学習を動機付けに利用した 前近代災害史料のクラウドソーシング翻刻

橋本　雄太（京都大学文学研究科）

近年，機械処理の適用が困難な大量のデータに対して，不特定多数のボランティアの協力をインターネット上で募り，分類やタギング等のタスクを実施するクラウドソーシングの手法が人文学分野でも一般的になりつつある．一方で，日本語の歴史資料の翻刻に同手法を適用するためには，①現代人には解読困難なくずし字と，②多数のボランティアを長期的に従事させる適切な動機付けの設計の２点が課題となる．本研究では，人文学コンテンツの学習サービスと，クラウドソーシングシステムを統合することによって，ボランティアの学習意欲を動機付けに利用するとともに，タスク実施スキルの向上を図る手法を提案する．また実際に，くずし字の学習サービスと前近代災害史料の翻刻システムを統合した Web システムを構築し，その有用性を評価する．

Integrating Online Learning with Crowdsourcing: Collaborative Transcription of Japanese Pre-modern Disaster Records

Yuta Hashimoto (Graduate School of Letters, Kyoto University)

The recent years have seen a wide range of acceptance of crowdsourcing techniques in academic fields. In humanities, crowdsourcing has become a major option for scholars, curators, and librarians to tag, classify, contextualize, or transcribe a large number of cultural heritages. However, it still remains an open question how to motivate properly the contributors of crowdsourcing projects in humanities domain. In this paper, the author propose a method for designing a crowdsourcing project that makes use of learning materials for motivating its contributors, as well as a crowdsourced transcription project of pre-modern disaster records following the method.

１．はじめに

近年，インターネット上で募った多数の市民に対して画像のテキスト化や分類などのタスクを依頼し，学術研究に活用する「クラウドソーシング」「シチズンサイエンス」と呼ばれる形態のプロジェクトが活発に実施されている．人文情報学分野では，哲学者 J. Bentham の大量の遺構を翻刻するロンドン大学のプロジェクト"Transcribe Bentham"[1]が成功事例として知られる．日本国内でも，OCR による自動テキスト化が困難な大量の歴史資料画像を翻刻する手段として，クラウドソーシング[1]が注目を集めている[2][3].

しかしながら，日本語の歴史資料，特にそのボリュームの大半を占める前近代の文献資料にクラウドソーシングの技法を適用した事例は，これまでのところ存在しない．その実施にあたっては，次の 2 点が主要な課題になると思われる．

第一は，タスクの難易度の問題である．たとえば，中世から活版印刷が普及していた欧州とは異なり，日本の前近代文献資料は，大部分が現在では使用されない「くずし字」で記述された木版本か書写本である．これら資料の解読能力を有する人間は，現在の日本人口の 1%にも満たないと言われている[6]．このため日本語文献資料を対象にした翻刻タスクの難易度は，欧米のそれと比較して著しく高い．

[1] なお，「クラウドソーシング」という用語は，2006 年に雑誌 Wired の編集者 Howe が，企業のアウトソーシングの一形態を表現する用語として初めて使用した造語である[4]．その後，学術分野でも大量の機械処理困難なデータを処理する手法として流通した．しかしながら，学術分野のクラウドソーシングには，多くの場合金銭による報酬が発生せず，コミュニティ事業的性質が強いことから，人文情報学分野でもクラウドソーシングという用語の利用には異論がある[5]．しかしながら，すでに広く流通している用語であることを考慮して，本稿では，「インターネットを介して参加する不特定多数のボランティアによって，機械処理困難な大量のタスクを実施すること」という意味でこの用語を利用する．

第二は，参加者の動機付けの問題である．クラウドソーシング型のプロジェクトを成功させるには，多数の参加者の適切な動機付けが不可欠である．しかし歴史資料を対象とした学術クラウドソーシングでは，金銭報酬のような外的動機を与えることは困難である．これまでの先行研究では，参加者相互の社会的接触の提供や娯楽的要素の導入など，様々な形態の動機付けが試みられている[7]．しかしながら，特に人文学分野のクラウドソーシングにおいて，他の方法より特別に優れた動機付けの方法は明らかになっていない．

