第 8 回日本語の歴史的典籍国際研究集会

「みをつくし」プロジェクト：AIくずし字認識研究の展開

カラーヌワット・タリン
Senior Research Scientist
Google Research, Brain team

発表内容は個人の見解であり、所属組織を代表するものではありません。

三國志傳
周憲王
救荒本草
趙氏医貫
金匱要略輯義
文化八年刊 全十冊
新訂
草木圖説
飯沼慾齋 編
明治七年刊
全廿冊
草木圖説
前篇 廿冊
安政三年刊
玉石志林
外科新説
明治七年刊
新訂増補
和蘭薬鏡
内科撰要

「トニカクカワイイ」小学館集英社プロダクション(2020)

『宇津保物語』（国文学研究資料）

くずし字

日本語の歴史的典籍の国際共同研究ネットワーク構築計画
（略称：歴史的典籍NW事業）

歴史的典籍22万6千点、コマ数では
2400万コマを撮影した。今後30万点の
画像を公開する予定。

大学共同利用機関法人　人間文化研究機構
国文学研究資料館
National Institute of Japanese Literature

くずし字データセット

江戸時代の古典籍 44 冊

6151 ページ

1,086,326 文字

4,328 文字種

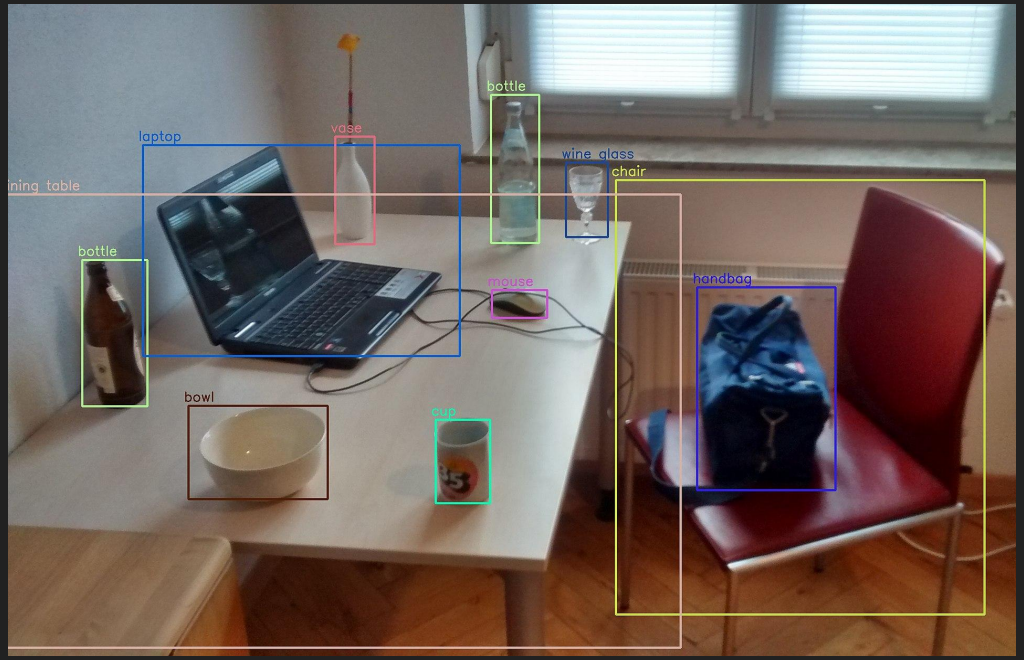
Unicode	Image	X	Y	Block ID	Char ID	Width	Height
U+81EA	100241706_00004_2	1852	1736	B0001	C0001	104	219
U+5E8F	100241706_00004_2	1816	2096	B0001	C0002	152	296
U+82E5	100241706_00004_2	1465	951	B0001	C0003	172	216
U+3044	100241706_00004_2	1495	1218	B0001	C0004	143	69
U+6642	100241706_00004_2	1465	1338	B0001	C0005	168	179
U+306E	100241706_00004_2	1497	1567	B0001	C0006	123	152
U+6C17	100241706_00004_2	1504	1754	B0001	C0007	145	200
U+5F37	100241706_00004_2	1479	2034	B0001	C0008	163	145
U+306B	100241706_00004_2	1521	2239	B0001	C0009	83	237

