クラウドソーシングとAI文字認識を駆使した「くずし字」資料の大規模テキスト化

橋本 雄太（国立歴史民俗博物館）

背景

◆ 莫大な点数が伝来する江戸時代以前の日本語文献資料

- 出版資料……カタログ化された資料で150万点超.
- 手書き資料…全国に20億点以上が保存されるという推計も.
- 人文学のみならず地震学や天文学においても重要な研究資料.

◆ デジタルテキスト化の必要性と課題

- デジタルアーカイブ化の進行…国文研30万点，NDL8万点
- 大多数の資料は全文検索できない．専門家が「くずし字」を目視で解読，解釈する必要がある.
- くずし字…江戸時代以前に使用されていた草書体漢字や平仮名の異体字（変体仮名）の総称．可読者人口は1万人を切るとも.

『元禄地震記』

「候」のくずし字用例
（出典：東京大学史料編纂所くずし字DB）

これまでの成果

◆ 翻刻作業の進捗

バージョン	V1	V2
運用期間	2017年～2019年3月	2019年7月～
翻刻が完了した史料	500/508点	1,800/3,469点
翻刻が完了した画像	8,411/8,925枚	66,512/138,995枚
参加登録者数	5,346人	3,426人
総入力文字数	625万字	3,343万字

◆ 翻刻文の品質

表 3.2　検証作業を通じて見つかった要修正箇所の件数（資料種別・タイプ別）

資料種別	合計文字数	誤読箇所		未読箇所		表記ゆれ箇所		合計	
		実数	比率 (%)	実数	比率 (%)	実数	比率 (%)	実数	比率 (%)
木版本	78,774	400	0.5	285	0.4	395	0.5	1,080	1.3
筆写本	21,901	210	1.0	105	0.5	132	0.6	447	2.0
合計	100,675	610	0.6	390	0.4	527	0.5	1,527	1.5

文字ベースで**98.5%の正確性**．内容把握や全文検索には十分な水準

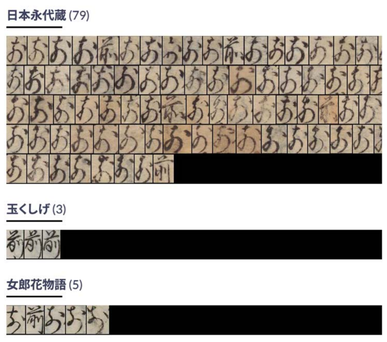
◆ データセット化

- 翻刻された4,000万字の史料を CC BY-SA でオープンデータ公開.
- https://github.com/yuta1984/honkoku-data
- awesome-japanese-nlp-resources に収録.

「くずし字」自動解読に向けた取り組み

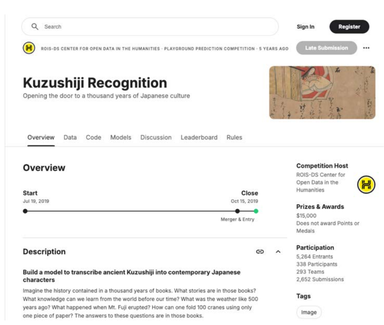
◆ 日本古典籍くずし字データセット (2016)

- 江戸時代の木版本から手動で採取した約4,000字種にわたる100万字のラベル付文字画像.
- このデータセットを活用した自動認識の研究が急速に進行する.



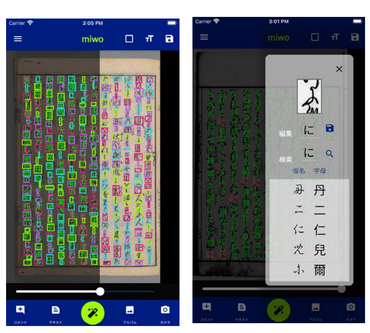
◆ Kaggleくずし字認識コンペ（2019）

- くずし字データセットを用いた文字認識モデルの精度を競うコンペティション.
- Center for Open Data Humanities主催.
- 世界中から294チームが参加.
- 人文学に関する初の Kaggle コンペ.



◆ AIくずし字認識アプリ「みを」(2021)

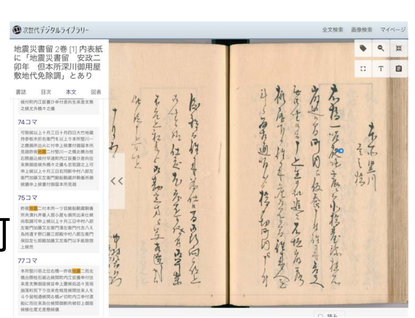
- Kaggleで上位成績を収めたモデルの構築手法を取り入れたくずし字認識アプリ.
- 開発：タリン・カラーヌワット氏（元Google/現Sakana AI）.
- 公開以来247万画像を認識.



翻刻データセットの活用

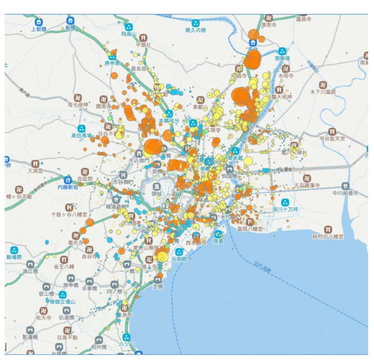
◆ 国立国会図書館古典籍OCR実験 （2022）

- 手書き文字認識モデルTrOCRをベースにくずし字認識のためのモデルを構築.
- みんなで翻刻データを加工して訓練に使用.
- 同館の古典籍8万点430万画像が全文検索可能に（F値の中央値0.92）.



