Vol.2022-CH-128 No.13
2022/2/19

デジタル源氏物語（AI 画像検索版）：くずし字 OCR と編集距離を用いた写本・版本の比較支援システムの開発

中村覚 [1]　田村隆 [1]　永崎研宣 [2]

概要：今日多くの機関から古典籍の写本・版本のデジタル画像が公開されている．しかし，これらの諸本からある一つの場面を確認したいとき，多数（『源氏物語』の場合，桐壺巻から夢浮橋巻まで約 2,000 枚）の画像から目当ての場面を探し出すにはコストがかかる．この課題に対して，本研究ではくずし字 OCR と編集距離を用いた写本・版本の比較支援システムを開発する．また『源氏物語』を対象とした実験を行い，本システムの有用性を検証する．

キーワード：源氏物語，くずし字 OCR

Digital Tale of Genji（Finding Digital Facsimiles Including Parallel Texts with AI）: Development of a Comparison Support System for Manuscripts and Printed Books Using Kuzushiji OCR and Editing Distance

SATORU NAKAMURA[†1]　TAKASHI TAMURA[†1]
KIYONORI NAGASAKI[†2]

Abstract: Today, digital images of manuscripts and printed books of classical books are available from many institutions. However, when one wants to check a certain scene from these manuscripts, it takes a lot of time to find the desired scene from many images. To solve this problem, we have developed a system that supports comparison between manuscripts and printed books using kuzushiji OCR and edit distance. To verify the usefulness of this system, we will conduct an experiment on "The Tale of Genji".

Keywords: The Tale of Genji, kuzushiji OCR

1. はじめに

1.1 背景と目的

今日多くの機関から古典籍の写本・版本のデジタル画像が公開されている．また IIIF に対応した画像公開も積極的に進められている．例えば源氏物語については，国立国会図書館，国文学研究資料館，京都大学，九州大学，東京大学，米国議会図書館などで IIIF 画像が公開されている．

しかし，これらの諸本からある一つの場面を確認したいとき，テキスト検索ができないことが一般的であり，多数の画像から目当ての場面を探し出すにはコストがかかる．源氏物語の場合，例えば国文学研究資料館では全巻（桐壺巻から夢浮橋巻まで）を一つのアイテム（IIIF マニフェスト）にまとめて公開しており，約 2,000 枚の画像から構成される．一方，巻毎にアイテム（IIIF マニフェスト）を公開している機関もあるが，この場合，平均約 40 枚の画像から各巻が構成される．これらの画像から，例えば「夕顔」という文字列が含まれる「校異源氏物語」の 743 頁の場面は、「九州大学 所蔵 源氏物語 古活字版」では 22 巻の（80 枚中の）52 枚目，「東京大学文学部国文学研究室所蔵本（国文学研究資料館提供）」では（2328 枚中の）866 枚目が該当する．[a]

この課題に対して，本研究では諸本画像の検索を支援するシステムを開発する．具体的には，くずし字 OCR と編集距離を用いた写本・版本の比較を支援する．また『源氏物語』を対象とした実験を行い，本システムの有用性を検証する．

1.2 関連研究

諸本の比較について，特に情報技術を活用した研究について取り上げる．写本の比較を対象とした研究として，齊藤[1]は仮名字母の出現傾向に基づき，類似する写本の探索を行なっている．また版本の比較を対象とした研究として，宮川[2]は匡郭間距離比較による版種弁別法を提案している．北本ら[3]は画像処理に基づく版本の差読を行うブックバーコーディング法を提案している．本研究では，くずし字 OCR を用いた諸本の検索を行う点で，目的・手法ともに差異がある．

1 東京大学
The University of Tokyo.
2 人文情報学研究所
International Institute for Digital Humanities.

a) なお「校異源氏物語」では、「夕かほ」「ゆふかほ」といった表記が他の頁に含まれる。

くずし字 OCR に関しては，カラーヌワットタリン[4]らがデータセットの公開と，くずし字認識サービスを展開している．中村ら[5]はこのくずし字認識サービスを一部利用し，源氏物語を対象として，『校異源氏物語』や『新編日本古典文学全集』の頁番号を付すことで諸本の横断検索を実現している．一方で，頁番号付与の作業は人による確認が欠かせず，少人数の作業体制では一度に多くの伝本を組み入れることは困難であった．この課題に対して，本研究は諸本の横断検索プロセスの自動化を試みている点に新規性がある．

