Digital Studies / Le champ numérique

Morris, James Harry and Antonia Karaisl. 2025. "Accuracy Isn't All: Testing KuroNet for Kanbun OCR." *Digital Studies/Le champ numérique* 15(1): 1–32. https://doi.org/10.16995/dscn.17257.

Open Library of Humanities

Accuracy Isn't All: Testing KuroNet for Kanbun OCR

James Harry Morris, National Institutes for the Humanities and National Museum of Japanese History, Japan, morrisjamesharry@gmail.com

Antonia Karaisl, Waseda University, Japan, antoniakaraisl@gmail.com

This paper tests the open-source OCR software KuroNet's performance on printed texts written in *kanbun*, comparing the results to other freely available off-the-shelf OCR solutions. *Kanbun*, a literary standard using Chinese characters and syntax to represent Japanese textual content, was employed widely throughout Japan from the classical into the modern period. Its peculiarities with regard to layout and characters present a challenge to standard OCR software that has not been tackled to date. KuroNet was developed for a different purpose, namely, to help decipher literature written in cursive Japanese characters; its idiosyncratic approach to irregular layouts, however, commends KuroNet for *kanbun* as well. The survey shows that due to its unique approach to *kanji* detection, KuroNet's output surpasses that of programs with higher transcription accuracy rates on pages with a difficult layout. The present paper provides the background to the study, compares the results between KuroNet and comparable OCR programs, and closes with an analysis of KuroNet's weaknesses, with recommendations for further improvements.

Cet article évalue les performances du logiciel open-source de Reconnaissance Optique de Caractères (ROC), KuroNet, sur des textes imprimés écrits en kanbun, en comparant les résultats à ceux d'autres solutions ROC disponibles gratuitement. Le kanbun, une norme littéraire utilisant des caractères et une syntaxe chinoise pour représenter du contenu textuel japonais, a été largement utilisé au Japon de la période classique à la période moderne. Ses particularités en matière de mise en page et de caractères posent un défi aux logiciels ROC standards, un défi qui n'a pas encore été relevé jusqu'à présent. KuroNet a été développé dans un autre but : aider à déchiffrer la littérature écrite en caractères cursifs japonais. Toutefois, son approche particulière des mises en page irrégulières le rend également pertinent pour le kanbun. L'étude montre que, grâce à son approche unique de la détection des kanji, les résultats de KuroNet surpassent ceux de programmes ayant des taux de précision de transcription plus élevés, notamment sur des pages à mise en page complexe. Le présent article fournit le contexte de l'étude, compare les résultats entre KuroNet et d'autres programmes de ROC comparables, et se termine par une analyse des faiblesses de KuroNet, accompagnée de recommandations pour des améliorations futures.

Digital Studies/Le champ numérique is a peer-reviewed open access journal published by the Open Library of Humanities. © 2025 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

OPEN ACCESS

1 Introduction

Throughout Japanese literary history, the writing convention of *kanbun* 漢文 (Classical Chinese writing) plays a curious role: its scope was both limited to educated audiences and geographically widened across borders due to its hybrid nature. *Kanbun* employed Chinese characters and syntax to represent Japanese textual content, utilizing a range of punctuation and diacritical marks to allow readers to reorder and augment the characters to conform to Japanese grammar. Amongst Japan's own populace, a certain level of education was required to be able to read and write *kanbun*. Meanwhile, the use of Chinese characters and syntax made it what Aliz Horvath terms "a sort of 'written lingua franca' in the Sinosphere," as it rendered Japanese textual content legible to a Chinese audience (Horvath 2022, 127). *Kanbun*'s *longue duree*, that is, from the classical period throughout the early modern period and beyond, and the variety of use cases make for a rich literary heritage in print and manuscript, large parts of which are preserved to date. The fact that Japanese high school students are still required to pass a *kanbun* exam upon entry to university pays testament to the continued importance of *kanbun* literacy. As Wang, Shimizu, and Kawahara note, however, in spite of this literary wealth and cultural relevance, the digital resources and tools available for *kanbun* documents remain scarce in comparison to other languages (Wang, Shimizu, and Kawahara 2023). Particularly, the lack of a bespoke OCR solution hinders the systematic exploration of *kanbun* sources at scale.

Part of the reason may be the hybrid nature of *kanbun* as a literary standard presenting Japanese textual content through Sinographic writing. *Kanbun*'s punctuation marks and diacritics prescribe where the sequence of characters needs to be switched in order to conform to Japanese syntax and provides prefixes and suffixes lacking in Chinese but needed in Japanese to clarify the grammatical relationships between the different words (Wang, Shimizu, and Kawahara 2023). This means that whilst a person trained in Classical Chinese may have been able to understand the content without resorting to these markers, a Japanese-literate Chinese-agnostic audience would rely on them to correctly render the meaning. In theory, either party would understand the same content from the same letters but mediated through a different language. This implies that virtually the same text can be used to represent two different languages at the same time. The reality, however, can be more complex. *Kanbun* is neither simply nor always Chinese text rendered grammatically Japanese. There are gradations making some forms of *kanbun* unique to Japanese as a language rather than Chinese as a writing standard. Throughout the period of adaptation, *kanbun* writing came to feature words and grammatical constructions unique to the Japanese language, even though written in Chinese characters and syntax (Morley 2022; Rabinovitch 1996).

Apart from "pure *kanbun*" (*junkanbun* 純漢文) which upheld Chinese norms in terms of syntax and semantics, divergent categories like *hentai kanbun* 変体漢文 (variant *kanbun*), *waka kanbun* 和化漢文 (Japanized *kanbun*) or *giji kanbun* 擬似漢文 (imitation *kanbun*), for example, could feature idiosyncratically Japanese elements. Not merely a problem of geography, orthographic conventions changed over time (Morley 2022). Furthermore, Elizabeth Oyler notes that at least during the Heian and medieval periods the boundaries between *kanbun* and *wabun* 和文 (Japanese writing) "were permeable and performative," and that the prevailing idea that writers saw themselves as choosing between Chinese *kanbun* and Japanese *wabun* reflects modern constructs rather than actual practice (Oyler 2006). Besides, the conception that *kanbun* was solely a language of the ruling classes, as has been traditionally postulated and popularly accepted, is blurred when we acknowledge that the use of diacritics and glosses were meant to make these texts accessible to more readers (Moretti 2020). Most notably, the transformations of Edo Japan's socioeconomic geography had a trickle-down effect on *kanbun* literacy. By initiative of the government, for example, numerous samurai were moved to the castle towns, which required many village headmen to acquire a level of *kanbun* fluency so as to be able to communicate in writing with their remotely living administrators (Sangawa 2017). From a geopolitical angle, meanwhile, scholars such as Atsuko Ueda have suggested that we reject the terms "Chinese" and "Japanese" in descriptors of *kanbun* due to their misleading nature and political connotations (Ueda 2008).

This linguistic fluidity and the variety of writing standards exacerbate the challenge of developing OCR solutions for *kanbun*, starting with the selection of a model: granted most OCR solutions classify models by language, there is no obvious choice for a hybrid writing standard like *kanbun*. As Horvath notes, attempting to tackle the issue through a solution for classical Chinese literature would be a start (Horvath 2022). Indeed, considerable efforts have been diverted towards developing OCR for premodern Chinese sources, with remarkable success (Sturgeon 2018; Yang, Jin, and Sun 2018). Meanwhile, open-source solutions remain rare and state-of-the-art commercial software, when applied to *kanbun*, tends to perform pitifully when confronted with the broad range of characters, layout peculiarities and the annotations inserted as reading aids (Horvath 2022). In particular, the frequent practice of inserting sections of double columns within a single line of text is a challenge to OCR systems that traditionally rely on line segmentation and layout analysis. In short, whilst an OCR solution for *kanbun* texts remains highly desirable, the particularities with regard to layout and letters remain a veritable problem to date.

The following article presents a series of experiments processing printed *kanbun* texts with a resource that may not be an obvious contender in the first place: the

open-source OCR software KuroNet (KuroNet 2020), trained for the recognition of Japanese *kuzushiji* (くずし字) or cursive characters. The following argument will explain the conceptual background, compare KuroNet to other freely available software, present the experiment, and discuss the results. The conclusion will comment on further outlook for using this technology on *kanbun* resources and further avenues for research.

2 Testing KuroNet on *kanbun*

2.1 Experiment description

KuroNet is designed as a system to help read books written in so-called *kuzushiji* characters, cursive forms of writing used in both handwritten and printed texts in pre-modern Japan. Pre-modern cursive writing is extremely hard to read for most contemporary Japanese users and presents an unusually difficult but poignant problem to solve: whilst most OCR systems render a text machine-readable that otherwise poses no problem to the human eye, KuroNet genuinely helps to unlock a literary heritage otherwise lost on a larger audience.

KuroNet's approach was first tabled by Tarin Clanuwat, Alex Lamb, and Asanobu Kitamoto in 2018 (Clanuwat, Lamb, and Kitamoto 2018), the application itself introduced in a conference paper in 2019 (Clanuwat, Lamb, and Kitamoto 2019), and further changes described in a follow-up paper in 2020 (Lamb, Clanuwat, and Kitamoto 2020). Throughout this series of papers, the authors outline the problems specific to *kuzushiji* that KuroNet seeks to tackle. Some of these problems relate to the cursive nature of *kuzushiji* writing, whereby characters are often connected or require the context of the preceding character to be correctly understood (Clanuwat, Lamb, and Kitamoto 2018). For most printed *kanbun* texts, these issues may be of lesser relevance. Yet several challenges regarding layout and character classification of *kuzushiji* and the solutions found for them, as outlined in the above papers, are relevant for the context of the experiment here: Firstly, many pre-modern books printed in *kuzushiji* feature characters written around illustrations, making the line segmentation process customary to many OCR engines a helpless exercise. Even where illustrations are not present, the reading order often does not align in regular lines—that is, text or blocks of text may be read not in sequential but seemingly random order. Differently from most OCR systems, the creators of KuroNet therefore chose to forego line segmentation and focus on identifying single characters on the page prior to classification (Lamb, Clanuwat, and Kitamoto 2020). KuroNet's text output, consequently, is not presented as a block of text in reading order as is customary in OCR software, but as a seemingly

randomly ordered set of characters. However, thanks to a bespoke interactive interface (the KuroNet Text Editor's Reading Order tool) the reader herself can manually choose the sequence of the characters in the text output (see **Figure 1**). Hence, rather than approaching the problem of highly irregular layouts technologically, KuroNet gives the user maximum flexibility when creating the text output. Kanbun, too, is a hazard to line segmentation, albeit due to quirks within an otherwise relatively regular outline: double lines (known as *warigaki* 割書 or *warichū* 割注, see **Figure 2**) within and glosses between the columns. What the experiment means to test, therefore, is whether the solution designed for specifically *kuzushiji* also responds well to the context of *kanbun*, and whether KuroNet's solution of flexibly choosing the order of the text output provides a relative advantage vis-à-vis alternative solutions.

