【招待講演】みんなで翻刻：歴史文献資料の市民参加型テキスト化プラットフォーム

橋本　雄太†

†国立歴史民俗博物館 研究部 〒285-0017 千葉県佐倉市城内町１１７
E-mail: †yhashimoto@rekihaku.ac.jp

あらまし　歴史文献資料を対象とした市民参加型の翻刻プラットフォーム「みんなで翻刻」（https://honkoku.org）を紹介する．
キーワード　歴史資料，デジタルアーカイブ，クラウドソーシング，文字認識

Minna de honkoku: A Crowdsourcing Platform for Large-scale Transcription of Historical Documents

Yuta Hashimoto†

† Research Department, National Museum of Japanese History　117 Jonaicho, Sakura City, Chiba, Japan
E-mail: †yhashimoto@rekihaku.ac.jp

Abstract This paper introduces Minna de honkoku, a crowdsourcing platform designed for large-scale transcription of historical Japanese documents.
Keyword　historical documents, digital archive, crowdsourcing, handwritten text recognition

1. はじめに

日本国内には，古文書，古記録，古典籍など，江戸時代以前に刊行または筆記された莫大な点数の歴史資料が保存されている．近年はこれら資料のデジタル化も急ピッチで進行しており，図書館や博物館，文書館が運営するデジタルアーカイブが多数公開されている．さらに，デジタルアーカイブにおける画像共有の国際標準 IIIF (International Image Interoperability Framework) [1]の普及によって，歴史資料画像の利活用は飛躍的に容易になった．

しかしながら，これら前近代の歴史資料の大多数はデジタルテキスト化されておらず，全文検索を適用することができない．そればかりか，資料の大多数は現代人にとって解読困難な「くずし字」で記述されているため，一般の人間にはその内容を把握することすら困難である．全文テキストの不在は，歴史資料の活用を阻む大きな足枷になっている．

本稿で紹介する『みんなで翻刻』（https://honkoku.org/）は，クラウドソーシングの手法で歴史資料をテキスト化し，その活用に繋げることを目的としたプロジェクトである．「翻刻（ほんこく）」とは歴史学の用語で，古文書や古典籍などの文字資料に書かれた文字を活字化し，再利用しやすい形式で再公開する作業のことを言う．一般に，翻刻は歴史学や国文学分野の研究者や学芸員の業務として考えられているが，こうした専門家の数は限られている．『みんなで翻刻』は，インターネットを介して多数のボランティアの参加を募ることで，一挙に歴史資料の翻刻を進めようという試みである．

このプロジェクトは 2017 年 1 月に開始し，本稿の執筆時点までに 3 年半ほど運用されてきた．すでに 7,500 名以上の参加者が『みんなで翻刻』上で歴史資料の翻刻に参加し，2,100 点以上の資料の翻刻が完了している．これまでに入力されたテキストの合計は 2,900 万字に達している．

基本的に『みんなで翻刻』は「人力」ベースのプロジェクトであるが，文字認識分野の AI 研究者との協働を通じて，AI により人間の翻刻作業をサポートする仕組みも提供しており，その意味で『みんなで翻刻』は，AI 技術の応用プロジェクトとも言える．

以下では，『みんなで翻刻』プロジェクトが企画された経緯と背景，システムの設計と実装，またシステム公開後の成果について述べる．

2. 開発の背景

2.1. 京都大学古地震研究会

『みんなで翻刻』は，筆者が大学院時代に参加した京都大学古地震研究会を中心に開発された．筆者も含

む研究会メンバーが他機関に異動したため，現在は国立歴史民俗博物館，東京大学地震研究所，京都大学古地震研究会の共同プロジェクトとして運営されている．

古地震研究会は防災研究と地震研究への応用を目的として，京都大学理学研究科を中心に歴史地震資料の解読にあたっている有志の研究グループである．主宰者の中西一郎氏（京都大学大学院理学研究科名誉教授）と加納靖之氏（東京大学地震研究所）はともに理学系の地震研究者であるが，研究会には日本史学，地理学，科学史，経済学，気象学，情報学といった分野を専攻する多彩なメンバーが参加している．

2012 年の研究会の活動開始から 2017 年までに，古地震研究会の活動により 13 万文字に及ぶ地震資料が翻刻された．しかしながら，それぞれが別に専門を持つ少数の研究者グループの手で翻刻できる資料の分量には限りがある．そこで，オンラインで多数の市民の協力を募り，膨大な点数が残されている地震資料を一挙に翻刻する計画が立てられたのである．