1.3 本論文の内容と公開済みシステムとの関係

1.1 で述べた目的に対して、筆者らは 2021 年 4 月 27 日に「デジタル源氏物語 AI 画像検索版」[6]を公開した。一方、本論文で述べる手法については、2022 年 1 月 25 日時点において、公開済みのシステムが採用している手法とは一部異なる点に注意されたい。本論文で述べる手法を採用した「デジタル源氏物語 AI 画像検索版」の改良版については、今後公開を行う予定である。

2. 提案手法

図 1 に提案手法の概要を示す．以下の 3 つのステップから構成される．

- 『校異源氏物語』の各頁テキストの作成
- くずし字 OCR を用いた諸本の見開き頁（画像）のテキスト作成
- 編集距離を用いた類似度の高い見開き頁の提示

以下では，これらの各ステップについて説明する．

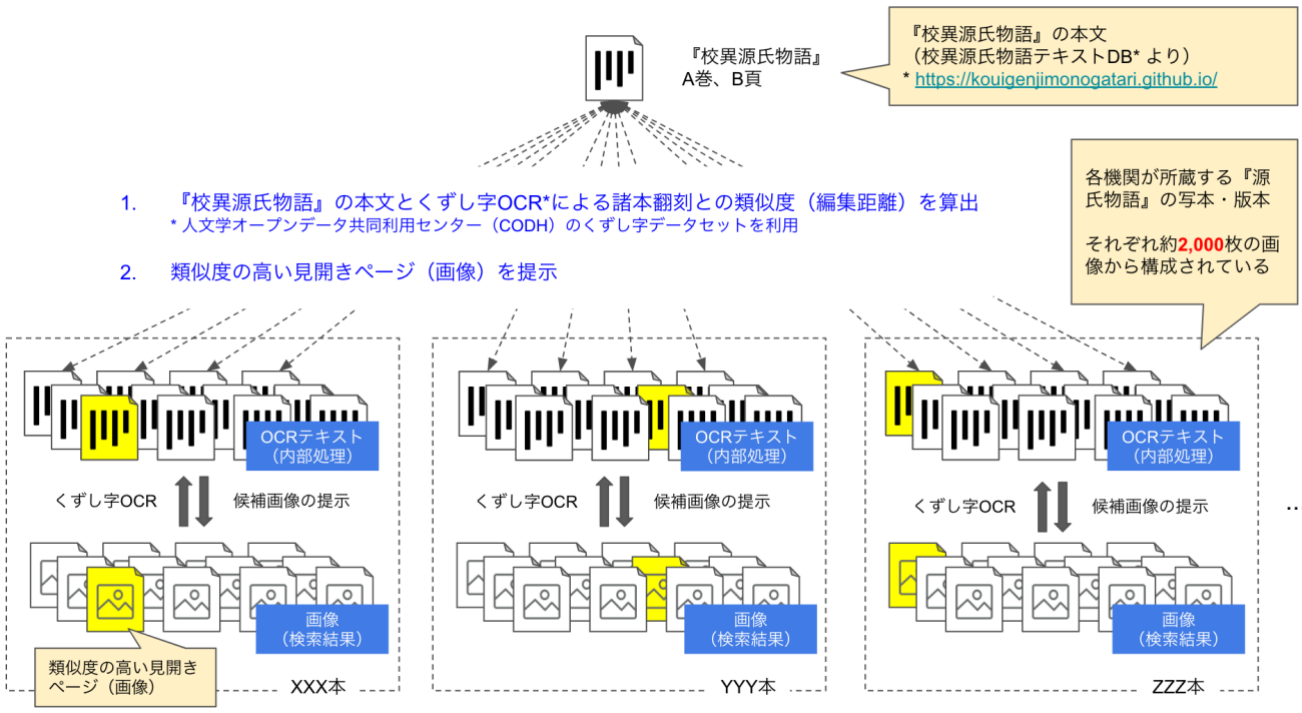

図 1　提案手法の概要

Figure 1　Overview of proposed methodology.

