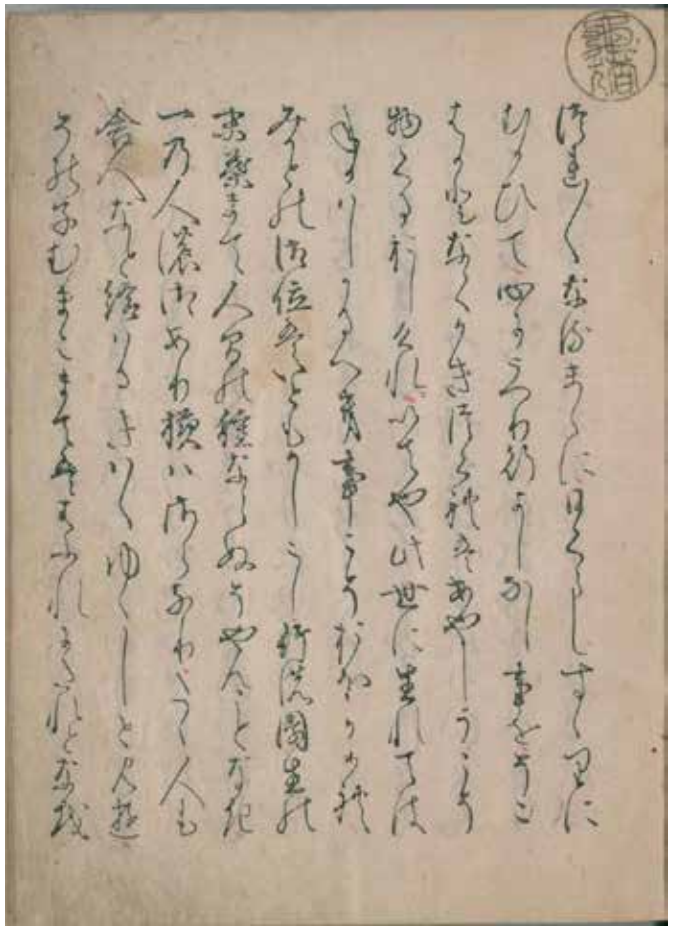
古活字版の活字ブロック自動分割・同定のためのAI分析基盤

北本 朝展（ROIS-DS人文学オープンデータ共同利用センター／国立情報学研究所）

カラーヌワット タリン（Sakana AI）

Yuxiao Li（EPFL／NIIインターンシップ）

古活字版とは？



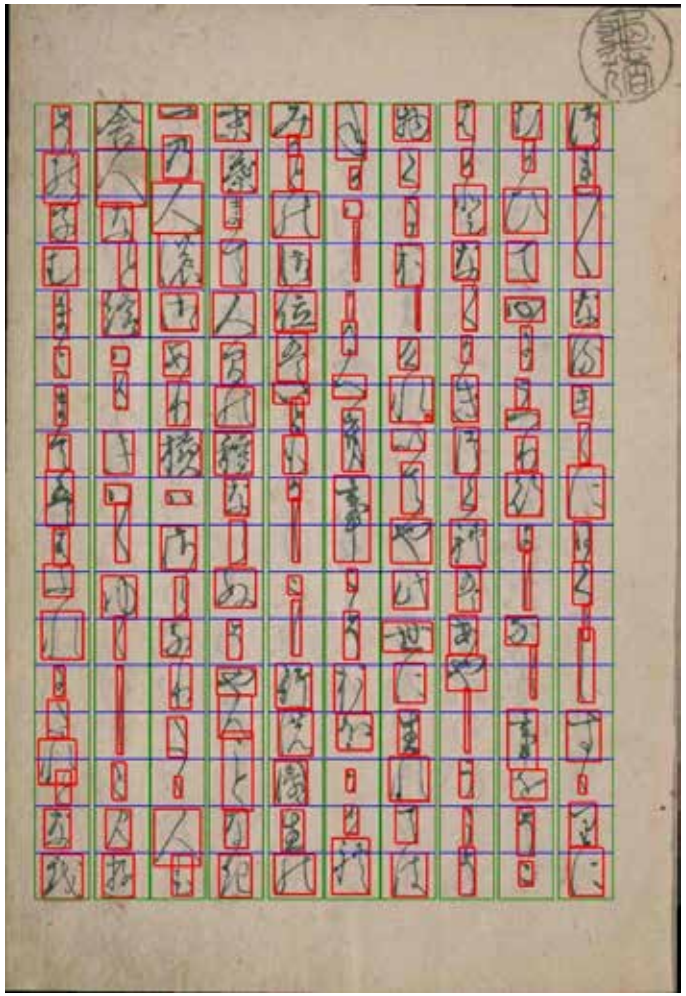
国立国会図書館デジタルコレクション

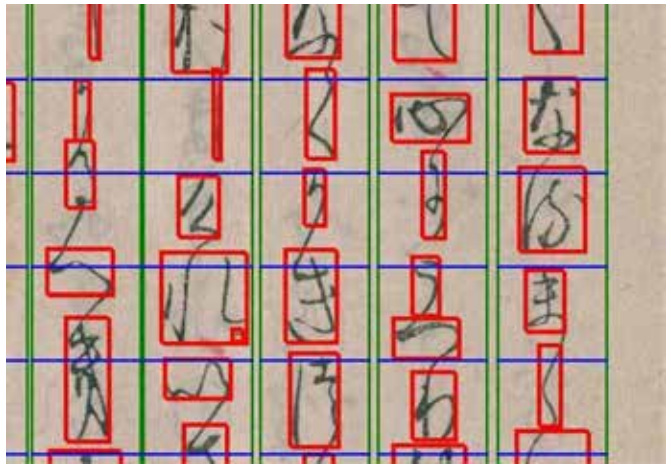
1. **古活字版**とは、16世紀末に西欧と朝鮮半島から伝えられた活字印刷の技術に基づき出版された本
2. 江戸時代初期の角倉 素庵（すみのくら そあん）は、京都嵯峨で出版業に関わった代表的存在
3. 古活字版「嵯峨本」は、日本の出版史上もっとも美しい書物の一つであり、日本の書物文化の粋

古活字版の情報解析の課題

1. 活字ブロック自動分割：文字の連なりから活字境界を推定する手法を開発
2. 活字ブロック自動同定：字形から同一活字を推定する手法を開発