本研究では，人文学分野のクラウドソーシングにおける上記の 2 課題を解消するための手法として，人文学コンテンツを対象にした学習支援サービスを組み込んだクラウドソーシングシステムの設計を提案する．また，実際にこの手法に基づき，前近代日本の歴史災害史料のクラウドソーシング翻刻システムを構築し，その有用性を評価する．

2．クラウドソーシングにおける学習コンテンツの活用

日本国内ではあまり大きな注目を受けることはないが，ここ数年の間にネットワークを介したオンライン学習は目覚ましい発展を遂げてきた．その代表はオンラインで提供される大規模講義の総称である MOOC（massive open online course）である．2011 年の段階では，MOOC を提供する大学は 1 大学のみに限られ，受講者の総数は 16 万人に過ぎなかった．しかし，2016 年度の段階では， MOOC 講義を提供する大学の数は 570 校を超え，受講者の総数は 3,500 万人にまで増大している[8].

また，モバイルマーケットにおいても，学習・教育アプリケーションは人気の高いコンテンツである．たとえば Google Play では，「教育」は「エンターテイメント」「ライフスタイル」等の他カテゴリを抑え，最も人気のあるカテゴリである．2014 年の段階で，Google Play 上では約 83,000 件の学習用アプリケーションが公開されている．

これらの数字は，「教育・学習」が多数の人々の関心を継続的に集めることのできるコンテンツであり，最近のテクノロジーの発達が，そうした関心に応えることに成功しつつあることを示している．タスク実施に多数の人々の参画を必要とするクラウドソーシングに，学習的要素を組み込むことは，ごく自然な発想であると思われる．

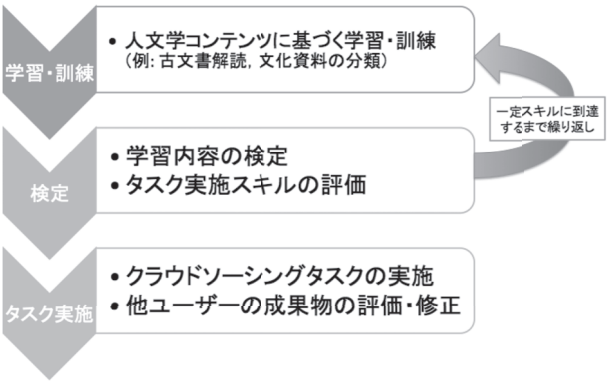

図 1 学習サービスを組み込んだ人文学クラウドソーシングの作業フロー

クラウドソーシングに学習要素を導入し，大きな成功を収めた事例として，Luis von Ahn（reCAPTCHA[1]の開発者）が CEO を務める語学学習サービスの"Duolingo"[2]が挙げられる．Duolingo は語学コンテンツを無料で提供するサービスであるが，一定レベルの習熟度に達した利用者に対して，Duolingo が他企業から受注した翻訳タスクを実施させることで事業収入を得ている[9]．2016 年までに Duolingo の登録ユーザー数は 1 億人を超え，コンテンツ翻訳に関して CNN や BuzzFeed と提携を結ぶなど，多大な成功を収めている．しかしながら，人文学資料を対象にしたクラウドソーシング事業において，同様の手法を採用した事例は見られない．

3．提案手法

本研究では，人文学資料を対象にしたクラウドソーシングにおいて，学習コンテンツをシステムに導入することで参加者の動機付けとスキル向上を促す手法を提案する．

本手法では，クラウドソーシングの参加者がタスク実施にあたるまでのフローを 3 段階に分割する（図 1）．第 1 段階では，タスクの実行に必要となる知識・技術を，関連する人文学コンテンツ（資料画像やテキストなど）をもとに学習・訓練する．例えばタスクの種類が古文書の翻刻であれば，くずし字の解読方法の習得をこの段階で行うことになる．

第 2 段階では，前段階で学習した内容の定着度を検定するテストを実施し，参加者がタスク実施に十分なスキルを有しているか判定を実施する．たとえば，絵画等の文化資料の分類を正しく実施