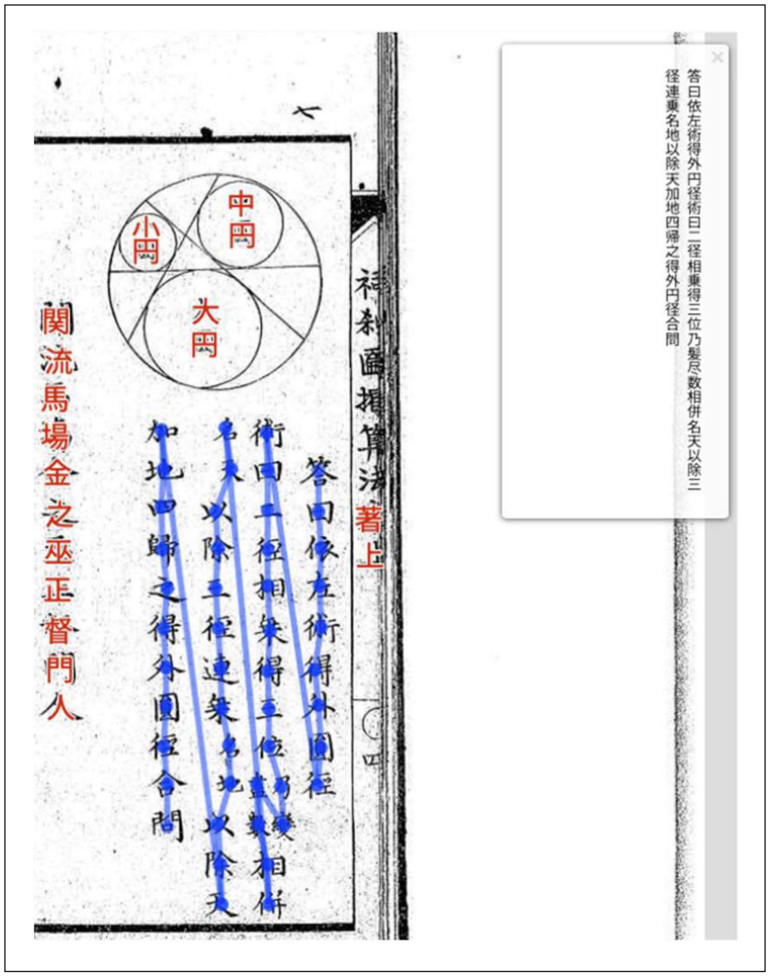

Figure 1: KuroNet's Reading Order tool being used to create a correctly ordered textual output.

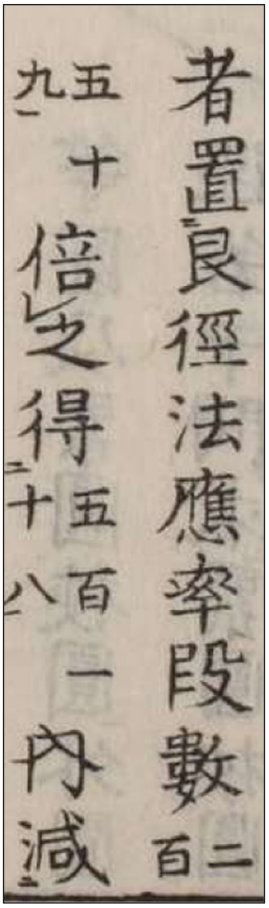

Figure 2: An example of double lines (*warigaki*) in *Shinpeki sanpō* 神壁算法, Vol. 1 (1789), *Koma* 12.

Apart from the layout problems, the wide variety of characters and their cursive forms present serious challenges for the recognition process for *kuzushiji* texts. The 2017 printing of the industry-standard *kuzushiji* dictionary, Kodama Kōta's *Kuzushiji yōrei jiten* (くずし字用例辞典), contains 6,406 characters, whilst the *Nihon kotenseki kuzushiji dēta setto* (日本古典籍くずし字データセット, henceforth *kuzushiji* dataset), that is, the dataset on which KuroNet's systems have been trained, contains 4,328 characters (CODH 2025a; Hashimoto 2025). Training neural networks for *kuzushiji* text recognition therefore already required a trade-off whereby the number of characters included in the training set is weighed off against the accuracy for recognizing each single one. The large number of characters—and particularly the long tail of rare characters—poses a problem in *kanbun* recognition, too. The question therefore is whether this trade-off made in the name of the *kuzushiji* task renders the dataset incommensurate with the classic *kanbun* set, or whether KuroNet is equipped to cover the range of *kanji* appearing in typical *kanbun* texts.

Additionally, *kuzushiji* texts frequently feature interlinear glosses as a guide for readers to pronounce the main text. KuroNet's authors specifically chose to forego these glosses and evaluate them as a false positive, whether recognized correctly or not (Lamb, Clanuwat, and Kitamoto 2020). Texts written or printed in *kanbun* for a Japanese audience rely on interlinear diacritics to aid the re-ordering of the text according to Japanese syntax. The question here is, therefore, whether KuroNet's choice to forego these characters is beneficial or a hindrance in the case of *kanbun* recognition.

The biggest question from the outset, however, was how KuroNet, trained to recognize the cursive and connected characters of *kuzushiji*, would deal with the variety of fonts used in printed and handwritten *kanbun* publications: on occasion characters appear in a more cursive format, but often enough, their appearance is spaced and regular. Moreover, KuroNet takes context into account for the recognition of single characters, albeit expecting texts written in Japanese syntax that combine *kanji* with *hiragana* (Clanuwat, Lamb, and Kitamoto 2018). In *kanbun*, meanwhile, the text is mainly composed of classical Chinese characters and follows Chinese syntax. The question is therefore whether the variety of characters found in *kanbun* is sufficiently covered, and whether KuroNet's contextual handling of single characters is counterproductive in the case of *kanbun*.

Experiments conducted by a team of researchers from Yamagata University show promise but also challenge when using the *kuzushiji* dataset for a set of *wasan* 和算 documents, Japanese mathematics books typically written in *kanbun*. Testing a LeNet network trained on the *kuzushiji* dataset also used for KuroNet, albeit stripped of classes with less than 20 examples and augmented in an effort to re-balance the dataset, shows the effect of such an imbalance: whilst 1,506 *kanji* classes were classified at a 73.10% accuracy rate, the correct classification rate across all characters was only 37.22%. The neural network clearly struggled to correctly classify characters underrepresented in the dataset (Diez et al. 2021). Although the *kuzushiji* dataset clearly outperformed the dataset for modern Japanese characters in this test, it did not produce a satisfactory output, nor is it clear how the same dataset performs for *kanbun* documents outside the range of *wasan*. In the current experiment, we therefore expanded the range of documents tested in terms of subject matter and type font. The focus was on testing KuroNet on as broad a range of fonts and topics as possible to see how it would respond to different contexts and to give some perspective on the type of material for which KuroNet could be used, and the scope.

We initially selected 12 *koma* (コマ, page images of a text—in most cases, images of a double-page spread) from 11 different texts. These were chosen for the diversity of their typeface, layout, and inclusion and exclusion of diacritics and glosses. Six texts

were related to the topic of *wasan*, whilst the other six texts included a broad range of topics such as geography, health, religion, and travel. Two *koma* were selected from the same publication, *Sangaku kōchi* 算學鉤致 (1819), albeit from sections using significantly different typefaces. Since KuroNet can only process images published in IIIF format, the choice was somewhat moderated though not significantly hampered by the question of what was readily available. After reviewing the results, we analyzed additional *koma* from six of the texts.

After OCR was performed on the selected images with KuroNet, the total number of transcribed characters was counted manually, as well as the total number of erroneously transcribed characters, the total number of extra characters that were transcribed (i.e., when a correctly transcribed character had an extra character inserted before or after it), and the total number of missing characters. This was cross-checked and corrected by another scholar. We then calculated a character error rate (CER)—the percentage of erroneously transcribed characters (excluding missing and extra characters) in a transcription. Glosses were excluded from the analysis, since the developers of KuroNet have designed the platform to ignore glosses and other annotations (Lamb, Clanuwat, and Kitamoto 2020), even though in some cases parts of glosses and diacritics were transcribed. We also chose to ignore text within illustrations or figures. KuroNet transcribes *kyūjitai* 旧字体 (historical character forms) as *shinjitai* 新字体 (simplified character forms), and although some people may make different transcription choices, we did not consider this to be erroneous, since it reflects the proper functioning of the application.

2.2 Other platforms

When we set out to conduct our experiments in early 2024, the number of accessible OCR platforms—open-access platforms that can be used without programming expertise—potentially applicable to *kanbun* texts was quite limited; however, as this paper underwent review, the field shifted. The National Diet Library's OCR platform, NDL *kotenseki* OCR (NDL古典籍OCR), which was previously available in various iterations through GitHub (most recently since November 2024 as NDL *kotenseki* OCR-Lite) (NDL Lab 2025), became available to a mass audience in May 2025 through the creation of a dedicated web platform—NDL *kotenseki* OCR-Lite Webban (NDL古典籍OCR-Lite Web 版) by Yuta Hashimoto (Hashimoto 2025). Additionally, generative AI has opened up new possibilities for extracting text from images. For the purpose of this experiment, we revisited KuroNet in the context of two other options to understand how the quality of the output fares in comparison. The chosen options were NDL *kotenseki* OCR-Lite

(henceforth NDL OCR) and Google's Gemini, comparing the results by error rate, missed characters, and extra characters transcribed. Whilst the results obtained from Google's Gemini Pro 2.5 and Flash 2.5 were unusable for most of the trial pages, NDL OCR performed very well, as was to be expected based on the National Diet Library's own research into the accuracy of its model (NDL Lab 2022). The results of NDL OCR are directly discussed in the paper, whilst the results from Google's Gemini were not of a comparable standard and cannot be replicated. They are therefore discussed in the appendix. Since the focus of our investigation was to identify a successful "off-the-shelf" solution with an easy-to-use interface, some potential alternatives like PaddleOCR (PaddleOCR 2025) were not considered here.