3. AI分析基盤：本と活字の関係を可視化し、分析と改良を支援する基盤を開発
4. 古活字の運用パターンなど、印刷史の謎を定量的かつ大規模に分析することを目指す

1. 活字ブロック自動分割





1. **仮定**：活字ブロックの大きさは規格化され、高さは単位高さの整数倍となる
2. **仮定**：組版はベタ組みで、活字の間にはスペースが入らない
3. 今回の対象の古活字版では、おおむね仮定が成立する

アルゴリズムの流れ

1. AIページ認識
2. AIくずし字認識＋文字矩形推定（RURI）
3. 画像傾き補正
4. 行認識・行幅推定・格子グリッド推定
5. 文字矩形と格子グリッドの交差判定
6. 連彫活字を含めた活字ブロック確定

古活字データセット

http://codh.rois.ac.jp/omt/dataset/

- 全36,869ブロック
- AIの正解率は94.5%
- 6文字以上の連彫活字は要検討（分割ミスの可能性）

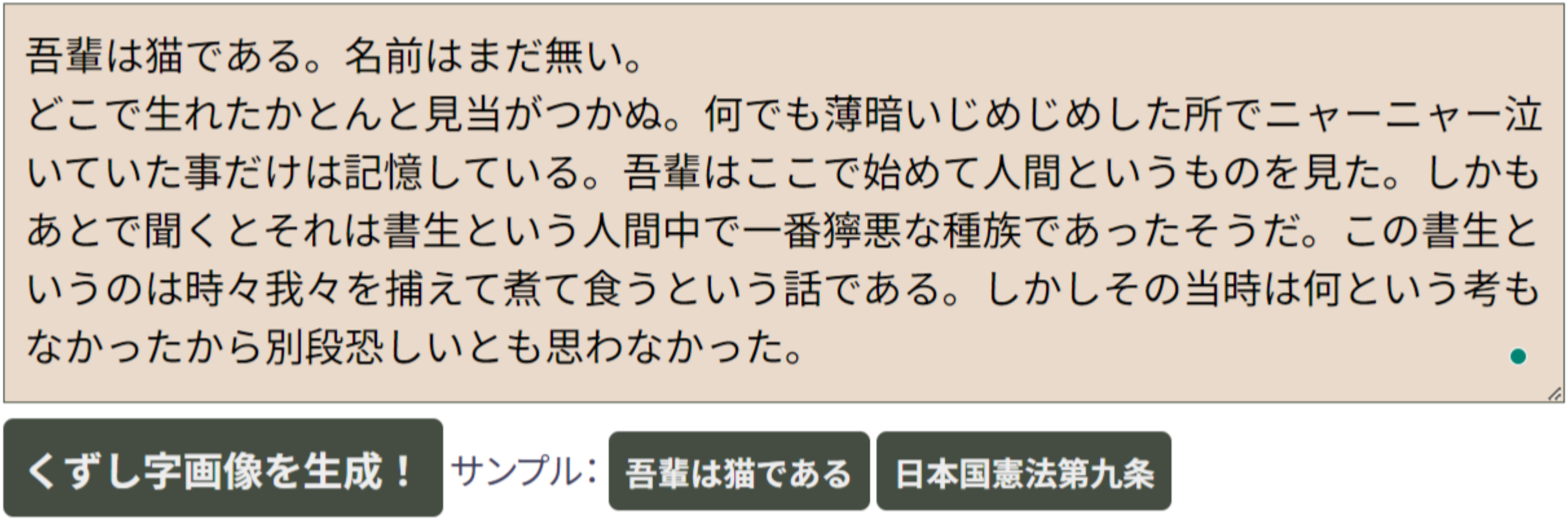
文字数	文字列	出現数
1	の	1536
2	なり	301
3	はかり	66
4	へからす	40
5	をのつから	9