もっとも『みんなで翻刻』の初期のアイデアは，京都大学キャンパスで開かれる定例の研究会に参加できない遠隔地の研究者と資料をオンラインで共有し，共同で翻刻を進めるためのプラットフォームを構築するというというものであった．ここからクラウドソーシング型のプロジェクトに転換したのは，以下に紹介する図書館情報学・人文情報学分野の既存プロジェクトに影響されたところが大きい．

学術におけるクラウドソーシングの利用はシチズンサイエンス（市民科学）とも呼ばれ，2000 年代後半から分野を問わず急速に普及した．2009 年にはシチズンサイエンスプロジェクトのポータルサイト Zooniverse [2]が開設され，日本でも情報学コミュニティを中心に学術クラウドソーシングプラットフォームの Crowd4U [3]が 2011 年に公開されている．

図書館情報学および人文情報学分野では，図書や歴史資料に書誌情報や全文テキストなどのメタデータを付与する作業に，クラウドソーシングがさかんに利用されてきた．たとえば 2008 年にオーストラリア国立図書館が開始した Australian Newspaper Digitization Project (ANDP) は，19 世紀から 20 世紀中葉にかけオーストラリア内で刊行された総計 2,000 万ページに及ぶ新聞記事のデジタル化プロジェクトである．ANDP は OCR 処理により新聞記事のフルテキストも提供したが，不鮮明な印刷のため多数の誤認識が発生していた．その修正を，オーストラリア国立図書館は同館の Web サービスを通じてボランティアに依頼したのである．このプロジェクトでは 2,600 万行の新聞記事がボランティアによって校正された[4].

英国のユニバーシティ・カレッジ・ロンドンが 2010 年に開始した Transcribe Bentham は，功利主義の提唱者として知られる哲学者 Jeremy Bentham（1748-1832）が遺した 4.4 万ページから成る未刊行の手稿群を，インターネットを通じて参加するボランティアの手によって全文翻刻する試みである[5]．ニューヨーク・タイムズなど複数の有力メディアに取り上げられたことから Transcribe Bentham は世界的な注目を集めることに成功し，2020 年 7 月時点で 2.8 万ページが同プロジェクトを通じて翻刻されたと報告されている[6].

学術由来のプロジェクトとは異なるが，日本国内では『青空文庫』が大きな成功を収めた[7]．これは編集者・著述家の富田倫生氏らの呼びかけにより 1992 年に開始したプロジェクトで，著作権が失効した近代以降の刊行作品をボランティアの手で翻刻し，XHTML やプレーンテキスト形式でインターネット公開している．2020 年 8 月時点でおよそ 1.6 万点の作品が青空文庫上で公開されており，多数の人々に愛用されている．

2.2. くずし字解読の問題

筆者も含む古地震研究会メンバーは，2016 年夏頃から上に挙げたクラウドソーシングの事例について学び，同様の手法が歴史地震資料の翻刻に適用できるか検討を始めた．翻刻作業にボランティアの力取り入れる上で，最大のネックになると考えられたのは，変体仮名や草書体漢字などの「くずし字」解読の問題である．

現在の日本語の筆記体系は，それぞれ 46 文字から成る平仮名と片仮名，これに漢字やアルファベットを組み合わせた文字で構成されるが，このシステムが成立したのは明治時代以降のことである．江戸時代以前には，現在使用されていない多数の平仮名が存在していた．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）．しかし江戸時代以前には，同じ「す」という音を表現するために「寿」「須」「数」などの異なる漢字を字母とする複数の字体が用いられていた（図 1 上）．こうした平仮名の異体字を「変体仮名」という．現在では，飲食店の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．このため江戸時代以前の文献を解読するためには，まず数十種類の変体仮名の読み方を習得する必要がある．

変体仮名に加えて，くずし字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（図 1 下）．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，変体仮名と同様に次第と草書体漢字も利用の場を失っていった．

