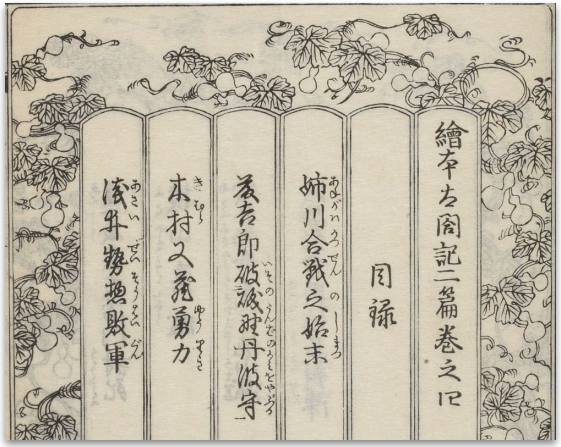
Experimenting With Generic Recognition Systems for Kuzushiji Documents: Furigana Extraction as a Use-Case

Hugo Scheithauer, Laurent Romary

Experimenting With Generic Recognition Systems for Kuzushiji Documents: Furigana Extraction as a Use-Case

Hugo Scheithauer, PhD Candidate, ALMAnaCH, Inria, Paris / École Pratique des Hautes Études, Paris (EPHE)

Laurent Romary, Directorate for Scientific Information and Culture, Inria, Paris



Excerpt from 太閤真顕記 , Edo period, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b8451517r

Kuzushiji Automatic Recognition: a (brief and recent) Research Timeline (1/2)

2016 — The National Institute of Japanese Literature (NIJL) releases the **Japanese Classical Books** (JCB) dataset, curated by the Center for Open Data in the Humanities (CODH).

2018 — Clanuwat et al., 2018, releases the **KMNIST** dataset, a pre-modern Japanese variant of the MNIST dataset.

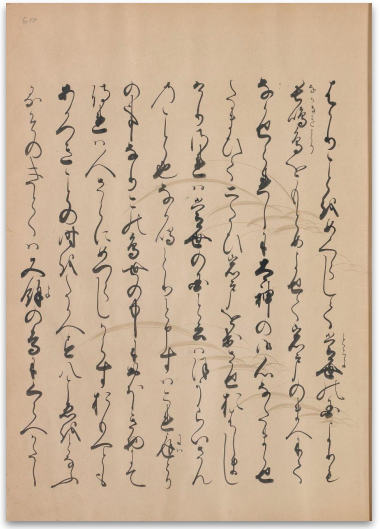
2019 — Clanuwat et al., 2019, releases **Kuronet** for Kuzushiji recognition, a highly efficient deep learning architecture trained on the JCB dataset.

T. N. Hung et al., 2019 introduced **PSNet** for Kuzushiji text segmentation.

The Kuzushiji Kaggle competition is launched by the CODH.

2021 — Lamb et al., 2021 tackles Kuzushiji **character ordering** with a deep recurrent sequence model.

2022 — The CODH releases a new cursive script recognition system, **RURI**, which improves on Kuronet and replaces it.



ほうらい山 , 1661-1681, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b8304433t

Kuzushiji (くずし字): style of cursive Japanese script used in pre-modern (before the early 20th century) written and printed texts.

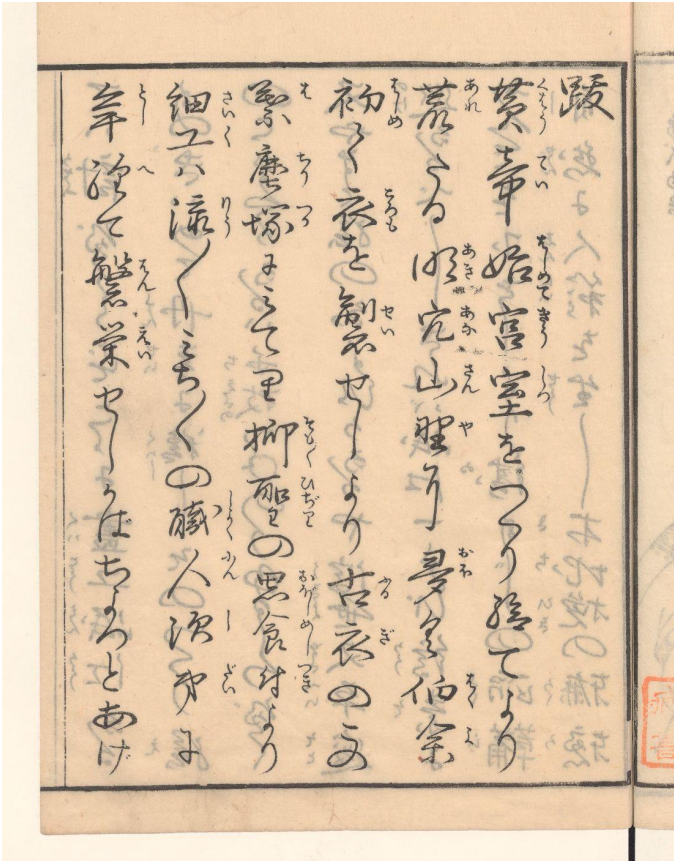
Kuzushiji Automatic Recognition: a (brief and recent) Research Timeline (2/2)

When Kuronet was introduced, Clanuwat et al. (2019) highlighted the intentional exclusion of **furigana** (phonetic annotations placed next to kanji, either in Katakana or Hiragana) due to the absence of labels for annotation in the JCB dataset.

The authors also stressed the need to **differentiate between the main text and annotations** in automatic processing of Kuzushiji documents.

Clanuwat et al., 2019, releases **Kuronet** for Kuzushiji recognition, a highly efficient deep learning architecture trained on the JCB dataset.

2019



Iroe shokunin burui, 大田南畝 , 1784, Edo period, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b531420548

Furigana Detection: Another Step Toward a Comprehensive Kuzushiji Transcription

→ To our knowledge, fine-tuning a system like Kuronet on new annotated data to accommodate furigana extraction is not available to the public.

→ Other custom deep learning architectures created for the Kuzushiji Kaggle competition in 2019 necessitate machine learning engineering expertise for customisation.

→ Furigana processing presents a **dual opportunity**:

- Its intrinsic value as another step toward **comprehensive Kuzushiji transcriptions**,
- And **testing generic systems for Kuzushiji recognition** (layout analysis, and automatic text recognition), as well as creating new shareable datasets for this task.
    - YOLO architecture for layout analysis and object detection.
    - Automatic text recognition engine Kraken.