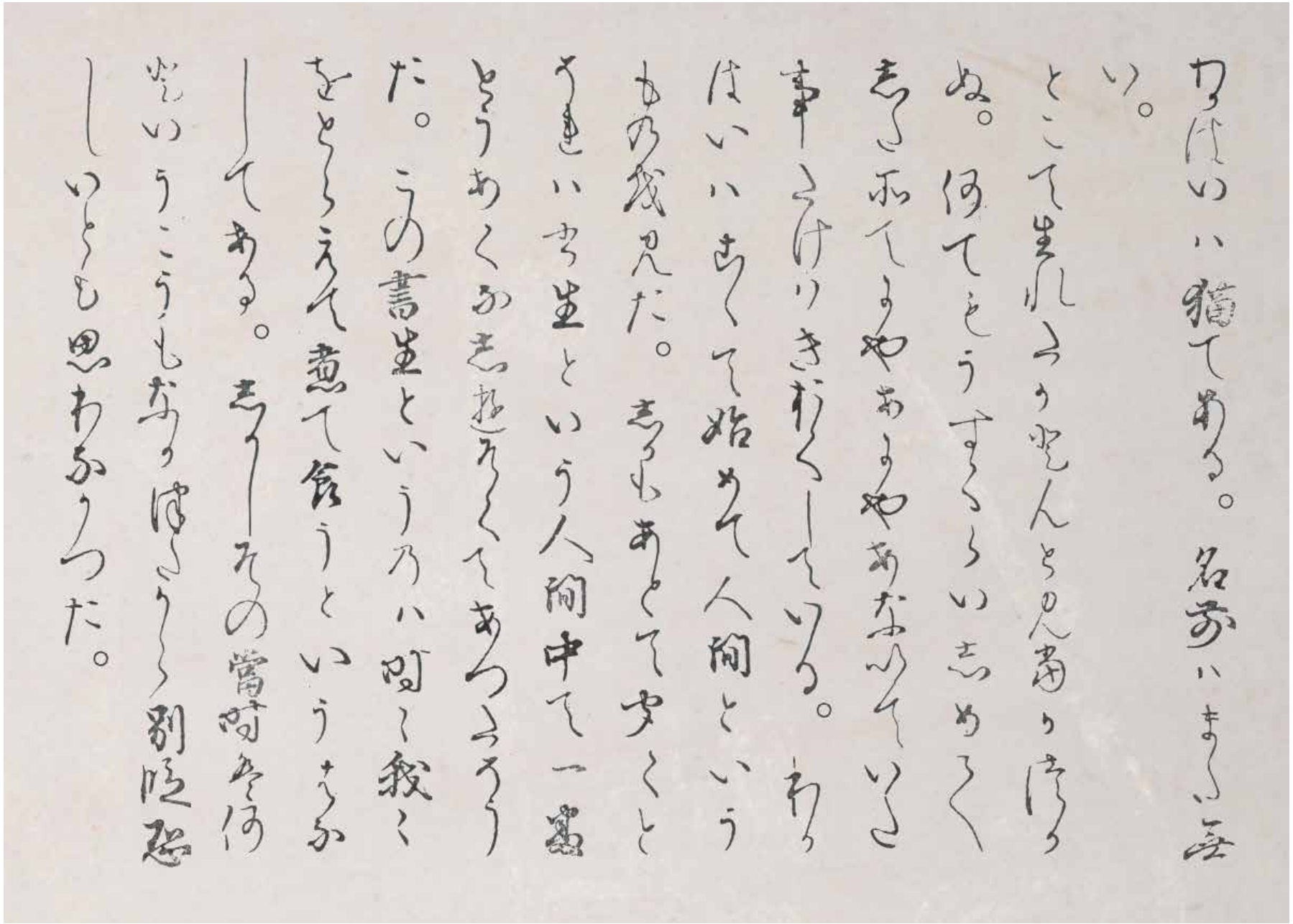
活字ブロックごとの文字（Unicode）・矩形座標（x1, y1, x2, y2）をCSV形式で出力

文字数	活字個数	活字サイズ
1	22451	1→21806, 2→642, 3→3
2	10425	1→85, 2→10015, 3→322, 4→3
3	3248	2→863, 3→2294, 4→90, 5→1
4	577	3→287, 4→278, 5→12
5	95	3→2, 4→51, 5→36, 6→5, 8→1
6	41	4→3, 5→21, 6→13, 7→4
7	11	5→3, 6→5, 7→3
8	11	6→2, 7→2, 8→5, 9→2
9	5	8→1, 9→2, 10→2
10	3	8→1, 9→2
12	2	11→1, 13→1