図 1上：変体仮名の「す」（出典：MJ 文字情報一覧表，CC BY-SA 2.1）．下：草書体で書かれた「前」の例．

3. システムの設計

3.1 学習をベースにしたモチベーション設計

くずし字解読の問題に対応するにあたって，筆者を含む古地震研究会メンバーが着目したのは「教育」であった．クラウドソーシングのシステムを，くずし字解読の教育サービスとしてデザインすることで，翻刻作業と参加者のスキル向上が同時に実現できると考えたのである．このアプローチに期待されるメリットは他にもあった．第一に，学習用コンテンツを無償公開することで，歴史や古典に高い関心をもつ多数の潜在的参加者の注目を我々のプロジェクトに惹きつけることができる．広報がプロジェクトの成否を左右するクラウドソーシングにとって，これは非常に都合がよい．第二に，参加者の学習意欲を，そのまま翻刻作業に参加するモチベーションに振り向けることができる．一般に学習は長期的プロセスであるから，これは参加者の継続的なプロジェクトへの参画に寄与する．

教育とクラウドソーシングを接続するアイデアは，言語教育サービスの Duolingo から着想したものであった[8]． Duolingo はクラウドソーシングの泰斗 Luis von Ahn が創設した，1 億人以上のユーザーを擁する世界最大の言語教育プラットフォームである．2020 年現在の Duolingo は広告やユーザー課金により事業収入を得ているが，その初期のビジネスモデルは次のようにクラウドソーシングに基づくものであった．Duolingo のユーザーのうち，一定の習熟度に達したと判断されたユーザーは，言語学習の一環としてニュース記事などのコンテンツの翻訳に招待される．このコンテンツは実は Duolingo が CNN や BuzzFeed などの提携企業から受託したものであり，Duolingo はユーザーによる翻訳結果を納品することで提携企業から翻訳料を得ていたのである．『みんなで翻刻』の設計は，Ahn が「一石二鳥」（twofer）と呼んだこのモデルを大いに参考にしている [9]．具体的には，以下の二つの形で教育がクラウドソーシングに組み込まれている．

第一は，『くずし字学習支援アプリ KuLA』との連携である[10]．KuLA は 2016 年に大阪大学文学研究科を中心とするチームによって開発された，初学者向けの学習アプリケーションである．iOS と Android の両プラットフォームに対応し，これまでに 14 万回のダウンロードを記録している．『みんなで翻刻』は KuLA が提供する全学習コンテンツを収録しており，参加者がくずし字解読の初歩を独習できるように学習セクションを設けている．ここでくずし字の読み方の基礎トレーニングを積んだ後，翻刻を通じて文字解読の実践的訓練に移行できるように設計されている．

一方で，初学者向けのアプリに過ぎない KuLA を利用しただけでは，くずし字を正確に翻刻することは難しい．そこで第二の施策として，参加者相互の添削システムを実装してある．『みんなで翻刻』の参加者は図 2 に示す翻刻エディターを利用し，画面右側に表示される資料画像の翻刻文を画面左のエディターに入力していく．翻刻文を保存すると，その内容は図 3 に示す「タイムライン」上に時系列順に表示され，全参加者に共有される．このタイムライン上では，誰でも他の参加者の翻刻を閲覧することができ，翻刻に誤りが含まれていた場合にはその場で修正を施すことができる．翻刻文が修正された際には，元の作業者に修正内容を示すフィードバックが送信されるので，元の作業者は誤読した文字や，自分では読めなかった文字の正解を確認することができる．仮に翻刻文の内容に自信が持てない際は，翻刻文の保存時に表示される「添削希望」のチェックをオンにすることで．上級者に明示的に添削を依頼することも可能である．こうして参加者が相互に翻刻を添削することで，参加者に協調的な学びの効果が得られると同時に，複数の参加者による翻刻文の多重チェックを促し，成果物の品質向上にもつなげることができる．