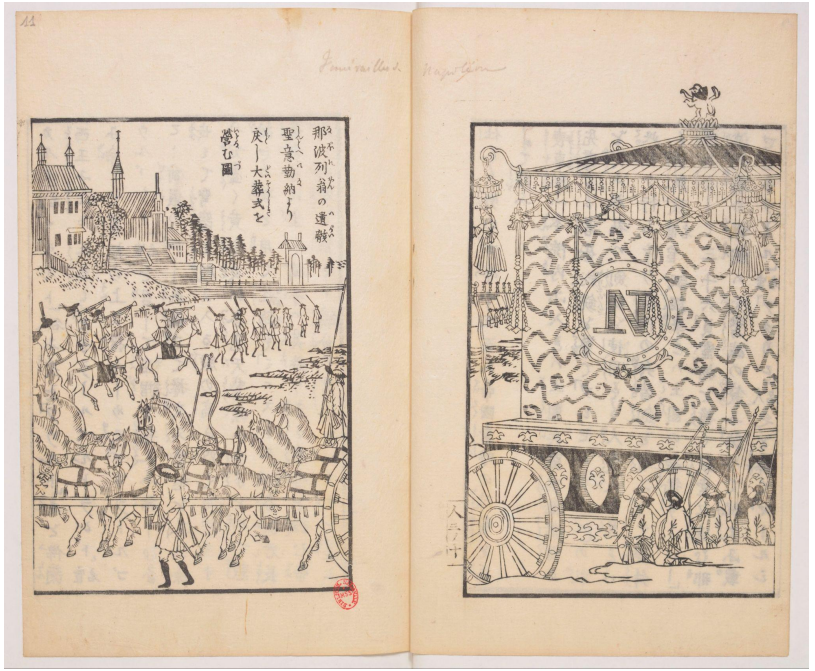
The Ancient Japanese Collections of the French National Library

→ The French National Library holds a large **collection of Japanese manuscripts and various documents spanning throughout the Edo period**. A lot of documents already digitized are available on its digital library, **Gallica**.

- 293 manuscripts,
- 1096 prints and photographs,
- 72 documents related to performing arts,
- 148 maps.

... And much more documents waiting to be digitized.

The only downside is that a lot of the metadata are often incomplete.



Japonais 615 (3), Edo period, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b10510296h

The Furigana-Kanji dataset: presentation (1/2)

→ **15 manuscripts** were sampled from the French National Library to constitute a first version of the Furigana-Kanji dataset.

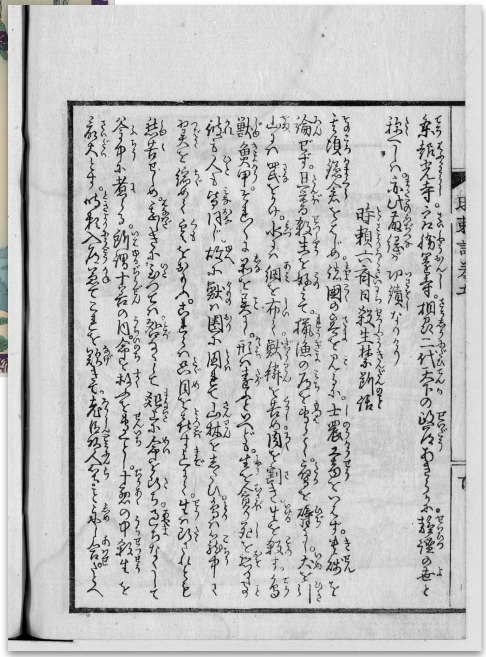
Archival Resource Key (ark)	Title	Romanized title	Alternative title	Author	Date	Annotated pages
ark:/12148/btv1b8304433t	ほうらい山	Hôrai san	N/A	N/A	1661-1681	12
ark:/12148/btv1b8451517r	太閤真顕記	Taikô shinkenki	N/A	N/A	Edo period	10
ark:/12148/btv1b10506231j	N/A	N/A	Japonais 5610	N/A	Edo period	20
ark:/12148/btv1b10510296h	N/A	N/A	Japonais 615 (3)	N/A	Edo period	13
ark:/12148/btv1b10511050r	隅田川両岸一覧 三巻 / 葛飾北斎画	Sumidagawa ryôgan ichi-ran 3 t. / Ill. Katsushika Hokusai	N/A	N/A	1800-1868	2
ark:/12148/btv1b10511093r	N/A	N/A	Japonais 632 (2)	N/A	Edo period	19
ark:/12148/btv1b10512765d	N/A	Arabia monogatari: kaikan kyôki	N/A	永峰秀樹 訳	1875	20
ark:/12148/btv1b10512768r	N/A	Fukushoku zukai	N/A	N/A	Edo period	29
ark:/12148/btv1b10525770x	N/A	Hisago gundan gojûyon-jô	N/A	歌川 芳艶	1864-1865	19
ark:/12148/btv1b10527650d	N/A	Kiyomizu tsuya mono-gatari	N/A	N/A	Edo period	14
ark:/12148/btv1b60002827	北条時頼記図会	Hôjô jiraiki zue	N/A	池田東籬	1788-1857	19
ark:/12148/btv1b60002879	N/A	Tôto shôkei ichiran	N/A	葛飾 北斎	1760-1849	10
ark:/12148/btv1b531420548	彩画職人部類	Iroe shokunin burui	N/A	大田南畝	1784	14
ark:/12148/btv1b6000280d	N/A	N/A	Hokusai manga / Katsushika	葛飾 北斎	1760-1849	25
ark:/12148/btv1b10509001k	傳神開手 北斎画式	Denshin kaishu Hokusai gashiki	N/A	葛飾 北斎	1825	9
ark:/12148/btv1b10523261c	花鳥写真図彙 / 北尾紅翠斎	Shashin kachō zue / Kitao Kōsuisai	N/A	北尾 重政	1805	8
ark:/12148/btv1b10523259j	花鳥写真図彙 / 北尾紅翠斎	Shashin kachō zue / Kitao Kōsuisai	N/A	北尾 重政	1805	9
						Total: 252

The Furigana-Kanji dataset: presentation (2/2)

Iroe shokunin burui, 大田南畝 , 1784, Edo period, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b531420548

ほうらい山 , 1661-1681, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b8304433t

Hisago gundan gojûyon-jô, 歌川 芳艶, 1864-1865, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b10525770x

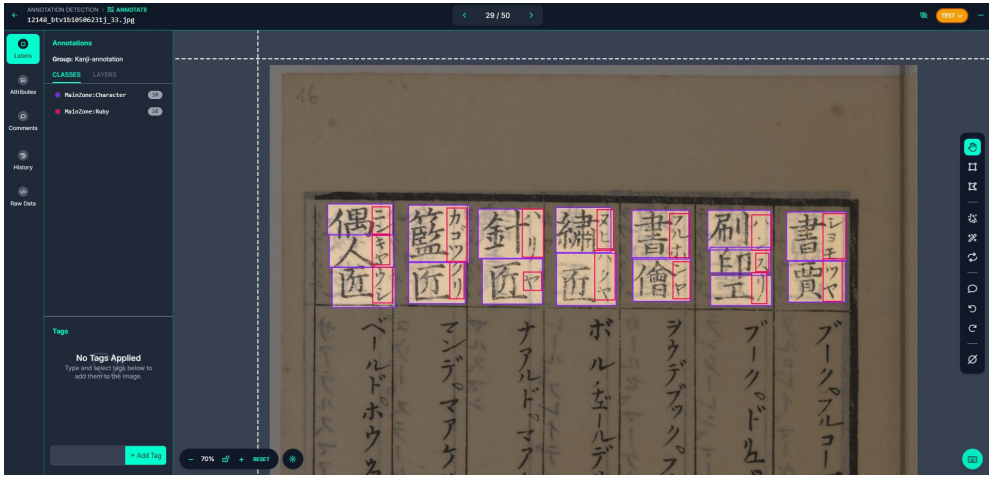


北条時頼記図会 , 池田東籬 , 1788-1857, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b60002827

Training YOLO for Furigana Detection in Object Detection Models

Based on the growing popularity and efficiency of the **YOLO architecture** in DH projects (Clérice, 2023 & Clérice et al., 2024), we decided to opt for an **object detection approach** for furigana detection.