2.1 『校異源氏物語』の各頁テキストの作成

諸本との比較のベースとなる『校異源氏物語』の各頁テキストについては，「校異源氏物語テキスト DB」[7]で公開されているテキストデータ（利用条件 : CC0）を利用した．

2.2 くずし字 OCR を用いた諸本の見開き頁（画像）のテキスト作成

諸本の見開き頁（半丁の場合を含む）のテキスト作成にあたっては，くずし字 OCR を利用した．この OCR モデルについて，1.3 で述べた 2021 年 4 月 27 日に公開したシステムでは、CODH（人文学オープンデータ共同利用センター）が提供する「KuroNet くずし字認識サービス」[8]を利用している．一方，「KuroNet くずし字認識サービス」は人手による OCR 処理の実行を前提としており，複数の諸本に対する OCR 処理の一括適用に課題があった．そこで，CODH が公開する「日本古典籍くずし字データセット」を利用した独自の OCR モデルを開発した．本研究では，以下の段階を踏む非 end-to-end の手法を用いている．

- 文字検出
- 文字分類

- （+α）読み順の自動推定

以下，それぞれについて述べる．

2.2.1 文字検出

モデルとしては，YOLOv5[9]を利用した．訓練データとしては，1024 x 1024 ピクセルの画像 2,243 枚を利用した．学習に使用した設定としては，エポック数 100，バッチサイズ 4，初期の重みとして yolov5x.pt を用いた．その結果，テストデータ（642 枚の画像）に対して，図 2 に示す結果を得た．P は Precision，R は Recall，mAP は mean Average Precision を示す．

P	R	mAP@.5	mAP@.5:.95
0.991	0.982	0.988	0.889

図 2　文字検出モデルの学習結果

Figure 2　Training results of the character detection model.

また、検出結果の例を図 4 上部に示す。Confidence score が 0.6 以下の場合は青、それ以上の場合は赤の矩形を示す。

2.2.2 文字分類

モデルとしては，ImageNet で学習済みの ResNet152 のファインチューニングを利用した。訓練データとしては，64 x 64 ピクセルの 1,753 クラスの文字画像 161,946 枚を利用した。学習に使用した設定としては，エポック数 50，バッチサイズ 64 を用いた。その結果を図 3 に示す。検証データ（訓練データの 15%）に対しては、約 90% の categorical_accuracy を示した。

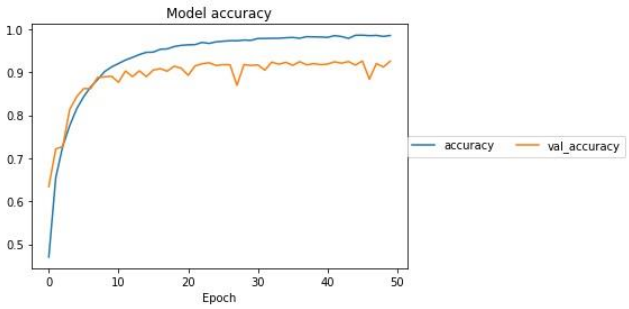

図 3　文字分類モデルの学習結果

Figure 3　Training results of the character classification model.

2.2.3 読み順の自動推定

2.2.2 で得られた結果に対して、読み順の自動推定を行う。文字行方向の2.2.1 で検出された Bounding Box に含まれるピクセル数の射影ヒストグラムを図 4 下部に示すように作成する。このヒストグラムから極大値・極小値などを算出し、読み順の自動推定を行う。