2.3 Results

The quality of the output differed widely across the different texts. Amongst our samples for OCR performed by KuroNet, 4 texts produced CERs of less than 10%, 4 produced CERs between 10% and 15%, and the remaining 4 produced CERs of over 20%. Our general assumption is that texts transcribed at 15% CER or less provide a useful output. As Sangiacomo and colleagues report, a corpus of transcriptions with 85–90% word accuracy can be used in text-mining projects; a word accuracy rate of 90% is "more than sufficient for performative semantic analysis" (Sangiacomo et al. 2022). Hence, whilst an editor might not consider a text of 15% CER useful, such texts, if procured at scale, can still be of analytical value. We furthermore assume that output with a CER of less than 5% can be manually corrected with an acceptable degree of effort, whilst an error rate of 15% or more is not considered immediately useful except for potentially identifying areas for systematic improvement of the OCR. Characters missed or inserted were recorded separately from the character error rate. Missed characters are measured by an omitted character rate, which is the percentage of omitted characters out of the total number characters. Extra characters are recorded as a simple number, but do not include additional characters that represent mistakenly transcribed diacritics or glosses. The results are displayed in **Table 1**, ordered by CER (when processed with KuroNet) in ascending order. The number of erroneous or missed characters is recorded in parentheses alongside the CER and omitted characters. We include some sample images in the appendix.

As can be observed in the table, NDL OCR performed better in terms of CER across all texts: 6 texts produced CERs of less than 5%; 5 texts produced CERs between 5% and 10%; and 1 text produced a CER of over 20%. This tracks well with the National Diet Library's own tests of its OCR models, which found an average F-score of 0.92 across

Text	*KuroNet* CER (No. of Errors)	*NDL OCR* CER	*KuroNet* Omitted Characters	*NDL OCR* Omitted	*KuroNet* Extra	*NDL OCR* Extra
Zoku shinpeki sanpō 続神壁算法 (1807), Koma 11	7.26% (22)	4.28% (13)	1 (0.33%)	2 (0.66%)	0	3
Sanpō kishō 算法奇賞 (1830), Koma 5	7.87% (21)	4.06% (11)	2 (0.74%)	0	0	52
Byōka yōron 病家要論, Vol. 1 (1695?), Koma 6	8.33% (11)	6.06% (8)	0	4 (3.03%)	0	4
Shinpeki sanpō 神壁算法, Vol. 1 (1789), Koma 12	8.38% (14)	6.63% (11)	2 (1.18%)	4 (2.41%)	0	7
Shusetsu yōjō ron 酒説養生論, Vol. 7 (1729), Koma 9	10.26% (16)	3.75% (6)	0	6 (3.75%)	0	12
Ha daiusu 破提宇子 (1620), Koma 3	11.95% (19)	4.4% (7)	1 (0.63%)	1 (0.63%)	0	3
Sangaku kōchi 算學鉤致, Vol. 1 (1819), Koma 10	12.02% (22)	5.4% (10)	1 (0.54%)	0	0	1
Nantōshi ro 南島志 呂 (1856), Koma 4	12.87% (48)	1.34% (5)	1 (0.27%)	1 (0.27%)	0	1
Sangaku kōchi 算學鉤致, Vol. 1 (1819), Koma 3	15.12% (13)	9.3% (8)	1 (1.16%)	1 (1.16%)	0	0
Chūzan denshinroku 中山伝信録, Vol. 1 (1721), Koma 16	27.29% (95)	0.83% (3)	13 (3.61%)	1 (0.28%)	4	0
Owari meisho zue 尾張名所図会, Vol. 1 (1844), Koma 10	32.31% (21)	7.69% (5)	0	0	2	0
Sanpō koren zen 算法瑚璉全 (1836), Koma 2	37.77% (51)	23.61% (34)	9 (6.25%)	1 (0.69%)	0	0

Table 1: Results.

the 3,028 images that were used to test the third iteration of the model (NDL Lab 2022). The efficacy of the OCR model is truly impressive. Looking at the results, NDL OCR is clearly the superior choice for documents written in a clear and consistent format, where the text is written in single columns. This is the case for *Sangaku kōchi* 算學鉤致 (1819), *Nantōshi* 南島志 (1856), *Chūzan denshinroku* 中山伝信録 (1721), *Owari meisho zue* 尾張名所図会 (1844), and *Sanpō koren* 算法瑚璉 (1836)—texts that KuroNet performed less well on.

In cases where the layout presents challenges, such as the presence of double columns within lines, the output from NDL OCR included a significant amount of noise that would obstruct a semantic analysis, text mining, and text editing. Briefly said, in these cases, the benefit of improved accuracy is outweighed by the issues caused by layout recognition (for example, a line being segmented and transcribed twice, or non-text being mistaken for characters). As will be discussed in the following, KuroNet does not suffer from the same problem, or at least not to the same degree. Whilst KuroNet's OCR model does warrant improvement, the layout handling commends it as a choice for *kanbun* texts—particularly when features such as *warigaki* are present. After a comparison between the results achieved by KuroNet and NDL OCR, we will focus on analyzing character classification issues with KuroNet to identify what kind of improvements could be made.

The counting system we employed is built on the specifications of KuroNet, a fact that might not be reflected immediately in the data. For example, KuroNet deliberately ignores diacritics; we therefore have not included diacritics in our analysis, although they are regularly transcribed by NDL OCR. Conversely, KuroNet ignores the colour and size charts used in scanning, as well as the inter-page titles that appear on the edge of texts; NDL OCR attempts to transcribe these pieces of information. We have ignored these parts of NDL OCR's transcriptions, but if they were included, they would, in some cases, significantly change the results. For example, *Zoku shinpeki sanpō* 続神壁算法 (1807) includes an additional 53 characters produced from transcribing the colour chart, whereas *Sanpō kishō* 算法奇賞 (1830) includes an additional 59 characters not included in our count, presumably caused by discoloration or noise on the page. Granted we already counted 52 additional characters besides, this seems exorbitant; it may be worth noting here that *Sanpō kishō* is the only pre-binarized scan in the experiment, with considerably more noise that might have caused these problems.

Two other issues seem to exist with NDL OCR. Firstly, historical and contemporary forms of *kanji* are used inconsistently. Whilst KuroNet transcribes all characters within their modern (simplified) forms, NDL OCR mixes historical and modern forms,

sometimes even within a single sentence. For example, in its transcription of *Sanpō koren*, NDL OCR included the following output:

無不由之也實為經疫治国也

A mix of character forms would be desirable if it reflected the original source, where such a variety may indeed be present; however, in this case, NDL OCR's output does not always reflect the actual forms in the text. If we were to transcribe the sample sentence (overlooking the mistake in NDL OCR's transcription) using modern character forms, 實 and 經 would be written as 実 and 経; if we were to adopt an approach using historical forms as they appear in the source, 為 and 国 would be written as 爲 and 國. As to the rest of the characters, the shape would remain the same throughout either transcription. Thus, a transcription consistent with the text itself would read as 無不由之也實爲經疫治國也, and one using modern character forms would be 無不由之也実為経疫治国也. This issue adds to processing time: either all characters have to be checked to ensure they are consistent with the transcription policy, or the entire output has to be changed into modern forms.

A considerably larger issue, particularly with regard to the mathematical texts in our sample, relates to layout analysis, and word and sentence order. Because NDL OCR uses line segmentation, it sometimes produces ordering errors. This is particularly noticeable with *warigaki* (double columns within lines) especially when there are multiple sections of *warigaki* in a single line or the *warigaki* crosses into a new line. For instance in *Shinpeki sanpō* (see **Figure 2**), a human eye would read the first number written in warigaki as 259 (二百五十九) split across two lines. NDL gives us 2 (二), a character error M, 50 (五十), a short line of text, part of the next double column—number 501 (五百一)—and then returns to the original double column to give the left-hand side—the number 9 (九)—before continuing with the sentence. In other words, whereas we would like to see the number 259, we actually receive an output of four numbers: 2, 50, 501, and 9. A similar ordering issue is seen with the next part of *warigaki* within the same sentence. **Table 2** is intended to make this easier to visualize. The table shows the expected output (the output in the correct order) and NDL OCR's output. The output is split into portions (lines) as identified by NDL OCR; in reality, this is a single line of text with two characters from the preceding line included at the beginning. We have removed diacritics and highlighted the output corresponding to the first part of *warigaki* in red and the output corresponding to the second section of *warigaki* in blue.

Portions of Text	Expected output	*NDL OCR*'s Output
1	二百	二M
2	五十九	五十
3	倍之得	倍之得
4	五百一	五百一
5	十八	九
6	内減	内減
7		十八

Table 2: Warigaki in Shinpeki sanpō.

When the text has a regularized line-based format, as is the case with most of the chosen texts, NDL OCR performs well with the output appearing in the correct order. For texts with a large amount of *warigaki* or irregular lines, however, the above issues emerge. The transcription for *Zoku shinpeki sanpō* contained 14 instances of incorrect or interrupted ordering; *Shinpeki sanpō* contained 7 instances; and *Sanpō kishō* contained 10 instances. These ordering issues are created by the inclusion of text from illustrations, the presence of extra characters, and the above-noted difficulties that materialize when *warigaki* is present. In addition, parts of NDL OCR's output such as excess characters may not be immediately identifiable in the output. Thus, a substantial amount of reordering is needed to get to the correct output. KuroNet's aforenoted Reading Order tool, which allows users to choose the sentence order themselves, means that issues such as this are rendered irrelevant. The editor can simply select the correct sequence themselves.

In other words, whilst a comparison of NDL OCR and KuroNet from the point of view of character error rate shows that NDL OCR has a consistently higher accuracy, the different way of layout handling and the free choice of output using the Reading Order tool means that KuroNet can be used to produce a cleaner output overall. In fact, editing the output from NDL OCR requires not only the correction of incorrect characters and the insertion of missing characters, but the removal of material that is not in the original image, as well as the reordering of the text. In other words, the true potential of KuroNet lies with its approach to character detection and its response to flexible layouts. This is a potentially more egregious problem for other sorts of pre-modern Japanese texts (such as *kusazōshi* 草双紙; i.e., popular illustrated literature), yet

for the particular situation presented by *kanbun*, the KuroNet approach actually turns out to be a convenient and time-saving choice for the text editor.