[1] URL: https://www.google.com/recaptcha/intro/index.html (accessed 2016-09-05)
[2] URL: https://ja.duolingo.com (accessed 2016-09-05)

できるか，テスト形式で確認することが考えらえる．タスク実施に十分なスキルを習得したと判定されるまでは，参加者はコンテンツの学習を繰り返しおこなう．

第 3 段階では，前段階の学習内容の検定の延長として，資料の翻刻や分類，タグ付けなど，クラウドソーシングの対象となる作業の遂行にあたる．タスクの実施を学習内容の検定として実施することで，作業者は自らの学習過程の一部としてタスクを実施することができる．ただし，あらかじめ正解が決められた検定とは異なり，資料翻刻や画像分類等のタスクの正解を事前に設定することはできない．そこで，タスク実施の内容の正しさについて，参加者間で評価と修正を実施する．これによって参加者は，作業内容についてフィードバックを受けることが可能になる．

以上が本研究で提案する人文学クラウドソーシング手法の概要である．このように，クラウドソーシングに学習を取り入れることで得られる利点は 3 つある．

第一は，参加者のスキル向上と検定が可能になることである．学習プログラムをシステムに組み込むことで，難易度の高い資料の解読等を，適切なスキルを有した参加者にアサインすることが可能になる．第二は，参加者の学習意欲を動機付けとして利用できることである．特に人文学の抱える歴史学や文学に関連するコンテンツは高い社会的関心を寄せられるものが多く，多数の参加者を集める上で有利にはたらく．第三は，学術・教育機関からの参加が見込める点である．一般に，大学等での人文学分野の基礎訓練は，実際の資料を演習等で扱う形式で行われる（古文書の翻刻など）．その教材としてクラウドソーシングシステムを提供することで，タスク実施を教育課程の一環として実施することが可能になる．

4．歴史災害史料のクラウドソーシング翻刻プラットフォーム

前節で述べた手法に基づき，前近代の日本語歴史災害史料を対象とするクラウドソーシング翻刻プラットフォームを構築した．以下では，この「災害史料翻刻システム」(http://honkoku.org) について，その目的と実装機能の概要を述べる．

4.1．目的と対象史料

災害史料翻刻システムの主な目的は，膨大な点数が残されている前近代の災害資料（特に地震関連史料）を翻刻し，Web 上で検索可能な形式で公開することで，災害研究や防災研究に貢献することである．

現在災害史料翻刻システムでは，東京大学地震研究所（以下，地震研）図書館が所蔵する，歴史災害史料のコレクション「石本文庫」の収録史料 140 点の画像を公開している．これらの史料を，歴史地震研究者と一般のボランティアの協働によって，全文翻刻することが本システムの当面の目標である．

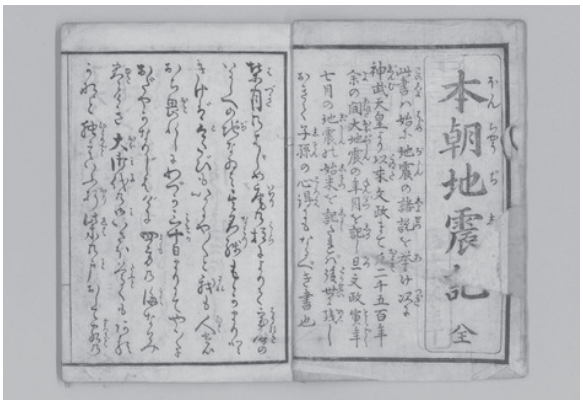

図 2　石本文庫収録史料の例『本朝地震記　全』

石本文庫は，地震研 2 代目所長であった石本巳四雄（1893-1940）の収集した，地震史料を中心とする歴史災害史料のコレクションである．コレクション全体の史料点数は 500 点を超えており，災害史料翻刻システムで公開している史料は，このうち地震研によりデジタル化された 140 点に限られる．石本文庫は，災害関係のかわら版や鯰絵など，多数の災害関連史料を含む．その大多数は江戸時代以前に刊行されたものであり，解読にはくずし字の読解能力が必要である（図 2）．