そあん（soan）

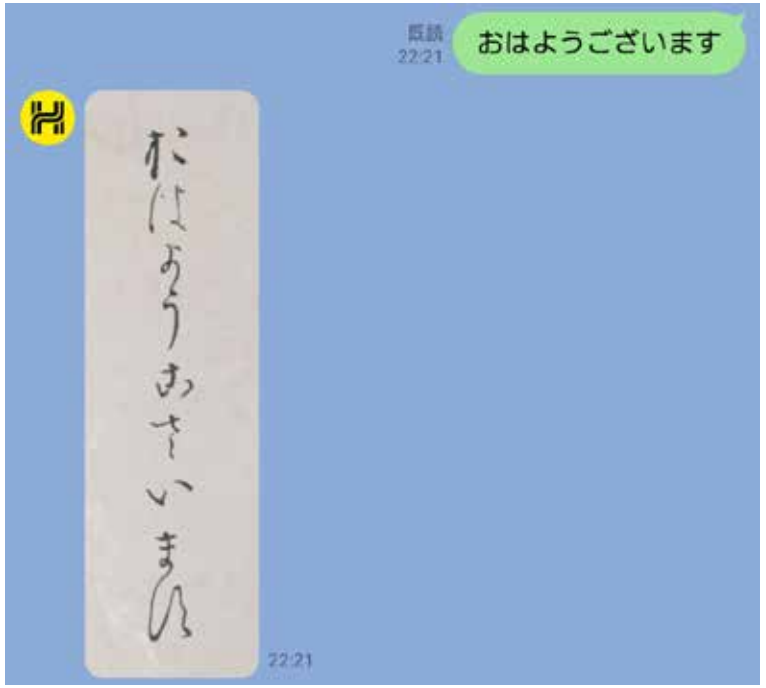
http://codh.rois.ac.jp/soan/



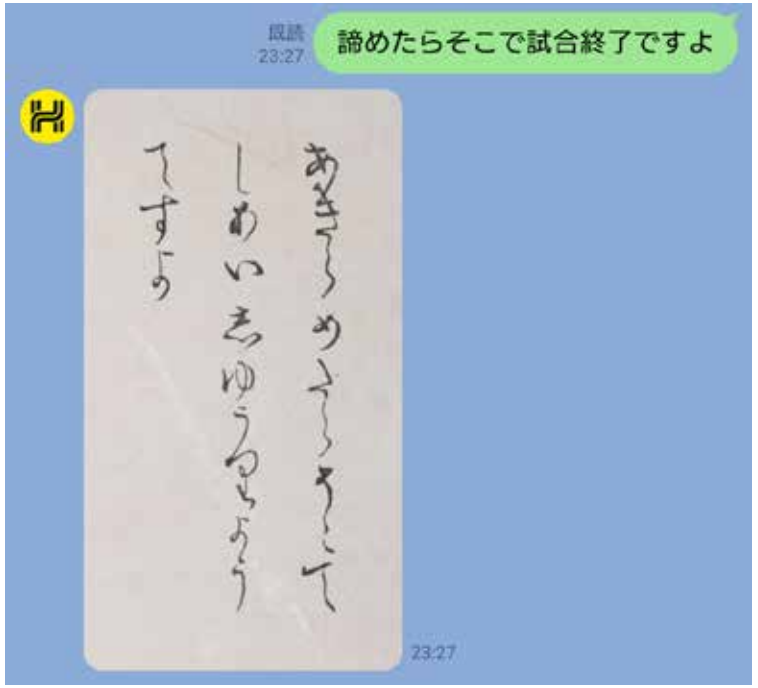


1. 任意の日本語テキストを、くずし字画像に変換可能
2. コラージュ技法：画像を断片化し、組み合わせ、合成する
3. **生成AI的要素はない**

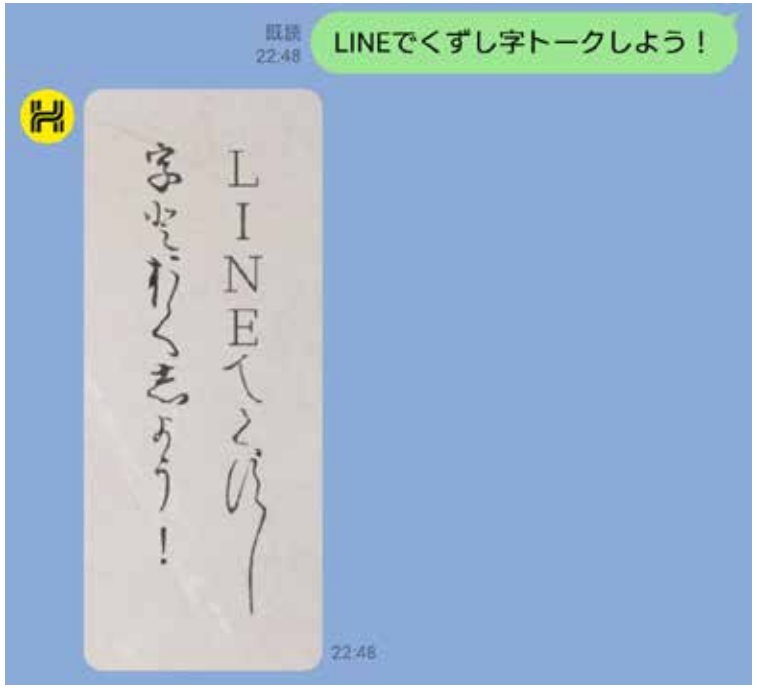
そあん（soan）ボット - LINEでくずし字トーク http://codh.rois.ac.jp/soan/line/