Unicode, x, y, w, h

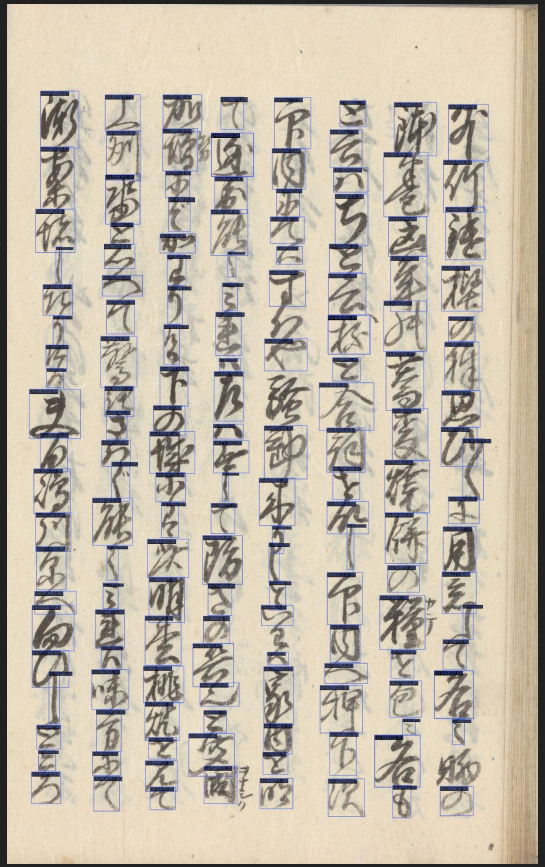
を U+3029, 512, 418, 56, 47

国文学研究資料館&国立国語研究所作成

くずし字認識の手法：物体検出の手法

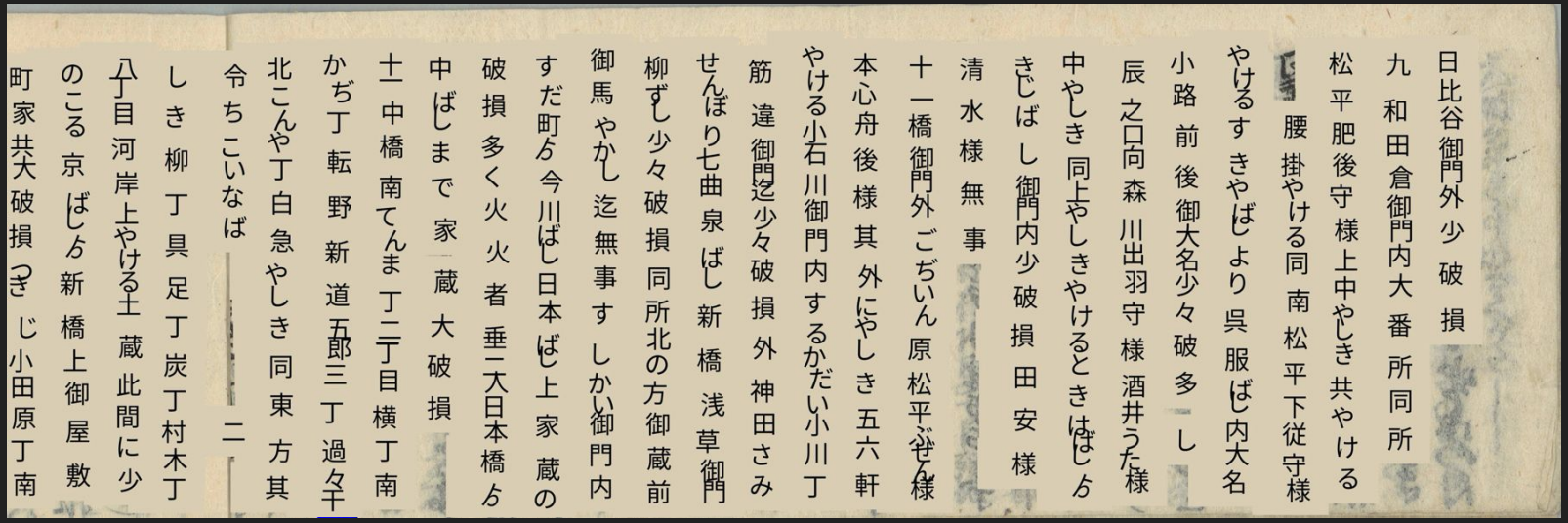


Wikipedia



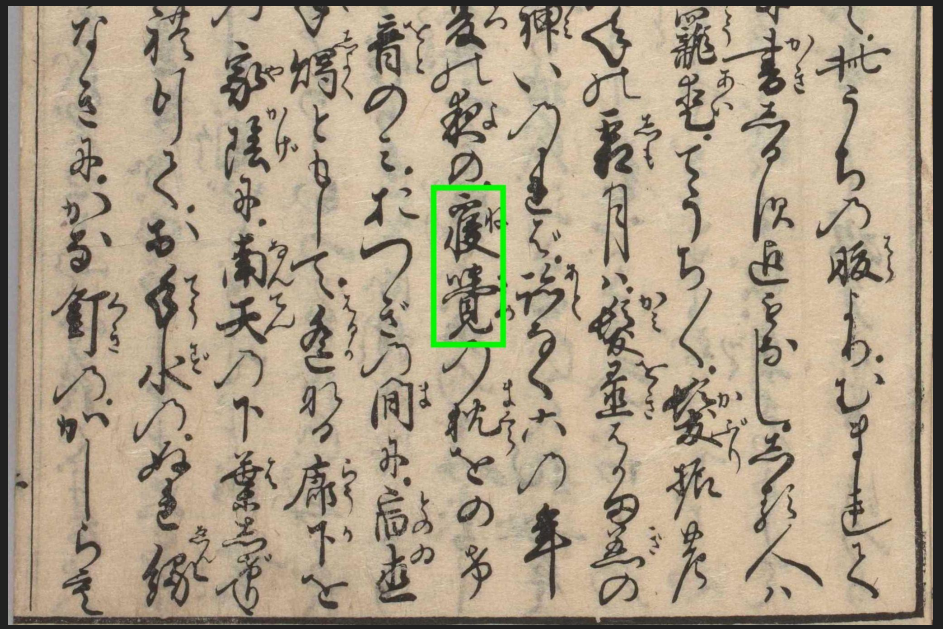
DetectoRS model with ResNet 50 backbone