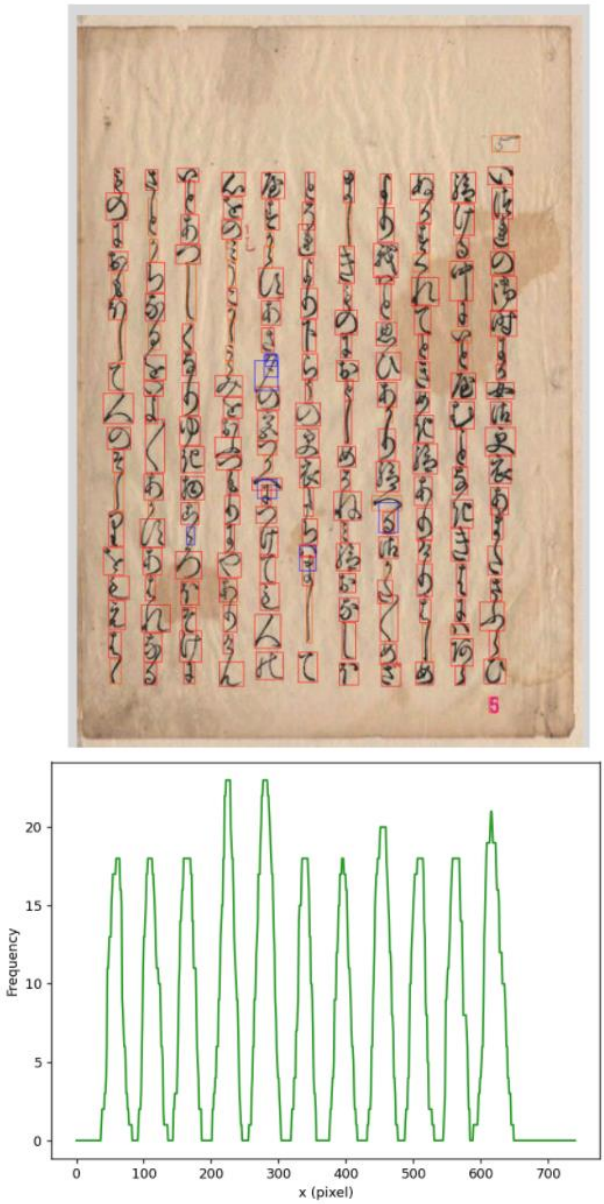

図 4　Bounding Box と射影ヒストグラムの例 [10]

Figure 4　Example of Bounding Box and Projection Histogram.

2.2.4 結果

KuroNet と開発したモデルの認識結果の比較結果を図 5 に示す．KuroNet の認識精度には及ばないものの，本研究が提案する編集距離を用いた類似度の算出プロセスにおいて適切に機能し得る認識精度であることを，後述するケーススタディにおいて確認している．

図 5　KuroNet と開発したモデルの認識結果の比較 [11] [12]
Figure 5　Comparison of recognition results between KuroNet and our model.

2.3 編集距離を用いた類似度の高い見開き頁の提示

2.1 と2.2 で得られたテキスト間の類似度を編集距離によって算出し,『校異源氏物語』の頁テキストを検索クエリとした際の，類似度の高い諸本の見開き頁を提示する．この提示方法について，後述する工夫を施している．図 6 とともに本手法を説明する．

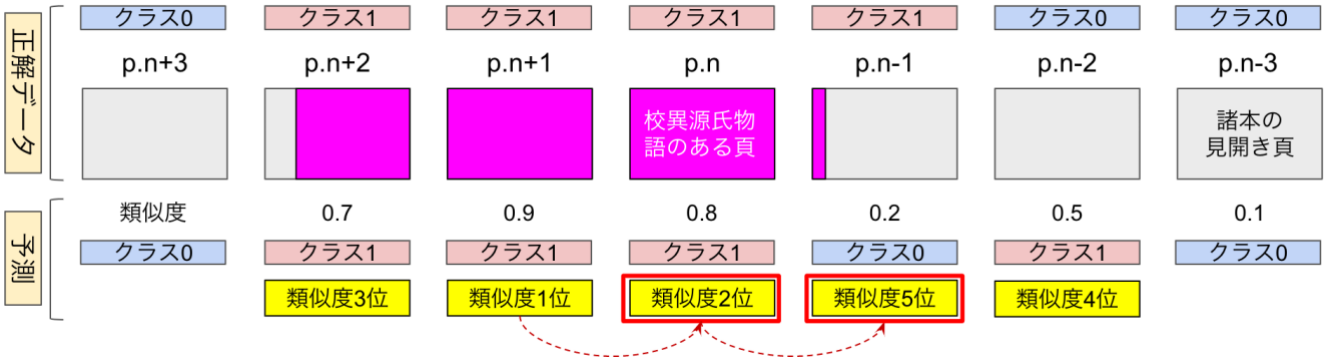

図 6　編集距離を用いた類似度の高い見開き頁の提示手法
Figure 6　A method for extracting similar pages using edit distance.