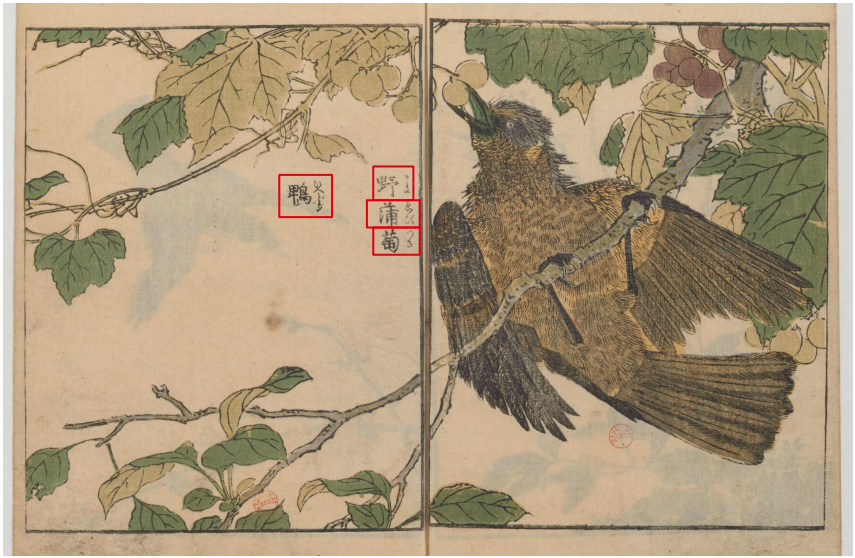
YOLO segments an image into zones based solely on **visual features**, assigning a label to each zone.



Screenshot of the Roboflow annotation online interface.

Effective Annotation Strategies for Furigana detection (1/3)

Sometimes, furigana can easily be matched with their kanji at a glance.

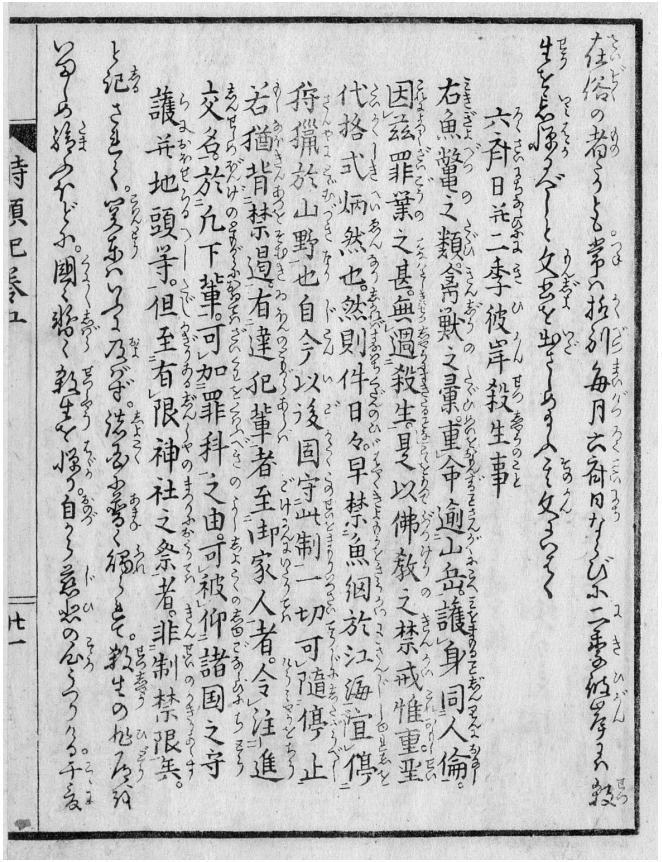


花鳥写真図彙/ 北尾紅翠斎, 北尾 重政, 1805, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b10523261c

Effective Annotation Strategies for Furigana detection (2/3)

.... But often they are not.

Hokusai manga / Katsushika, 葛飾 北斎, 1760-1849, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b6000280d

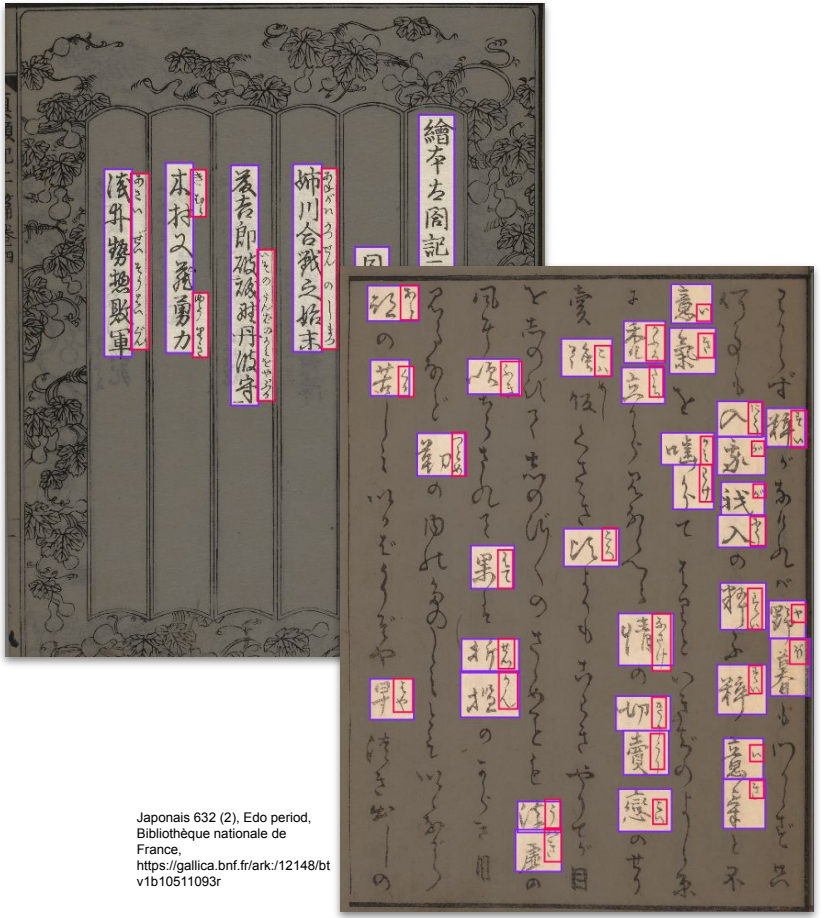


北条時頼記図会 池田東籬, 1788-1857, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b60002827

Effective Annotation Strategies for Furigana detection (3/3)

To evaluate the performance of the YOLO object detection model for furigana detection, we employed **two layout annotation schemes**:

1. **Annotating the main text and furigana independently**, allowing for differentiation between them. Schema inspired from Bjerregaard et al., 2022.
2. **Nesting the furigana annotations within the kanji annotations**, enabling both the distinction between the main text and furigana and the association of furigana with its corresponding kanji. The objective is to **detect only kanji annotated with furigana**.