くずし字認識の手法：物体検出の手法



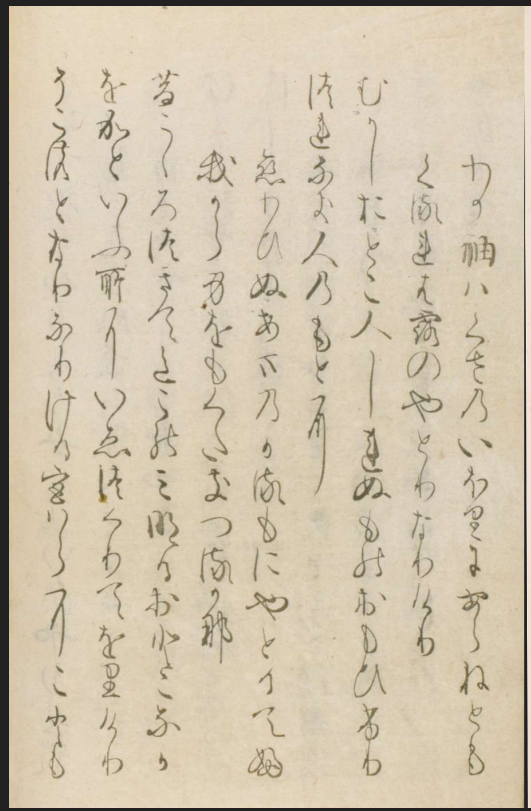
画像：みんなで翻刻プロジェクト

くずし字認識の手法：物体検出の手法

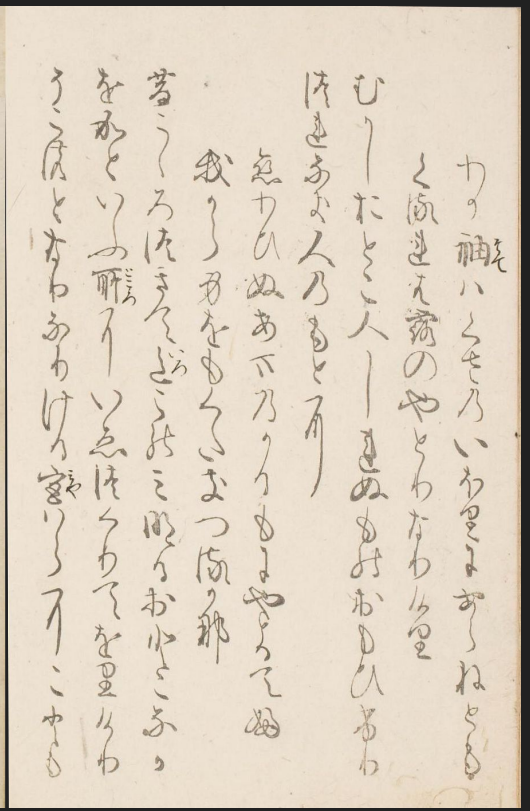


くずし字認識研究の展開：古活字の研究

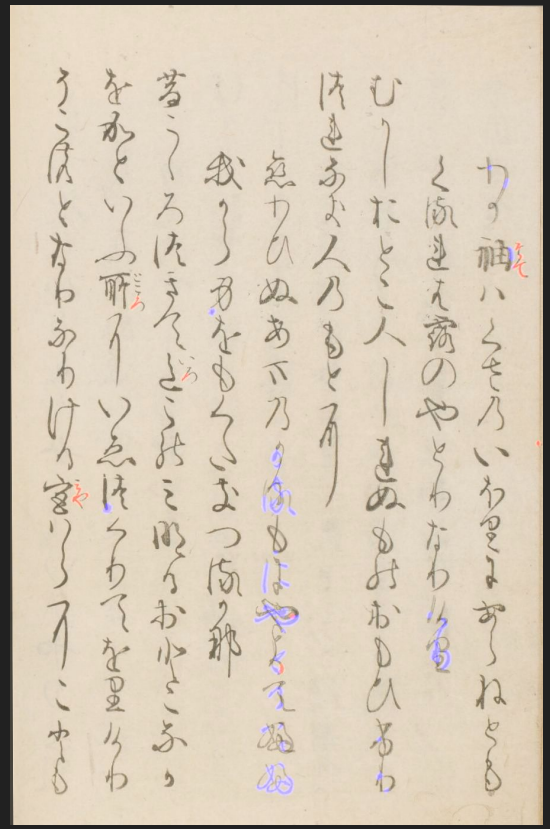
A



B

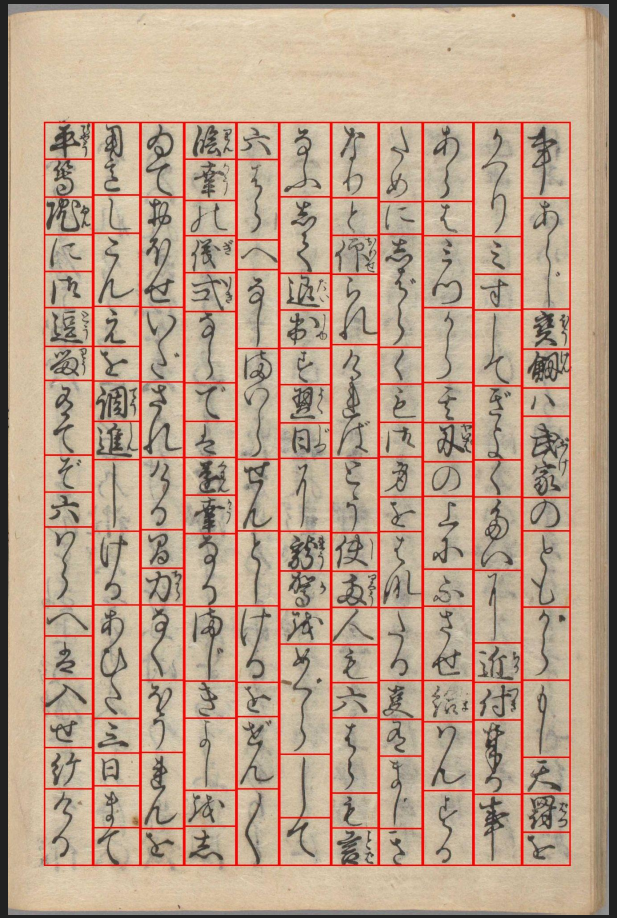