多くの場合において，校異源氏物語の頁の文字数が，諸本の見開き頁あたりの文字数を上回る．図 4 では,「校異源氏物語のある頁」として示されているように，校異源氏物語の頁テキストが，諸本の複数頁（p.n-1 から p.n+2 まで）に跨っている例を示す．

そのため,『校異源氏物語』の頁（以下，K 頁）のテキストと，諸本の見開き頁（以下，S 頁）全てを比較した際，図 4 上部に示すように，K 頁のテキストを含む S 頁（クラス 1）と K 頁のテキストを含まない S 頁（クラス 0）の 2 種類に分けられる．

このような関係を持つK 頁と S 頁のテキストの類似度を

編集距離によって算出する．その後，大津の二値化処理を用いて，S 頁を 2 クラス（クラス 1 と 0）に分類する．この時，p.n-1 のように，S 頁が K 頁のテキストのごく一部を含む場合，誤ってクラス 0 に割り当てられる場合がある．また，p.n-2 のように，テキストの類似性により，誤ってクラス 1 に割り当てられる場合もある．ただし後者の場合，p.n や p.n+1 のように K 頁のテキストの大部分を含む S 頁と比較して，類似度が小さくなることが想定される．

そしてクラス 1 に割り当てられた S 頁群について，類似度が最も高い S 頁（図 4 の場合，p.n+1）から連続する前頁について，最も頁数が小さくなる頁（p.n）を取得する．さらに，クラス 0 に割り当てられたその前の頁（p.n-1）を取得し，これらの 2 頁を結果として取得する．この流れを図の赤点線で示す．

これにより，K 頁のテキストの開始位置を含み得る S 頁を提示する．本手法を実際の画像に当てはめた結果を図 7 および図 8 に示す．図中には諸本の複数の頁を示し，校異源氏物語のある頁の範囲を紫色の線で示している．また，編集距離の算出により，最も類似度が高いと判定された頁を黄色で示す．