石本文庫収録史料のうち，一部はすでに翻刻が刊行されている．しかし，これらの史料は現在インターネット上で公開されていない．また，刊行されている史料でも，多数の省略がなされるなど文献学的な不備のあるものが多い．そのため，石本文庫史料の全文を翻刻し，①テキスト検索可能な形式で，また②史料画像と突き合わせて検証加納な形式でインターネット公開することで，歴史災害研究や防災研究に寄与することが期待される．

なお，史料画像データは地震研図書館から提供を受け，IIIF（International Image Interoperability Framework）対応の画像サーバーで配信している．したがって，IIIF に対応した画像ビューワを利用することで，地震研の Web サイトや災害史料翻刻システムを経由せずとも閲覧することが可能である．2016 年 11 月現在，史料画像データは京都大学防災研究所の Web サーバー上でホストされている．

4.2．文字学習セクション

前近代の日本語史料の翻刻には，くずし字の解読能力が必要である．そこで，災害史料翻刻システムでは前節で述べた手法に基づき，くずし字の学習支援システムを組み込むことで，翻刻作業者の動機付けとタスク実施に必要なスキルの向上をはかっている．このため，災害史料翻刻システ

ムは（1）くずし字解読能力の習得と検定を実施する文字学習セクションと，（2）実際に史料の翻刻を実施する翻刻セクションの 2 パートから構成されている．まず，文字学習セクションの概要を述べる．

文字学習セクションでは，くずし字を構成する代表的な変体仮名 102 種，草書体漢字 176 種類について，江戸時代の木版本から採取した約 3,000 枚の文字画像データを収録する．これらの学習用データは，大阪大学文学研究科を中心に開発された『くずし字学習支援アプリ KuLA』[1]の収録データを再利用したものである．

また，くずし字の判読能力の検定のため，上記の画像データの読み方を回答するテストを実施することができる．

災害史料翻刻システムでは，OAuth2 プロトコルを利用して，ユーザーの認証を実施する．テストの実施成績はユーザー情報に紐付けて記録され，この数値を元にくずし字の判読能力が評価される．

4.3. 史料翻刻セクション

史料翻刻セクションは，作業者が史料画像をもとに翻刻を実施するための諸機能を提供する，クラウドソーシング翻刻の中心となるモジュールである．

史料リスト画面（図 4 左）では，災害史料翻刻システムで公開されている石本文庫史料の一覧が表示される．各史料には史料タイトルとサムネイルの他，①翻刻の難易度，②翻刻作業の進捗状況，③キーワード（地震研によって付与されたもの）などの付加情報が表示される．また史料一覧画面では，現在作業者が編集中の史料一覧や，ブックマーク登録した史料の一覧も表示される．この画面で史料を選択すると，個別の史料の翻刻画面に遷移する（図 4 右）．

翻刻画面は，史料画像を表示する画像ビューワと，翻刻テキストを編集する多機能エディタから構成される．画像ビューワは JavaScript 製のオープンソースライブラリの OpenSeadragon を利用しており，画像の拡大縮小や移動をサポートする．

翻刻テキストのエディタは，①閲覧，②入力，③ノート，④編集履歴，⑤ヘルプの 5 つのタブから構成される．翻刻作業者は，画像ビューワの表示をもとに入力タブでテキストを入力する．

前近代の日本語史料をテキスト化するにあたって重要な問題のひとつは，ルビや割書きなど日本語特有の記法の処理の問題である．とりわけ，情報処理の知識を持たない参加者を前提とするクラウドソーシング翻刻では，XML を使用したマークアップを依頼することは難しい．可能な限り参加者の負荷を軽減するような，簡便な入力方法を用意する必要がある．災害史料翻刻システムでは，ルビや割書きなどの特殊記法を，プレーンテキストとして直観的に入力することを可能にするために，Markdown 等で採用されている簡易入力記法のアプローチを採用した．たとえば，ルビ文字を入力するためには，連続する漢字の直後に全角括弧を入力し，その中にルビ文字を入力すればよい．さらにこの作業を簡易にするため，これら入力専用の特殊記法を一括で入力するためのショートカットを用意した．閲覧タブで翻刻テキストを表示する際には，入力テキストを正規表現でパースし，適切にスタイリングした形式で表示する（図 5）．