A + B



嵯峨本『伊勢物語』、CODH公開したvdiff.jsを利用

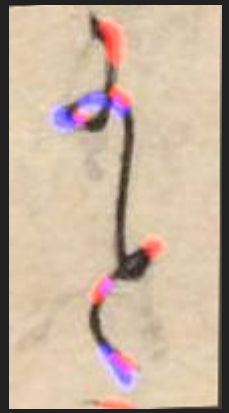
くずし字認識研究の展開：古活字の研究



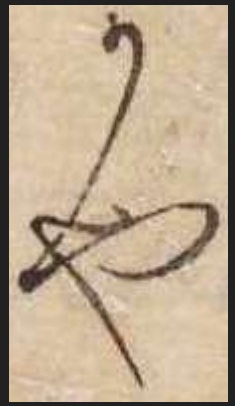
くずし字認識研究の展開:古活字の研究

A

B

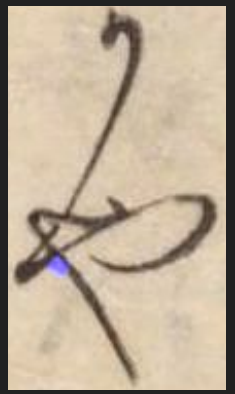


A + B



A

B



A + B

A： 内閣文庫蔵:嵯峨本第一種本『徒然草』
B： 内閣文庫蔵:嵯峨本第四種本『徒然草』

「みをつくし」プロジェクト

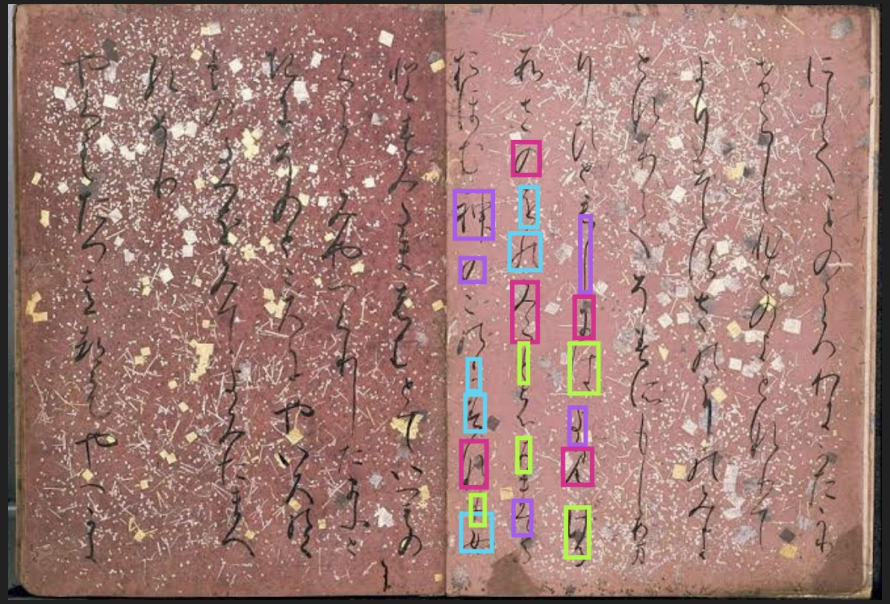
みをつくし（澪標）

- 『源氏物語』第 14 巻「みをつくし」（澪標）。
- 「みを（船の水路）を示すために立ててある杭」の意。
- 「身を尽くし」の掛詞。
- 「みをつくし」が人々の水先案内となるように、「みをつくし」プロジェクトがくずし字資料を読むための道案内となることを目指している。