Turning to KuroNet, the results show that the platform is well capable of dealing with one of the major challenges of *kanbun* text recognition: *warigaki*. Although not every single character in every double column in all the test pages was correctly recognized, the detection of these characters generally did not pose a problem. KuroNet's irregularized training set, which makes the system unusually robust for characters of different size, together with its character identification system, is presumably the crucial factor in this process: from all aspects, it can parse the irregular layout and the small size of the characters in the double columns without a problem. For example, the *Zoku shinpeki sanpō* contains 3 sections of double-column text with a total of 28 characters. Within this double-column text, KuroNet omitted only 1 character (一) and made 3 erroneous transcriptions (寸 to 北, 釐 to 煮, 釐 to 蚕). Two of these errors concerned the same character (釐), which does not appear in the dataset. *Shinpeki sanpō*, Vol. 1 provides another good example of KuroNet's successful transcription of double columns. The text features 6 sections of double-column writing consisting of 5 characters each. Of these 30 characters, 28 were transcribed correctly and only 2 (an 一 and a 十) were omitted. Both *Zoku shinpeki sanpō* and *Shinpeki sanpō* primarily feature numbers within their double columns, which makes it tempting to suggest that the simplicity of many of these characters contributed to KuroNet's successful transcription. But even in cases where more complicated characters were featured, there was a high level of success. In *Sanpō kishō*, there were 3 sections of double-column text with a total of 16 characters, all of which were transcribed with no errors. Overall, whilst we record occasional lapses due to gaps in the dataset and the use of unusual fonts (which we discuss below), some of the layout problems associated with *kanbun* are very gracefully handled.

Five of the texts showed promising results with CERs of under or, in one case, very close to 10%. We decided to test additional *koma* from these texts in order to determine whether our results from a single *koma* would be consistent across a larger number of images. We also decided to conduct additional tests on *Nantōshi* because of its typeface, although it performed less well than the other texts chosen for this second round. We found that KuroNet was able to perform at very high accuracies for some of the texts, with less than 5% CER for 6 images and 5–10% CER for 10 images. The results are displayed in **Table 3**; some images of the results are included in the appendix.

Text	CER (No. of Errors)	MCR (No. Missed)	Extra
Zoku shinpeki sanpō, 12	4.28% (13)	0	0
Zoku shinpeki sanpō, 13	4.75% (15)	0	0
Zoku shinpeki sanpō, 14	6.74% (19)	2 (0.7%)	0
Zoku shinpeki sanpō, 15	7.57% (24)	1 (0.31%)	0
Zoku shinpeki sanpō, 16	7.43% (26)	0	0
Sanpō kishō, 6	1.23% (3)	0	0
Sanpō kishō, 7	6.72% (18)	1 (0.37%)	0
Sanpō kishō, 8	2.01% (4)	0	0
Sanpō kishō, 9	1.52% (3)	0	0
Sanpō kishō, 10	4.87% (11)	1 (0.44%)	0
Byōka yōron, 5	11.66% (7)	0	2
Byōka yōron, 7	12.21% (16)	1 (0.76%)	0
Byōka yōron, 8	11.9% (15)	0	0
Shinpeki sanpō, 13	8.47% (25)	5 (1.66%)	0
Shinpeki sanpō, 14	6.66% (22)	0	0
Shinpeki sanpō, 15	6.27% (18)	0	0
Shinpeki sanpō, 16	8.33% (21)	0	0
Shinpeki sanpō, 17	9.09% (19)	0	0
Shusetsu yōjō ron, 8	11.27% (8)	0	0
Shusetsu yōjō ron, 10	12.82% (20)	0	2
Shusetsu yōjō ron, 11	14.29% (8)	0	0
Nantōshi ro, 5	9.28% (35)	0	0
Nantōshi ro, 9	13.2% (61)	9 (1.91%)	0
Nantōshi ro, 10	16.35% (68)	5 (1.19%)	0
Nantōshi ro, 11	16.66% (59)	4 (1.12%)	0

Table 3: Additional results.

These additional results further corroborate that KuroNet can be used to transcribe works like *Zoku shinpeki sanpō*, *Sanpō kishō*, and *Shinpeki sanpō* with relatively high accuracy. In fact, according to our aforenoted criteria, all of these texts may prove useful for text-mining projects. Taking all transcriptions from the first and second trials into account, we see a total character error rate of 6.36% (199 errors) for *Zoku shinpeki sanpō*;

4.29% (60) for *Sanpō kishō*; 7.73% (119) for *Shinpeki sanpō*; 10.9% (49) for *Byōka yōron* 病家要論 (1695); 11.85% (52) for *Shusetsu yōjō ron* 酒説養生論 (1729); and 13.67% (271) for *Nantōshi*; or a 9.77% (750) CER for all these texts. It is interesting to note that the highest performing texts are all related to mathematics, whilst those on other topics did not perform as well. It seems likely that the accuracy of the transcriptions for *Zoku shinpeki sanpō*, *Sanpō kishō*, and *Shinpeki sanpō* is related to their typefaces and layout, but possibly also a limited and formulaic set of characters. Although the typefaces for each text are different, they are clearly written often in well-spaced, printed (i.e., non-cursive) characters. This can be seen in the following renderings (**Figure 3**) of the phrase 今有如図, found in the 3 texts from right to left (*Zoku shinpeki sanpō*, *Shinpeki sanpō*, *Sanpō kishō*). In the case of other texts, such as *Nantōshi*, lower accuracy rates might be related to the prevalence of variant characters (*itaiji* 異体字), which may not feature in KuroNet's dataset, amongst other factors.

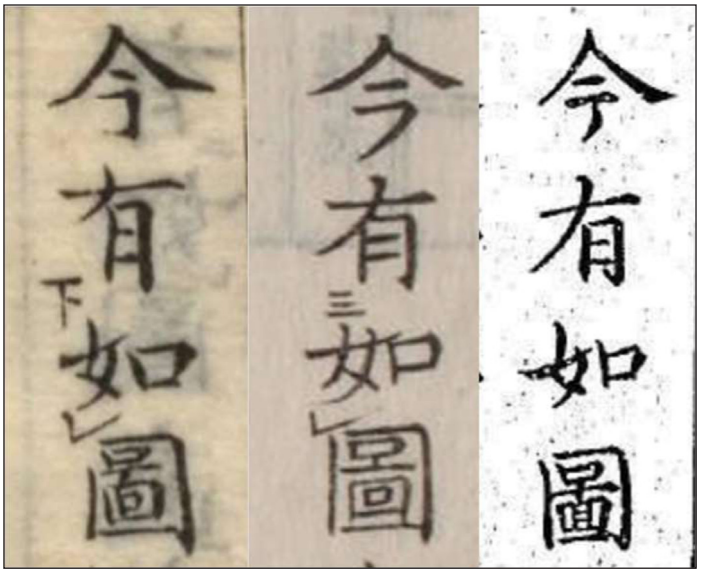

Figure 3: Sample typefaces.

For the low-performing texts, the separate metrics warranted some interpretation to gauge what the problem might be. With regard to omissions, characters can be skipped for several reasons. Firstly, the character might not have been detected in the first place. Secondly, the character has been detected but could not be classified by KuroNet and was consequently skipped (see discussion in Lamb, Clanuwat, and Kitamoto 2020). If a character is not classified correctly, two issues can be the reason: either the character is not present in the dataset, or the character is present in the dataset, but its shape (e.g., font, cursivity) is not well represented. In the following, we offer some interpretations for the materials tested.

Sanpō koren has the highest rate of omitted characters. Looking at the sample *koma* (**Figure 4**), the layout is regular, and the characters are well spaced. It is therefore highly unlikely that characters have not been detected, but rather that they were skipped because they could not be classified. All of the 9 missing characters (可, 下, 賦, 治, 國 [国 in its *shinjitai* form], 所, 心, 爲 [為 in its *shinjitai* form], and 名) feature in KuroNet's dataset. Furthermore, all are *jōyō kanji* 常用漢字 (regular-use *kanji*), or in other words, are amongst the characters that people within the modern Japanese schooling system must learn to read—meaning, they are not particularly rare or unusual. We suspect that both the high CER (37.77%) and the number of omitted characters are linked to the text's unusually broad typeface.

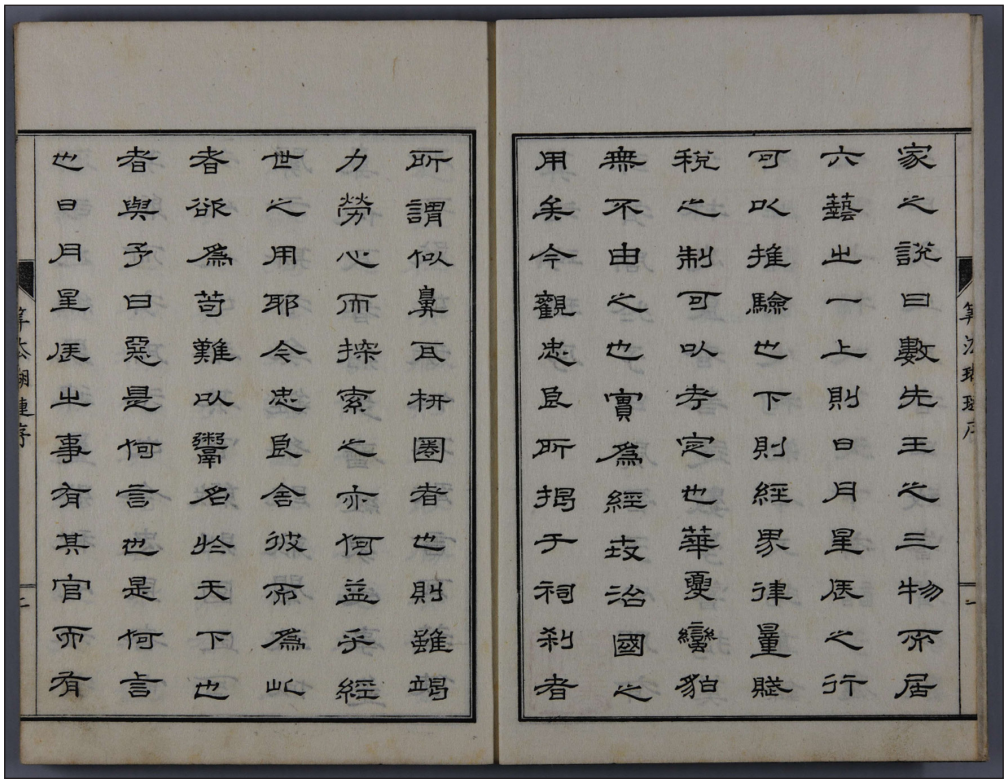

Figure 4: Sample of the typeface used in *Sanpō koren*.