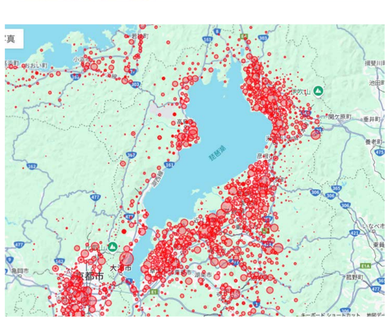
◆ 安政江戸地震史料の構造化（2023-）

- 史料中の地名や被害記述をアノテーションするWebシステム「みんなで注釈」を開発.
- 安政江戸地震（1855）関連の史料十数点に適用し，被害状況を可視化・地理データ化.
- NAIST大内研にて Location Referring Expressions データセットとしても活用.



◆ 「天保郷帳」の地理データ化（2023-）

- 天保郷帳…幕府の命で天保年間に作成された全国の村々の郷帳．各村の生産高を含む.
- 村名を抽出し「日本歴史地名大系データセット」の緯度経度情報を付与.
- 生産高の地理的分布の可視化が可能に.



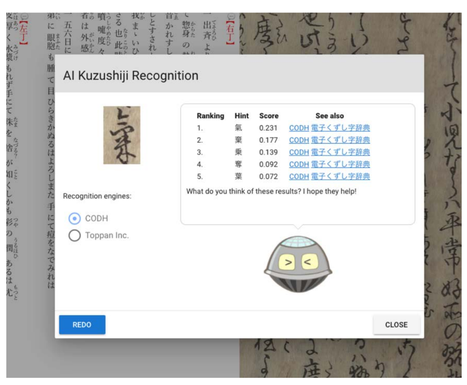
クラウドソーシング＋AIを用いた大規模テキスト化

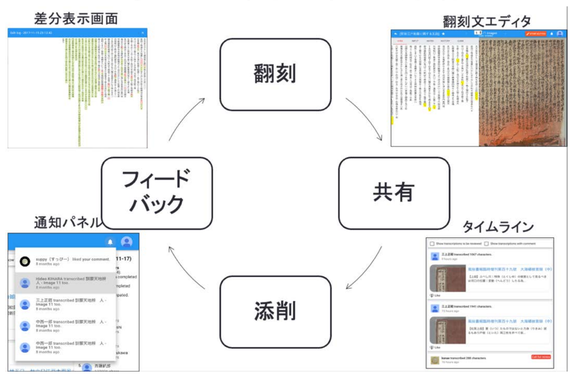
◆ みんなで翻刻（2017-）

- URL: https://honkoku.org/
- 歴史資料の市民参加型翻刻プラットフォーム.
- 翻刻…歴史学用語で史料を活字に起こすこと.
- IIIF (International Interoperability Framework) API を介して任意のデジタルアーカイブから翻刻対象をインポート可能.

◆ クラウドソーシングのデザイン

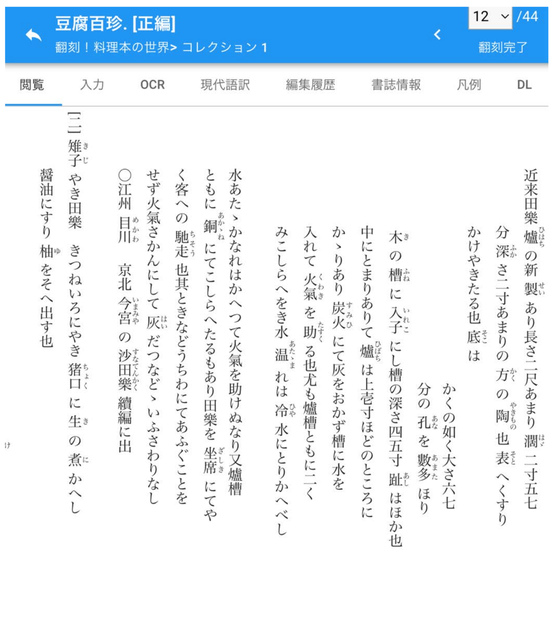
- AI文字認識による初学者サポート＋参加者間の相互添削システム



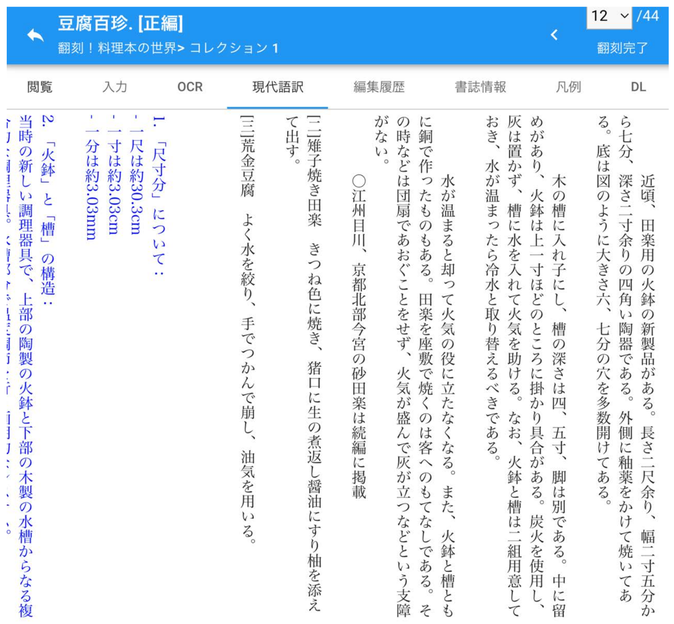


生成AIの適用

- Claude 3.5 の API を利用し，近世の料理本翻刻プロジェクトの史料に対して現代語訳と注釈を実験的に生成



▲翻刻文（『豆腐百珍』）



▲Claude が生成した現代語訳と注釈