太閤真顯記 , Edo period, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b8451517r

Japonais 632 (2), Edo period, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b10511093r

Layout Annotation Using the SegmOnto Controlled Vocabulary

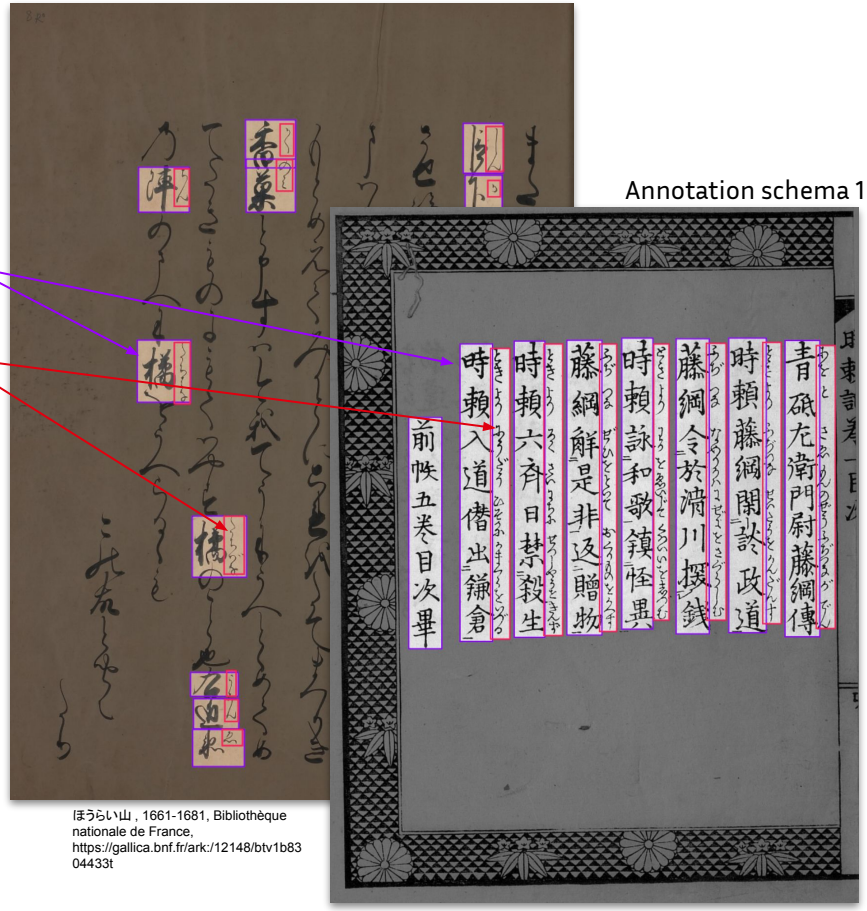
https://segmonto.github.io/

MainZone:Character

MainZone:Ruby

Annotation schema 2

Annotation schema 1



ほうらい山 , 1661-1681, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b8304433t

北条時頼記図会 , 池田東籬 , 1788-1857, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b60002827

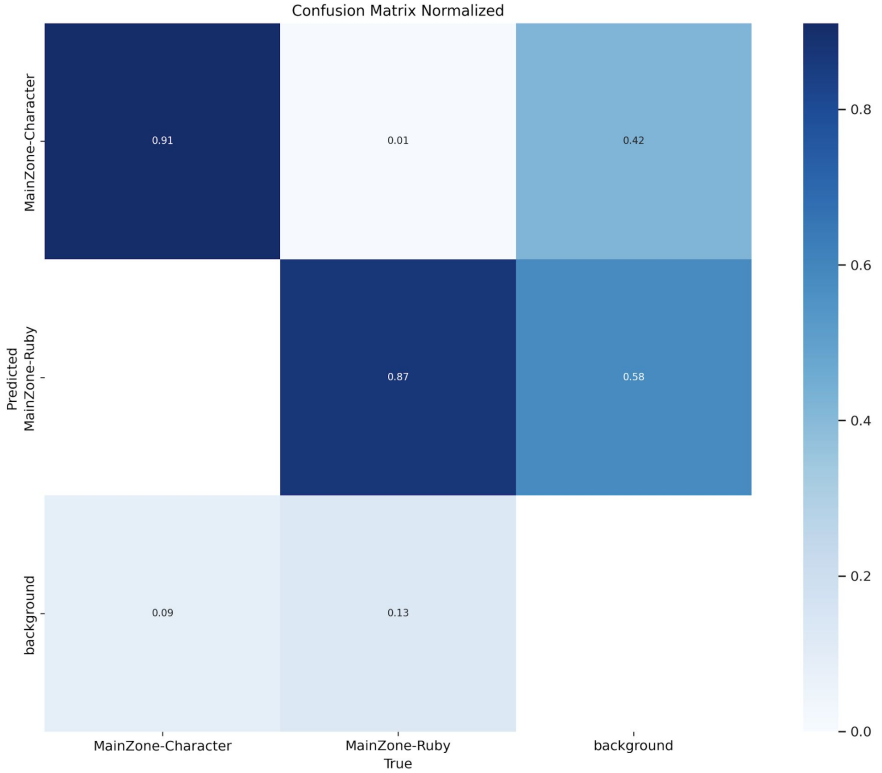
→ The **Kanji-Furigana** dataset was annotated following the **SegmOnto controlled vocabulary**, designed for **describing the content of books and manuscripts**.

→ Using a controlled vocabulary ensures that the dataset follows the **FAIR principles**.

Model Performance for Detecting Unlinked Text and Furigana (1): Quantitative Analysis

mAP	Precision	Recall
92.3%	94.1%	86.8%

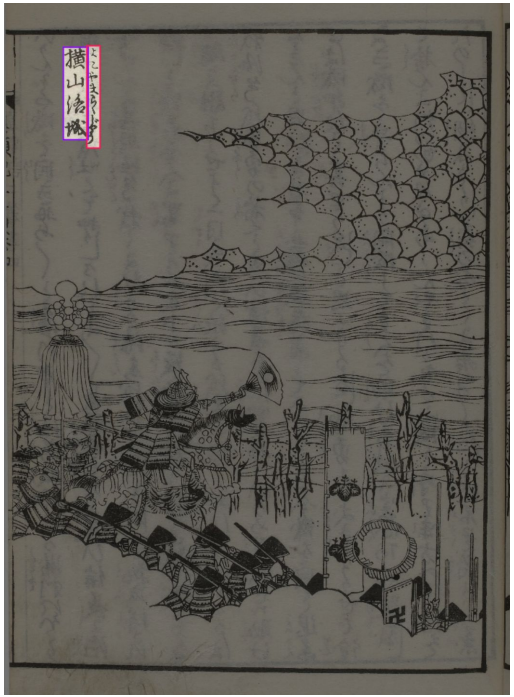
Very good scores can be achieved with a **relatively small and diverse dataset**.



The model can be visualized and tested online at:
https://app.roboflow.com/jadh2024/furigana-detection-unlinked/visualize/3.
To be uploaded on GitHub at:
https://github.com/HugoSchtr/Kanji-Furigana-Dataset

Model Performance for Detecting Unlinked Text and Furigana (1): Qualitative Analysis (1/2)