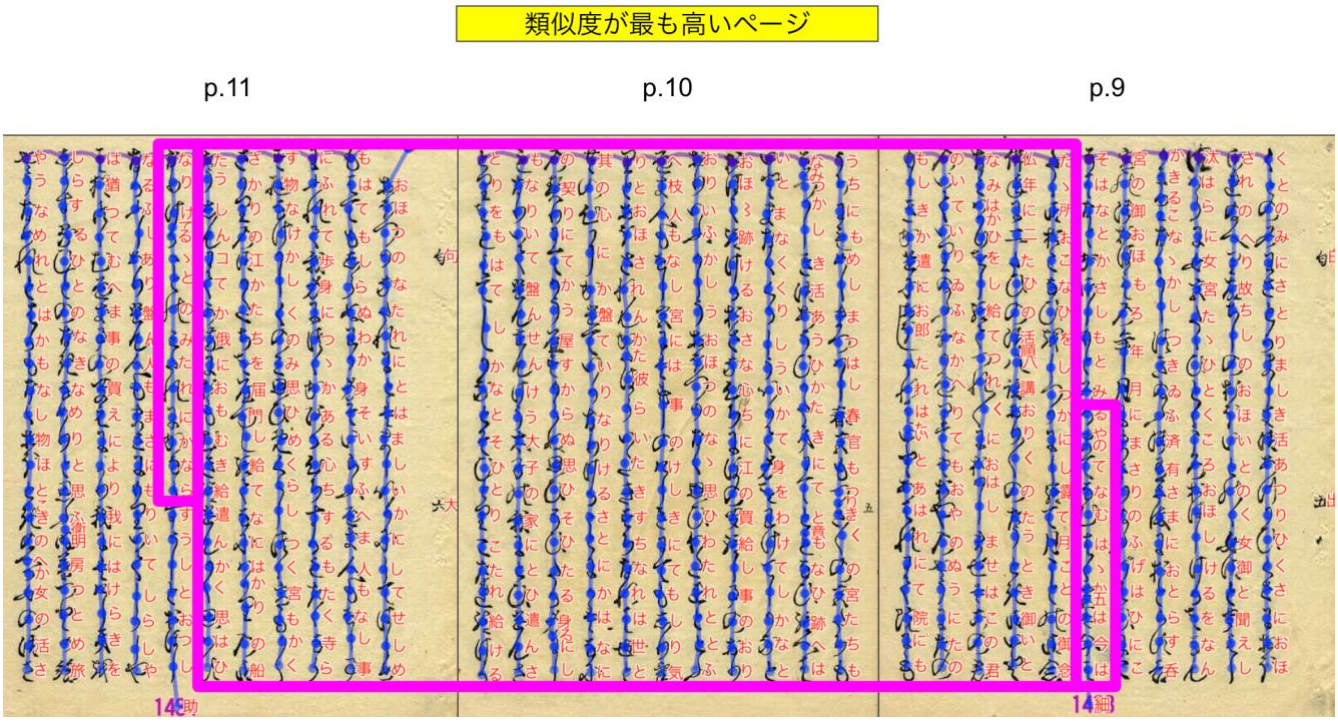

図 7　校異源氏物語と諸本の頁の対応例 1 [13]

Figure 7　Example 1 of correspondence between the Tale of Genji and the pages of various books.

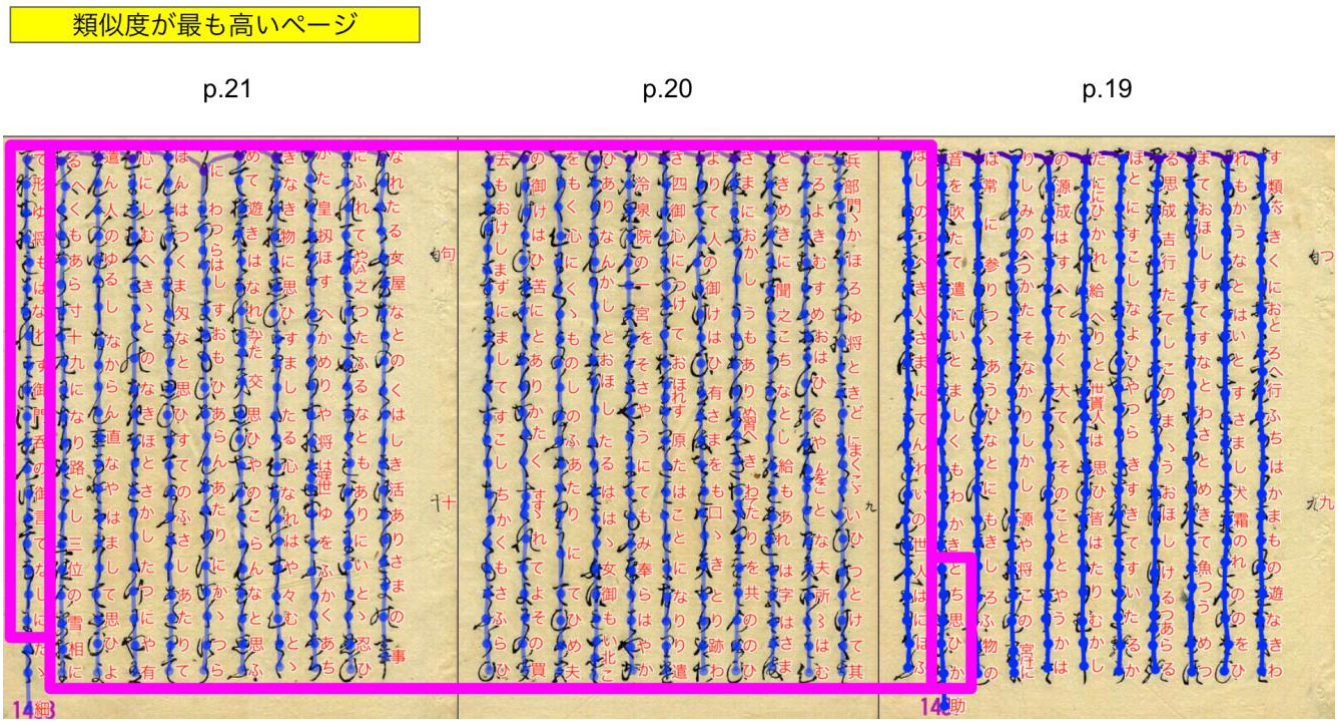

図 8　校異源氏物語と諸本の頁の対応例 2 [13]

Figure 8　Example 2 of correspondence between the Tale of Genji and the pages of various books.