In the case of *Chūzan denshinroku*, 13 characters were omitted; however, 6 of these are close to the binding and warped in the scan. In 95 instances, a character was mistranscribed. Some of these instances are made up of the same character mistranscribed repeatedly, even though it is included in the dataset. For example, the character 兩 was mistranscribed 3 times (twice as 雨 and once as 間), although correctly transcribed (as the *shinjitai* 両) once. Similarly, 號 (the historical version of 号) was incorrectly transcribed 3 times as 焼, 驪, and 驍. In these cases, it seems

that KuroNet struggled with recognizing the characters' historical forms. In other cases, however, typeface appears to have played a role, for example, when the form of the character did not radically change over time. The character 船, for example, is mistranscribed in 7 out of 8 instances, whilst 舟 is mistranscribed in 4 out of 6 instances.

Conversely, in some cases, characters are mistranscribed because they are not included in the dataset. In *Chūzan denshinroku*, the character 艙, which features 10 times, is not amongst the dataset and hence consistently mistranscribed. This also puts the error rate for this sample in perspective: 10.53% of all erroneous transcriptions are made up of one single character excluded from the dataset. In this particular sample, the text in double columns was well detected, but not always successfully classified. Of 33 characters across two double-column sections, 2 were omitted (臣, which is in the dataset, and 兢, which is not). In addition, 9 characters were erroneous (按 to 横, 宋 to 米, 舟 to 身, 二 to 一, 皆 to 昔, 字 to 宇, 舟 to 月, 八to の, 弁 to 舟), although each of these feature in the dataset. Finally, 1 extra character was added where the character 二 was transcribed twice as both 二 and 一 and thus overlapping in the output image. The same error features an additional 3 times elsewhere in this text. One factor contributing to the issues above may have been the lower quality of the scan available for this text (see **Figure 5**).

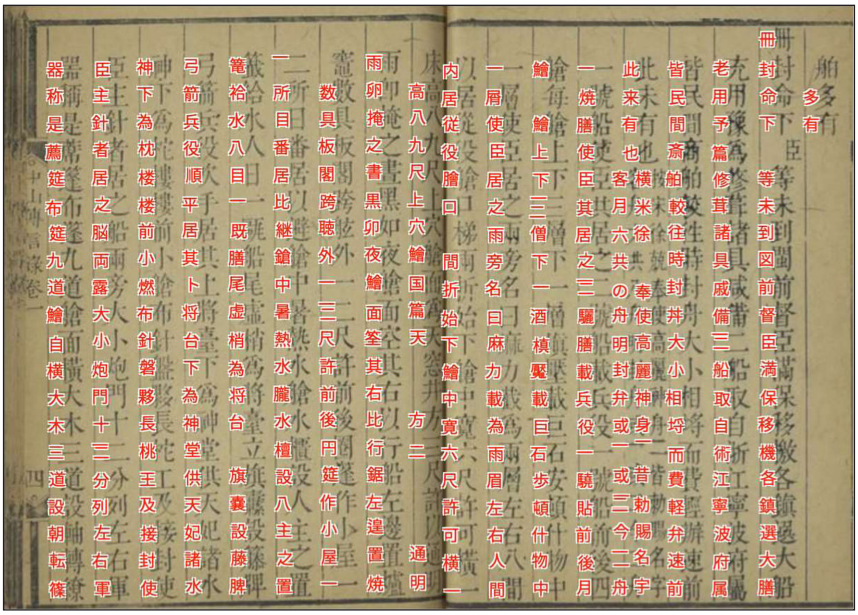

Figure 5: KuroNet's transcription of *Chūzan denshinroku*.

Issues related to historical forms can also be observed in *Zoku shinpeki sanpō*. For instance, the *kyūjitai* 徑 (the historical form of 径) is correctly transcribed 3 times, but is mistranscribed as 経 (historical form 經) on 3 occasions. In both correct and incorrect

transcriptions, the right-hand part of the *kanji* has been correctly recognized; however, where mistranscribed, the left-hand radical of the *kanji* is incorrectly recognized as 糸 rather than 彳. Looking at the cursive forms for these radicals, we can observe that some variants look very similar. The *kuzushiji* dataset features almost exclusively such cursive forms for these two characters (CODH 2025b; CODH 2025c), whilst less cursive varieties that are easily distinguishable are close to nonexistent. For example, in the case of 径, only a single less cursive variety is included. Furthermore, there are only 6 images in the dataset for 径 but 123 for 経 (CODH 2025b; CODH 2025c). This imbalance may be a compounding factor for the bias towards 経—and confirm the observations made in the paper by the Yamagata University researchers mentioned above. This seems plausible as an explanation of why KuroNet might confuse these components (see **Figure 6**) even though the characters can be clearly distinguished here (in their printed forms) by a human reader. The same character is mistranscribed in the same way in other texts such as *Shinpeki sanpō*.

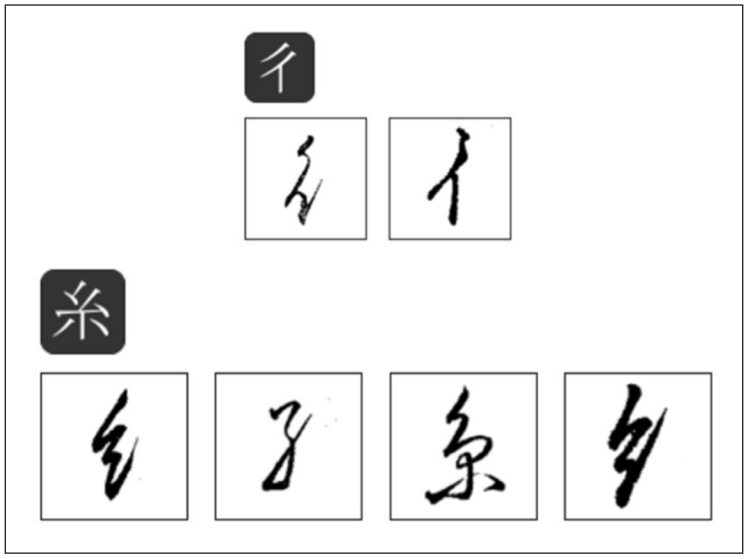

Figure 6: Cursive radicals (Kaji 2008).

In some cases, the dataset's bias towards cursive characters also seems to have an effect on recognition. For example, the character 乗 (here in variant form as 乘, which frequently appears in *wasan* texts) is consistently misrecognized despite its being printed and therefore legible for human readers. The dataset includes 144 images for this character, including samples rendered in this variant form (CODH 2025d), but most of these do not resemble the printed forms found in our texts. In other instances, KuroNet inserts characters that do not usually feature in *kanbun*. In the *Sangaku kōchi*, *koma* 3, we can observe issues with the transcription of the common character 之, including its

transcription as 2 characters (こゝ). Similar issues appear in *Owari meisho zue*, where the character is transcribed as both し (this *kana* utilizes 之 as its base character or *jibō* 字母) and ミ. However, this sort of error is not very common and was only observed in these instances.

As noted, KuroNet is trained to ignore glosses. Presumably due to the same factors that denote its success with double columns, *kanbun* diacritics were correctly recognized in some of the test samples, although not consistently in any single one. For example, in *Shusetsu yōjō ron*, 30 such characters were transcribed; in *Byōka yōron*, there were 15; in *Ha daiusu*, 破提宇子 (1620) there were 11 (see **Figure 7**); and in *Owari meisho zue*, there were 2. Whilst it is not possible to make a clear judgment call in this regard, it seems that with the availability of more training data with labelled glosses, KuroNet would be well capable of dealing with such diacritics gracefully. As the authors themselves note, the decision to exclude annotations from the current system bowed to economical more than technical factors (Clanuwat, Lamb, and Kitamoto 2019). OCR-processing and proofreading *kanbun* documents with the help of KuroNet might pave a way towards procuring such labelled training material in a more cost-effective way. NDL OCR does transcribe diacritics, although we have not explored the accuracy of the transcription. The diacritics are displayed in the *koji* (Hashimoto and Miyagawa 2025) mark-up language, which may aid analysis by making diacritics easily identifiable to a computer, but may require some processing, depending on the use case.

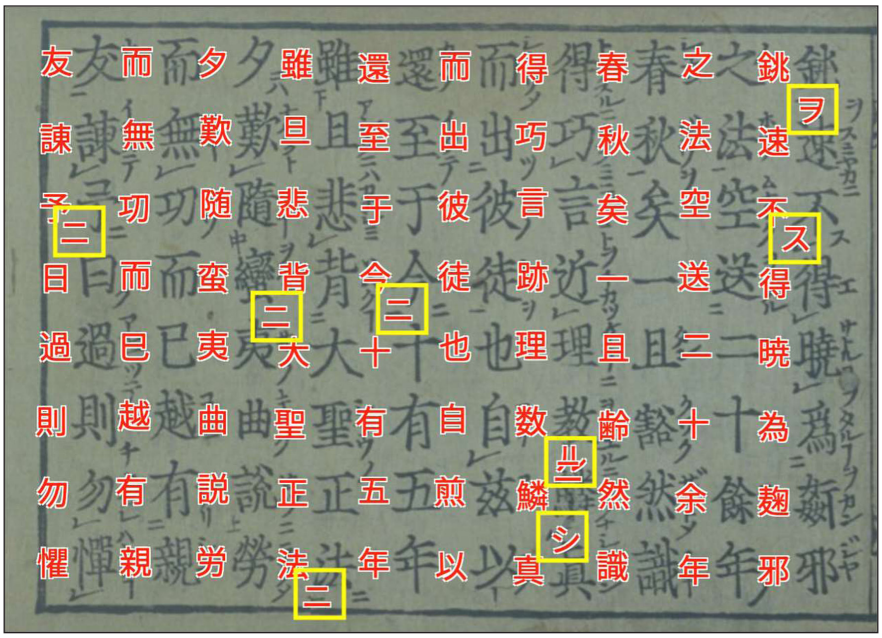

Figure 7: Diacritics (in yellow) recognized in part of the KuroNet output for *Ha daiusu*.