日常のあいさつや好きな言葉など、何でもくずし字に変換できます



漢字が表示できないときは、読みを推定してひらがなを表示します



漢字とひらがなだけでなく、カタカナやアルファベットにも対応します

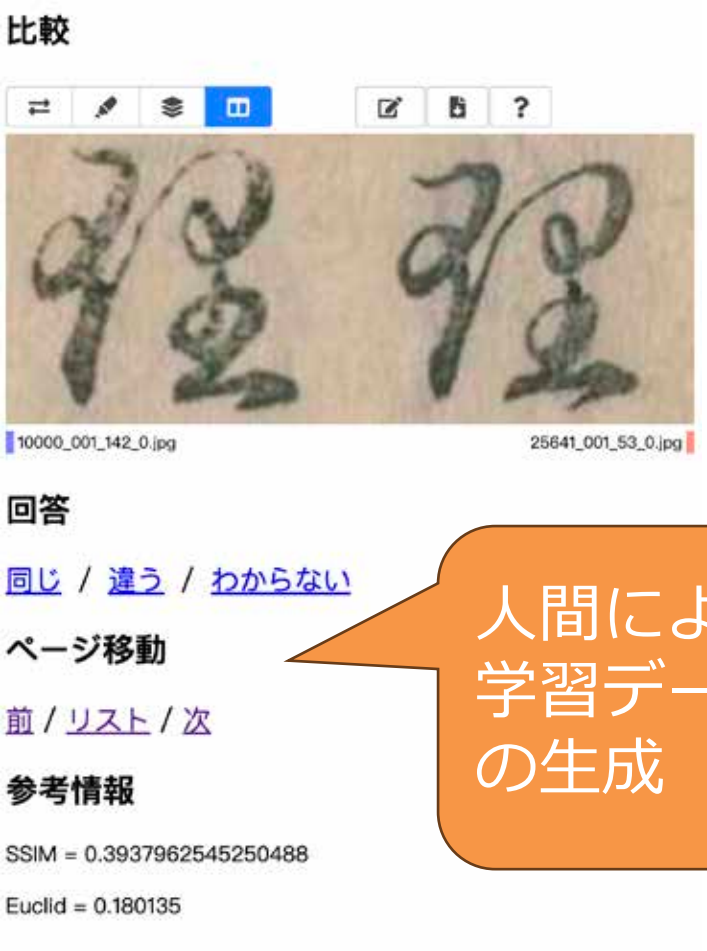
2. 活字ブロック自動同定

1. 同一文字列（＝同一字母文字列）に対して、複数の字形が存在する
2. 字形の類似度に基づき、類似した字形同士をまとめる「クラスタリング」を行う
3. 同一字母文字列に対して、クラスタ数とブロック数の統計情報を計算する
4. 本での出現場所と、活字ブロックのクラスタとの関係をリンクしてIIIF Curation Viewerに表示