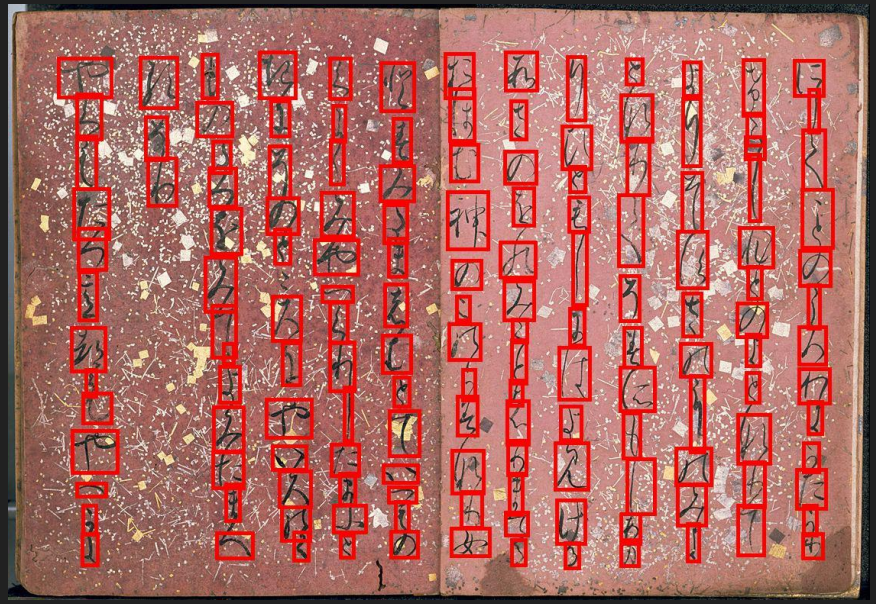
出典：Wikipedia

新AIくずし字認識モデル：RURI

東博蔵：元永本古今集



旧モデル

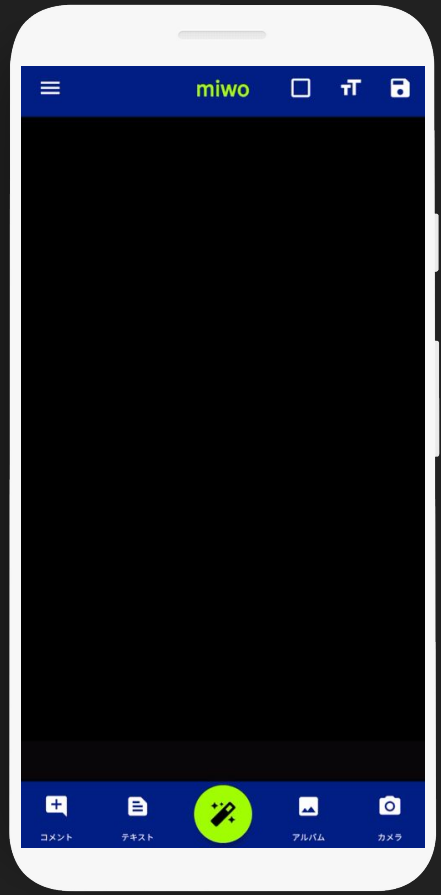


RURIモデル

「みを」

「みを」くずし字認識アプリ

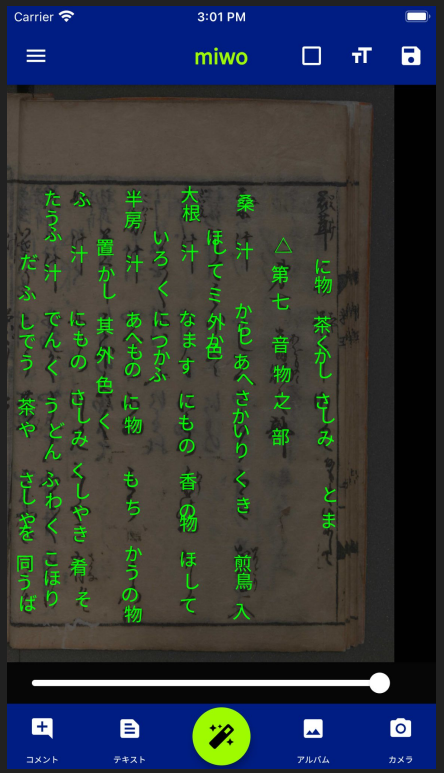
- 画像からくずし字を認識し、
  現代日本語文字に変換する。
- 2021 年 8 月 30 日にリリース。
  現在、10万回以上ダウンロード。
- iOS、Android両方リリースした。Webアプリとデスクトップアプリもリリース可能。
- 現在まで認識した画像の枚数は
  100万枚以上。