Characters within illustrations were not included in the analysis since KuroNet is specifically trained to omit them (Clanuwat, Lamb, and Kitamoto 2019). Still, some of them were identified by KuroNet in the test pages regardless and also transcribed correctly on occasion. In *Sanpō kishō*, 13 of 26 characters included in the figure were missing. In *Zoku shinpeki sanpō*, 12 of 18 characters were missing in one of the figures, whilst 6 of 8 characters were missing for the second figure with only one mistranscription. Finally, whilst all the characters in the figure in *Shinpeki sanpō* were transcribed, there was a CER of 24.52% (38), as well as an extra 9 characters featuring in the output. The *koma* from the *Shinpeki sanpo* featuring images of *sangi* (calculation rods typically depicted as vertical and horizontal lines) had KuroNet interpreting a few of the *sangi* images as characters, presumably thanks to their regular format (see **Figure 8**). NDL OCR is trained to transcribe annotations within illustrations. Whether it is beneficial to include these characters within a final transcript depends on the use case; for example, it would not help to include them in a text-mining corpus or in a transcript used for a text edition. Within the scope of this article, we are therefore not considering the transcription of these annotations in our analysis.

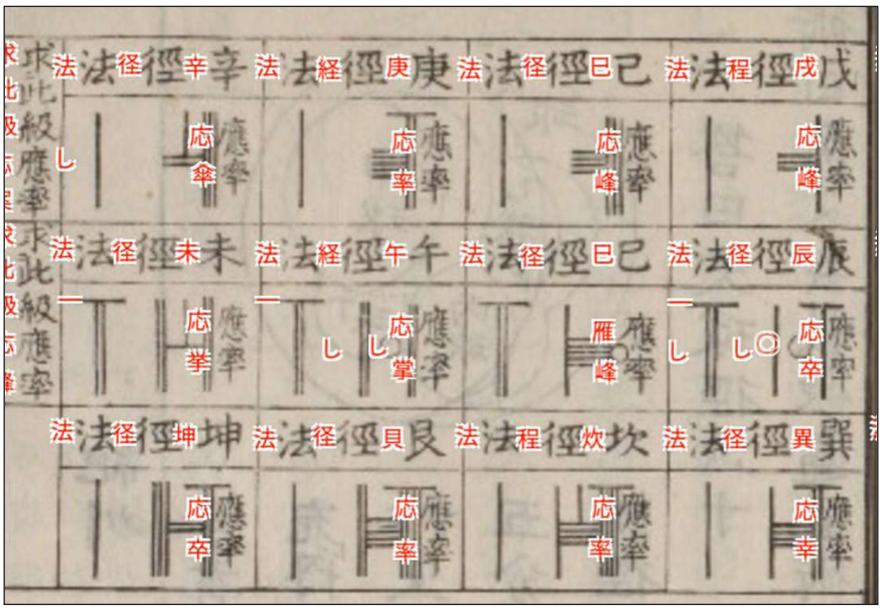

Figure 8: Transcription of an illustration from *Shinpeki sanpō*, parts of which are incorrectly recognized as characters.

2.4 Discussion

Overall, the accuracy of KuroNet on the *kanbun* samples is promising: 4 texts had CERs below 10%, and 8 had CERs below 15%. Additional trials produced 6 samples with CERs

below 5%, 16 below 10%, and an overall average CER (for all texts upon which more than one trial was performed) of 9.77%. The texts that performed the best were thematically mathematical, and were written in consistent and printed script, but interestingly also had layouts that included elements that platforms using line segmentation struggle with such as *warigaki*. Given the accuracy of the output, it is clear that KuroNet could be useful for creating data for those interested in text mining or simply doing keyword searches depending on the chosen texts. An error rate below 10%, moreover, is a significant step towards creating ground truth at lower cost. The model performs particularly well at *kanji* detection in the main text, even where *warigaki* are present (excluding illustrations and glosses); very few are omitted, and the majority conceivably due to a failure to classify rather than detect the character. As the authors of KuroNet themselves note, the open availability of the system also functions as an invitation for feedback. Using the system on a wider variety of sources could also help enlarge the shared data pool that could be made publicly available. In that sense, the above experiments are an exhortation more than anything to take advantage of an open-source, easy-to-use system that outperforms other available systems on a number of *kanbun* texts, even if the output is not yet perfect.

Particularly important is KuroNet's ability to deal with unusual layout features and its interface which allows users to decide the order of the output. These functions help to speed up text-editing processes. In comparison, NDL OCR provides a more accurate character recognition, but errors in the output, such as changes to reading order and the insertion of additional characters, particularly with mathematical and other *kanbun* texts, which make extensive use of double lines, make text editing an arduous task. By adopting an approach to layout that focuses on characters rather than lines, KuroNet gracefully deals with some of the layout features that make *kanbun* a challenge to OCR.

Given its general efficacy, there are still some caveats. KuroNet's output is in *shinjitai* (i.e., the modern forms of Japanese characters). There is ongoing discussion about best transcription practices grounded in debates about human and machine readability on the one hand, and fidelity to the original source on the other (Kawahira 2015). KuroNet is designed to make sources readable, not to produce diplomatic transcriptions. Nevertheless, its features will make it more useful to those in certain fields. For example, the use of *shinjitai* means that the raw output will likely be most useful to those who use *shinjitai* within their transcriptions. For those who favour fidelity to the original using the character forms as they appear in the text, the raw output requires some human or digital processing. The use of various conversion tools available online, such as Gaoshukai's *Shinjitai to kyūjitai sōgo henkan tsūru*

新字体と旧字体相互変換ツール (Gaoshukai 2020) is one possibility that would allow modern forms of *kanji* to be automatically, but indiscriminately, replaced with their historical counterparts or vice versa. KuroNet's use of modern character forms also limits some of the applications of the data. A scholar would not, for example, be able to compare the use of different character forms. Intimately related to this discussion is the fact that a *shinjitai* output is only useful for those working with the Japanese language. It would not be useful for Sinologists since Chinese has developed different modern character versions from Japanese.

Secondly, the fact that diacritics are transcribed sometimes, but not always, poses a conundrum. Whilst this demonstrates on one hand KuroNet's theoretical capacity for handling such characters, the appearance of merely a fraction of them creates an editing burden for the moment where the few are better deleted than supplemented with the missing diacritics. This brings us to the question of transcribing diacritics in general. As to the modern transcription of *kanbun* texts, this can take many shapes and forms. Often enough, the diacritics are left out, and the text merely transcribed as the Chinese characters. If anything, this signals the lack of an orthographic standard for *kanbun* across disciplines and use cases. Granted, some promising work has been done training language models to turn Chinese textual sources into *kanbun* together with diacritics (Wang, Shimizu, and Kawahara 2023), and NDL OCR appears to recognize these sorts of glosses well. Again, what a desirable output should actually look like might still need to be clarified before a functioning OCR system for *kanbun* can be tackled. Whilst diacritics are an important issue for human readers and for those wishing to create diplomatic transcriptions, for those interested in text mining, which we believe is the primary utility of text output from KuroNet, this needn't cause issues, since diacritics are unlikely to be included in a computerized analysis of the text.

3 Conclusion

KuroNet as a tool was designed to unlock Japanese writing that is extremely hard to decipher for the modern reader. The challenge of *kanbun* is related and yet different when it comes to the matter of legibility: even where the letters are clearly printed, it is not straightforward to read for modern Japanese readers, whether they are familiar with the *kanbun* reading aids or not. The content is fairly accessible to readers of classical Chinese, however, who can forego the transformation achieved by means of the reading aids and take the base text as their point of departure. Making *kanbun* literature accessible to a modern readership is therefore not a matter of untangling illegible characters—it is a matter of transforming a legible text into intelligible

content. OCR could crucially pave the road towards such a transformation, be that by language modelling, machine translation or else.

We hope that the above experiments serve to open a discussion beyond accuracy when it comes to the assessment of OCR software. In the digital humanities, it is often the case that parallel efforts are expended on the same problem. The result is duplication, a waste of resources and information silos. In the spirit of Horvath's call for collaboration, the above experiments mean to show that when it comes to *kanbun*, significant pieces of the OCR puzzle already exist, even though advertised under a different functionality. KuroNet is not designed for *kanbun*. But the solutions for problems within its own context also handle *kanbun*-related challenges well, with particular reference to irregularities of layout. In that sense, it could provide a good basis for a renewed effort to tackle *kanbun* OCR.

The main shortcoming of KuroNet may be the restricted cast of characters and their limited representation of non-cursive characters in the dataset. For some, the inability to use the platform to transcribe diacritics may also prove problematic. Yet granted KuroNet can be used to create transcriptions at an accuracy above 85–90% for many texts, it could significantly speed up the creation of ground truth and samples to feed back into the *kuzushiji* dataset. This could widen the scope to handle *kanbun* texts, as well as ancient Chinese manuscripts. Alternatively, looking at the excellent error rate of the NDL OCR model, the prospect of expanding the ground truth pool for KuroNet in the same way holds significant promise. In such a case, it would be wise to learn from the NDL OCR experience, making sure that the characters are transcribed consistently, be that historical or modern. Conversely, some of the editorial challenges posed by line segmentation errors in NDL OCR might be mitigated by KuroNet-style character detection and a Reading Order tool.

On a final note, part of the challenge of *kanbun* OCR is beyond the scope of technology (i.e., the development of orthographic standards). The development of a functioning OCR solution and an orthographic standard as a necessary prerequisite, however, may present a good impulse to formulate just that.

Appendix

Gemini and OCR

Our use of Gemini was motivated by the comments of the reviewers but was also partially inspired by a short article by Eric H. C. Chow exploring the use of Gemini for Chinese OCR (Chow 2025), which suggested that the platform holds some promise. Nevertheless, we found ourselves quite disappointed with Gemini's results.

We used the simple prompt "Please perform OCR on the following page." In some cases, such as when used with *Sanpō kishō*, one of the documents that was transcribed with a high level of accuracy with KuroNet, Gemini informed us:

> I am unable to perform an automated OCR (Optical Character Recognition) on the image you provided directly. However, I can describe the image and attempt to transcribe some of the visible text for you. [...] If you need a complete and accurate transcription of the text, you might consider using a dedicated OCR software or service that specializes in older Japanese documents.

For other texts, Gemini produced transcriptions, but it is difficult to assess the accuracy. Generally speaking, Gemini Flash 2.5 did not produce results that we could work with, so here we focus on the results produced by Gemini Pro 2.5.

In some cases, Gemini Pro seems quite accurate. For example, when Gemini Pro performed OCR on *koma* 12 of *Shinpeki sanpō*, there were only 6 character errors compared to KuroNet's 14. However, Gemini also introduced issues that were not present in KuroNet's transcription. Gemini missed 4 characters (KuroNet = 2), introduced 3 extra characters (KuroNet = 0) and reproduced parts of the text (usually numbers written in half-width double lines) in the wrong order, leading to 8 misordered characters (KuroNet = 0). In other words, although the character error rate was better than KuroNet, other issues led to the creation of an overall less accurate transcription.