太閤真顕記, Edo Period, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b8451517r

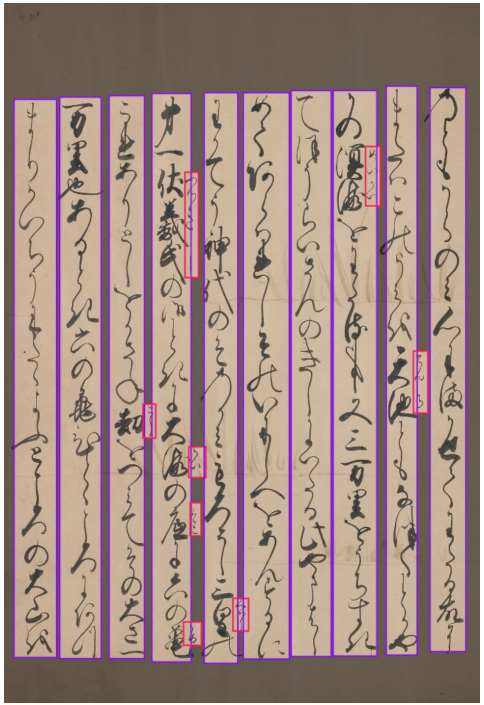


Ground truth

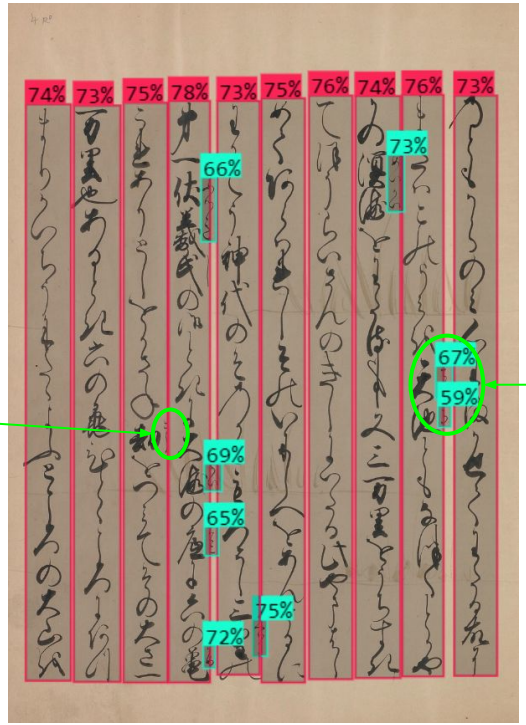
Prediction

Model Performance for Detecting Unlinked Text and Furigana (1): Qualitative Analysis (2/2)

ほうらい山, 1661-1681, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b8304433t



Ground truth



Prediction

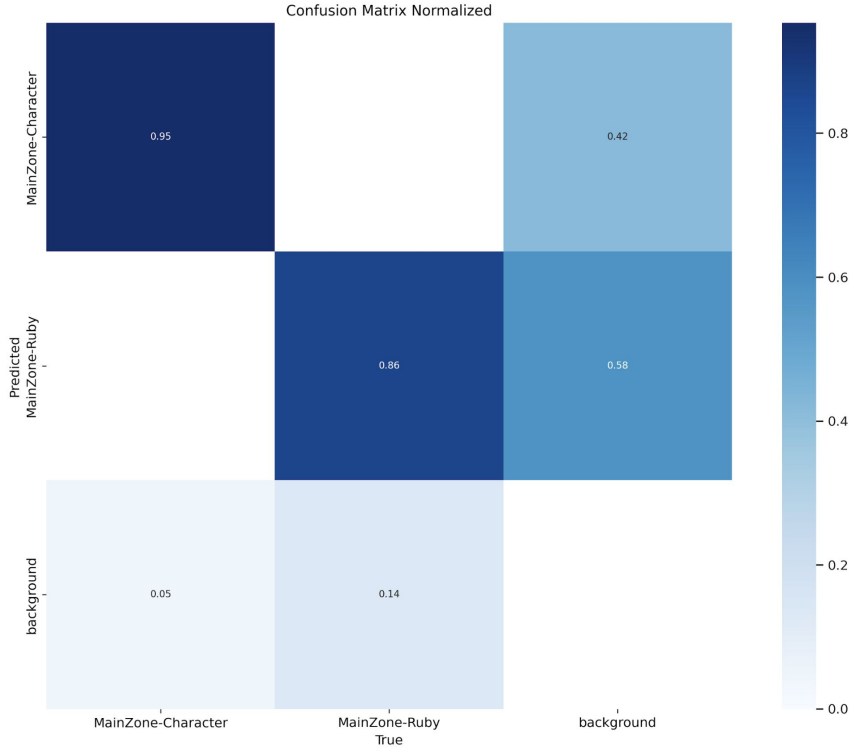
Missing furigana instance

Split error (but that's not very important as long as we can differentiate between the main text and the furigana)

Model Performance for Detecting Linked Kanji and Furigana (2): Quantitative Analysis

mAP	Precision	Recall
94.2%	91.9%	90.9%

Again, **very good scores** (even better than the previous experiment) can be achieved with a **relatively small and diverse dataset**.



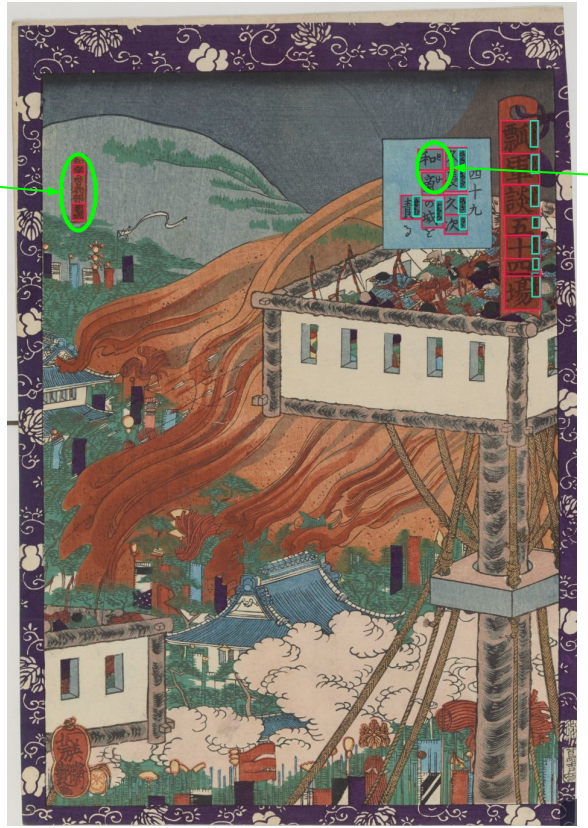
The model can be visualized and tested online at: https://app.roboflow.com/jadh2024/annotation-detection/visualize/15.
Uploaded on GitHub: https://github.com/HugoSchtr/Kanji-Furigana-Dataset

Hisago gundan gojûyon-jô, 歌川 芳艷, 1864-1865, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b10525770x

Model Performance for Detecting Linked Kanji and Furigana (2): Qualitative Analysis (1/2)