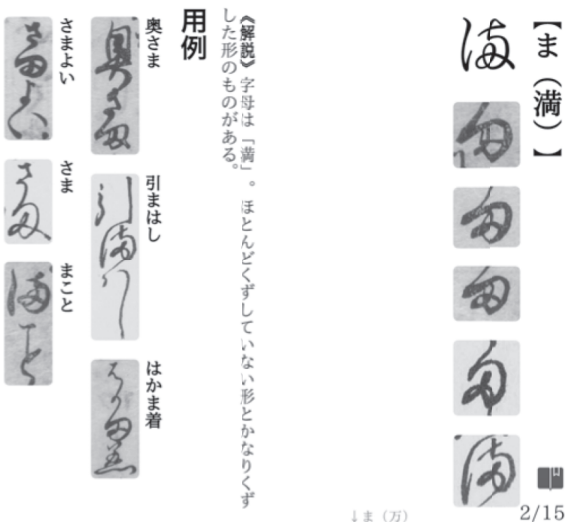

図 3 文字学習セクション

また，大学の演習・ゼミ等において多人数で同時に翻刻を実施する場合を想定して，本システムは翻刻テキストのリアルタイム編集機能をサポートしている．翻刻対象の史料画像の翻刻を開始する際に，「ライブ翻刻」のオプションを有効にすると，編集内容が他の参加者の画面にもリアルタイムに反映される．この機能を利用することで，くずし字解読の習熟者が未習熟者に演習形式で解読法を指導することが可能になる．さらにエディタに附属するチャット機能を利用することで，遠隔地においても多人数の共同翻刻を可能にしている．

[1] iOS 版:
https://itunes.apple.com/jp/app/id1076911000
Android 版:
https://play.google.com/store/apps/details?id=yuta.hashimoto.kula&hl=ja