Whilst Gemini performed comparably to KuroNet on some other texts—for example, for *koma* 6 of *Byōka yōron*, Gemini's transcription contained 16 erroneous characters (KuroNet = 11) and 1 extra (KuroNet = 0)—it generally did not perform well. In several cases, the errors became completely unquantifiable. Here we use an example from *Zoku shinpeki sanpō*, *koma* 11. For the first 9 lines of the text, Gemini's transcription contained 13 erroneous characters (KuroNet = 5) and 3 extra characters (KuroNet = 0), and missed 5 characters (KuroNet = 0). Then, in the tenth line, the platform transmogrified the text into something very different. Here is a human transcription of the tenth line matching the style of Gemini's output (i.e., without diacritics and with characters displayed in their *kyūjitai* character forms):

> 乗天冪及地寄左列天加地乗等圓徑得數自之與寄左

Between line 9 and the start of the transcription for line 10, Gemini's output contains a flurry of 16 random characters that don't seem to correspond to any part the text—though two of these characters (方面) do appear in the diagram on the page. This additional 16-character sentence appears as follows:

> 爲天元一乘方面平得數内自乘減去左

Following this we have Gemini's transcription for line 10. It contains 31 characters rather than the anticipated 22. We have underlined what we would consider to be extra characters and highlighted those that we consider to be character errors in bold. One can observe that part of the sentence repeats itself within the transcription:

> 乘天元冪及地寄左列天元冪及地寄左列天**元**地乘等圓徑得數自**乘減去**左

Although we do not intend to give a full analysis of this or similar errors, the large presence of extra characters (something we generally do not see when using KuroNet) likely points to the limitations of using generative AI for the purposes of OCR. The platform appears to be generating (or hallucinating) additional text based on what it "expects" within a sentence rather than the reality of what is on the page. Due to these issues, as well as the fact that these results cannot be reliably reproduced, we didn't believe that it was fruitful to continue our trials with Gemini.

Sample outputs from KuroNet

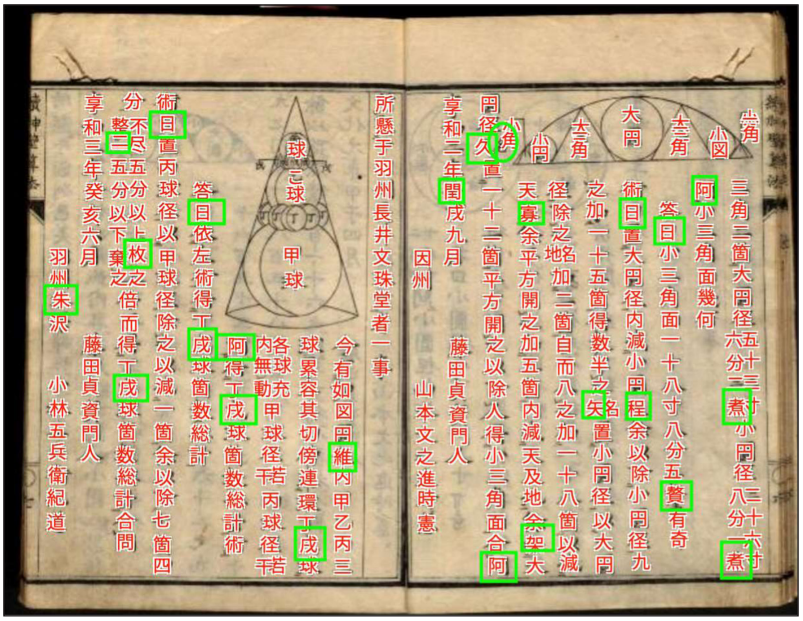

Figure A1: *Zoku shinpeki sanpo*, *Koma* 15 (character errors are highlighted in green).

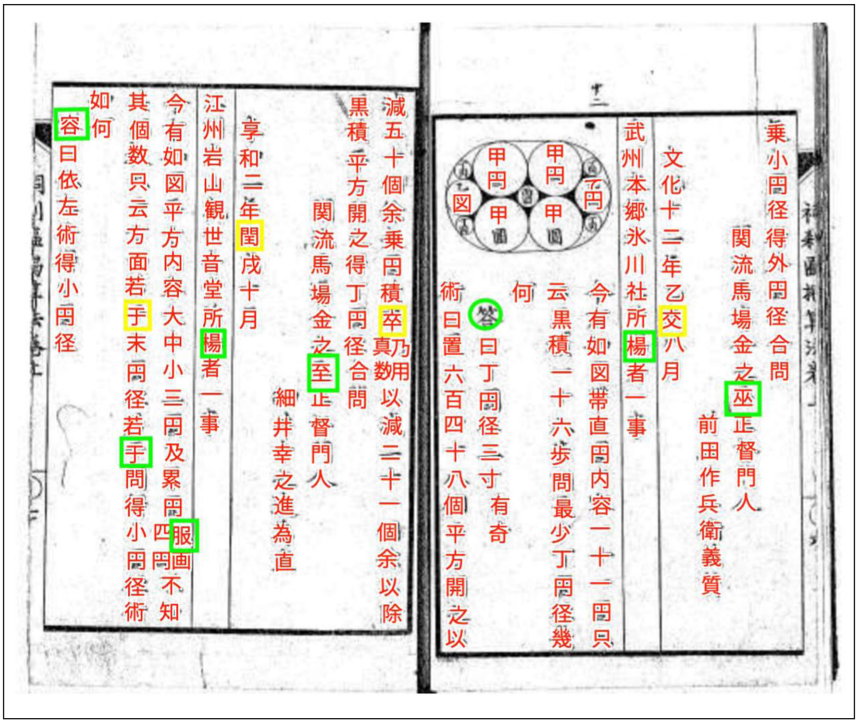

Figure A2: *Sanpō kishō*, *Koma* 10 (character errors are highlighted as green or yellow squares, missing characters as green circles).

Figure A3: *Byōka yōron*, *Koma* 8 (character errors are highlighted in green, transcribed diacritics in blue).

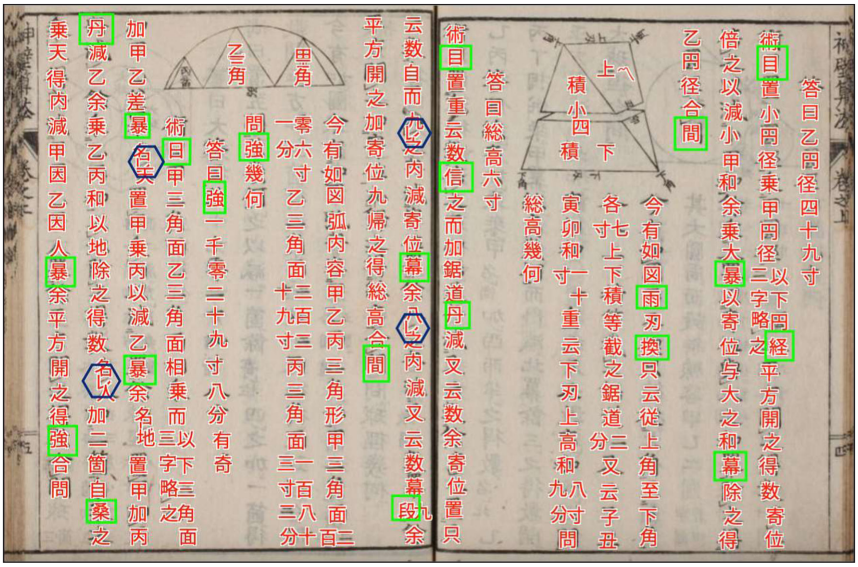

Figure A4: *Shinpeki sanpō*, *Koma* 14 (character errors are highlighted in green, transcribed diacritics in blue).

Acknowledgements

This study was funded by a KAKENHI Early Career Researchers Grant (Project number 24K15964, 2024–2026) and a DNP Foundation for Cultural Promotion of Graphic Culture Research Grant (Project: "*Kirishitan-ban* in the Digital Age: A Study of the Opportunities and Limitations of Applying Digital Methods to *Kirishitan-ban*"). We would also like to express our thanks to Joseph Bills and Yang Tianle for their help with analyzing OCR results.

Competing interests

The authors have no competing interests to declare.

Contributions

Authorial

We view ourselves as equal partners within the authorship of this paper. Authorship in the byline is by magnitude of contribution unless designated as "equal." Author contributions, described using the NISO (National Information Standards Organization) CrediT taxonomy, are as follows:

Author names and initials:

James Harry Morris (JM)
Antonia Karaisl (AK)

Authors are listed in descending order by significance of contribution. The corresponding author is JM.

Conceptualization: Equal
Data Curation: Equal

Formal Analysis: JM, AK
Funding Acquisition: AK, JM
Investigation: JM, AK
Methodology: Equal
Writing – Original Draft: Equal
Writing – Review & Editing: Equal

Editorial

Section Editor

Frank Onuh, The Journal Incubator, University of Lethbridge, Canada
Davide Pafumi, The Journal Incubator, University of Lethbridge, Canada

Copy and Layout Editor

A K M Iftekhar Khalid, The Journal Incubator, University of Lethbridge, Canada

Copy and Production Editor

Christa Avram, The Journal Incubator, University of Lethbridge, Canada

Bibliographical Data for Test Images

Baba Seitō (馬場正統). 1830. *Sanpō kishō* 算法奇賞 [also known as *Shisatsu henkei sanpō* 祠刹匾掲算法]. Collection of Tohoku University Library 東北大学附属図書館所蔵 accessed through National Book Database 国書データベース. Accessed August 20, 2025. https://kokusho.nijl.ac.jp/biblio/100234438/1?ln=ja.

Fujita Sadasuke (藤田貞資), and Fujita Yoshitoki (藤田嘉言). 1789. *Shinpeki sanpō* 神壁算法, Vol. 1. Kyoto University Rare Materials Digital Archive 京都大学貴重資料デジタルアーカイブ. Accessed August 20, 2025. https://rmda.kulib.kyoto-u.ac.jp/item/rb00028551.

Fujita Sadasuke (藤田貞資), and Fujita Toshitoki (藤田嘉言). 1807. *Zoku shinpeki sanpō* 続神壁算法. Collection of Hokkaido University Library 北海道大学附属図書館所蔵 accessed through National Book Database 国書データベース. Accessed August 20, 2025. https://kokusho.nijl.ac.jp/biblio/100345946/1?ln=ja.