字形の類似度

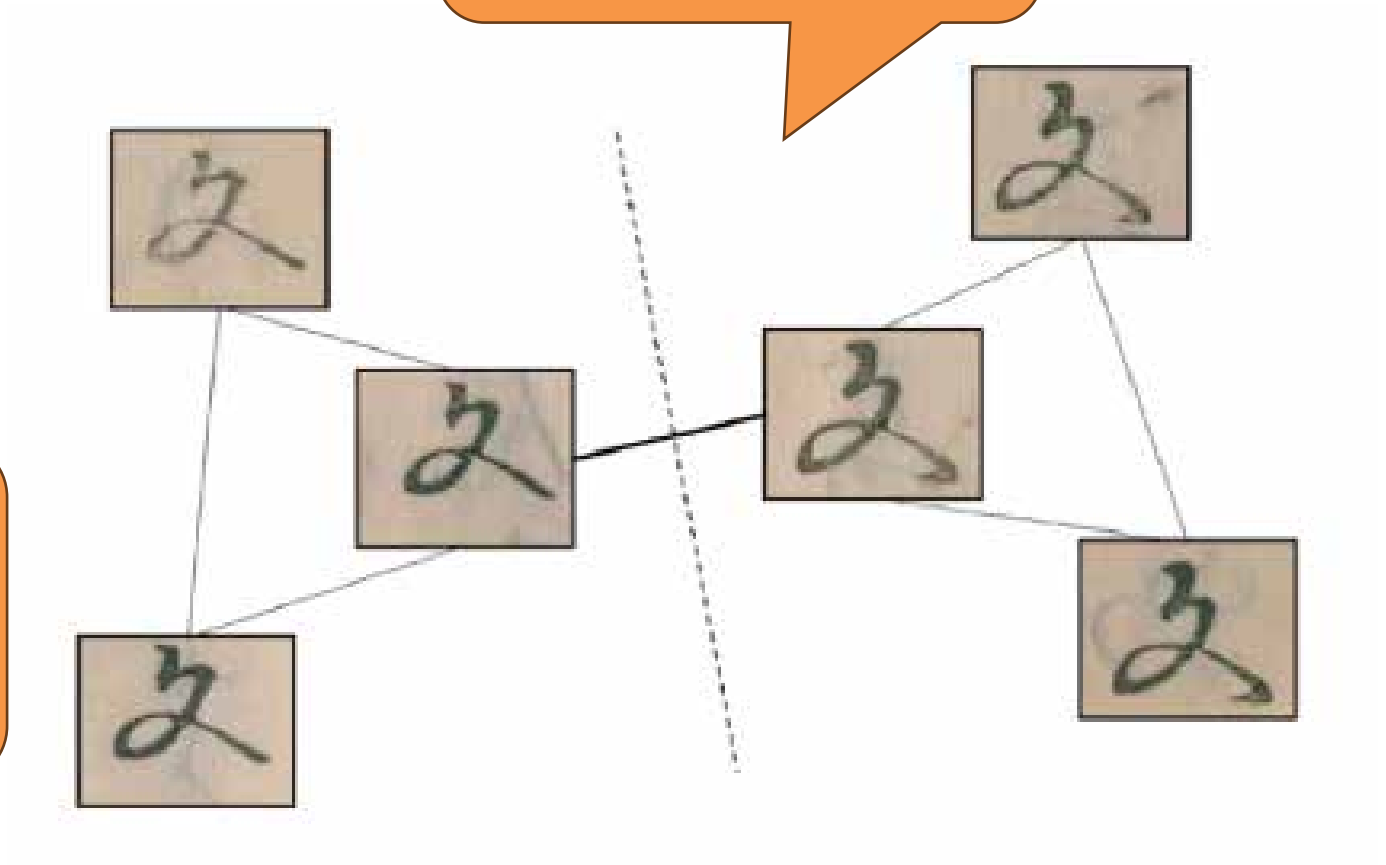
1. Structural Similarity Index Measure（SSIM）：画像の知覚的な類似度を計算
2. Siamese Network：画像のペアから距離を学習（ノイズに強い類似度を学習）
3. Random Forest：一部のデータに人間が正解を付与し、半教師あり学習を行い、2つの類似度を統合
4. Community Detection：類似度グラフに基づき、凝集度が高いクラスタを検出