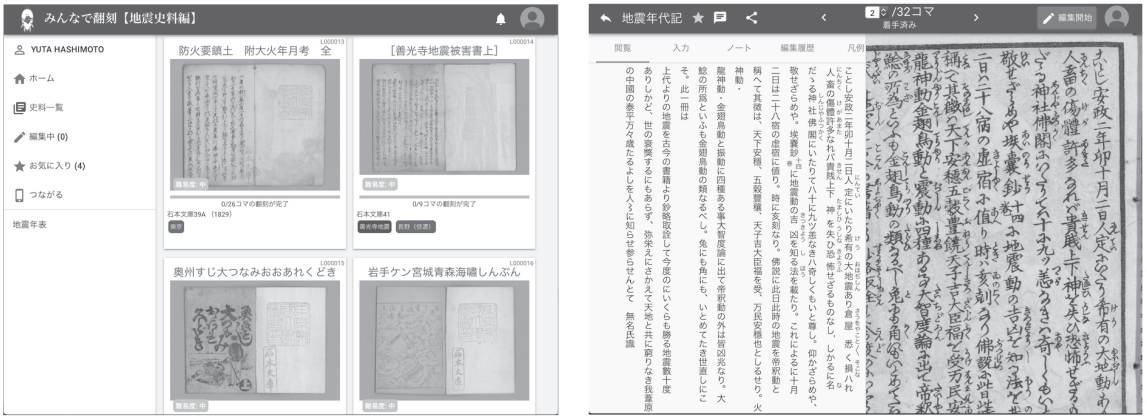

図 4（左）史料リスト画面（右）史料翻刻画面

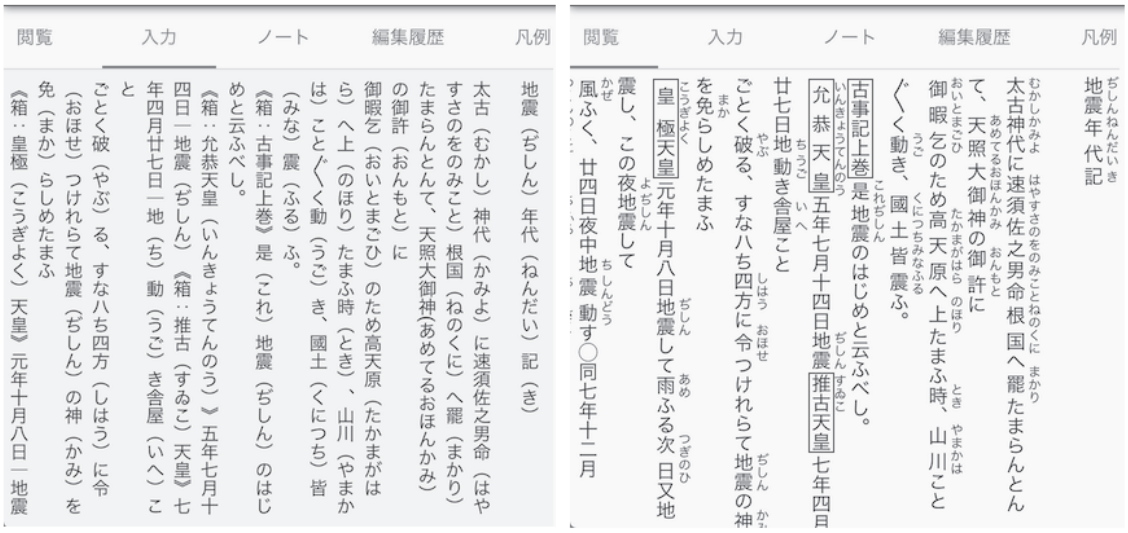

図 5　（左）翻刻テキストの入力（右）HTML 上での出力結果

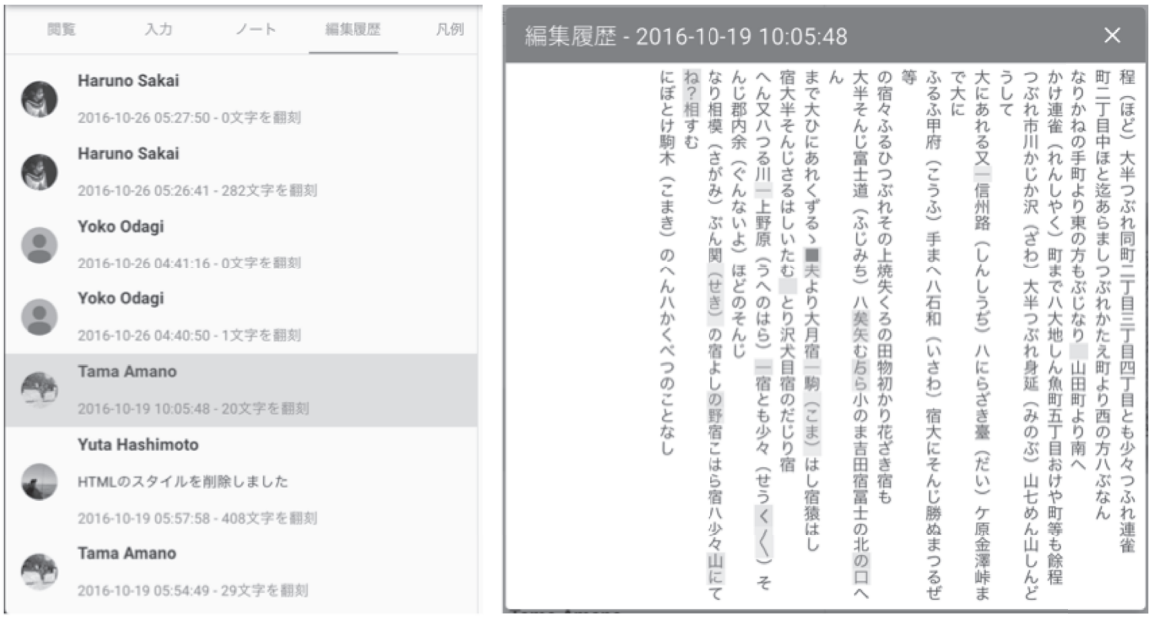

図 6　翻刻テキストの編集履歴表示（赤：削除箇所，緑：追加箇所）

なお，翻刻テキストを多人数で共同編集する場合，別々の作業者の編集内容が衝突する可能性がある．これを防ぐため，本システムでは作業者の翻刻作業中は，同一画像の翻刻テキストを禁止する排他制御を導入している．しかしながら，悪意を持った編集者により，翻刻テキストが消去あるいは改悪されることも想定される．このため，翻刻テキストの全編集履歴は保存され，管理権限を持つ参加者によって過去の履歴を復元する機能を実装した．また，過去の編集履歴の差分を表示することも可能である（図 6）．

現段階では実装が完了していないが，翻刻されたテキストは HTML/PDF/TEI の各フォーマットに変換され，認証無しでアクセス可能な Web ページ上で，史料画像と併せてパブリックドメイン・コンテンツとして公開される予定である．

5．おわりに

本稿では，人文学資料を対象にしたクラウドソーシング・プロジェクトの設計方法として，人文学コンテンツを対象にした学習サービスをクラウドソーシングに組み込む手法を提案した．

また，この手法に基づき，前近代日本の歴史災害資料をクラウドソーシングによって翻刻する Web システムの諸機能について紹介した．

歴史災害資料のクラウドソーシング翻刻システムは，2016 年 11 月現在では一般に公開されていない（同年 12 月 1 日を目途に公開を予定している）．そのため本稿で提案した手法とシステムの有用性は，いまだ検証・評価できていない．今後，システム公開後の運用状況をもとに，別途成果を報告する予定である．

一方で，国文学研究資料館の大規模デジタル化プロジェクトを始めとして，前近代の日本語資料のデジタル化は昨今急速に進行しつつあり，これらの資料をテキスト化し活用に繋げる必要は日々高まっている．

画像認識技術を利用したくずし字認識が実用レベルの精度に達するまでには，いまだ相当の時間を要すると考えられる．また，同様の技術が実用化されたとしても，人間によるチェック作業は不可欠である．したがって，前近代の史料を解読可能な人材を養成し，これらの人々の能力を有効に活用するシステムの構築は，人文科学とコンピューター分野においても重要な課題を形成すると考えられる．

参考文献

1) Causer, T. and Wallace, V: Building A Volunteer Community: Results and Findings from Transcribe Bentham, Digital Humanities Quarterly, Vol.6, No.2 (2012)〈http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html〉.