Habian (ハビアン). 1620. *Ha daiusu* 破提宇子. Kyoto University Rare Materials Digital Archive 京都大学貴重資料デジタルアーカイブ. Accessed August 20, 2025. https://rmda.kulib.kyoto-u.ac.jp/item/rb00012881.

Ishiguro Nobuyoshi (石黒信由). 1819. *Sangaku kōchi* 算學鉤致, Vol. 1. Kyoto University Rare Materials Digital Archive 京都大学貴重資料デジタルアーカイブ. Accessed August 20, 2025. https://rmda.kulib.kyoto-u.ac.jp/item/rb00028475.

Itō Genjo (伊藤玄恕). ca. 1695. *Byōka yōron* 病家要論, Vol. 1. Kyoto University Rare Materials Digital Archive 京都大学貴重資料デジタルアーカイブ. Accessed August 20, 2025. https://rmda.kulib.kyoto-u.ac.jp/item/rb00019887.

Kobayashi Tadayoshi (小林忠良). 1836. *Sanpō koren zen* 算法瑚璉 全. Nagano Prefectural Library 県立長野図書館 accessed through Shinshu Digital Commons 信州デジタルコモンズ. Accessed August 20, 2025. https://www.ro-da.jp/shinshu-dcommons/library/02BK0103231148.

Minamoto no Kimmi (源君美) (Arai Hakuseki 新井白): Nantōshi ro 南島志 呂 (1856). Ryukyu/Okinawa-Related Materials Digital Special Collections, University of the Ryukyus. Accessed August 20, 2025. https://shimuchi.lib.u-ryukyu.ac.jp/collection/sakamaki/hw51602/1.

Moribe Shōkei (守部正稽). 1729. *Shusetsu yōjō ron* 酒説養生論, Vol. 7. Kyoto University Rare Materials Digital Archive 京都大学貴重資料デジタルアーカイブ. Accessed August 20, 2025. https://rmda.kulib.kyoto-u.ac.jp/item/rb00003007.

Okada Kei (岡田啓). 1844. *Owari meisho zue* 尾張名所図会, Vol. 1. BnF (Bibliothèque nationale de France) Gallica. August 20, 2025. https://gallica.bnf.fr/ark:/12148/btv1b105080034.

Xu Baoguang (徐葆光) (J. Jo Hokō). 1721. *Chūzan denshinroku* 中山伝信録, Vol. 1. Ryukyu/Okinawa-Related Materials Digital Special Collections, University of the Ryukyus. Accessed August 20, 2025. https://shimuchi.lib.u-ryukyu.ac.jp/collection/sakamaki/hw78201/1.

References

Chow, Eric H. C. 2025. "Evaluating LLMs for Linked Data Extraction from Chinese Texts: A Comparative Analysis." *The Digital Orientalist*, January 24. Accessed August 3. https://digitalorientalist.com/2025/01/24/evaluating-llms-for-linked-data-extraction-from-chinese-texts-a-comparative-analysis/.

Clanuwat, Tarin, Alex Lamb, and Asanobu Kitamoto. 2018. "End-to-End Pre-Modern Japanese Character (Kuzushiji) Spotting with Deep Learning." In *Proceedings of IPSJ SIG Computers and the Humanities Symposium (Jinmonkon 2018)*, 15–20. Accessed August 19, 2025. https://ipsj.ixsq.nii.ac.jp/records/192436.

———. 2019. "KuroNet: Pre-Modern Japanese Kuzushiji Character Recognition with Deep Learning." In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, edited by Lisa O'Conner, 607–614. Accessed August 19, 2025. https://doi.org/10.1109/ICDAR.2019.00103.

CODH (Center for Open Data in the Humanities). 2025a. "日本古典籍くずし字データセット." Accessed August 19. http://codh.rois.ac.jp/char-shape/.

———. 2025b. "「経」（U+7D4C） 日本古典籍くずし字データセット." Accessed August 19. http://codh.rois.ac.jp/char-shape/unicode/U+7D4C/.

———. 2025c. "「径」（U+5F84） 日本古典籍くずし字データセット." Accessed August 19. http://codh.rois.ac.jp/char-shape/unicode/U+5F84/.

———. 2025d. "「乗」（U+4E57） 日本古典籍くずし字データセット." Accessed August 19. http://codh.rois.ac.jp/char-shape/unicode/U+4E57/.

Diez, Yago, Toya Suzuki, Marius Vila, and Katsushi Waki. 2021. "Automatic Processing of Historical Japanese Mathematics (Wasan) Documents." *Applied Sciences* 11 (17): 8050. Accessed August 19, 2025. https://doi.org/10.3390/app11178050.

Gaoshukai. (2015) 2020. "Shinjitai to kyūjitai sōgo henkan tsūru 新字体と旧字体相互変換ツール." Accessed August 19, 2025. https://www.gaoshukai.com/lab/0039/.

Hashimoto, Yuta. 2025. “NDL kotenseki OCR-Lite Webban NDL古典籍OCR-Lite Web版.” Accessed August 19. https://ndlkotenocr-lite-web.netlify.app/.

Hashimoto, Yuta, and Shinya Miyagawa. 2025. “Koji 日本語史料のための軽量マークアップ言語.” Accessed August 19. https://koji-lang.org/.

Horvath, Aliz. 2022. “Digital Brush Talk: Challenges and Potential Connections in East Asian Digital Research.” In *Global Debates in the Digital Humanities*, edited by Domenico Fiormonte, Sukanta Chaudhuri, and Paola Ricaurte, 127–140. Accessed August 19, 2025. https://muse.jhu.edu/pub/23/oa_edited_volume/chapter/3144060.

Kaji, Yoshiyuki. 2008. “Komonjo nabi 古文書なび. Bushu no kuzushiji 部首のくずし字.” Accessed August 19, 2025. http://komonjo.rokumeibunko.com/binran/bushu01.html.

Kawahira, Toshifumi. 2015. “Paneru disukasshon 1 ‘Hokoku no mirai’ gaiyō パネルディスカッション1「翻刻の未来」概要.” *Kinsei Bungei* 近世文藝 101, 49–58. Accessed September 11, 2025. https://www.jstage.jst.go.jp/article/kinseibungei/101/0/101_49/_article/-char/ja/.

KuroNet. 2020. KuroNet kuzushiji ninshiki sābisu (KuroNetくずし字認識サービス). Center for Open Data in the Humanities. Accessed September 10, 2025. https://mp.ex.nii.ac.jp/kuronet/.

Lamb, Alex, Tarin Clanuwat, and Asanobu Kitamoto. 2020. “KuroNet: Regularized Residual U-Nets for End-to-End Kuzushiji Character Recognition.” *SN Computer Science* 1 (177), 1–15. Accessed August 19, 2025. https://doi.org/10.1007/s42979-020-00186-z.

Moretti, Laura. 2020. *Pleasure in Profit: Popular Prose in Seventeenth-Century Japan*. Columbia University Press.

Morley, Brendan Arkell. 2022. “Kanbun, Kundoku, and the Language of Literary Sinitic: Terminological Issues in the Study of Sinography in Japan.” *Japanese Language and Literature* 56 (2): 329–354. Accessed August 19, 2025. https://doi.org/10.5195/jll.2022.237.

NDL Lab. 2022. “Kotenseki shiryō no OCR tekisutoka jikken 古典籍資料のOCRテキスト化実験(令和4年度〜).” National Diet Library. Accessed August 19, 2025. https://lab.ndl.go.jp/data_set/r4_koten/.

——— (@ndlkotenocr-lite). 2025. “NDL古典籍OCR-Liteアプリケーションのリポジトリ.” National Diet Library. GitHub. Accessed August 19. https://github.com/ndl-lab/ndlkotenocr-lite.

Oyler, Elizabeth. 2006. *Swords, Oaths, and Prophetic Visions: Authoring Warrior Rule in Medieval Japan*. University of Hawaii Press.

PaddleOCR (@PaddleOCR). 2025. “PaddleOCR 3.0: Text Recognition & Doc Parsing Toolkit.” GitHub. Accessed August 19. https://github.com/PaddlePaddle/PaddleOCR.

Rabinovitch, Judith N. 1996. “An Introduction to Hentai Kambun [Variant Chinese], a Hybrid Sinico-Japanese Used by the Male Elite in Premodern Japan.” *Journal of Chinese Linguistics* 24 (1): 98–127. Accessed August 19, 2025. https://www.jstor.org/stable/23753995.

Sangawa, Kristina Hmeljak. 2017. “Confucian Learning and Literacy in Japan’s Schools of the Edo Period.” *Asian Studies* 5 (2): 153–166. Accessed August 19, 2025. https://doi.org/10.4312/as.2017.5.2.153-166.

Sangiacomo, Andrea, Hugo Hogenbirk, Raluca Tanasescu, Antonia Karaisl, and Nick White. 2022. “Reading in the Mist: High-Quality Optical Character Recognition Based on Freely Available Early

Modern Digitized Books." *Digital Scholarship in the Humanities* 37 (4): 1197–1209. Accessed August 19, 2025. https://doi.org/10.1093/llc/fqac014.

Sturgeon, Donald. 2018. "Large-Scale Optical Character Recognition of Pre-Modern Chinese Texts." *International Journal of Buddhist Thought & Culture* 28 (2): 11–44. Accessed August 19, 2025. https://doi.org/10.16893/IJBTC.2018.12.28.2.11.

Ueda, Atsuko. 2008. "Sound, Script, and Styles: Kanbun kundokutai and the National Language Reforms of 1880s Japan." *Review of Japanese Culture and Society* 20 (December): 133–156. Accessed August 19, 2025. https://www.jstor.org/stable/42800998.

Wang, Hao, Hirofumi Shimizu, and Daisuke Kawahara. 2023. "Kanbun-LM: Reading and Translating Classical Chinese in Japanese Methods by Language Models." In *Findings of the Association for Computational Linguistics (ACL2023)*, edited by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, 8589–8601. Association for Computational Linguistics. Accessed August 19, 2025. https://doi.org/10.18653/v1/2023.findings-acl.545.

Yang, Hailin, Lianwen Jin, and Jifeng Sun. 2018. "Recognition of Chinese Text in Historical Documents with Page-Level Annotations." In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, edited by Juan E. Guerrero, 199–204. Accessed August 19, 2025. https://doi.org/10.1109/ICFHR-2018.2018.00043.