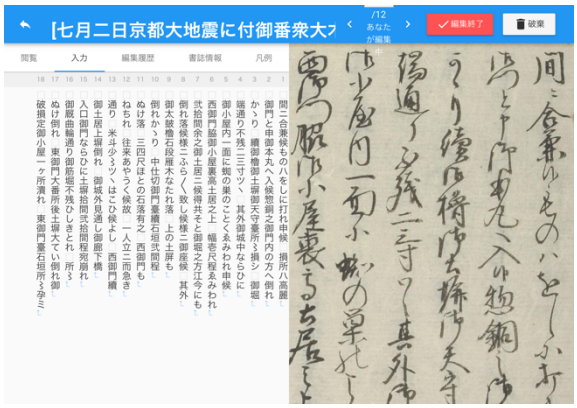

図 2 翻刻文エディター

図 3 参加者の活動を表示するタイムライン

3.2 文字認識 AI によるサポート

2017 年 1 月に公開された『みんなで翻刻』の最初のバージョン（バージョン 1）は，完全に「人力」ベースのシステムであった．ところが，この時期から AI によるくずし字認識の研究が急速に発展を始めた.『みんなで翻刻』も AI 研究者との協働を開始し，2019 年に公開したバージョン 2 では AI による文字認識機能の提供を開始した．この変化を引き起こしたのは，国立情報学研究所と国文学研究資料館が 2016 年 11 月に共同公開した『日本古典籍くずし字データセット』である[10]．これは文字単位でくずし字の画像を切り出し，Unicode のラベルを付したデータセットで，2023 年 3 月時点で約 4,000 文字種にわたる 100 万文字を収録している．AI 研究の現場では頻繁に見られる現象であろうが，容易に利用できるオープンなデータセットが登場したことで，多数の研究者がこの分野の研究に参画を始めた．2019 年には，くずし字認識の Kaggle コンペティションも開催された[11]．古典籍の複雑なレイアウトを解析可能な AI も登場し，くずし字認識 AI の性能は日進月歩の勢いで発展を続けている．

バージョン 2 の『みんなで翻刻』では，2 種類のくずし字認識 AI を提供している．ひとつは人文学オープンデータ共同利用センター（CODH）のタリン・カラーヌワット氏が開発した KogumaNet である[12]．これは Tensorflow.js で動作する軽量なモデルで，Web ブラウザ上で動作が完結する．もう一方は凸版印刷株式会社が開発したくずし字認識 AI で，こちらは同社が提供する Web API を介して利用している．いずれの AI も，上述のくずし字データセットを用いている．

『みんなで翻刻』上では，くずし字 AI は次のように利用される．翻刻エディター上で，作業者が資料画像中のくずし字の領域を指定すると，図 4 に示すように AI が推定した文字の読み方がスコア付きで表示される．したがって AI はあくまで読み方の候補を提示するだけで，最終的な読み方を決定するのは人間の作業者である．

『みんなで翻刻』における AI の役割は，人間の作業者を置き換えることではなく，あくまで人間の作業者をサポートすることである．特にくずし字解読に熟達していない初学者の参加者にとって，古文書や古典籍に書かれた文字を正確に翻刻することは非常に難しい．こうした参加者にとっては，AI がいわば「補助輪」となって，文字の読み方の候補を提示するだけでも大きな助けとなる．もちろん AI が必ずしも正しい候補を示すとは限らないが，AI の推定を人間が文脈や背景知識にもとづく判断で補うことで，初学者でも相当正確にくずし字を翻刻できることが経験的に明らかになりつつある．

図 4 「鹿」のくずし字の認識結果．下に見えるのはマスコットの「そらまる」．

4. 翻刻結果の教師データとしての活用

『みんなで翻刻』上で入力された 2900 万文字に及ぶ翻刻テキストは，Creative Commons BY-SA ライセンスで提供されており，出展を示す限りにおいて自由に利用可能である．翻刻テキストは API[13]から取得できるほか，全文を一括ダウンロードできる github レポ

ジトリ[14]を設けている．ここ最近は，機械学習のためのデータセットとして注目を受けることが多く,『みんなで翻刻』データセットを用いた AI 研究者との共同研究も進みつつある．

たとえば国立国会図書館の実験的検索サービスを運営する「次世代デジタルライブラリー」にて 2022 年末に古典籍資料を対象とした OCR テキスト化実験[15]が実施され，国立国会図書館が所蔵する古典籍資料のうち 6 万点程度の全文検索が可能になった．この OCR プログラムの訓練には『みんなで翻刻』の翻刻テキストが利用されている．

また,『みんなで翻刻』における AI とクラウドソーシング技術の統合についての研究計画を 2022 年に Google の研究助成プログラムに提出したところ，3 万ドルの研究助成を得ることができた．この資金を用いて，今後さらに AI 技術との結びつきを深める研究を進めていく予定である．

5. おわりに

本稿では，前近代の歴史資料を対象としたクラウドソーシング翻刻プロジェクト『みんなで翻刻』について，開発の経緯と背景，教育を組み込んだクラウドソーシングの設計と実装，またシステム公開後の成果について述べた．