字形の類似度



人間による学習データの生成

コミュニティ検出の概念図

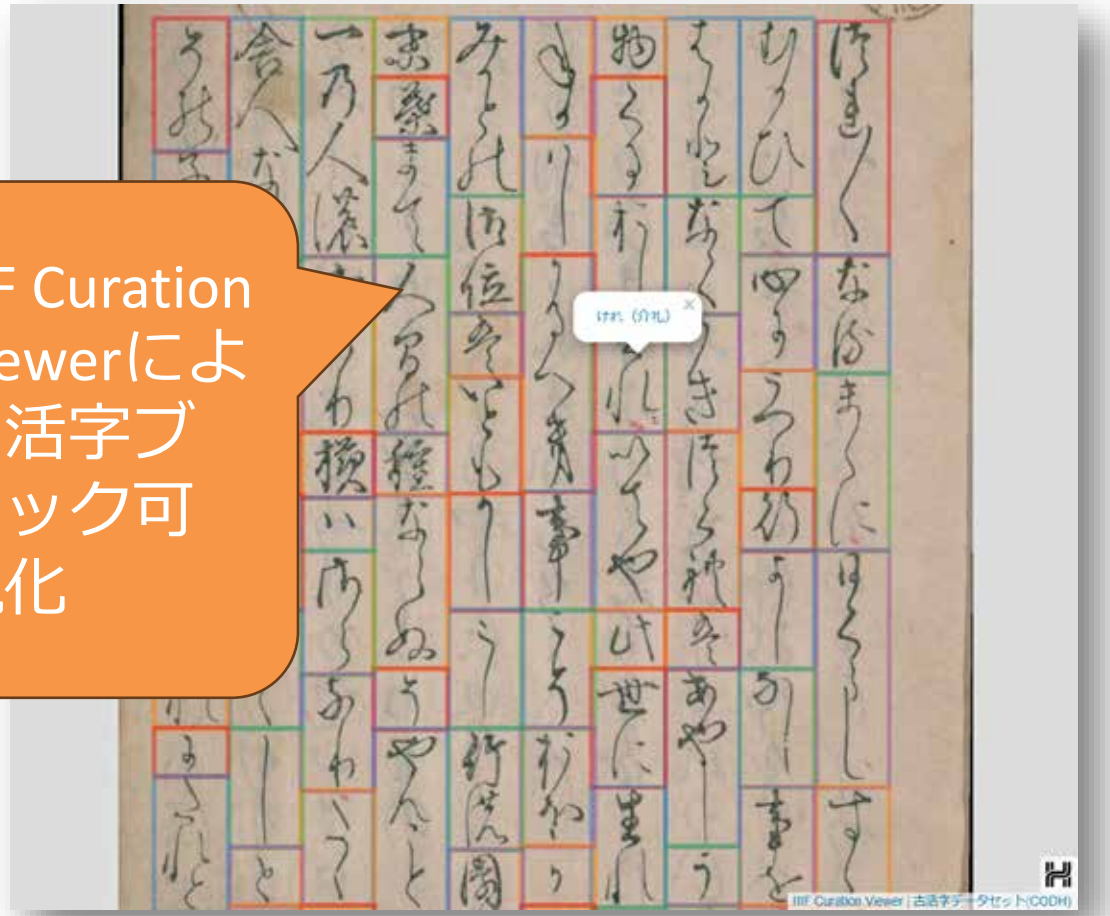


古活字ブロック分析

https://codh.rois.ac.jp/omt/block/

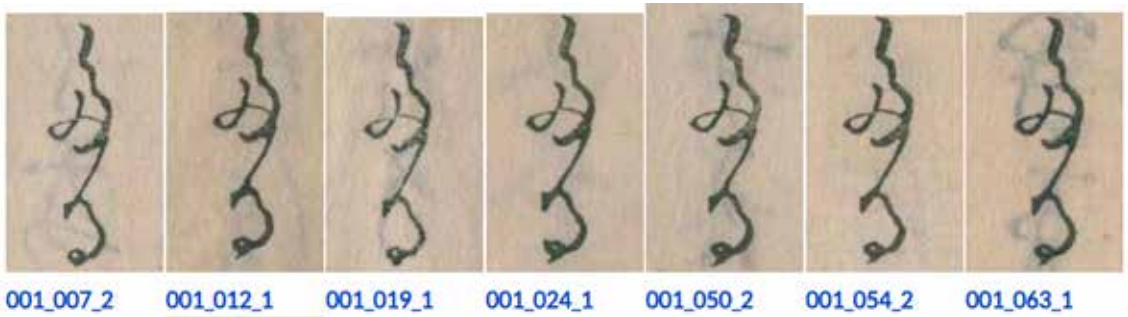
字母ごとに集計したクラスタ数とブロック数の統計情報

IIIF Curation Viewerによる活字ブロック可視化



字母	翻字	クラスタ数	ブロック数
能	の,能	235	762
遠	を,遠	150	715
越	を,越	138	334
天	て,天	133	933
乃	の	109	491
毛	も,毛	96	509
尓	に	90	890
人	人	83	509
幾	き,幾	74	391
濃	の,濃	69	182
仁	に,仁	65	364
比	ひ,思ひ,比	64	150
也	や,也	62	282
物	物	56	146
登	と,登	55	294
八	は,ハ,八	51	713