Ground truth

Hallucinated kanji annotated with furigana

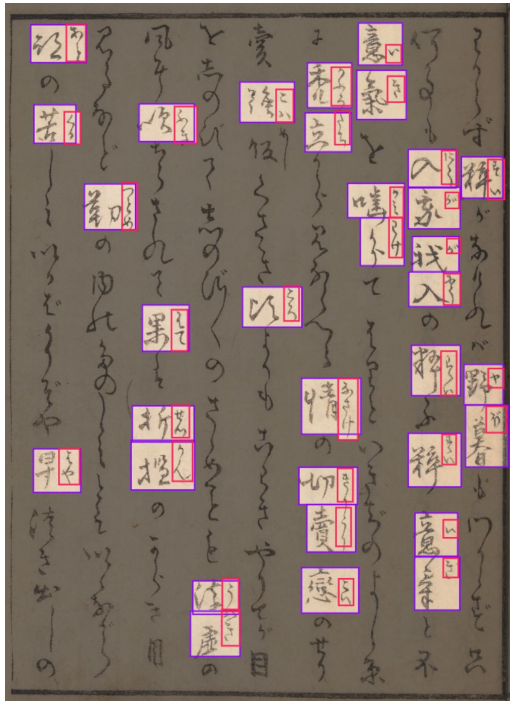


Missed furigana annotations

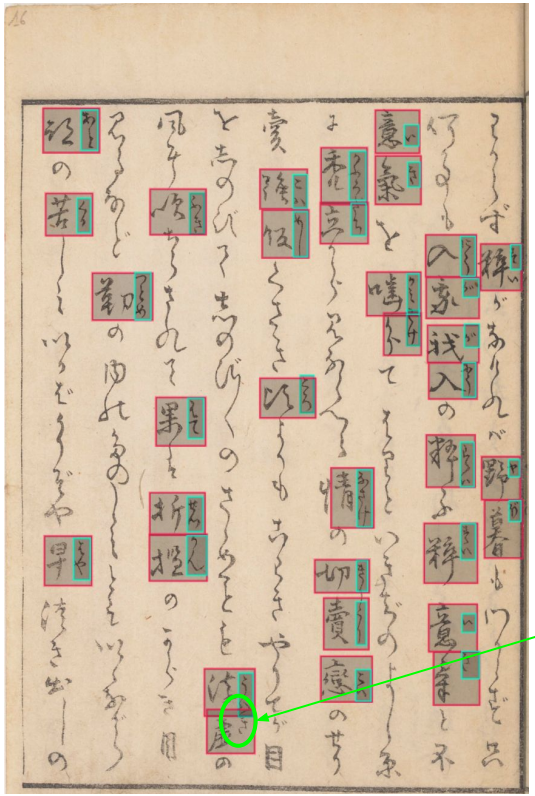
Prediction

Model Performance for Detecting Linked Kanji and Furigana (2): Qualitative Analysis (2/2)

Japonais 632 (2), Edo period, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b10511093r



Ground truth



Prediction

Missed furigana annotation, but near perfect detection.

Flaws in Object Detection for Linked Kanji and Furigana

→ The dataset includes manuscripts and illustrated manuscripts but **lacks highly complex layouts** (Clanuwat et al., 2019).

→ The second annotation schema assumes that each kanji works separately, which is not always the case: it does not take into account complex linguistic phenomena such as **Jukujikun (熟字訓)** as object detection models are based on visual features, not textual features.

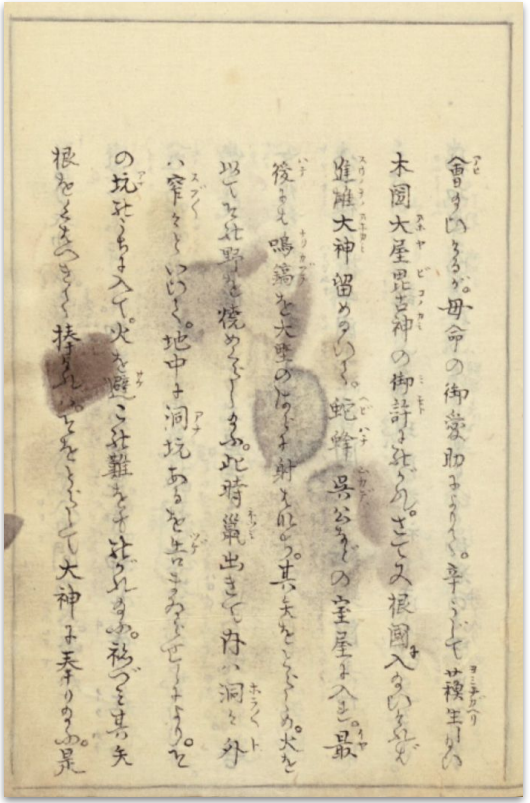
→ The second annotation schema works well for documents where kanji and furigana are distinctly separated. However, when **furigana is written cursively across multiple kanji, it becomes difficult to determine whether the model splits the furigana based on subtle annotated visual writing features or arbitrarily**.

Figures taken from Clanuwat et al., 2019. The layout does not lay out the text in a clear sequential order and regular columns.

Expanding the Furigana-Kanji dataset

- **Sample more documents from the French National Library** to have a more diverse dataset.
- Try to reach out to curators to complete the documents' metadata.
- **Explore and sample documents from other French heritage institutions**

→ Expanding the dataset will help determine **whether a generic Furigana detection model can be developed or if domain-specific factors play a critical role in models' performance.**



当家秘歌集 , 1762, Bibliothèque universitaire des langues et civilisations, https://bina.bulac.fr/s/bina/ark:/73193/b3r2dm

Kraken for Kuzushiji Recognition: Evaluating a Generic Text Recognition System (1/2)

- **Kraken is a generic, language-agnostic and all-in-one text recognition engine.**
- It features trainable **layout analysis** (for text regions and lines) and **character recognition** models that support all Unicode scripts.
- Kraken has demonstrated satisfactory results on historical Chinese documents (18th-20th century) (Brisson et al., 2023) with challenges similar to Kuzushiji recognition, such as unbalanced class distributions, which encouraged us to experiment with it.

Kraken for Kuzushiji Recognition: Evaluating a Generic Text Recognition System (2/2)

What Kuzushiji datasets do we have to experiment with Kraken?

→ The **KMNIST dataset** (Clanuwat et al., 2018).

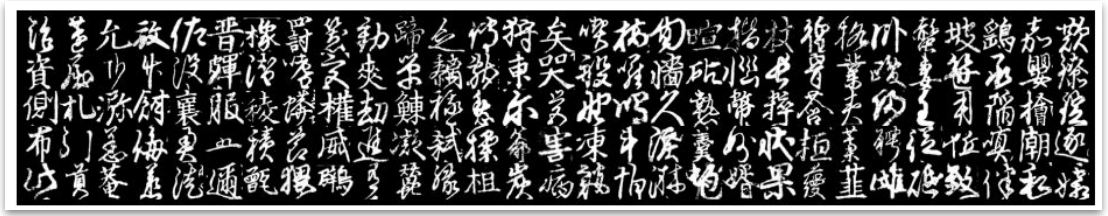
It comprises three subsets:

- **Kuzushiji-MNIST**: 10 classes of hiragana characters.
- **Kuzushiji-49**: 49 different hiragana characters.
- **Kuzushiji-Kanji**: 3832 unbalanced classes of kanji characters.

Hiragana	Unicode	Samples	Sample Images
お (o)	U+304A	7000	
き (ki)	U+304D	7000	
す (su)	U+3059	7000	
つ (tsu)	U+3064	7000	
な (na)	U+306A	7000	

Hiragana	Unicode	Samples	Sample Images
は (ha)	U+306F	7000	
ま (ma)	U+307E	7000	
や (ya)	U+3084	7000	
れ (re)	U+308C	7000	
を (wo)	U+3092	7000	

Taken from Clanuwat et al., 2018. The 10 classes of the Kuzushiji-MNIST



Taken from Clanuwat et al., 2018. Examples of some of the 3832 classes in the Kuzushiji-Kanji subset.

