文科学とコンピュータ研究会. 2016, p. 19-24. http://id.nii.ac.jp/1001/00176180/, (参照 2022-01-04).
小島朋佳, 植木一也. 特集, 画像技術の実利用：くずし字の翻刻に向けたディープラーニングの活用と分析. 精密工学会誌. 2019, 85(12), p. 1081-1086. https://doi.org/10.2493/jjspe.85.1081, (参照 2022-01-04).
橋本雄太. 特集, 図書館情報学とAIの新展開：AI文字認識とクラウドソーシングを組み合わせた歴史資料の大規模テキスト化. 人工知能. 2020, 35(6), p. 754-760. https://doi.org/10.11517/jjsai.35.6_754, (参照 2022-01-04).
北本朝展. 日本古典籍くずし字データセットとAIくずし字認識. 現代の図書館. 2021, 59(2), p. 102-108.

［受理：2022-02-15］

Hashimoto Yuta

Digital Technologies to Support Reading Historical Japanese Documents Written in *Kuzushiji*