クラスタリング結果「多留」

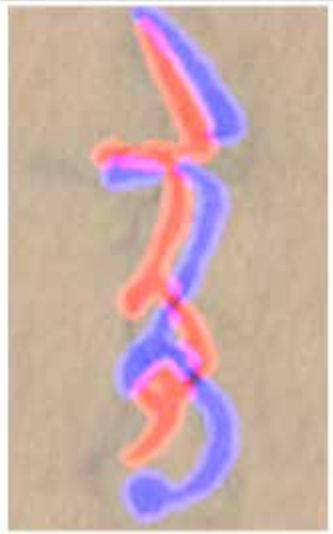


Vdiff.jsによる差読（Differential Reading）

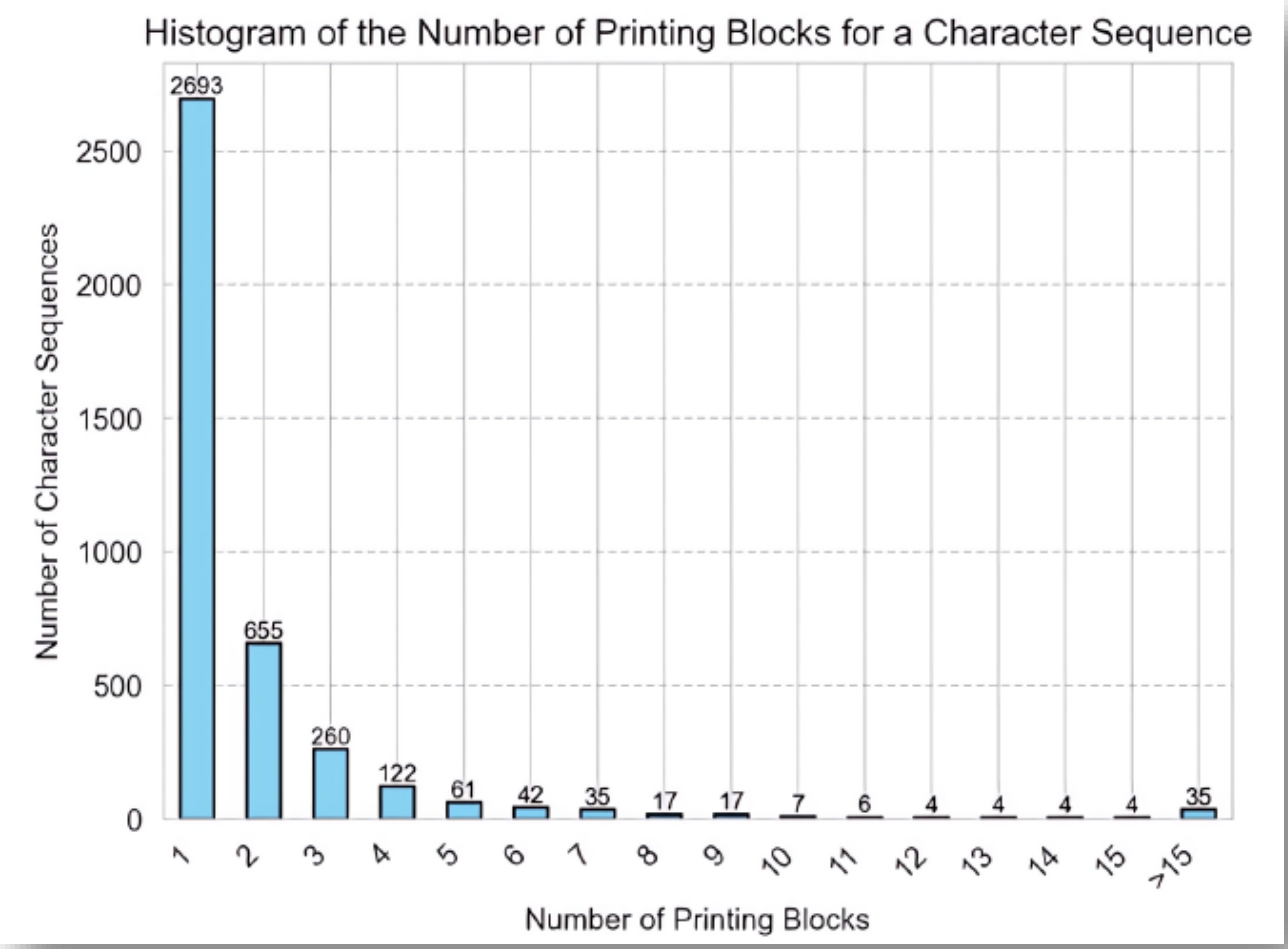
https://codh.rois.ac.jp/differential-reading/

同一活字の場合
ぴたりと重なる

同一活字でない場合
両者の字形の差異がわかる



統計情報の分析



1. 文字列ごとに何種類の活字があるか？
2. 活字は何回再利用されているか？
3. 活字の初出と印刷順序の関係は？
4. 異なる本での活字の再利用の状況は？

活字ブロック自動同定の課題

1. 字形の類似度は、画像の全体を見るため、字形の違いが細部のみにある場合は区別しづらい
2. 字形が単純な文字（「一」など）は類似度の計算が難しい
3. 墨のかすれや裏写りなど、ノイズに対する耐性を高める必要がある
4. **専門家がすでに有する知見を、類似度計算にもっと埋め込んでいく必要がある**

3. AI分析基盤

1. 可視化環境：本と活字の空間的・順序的な関係をビジュアルに閲覧
2. アナリティクス基盤：頻度などの統計情報などをダイナミックに閲覧
3. 対話型インタフェース：自動分析の誤りを修正し、それを新たな学習データに追加
4. これからAI分析基盤の研究を開始し、成果をまとめる予定

まとめ

1. 古活字版の情報解析の3課題：活字ブロック自動分割、活字ブロック自動同定、AI分析基盤
2. 活字ブロック自動分割の成果として、「古活字データセット」および「そあん」を公開した
3. 活字ブロック自動同定の成果として、同一文字列に何種類の活字があるか、などの統計情報を計算できた
4. 今後は専門家のフィードバックを取り込み、研究に有用なAI分析基盤を構築したい