Evaluating Kraken on the KMNIST Dataset

Train Set	Test Set	Kraken Accuracy (%)
Kuzushiji-MNIST (train)	Kuzushiji-MNIST (test)	90.8
Kuzushiji-49 (train)	Kuzushiji-49 (test)	86.5
Kuzushiji-Kanji (train, 80%)	Kuzushiji-Kanji (test, 20%)	48.9

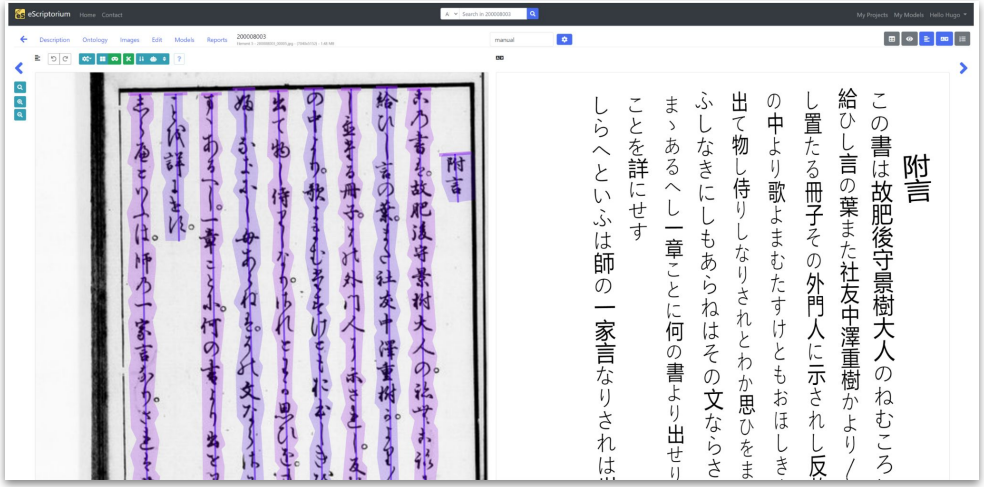
→ Kraken was trained on the **official split for both Kuzushiji-MNIST and Kuzushiji-49** subsets. For the Kuzushiji-Kanji subset, **we implemented a balanced split to ensure the presence of all classes** in the train and test sets.

→ **Kraken is optimized for line-based training data**, which can explain the relatively low scores achieved on the Kuzushiji-49 subset, and the mediocre accuracy on the Kuzushiji-Kanji subset.

→ We hope to provide a **baseline for future evaluations of Kraken on Kuzushiji recognition**.

eScriptorium (PSL - SCRIPTA): a GUI for Automatic Text Recognition Projects

- GUI for **transcribing** textual documents, **creating training sets**, training **segmentation and transcription models**, exporting dataset in standardized formats (XML ALTO), etc.
- **Project management tool**
- eScriptorium uses the language-agnostic transcription engine **Kraken** (Benjamin Kiessling, PSL)



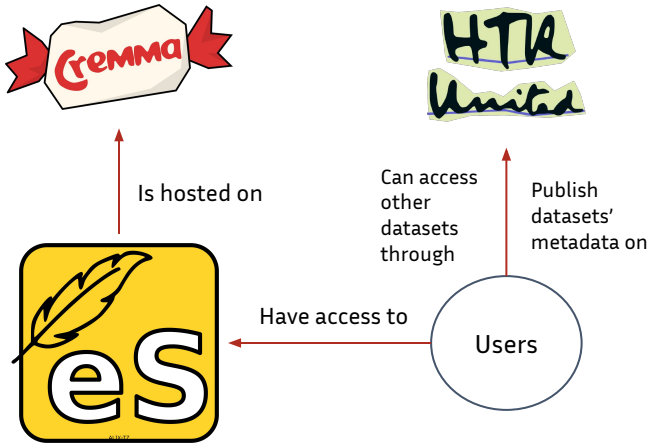
An annotated page of 歌学提要 (doi:10.20730/200008003) (国文研書誌 ID / Kokubunken Bibliography ID: 200008003, from the JCB dataset.

eScriptorium as an Automatic Text Recognition Infrastructure

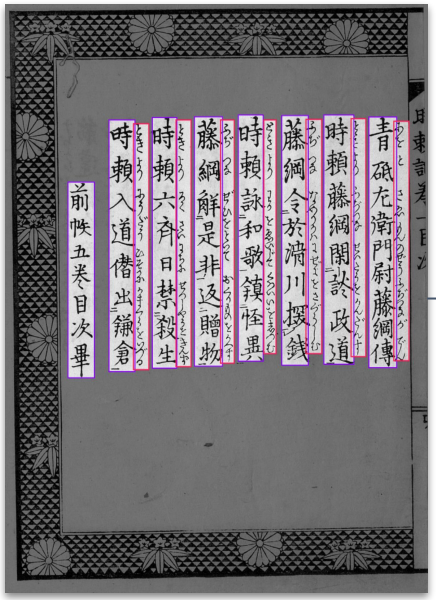
→ **CREMMA** (Consortium for the handwritten text recognition of ancient materials) is an **infrastructure** is led by Inria and the École nationale des chartes (Jean-Mabillon Center). It is funded by the DIM "Ancient and Heritage Materials."

→ Inria provides an access to an instance of eScriptorium hosted on a **powerful server** (GPUs & storage).

→ In exchange, users engage themselves to make their training data open and to publish their metadata on the **HTR-United** catalog.



Incorporating Furigana Detection in Kuzushiji Documents Digital Editions



北条時頼記図会，池田東籬，1788-1857, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b60002827

Automatic Text Recognition with Kraken

XML ALTO export

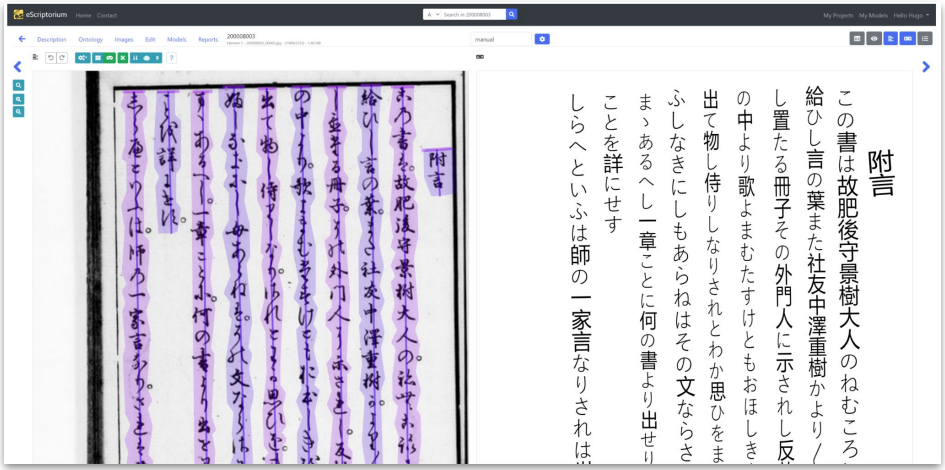
Semi-automatic transformation into TEI based on text zones coordinates and transcribed text

ALTO XML ensures the creation of **standardized interoperable datasets** by being easily shareable, and acts as a pivot format to convert layout analysis and recognition results into TEI

Training Kraken with the JCB dataset (1/2)

→ The **JCB dataset** is also a significant resource to further our experiment with Kraken.

→ Some work is required to convert the JCB dataset into a format usable for training Kraken (**ALTO XML**), as it is currently redistributed as CSV files. But doing so would allow to have an **ALTO XML line-level annotated dataset** for training line detection and Kuzushiji recognition Kraken models.