3. ケーススタディ

『校異源氏物語』の頁に対応する諸本の見開き頁を正しく検出することができるかを確認し，提案手法の有用性を評価する．

3.1 内容

対象資料として，以下の 4 つを選定した．

- 個人蔵 源氏物語 無跋無刊記整版本 [10]
- 九州大学文学部 所蔵 源氏物語 古活字版 [13]
- 東京大学本（東京大学総合図書館所蔵）[14]
- 『湖月抄』鵜飼文庫（国文研所蔵）[15]

これらを選定した理由として，既存研究[5]において，上記資料群の見開き頁に『校異源氏物語』の頁数を人手で付与済みであり，これを正解データとして利用可能であるためである．

3.2 結果

結果を図 9 に示す．『湖月抄』については，正解データの作成が 23 巻までとなっている点に注意されたい．

各対象資料について，99%以上の精度で『校異源氏物語』の頁に対応する諸本の見開き頁を正しく検出することができた．

検出が失敗した箇所を赤色で示す．検出に失敗した理由として，「くずし字 OCR の認識誤り」「読み順の自動推定誤り」「編集距離を用いた類似度の高い見開き頁の提示誤り」など，複数の要因が挙げられる．

巻数	頁数	正解した頁数			
		東大本	九大・古活字版	九大・無跋無刊記整版本	湖月抄
1	24	24	24	24	23
2	45	45	45	45	45
3	11	11	11	11	11
4	46	46	46	46	46
5	45	45	45	45	43
6	30	30	30	30	30
7	27	27	27	27	27
8	10	10	10	10	10
9	44	44	43	44	44
10	47	47	47	47	47
11	4	4	4	4	4
12	42	42	42	42	42
13	38	38	38	38	38
14	32	32	32	32	32
15	23	22	23	23	23
16	5	5	5	5	5
17	18	18	18	18	18
18	20	20	20	20	20
19	29	29	29	29	29
20	20	20	20	20	20
21	48	48	48	48	48
22	39	39	39	39	39
23	14	14	14	12	14
24	20	20	20	20	-
25	19	19	19	19	-
26	21	21	21	21	-
27	4	4	4	4	-
28	17	17	17	17	-
29	27	27	27	27	-
30	14	14	14	14	-
31	35	35	35	35	-
32	18	18	18	18	-
33	23	20	23	23	-
34	96	96	96	96	-
35	97	96	97	97	-
36	38	38	38	38	-
37	18	18	18	18	-
38	14	14	14	14	-
39	67	67	67	67	-
40	18	10	18	16	-
41	21	21	21	21	-
42	13	13	13	13	-
43	12	12	12	12	-
44	39	39	39	39	-
45	36	36	36	36	-
46	35	35	35	35	-
47	84	84	84	84	-
48	18	18	18	18	-
49	88	86	88	88	-
50	60	60	60	60	-
51	67	67	67	67	-
52	54	54	54	54	-
53	62	62	62	62	-
54	16	16	16	16	-
合計（コマ数）	1812	1797	1811	1808	658（／661）
合計（%）	-	99.45	99.94	99.78	99.55

図 9　ケーススタディの結果

Figure 9　Results of the case study.

4. 開発したシステム

2 の提案手法に基づき開発したシステムを図 10 に示す．利用者は，図左に示す検索画面から，校異源氏物語のテキストデータに対して検索を行い，目的とする場面を検索する．図では，「夕顔」をクエリとして検索している例を示す．

その結果，校異源氏物語の 743 頁が結果として得られ，図右に示す詳細画面に遷移する．本画面において，システムに登録済みの諸本から，類似度の高い見開き頁が最大 2 件ずつ表示される．参照可能な資料群としては，3 で挙げた 4 つに加えて，国立国会図書館所蔵本や京都大学所蔵本など，不定期にデータ追加を進めている．2022 年 1 月 24 日時点においては，多い巻で 20 件以上の諸本を比較することができる．

図 10　開発したシステム

Figure 10　Developed System.