GOOD DESIGN AWARD 2022

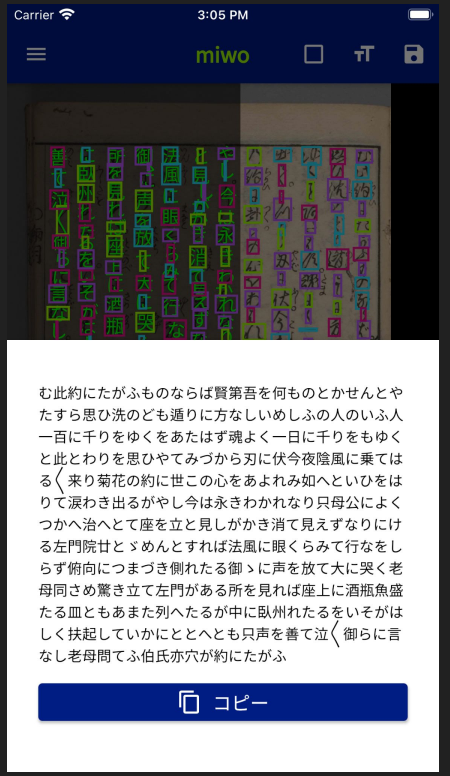
2022年10月7日にグッドデザイン賞（システム・サービス部門）を受賞した。

「みを」アプリの主な機能

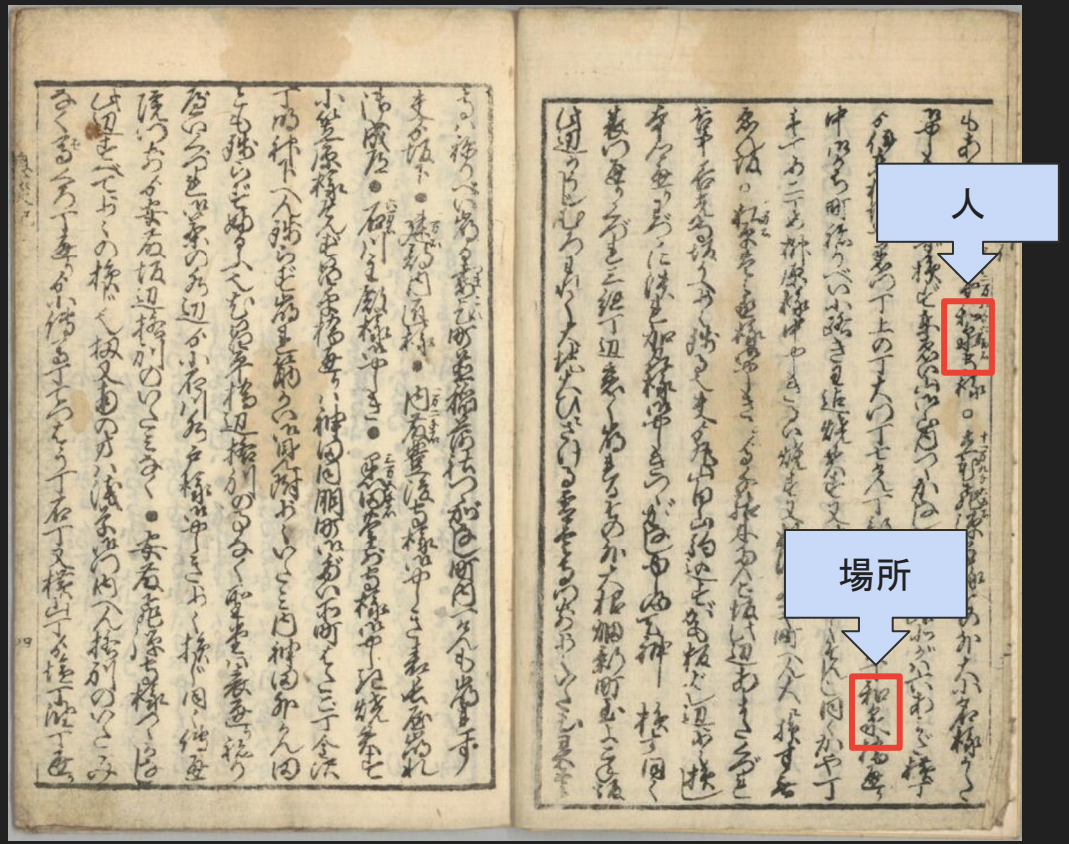

Carrier
3:01 PM
miwo
コメント
テキスト
アルバム
カメラ


Carrier
3:01 PM
miwo
コメント
テキスト
アルバム
カメラ


Carrier
3:01 PM
miwo
編集
に
検索
に
仮名
字母
丹
二
仁
兒
爾
コメント
テキスト
アルバム
カメラ


Carrier
3:05 PM
miwo
む此約にたがふものならば賢第吾を何ものとかせんとやたすら思ひ洗のども通りに方なしいめしふの人のいふ人一百に千りをゆくをあたはず魂よく一日に千りをもゆくと此とわりを思ひやてみづから刃に伏今夜陰風に乗てはる〳〵来り菊花の約に世この心をあよれみ如へといひをはりて涙わき出るがやし今は永きわかれなり只母公によくつかへ治へとて座を立と見しがかき消て見えずなりにける左門院廿とゞめんとすれば法風に眼くらみて行なをしらず俯向につまづき倒れたる御ゝに声を放て大に哭く老母同さめ驚き立て左門がある所を見れば座上に酒瓶魚盛たる皿ともあまた列へたるが中に臥れたるをいそがはしく扶起していかにととへとも只声を善て泣〳〵御らに言なし老母問てふ伯氏亦穴が約にたがふ
コピー