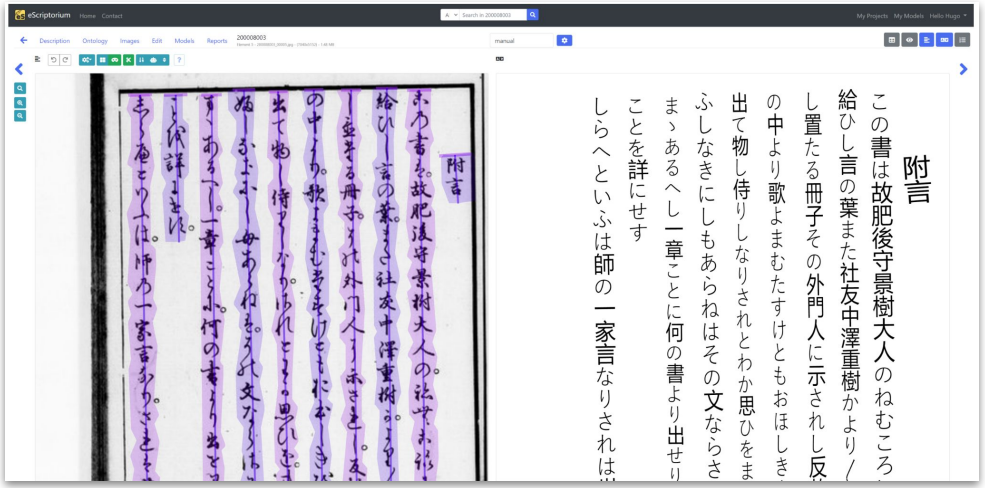


An annotated page of 歌学提要 (doi:10.20730/200008003) (国文研書誌 ID / Kokubunken Bibliography ID: 200008003, from the JCB dataset.

Training Kraken with the JCB dataset (2/2)

However, Clanuwat et al. (2019) already highlighted that line-based recognition models are not ideal for Kuzushiji documents, as these texts often lack a consistent layout, which complicates line ordering and recognition.

Experiments with Kraken's new line ordering model need to be conducted.



An annotated page of 歌学提要 (doi:10.20730/200008003) (国文研書誌 ID / Kokubunken Bibliography ID: 200008003, from the JCB dataset.

Conclusion

- Detecting furigana is **easily achievable** with today's object detection technology, even with relatively little amounts of annotated diverse data.
- The questions are: what is the most efficient layout annotation schema for furigana processing in the context of archiving and editing Kuzushiji documents? How can we incorporate furigana processing in edition pipelines?
- Furigana detection experiments also pave the way to broader experiments on **Kuzushiji documents layout analysis**.
- Work still need to be done on Kraken for Kuzushiji recognition, but Kraken and eScriptorium offers an opportunity to create an infrastructure for creating and sharing new Kuzushiji FAIR datasets to the wider research community.

Thank you! All questions and feedback are welcomed!

花鳥写真図彙 / 北尾紅翠斎 , 北尾 重政, 1805, Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/btv1b10523261c

References

- T. Clanuwat, A. Lamb, A. Kitamoto, KuroNet: Pre-Modern Japanese Kuzushiji Character Recognition with Deep Learning, 2019. URL: http://arxiv.org/abs/1910.09433. doi:10.48550/arXiv.1910.09433, arXiv:1910.09433 [cs].
- Y. Hashimoto, Y. Iikura, Y. Hisada, S. Kang, T. Arisawa, A. Okajima, T. Yada, R. Goyama, D. K.-B, The Kuzushiji Project: Developing a Mobile Learning Application for Reading Early Modern Japanese Books, 2016. URL: https://www.semanticscholar.org/paper/The-Kuzushiji-Project%3A-Developing-a-Mobile-Learning-Hashimoto-Iikura/394041efdd271fa54588516d93b46e48ff15ec61.
- K. Ueki, T. Kojima, Survey on Deep Learning-based Kuzushiji Recognition, 2020. URL: http://arxiv.org/abs/2007.09637. doi:10.48550/arXiv.2007.09637, arXiv:2007.09637 [cs].
- G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YOLO, 2023. URL: https://github.com/ultralytics/ultralytics.
- B. Kiessling, Kraken - a Universal Text Recognizer for the Humanities, 2019. URL: https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/Z9G2EX. doi:10.34894/Z9G2EX.
- T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, D. Ha, Deep Learning for Classical Japanese Literature, 2018. URL: http://arxiv.org/abs/1812.01718. doi:10.20676/00000341, arXiv:1812.01718 [cs, stat].
- T. N. Hung, T. N. Cuong, K. Asanobu, N. Masaki, Segmenting Text in Japanese Historical Document Images using Convolutional Neural Networks, 2019 2019 (2019) 253–260. URL: https://cir.nii.ac.jp/crid/1050855522047807616.
- A. Lamb, T. Clanuwat, S. Han, M. Bober-Irizar, A. Kitamoto, Predicting the ordering of characters in Japanese historical documents, 2021. arXiv:2106.06786.
- N. K. Bjerregaard, V. Cheplygina, S. Heinrich, Detection of Furigana Text in Images, 2022. URL: http://arxiv.org/abs/2207.03960. doi:10.48550/arXiv.2207.03960, arXiv:2207.03960 [cs].
- N.-M. Sven, R. Matteo, Page Layout Analysis of Text-heavy Historical Documents: a Comparison of Textual and Visual Approaches, 2022. URL: http://arxiv.org/abs/2212.13924. doi:10.48550/arXiv.2212.13924, arXiv:2212.13924 [cs].
- T. Clérice, J. Janes, H. Scheithauer, S. Bénière, L. Romary, B. Sagot, “Layout Analysis Dataset with SegmOnto”. In: DH2024 - Annual conference of the Alliance of Digital Humanities
- Organizations. ADHO. Washington DC, United States, Aug. 2024. URL: https://inria.hal.science/hal-04513725.
- T. Clérice, You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine, Journal of Data Mining & Digital Humanities Historical Documents and... (2023) 9806. URL: http://arxiv.org/abs/2207.11230. doi:10.46298/jdmdh.9806, arXiv:2207.11230 [cs].
- S. Gabay, J.-B. Camps, A. Pinche, C. Jahan, SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more), in: 1st International Workshop on Computational Paleography (IWCP@ICDAR 2021), Lausanne, Switzerland, 2021. URL: https://hal.science/hal-03336528.
- A. I. Solis, J. Zarkovacki, J. Ly, A. Atyabi, Recognition of Handwritten Japanese Characters Using Ensemble of Convolutional Neural Networks, 2023. URL: http://arxiv.org/abs/2306.03954. doi:10.48550/arXiv.2306.03954, arXiv:2306.03954 [cs].
- P. A. Stokes, B. Kiessling, D. S. B. Ezra, R. Tissot, E. H. Gargem, The eScriptorium VRE for Manuscript Cultures, Classics@ Journal (2021). URL: https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/.
- C. Brisson, F. Constant, M. Bui, Chinese Historical documents Automatic Transcription (CHAT) models, 2023. URL: https://doi.org/10.5281/zenodo.8383732. doi:10.5281/zenodo.8383732.
- A. Chagué, T. Clérice, Sharing HTR datasets with standardized metadata: the HTR-United initiative, 2022. URL: https://inria.hal.science/hal-03703989.
- Z. Shen, K. Zhang, M. Dell, A Large Dataset of Historical Japanese Documents with Complex Layouts, 2020. URL: http://arxiv.org/abs/2004.08686. doi:10.48550/arXiv.2004.08686, arXiv:2004.08686 [cs].

Bonus: check the quality of your data when training models!

Furigana object detection model - 2nd annotation schema

mAP	Precision	Recall
69.2%	77.0%	68.1%

→

mAP	Precision	Recall
94.2%	91.9%	90.9%

“Dirty”, not proofread, version of the dataset

Same dataset, but clean