2) 増井ゆう子，山本和明：国文学研究資料館・日本語の歴史的典籍のデータベース構築について，情報の科学と技術，Vol.65，No.2，pp.169-175（2015）.

3) 永崎研宣：翻デジ 2014，入手先〈http://lab.ndl.go.jp/dhii/omk2/〉.

4) Howe, J.: The rise of crowdsourcing, Wired magazine 14.6 (2006): 1-4.

5) Dunn, S., and Hedges, M.: Crowd-sourcing Scoping Study. Engaging the Crowd with Humanities Research. Centre for e-Research, King's College London, (2012), 入手先〈http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/12/Crowdsourcingconnected-communities〉（参照 2016-11-02）.

6) Oomen, J., and Aroyo, L: Crowdsourcing in the cultural heritage domain: opportunities and challenges, Proceedings of the 5th International Conference on Communities and Technologies, pp.138-149, (2011).

7) 中野三敏：和本のすすめ，岩波書店（2012）.

8) Online Course Report: State of the MOOC 2016: A Year of Massive Landscape Change For Massive Open Online Courses, 入手先〈https://www.onlinecoursereport.com/state-of-the-mooc-2016-a-year-of-massive-landscape-change-for-massive-open-online-courses/〉（参照 2016-11-02）.

9) von Ahn, L.: Duolingo: learn a language for free while helping to translate the web, *Proceedings of the 2013 international conference on Intelligent user interfaces*, ACM, pp.1-2, (2013).