5. 考察

今後の取り組みとして，他の作品への展開が挙げられる．この展開にあたっては，「第三者がデータを作成できる環境」と「作成されたデータを集約するプラットフォーム」が必要となる．前者については，Google Colaboratory を利用したデータ作成環境を構築している．具体的には 2 で述べた手法に基づき，源氏物語や他の作品を対象としたデータの作成が可能な環境を構築している．作成されたデータは GitHub で共有する仕組みを採用している．後者については，

GitHub に集約されたデータを呼び出すことで，複数の作品を対象に，4 で示した機能を提供するシステムを開発している．「伊勢物語」「栄花物語」「大鏡」などを対象に，試験環境の構築を進めている．本システムについては，改めて報告の機会を得たい．

一方，他の作品への展開を進める上で，「校異源氏物語」のようなベースとなるテキスト（例：新編日本古典文学全集）の作成が課題となる．著作権を侵害しない範囲でのテキストの作成と利用の方法について今後検討したい．

6. おわりに

本研究ではくずし字 OCR と編集距離を用いた写本・版本の比較支援システムを開発した．『源氏物語』を対象とした実験を行い，本システムの有用性を検証した．具体的には，4 つの諸本を対象として，『校異源氏物語』に対応する諸本の見開き頁について，99%以上の精度で正しく抽出できることを確認した．今後は，他の作品への展開を視野に入れた取り組みを行う．

謝辞 本研究は JSPS 科研費 19K20626 および東京大学 FSI 事業「データ駆動型歴史情報研究基盤の構築」の助成による成果の一部です．

参考文献

[1] 齊藤鉄也．仮名字母の出現傾向が類似する鎌倉時代書写の源氏物語写本の探索，じんもんこん 2021 論文集, Vol. 2021, pp.162-169, 2021.
[2] 宮川真弥，覆刻版における版面拡縮現象の具体相 ： 匡郭間距離比較による版種弁別法確立のために，斯道文庫論集 = Bulletin of the Shidô Bunko Institute, Vol.53, pp.231-296, 2018.
[3] 北本朝展，藤實久美子，本間淳．ブックバーコーディング法: 版本の差読に基づく「武鑑全集」の網羅的な解析に向けて，じんもんこん 2021 論文集, Vol. 2021, pp.268-275, 2021.
[4] カラーヌワットタリン，北本朝展．くずし字認識の進化とサービス化の展開，じんもんこん 2020 論文集, Vo.2020, pp.3-10, 2020.
[5] 中村覚，田村隆，永崎研宣．源氏物語本文研究支援システム「デジタル源氏物語」の開発における IIIF・TEI の活用，研究報告人文科学とコンピュータ（CH），Vol. 2020-CH-124, No.2, pp.1-7, 2020.
[6] デジタル源氏物語 AI 画像検索版, https://genji-ai.web.app/, (参照 2022-01-25).
[7] 校異源氏物語テキスト DB, https://kouigenjimonogatari.github.io/, (参照 2022-01-24).
[8] KuroNet くずし字認識サービス（AI OCR），http://codh.rois.ac.jp/kuronet/, (参照 2022-01-24).
[9] YOLOv5, https://github.com/ultralytics/yolov5, (参照 2022-01-24).
[10] 個人蔵 源氏物語 無跋無刊記整版本, http://hdl.handle.net/2324/411265, (参照 2022-01-24).
[11] 『日本古典籍くずし字データセット』（国文研所蔵／CODH 加工） doi:10.20676/00000340
[12] 源氏物語（国文学研究資料館所蔵），https://kotenseki.nijl.ac.jp/biblio/200010454/, (参照 2022-01-26).
[13] 九州大学文学部 所蔵 源氏物語 古活字版, http://hdl.handle.net/2324/411193, (参照 2022-01-24).
[14] 東京大学本（東京大学総合図書館所蔵），https://iiif.dl.itc.u-tokyo.ac.jp/repo/s/genji/page/list, (参照 2022-01-24).
[15] 『湖月抄』鵜飼文庫（国文研所蔵），https://kotenseki.nijl.ac.jp/biblio/200018258, (参照 2022-01-24).