デジタル画像化された歴史資料の公開は近年急ピッチで進められてきたが，その活用についてはまだ技術の発展が追いついていないように思われる．特に日本語で書かれた歴史資料の場合，くずし字の解読が資料の活用を阻む高いハードルとなってきた．

『みんなで翻刻』の成果のひとつは，くずし字で書かれた歴史資料の翻刻のように，非常に高度なスキルを要求されるタスクに対しても，クラウドソーシングの手法が有効であると実証したことである．AI によるくずし字認識技術の急速な発展を通じて，非専門家が歴史資料の翻刻に参加する障壁はさらに低下しつつある．くずし字認識 AI の性能は，翻刻の支援ツールとしてすでに実用レベルに達している．今後，データセットの拡張とアルゴリズムの改善を通じて，さらに向上を続けるだろう．

こうしたクラウドソーシングや AI の発展は，一見「専門家」の牙城を切り崩す運動のようにも見える．実際，最近のくずし字認識 AI の発展をふまえて，「古文書解読を生業にしていた専門家の地位が危機にさらされるのではないか」といった論調の意見がしばしば表明されることがあった．しかしながら，筆者はこの見解には与しない．というのも，歴史資料の内容を適切に把握するには，くずし字を読む能力だけでは不十分であり，現代とは異なる語彙や生活習慣，時代背景についての専門的な知識と分析能力が欠かせないからだ．こうした知識やスキルを，現在の AI により代替することは難しい．また，ボランティアが入力した翻刻文や，AI が出力した翻刻文の正確性を検証するにあたっても，くずし字解読に熟達した専門家の協力が不可欠である．こうした理由から，クラウドソーシングや AI によって歴史資料のテキスト化が進むことで，むしろ専門家の能力が必要とされる機会は増えるのではないかと筆者は予想している．大量のテキストが翻刻されても，その意味を適切に解釈できる人間がいなくては活用が難しいからだ．

特定分野の訓練を積んだ少数の専門家と，インターネットを通じて連携する多数のアマチュア，そして人間の判断を部分的に肩代わりする AI という構図は，歴史資料の翻刻の世界のみならず，多くの分野に共通して見られるようになった．この三つのアクターを効果的に連携させる仕組みの構築が，学術を含む今後の知的生産のあり方を左右する鍵になるのではないだろうか．

文　献

[1] International Image Interoperability Framework. https://iiif.io/.

[2] Zooniverse, https://www.zooniverse.org/.

[3] Crowd4U, https://crowd4u.org/.

[4] Holley, R.: Crowdsourcing: How and Why Should Libraries Do It?, D-Lib Magazine, Vol. 16, No. 3/4, (2010).

[5] Moyle, M: Manuscript Transcription by Crowdsourcing: Transcribe Bentham, LIBER Quarterly, Vol. 20, No. 3-4 (2011).

[6] Transcription Update – 21 July 2020 | UCL Transcribe Bentham, https://blogs.ucl.ac.uk/transcribe-bentham/2020/07/21/transcription-update-21-july-2020/

[7] 大久保ゆう：クラウドソーシングを先取りした青空文庫の軌跡 -ボランティアによる電子ライブラリ活動-，情報処理 55 (5), 470-474，(2014).

[8] Duolingo, https://duolingo.com/

[9] Mayer-Schönberger, V., Cukier, K., Learning with Big Data: The Future of Education. Houghton Mifflin Harcourt. pp. 9–10. ISBN 978-0-54435550-7 (2014).

[10] くずし字学習支援アプリ KuLA, https://kula.honkoku.org/.

[11] 北本朝展：Kaggle くずし字認識 ―世界規模の人文系コンペ開催への挑戦―，人工知能, Vol. 35, No. 3, pp.366-376 (2010)

[12] KogumaNet くずし字認識サービス，http://codh.rois.ac.jp/kuzushiji-ocr/#kogumanet

[13] みんなで翻刻 API, https://wiki.honkoku.org/doku.php?id=api.

[14] みんなで翻刻データセット，https://github.com/yuta1984/honkoku-data

[15] 国立国会図書館次世代デジタルライブラリー：古典籍資料の OCR テキスト化実験，https://lab.ndl.go.jp/data_set/r4ocr/r4_koten/