「つくし」プロジェクト



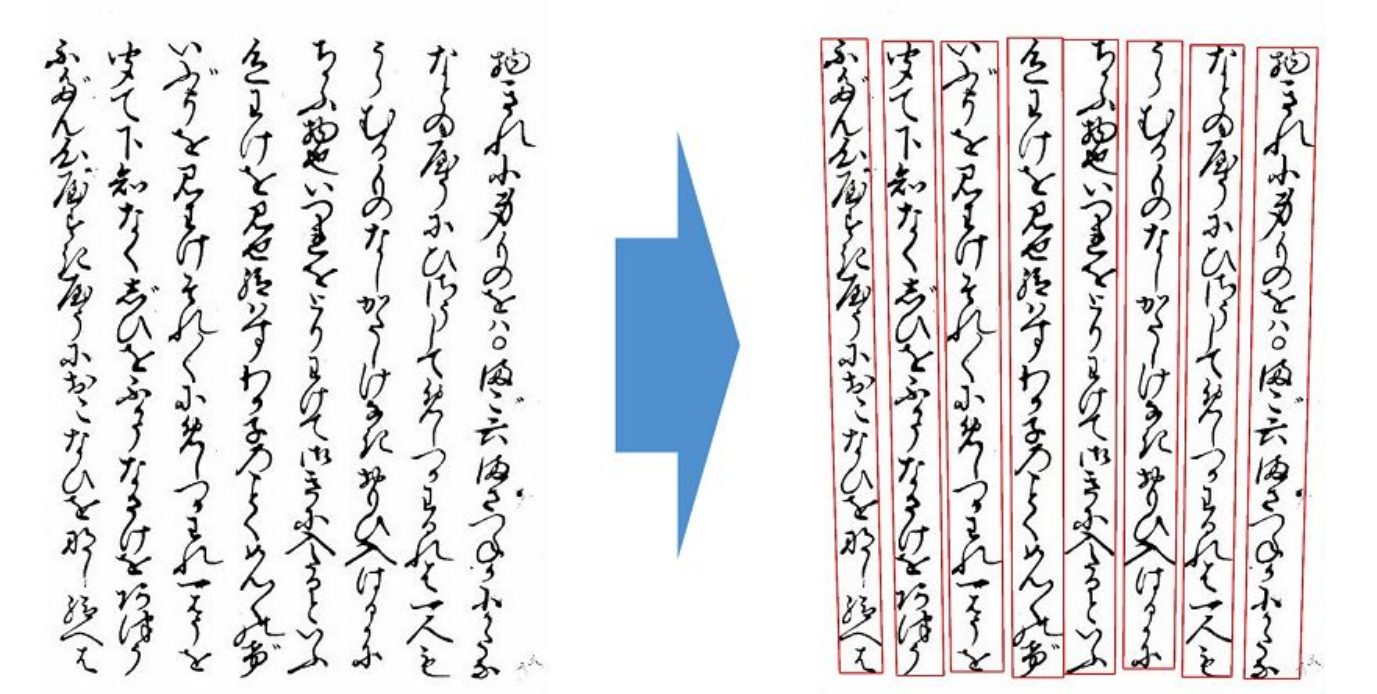
和泉 🔍

- くずし字 認 識を使って新しい研究方法を開拓する。
- 検索機能
  - 全文検索
  - ユニバーサル検索
    - 画像
    - 地図
    - ナレッジグラフ
- 非文字情報を分析する方法。
- 古典籍を探し尽くす。

東大地震研蔵:『安政息災記』

くずし字認識の手法:くずし字行領域認識

SIGNATE　凸版印刷株式会社 くずし字認識チャレンジ①　くずし字　行領域認識アルゴリズム作成



出典:人文学オープンデータ共同利用センター(http://codh.rois.ac.jp/)
『日本古典籍くずし字データセット』に含まれる『吉利支丹物語』(国文学研究資料館所蔵、doi: 10.20730/200006665、
CC BY-SA(https://creativecommons.org/licenses/by-sa/4.0/)にて配布、新日本古典籍総合データベースより)を加工した画像
加工内容:行領域を4点ポリゴンで囲って赤枠で表示