

テクノロジーとみんなで翻刻



KogumaNet

# AIとみんなで翻刻

カラーヌワット・タリン  
特任助教

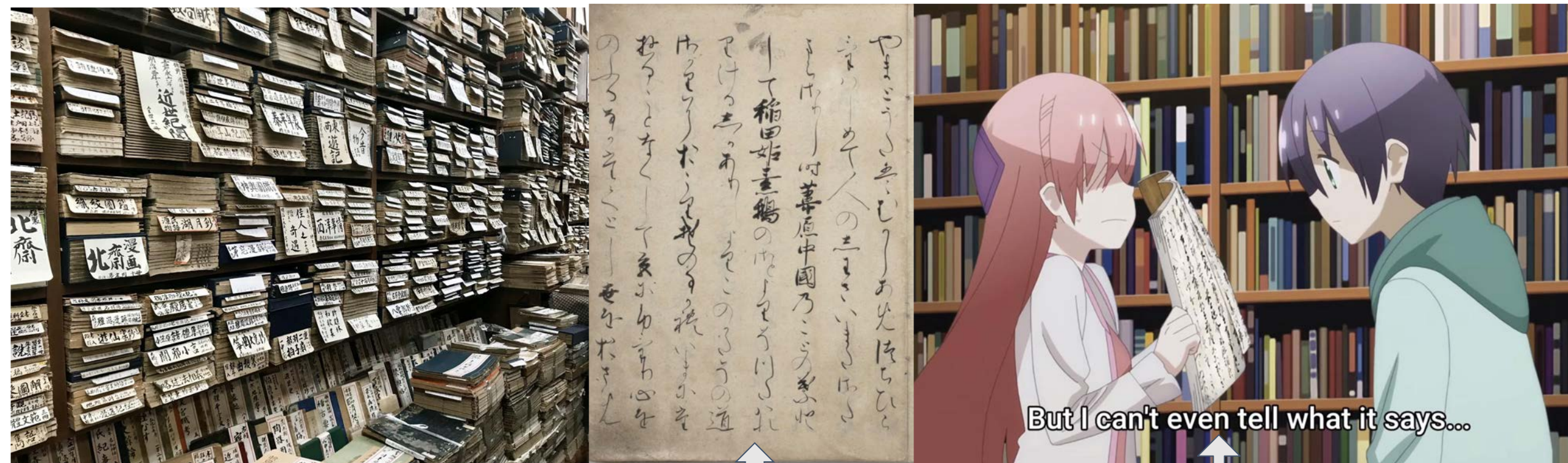
ROIS-DS人文学オープンデータ共同利用センター  
国立情報学研究所



# 自己紹介

- 早稲田大学文学研究科博士後期課程修了（文学）。
- 専門は中世『源氏物語』の注釈書。
- 2017年、東京大学大学院総合文化研究科広域システム科学系で特別研究学生としてディープラーニング研究開始。
- 2018年、ROIS-DS人文学オープンデータ共同利用センター(CODH)、国立情報学研究所に就職。
- 2018～2019年くずし字認識モデル、KogumaNet, KuroNetを開発。

- くずし字資料は1億点以上が現存している。
- くずし字の読める人が少ない。

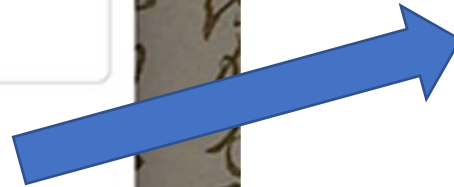
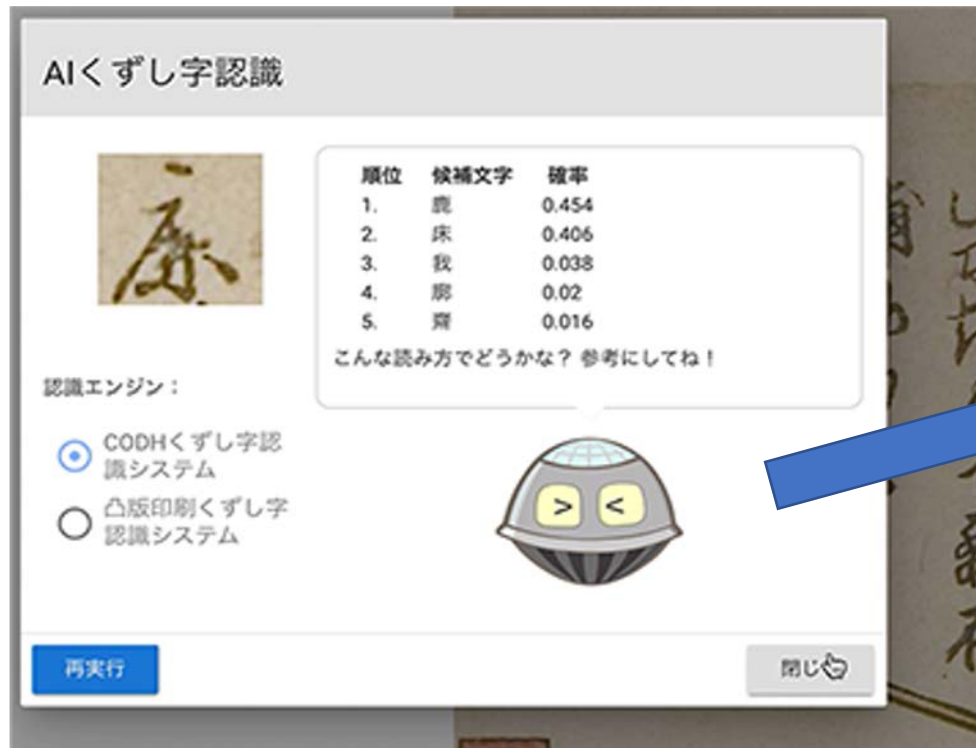


Tonikaku Kawaii (2020)

国文研蔵『新古今和歌集』



# AIを用いてくずし字を読む: くずし字一文字認識 (KogumaNet)

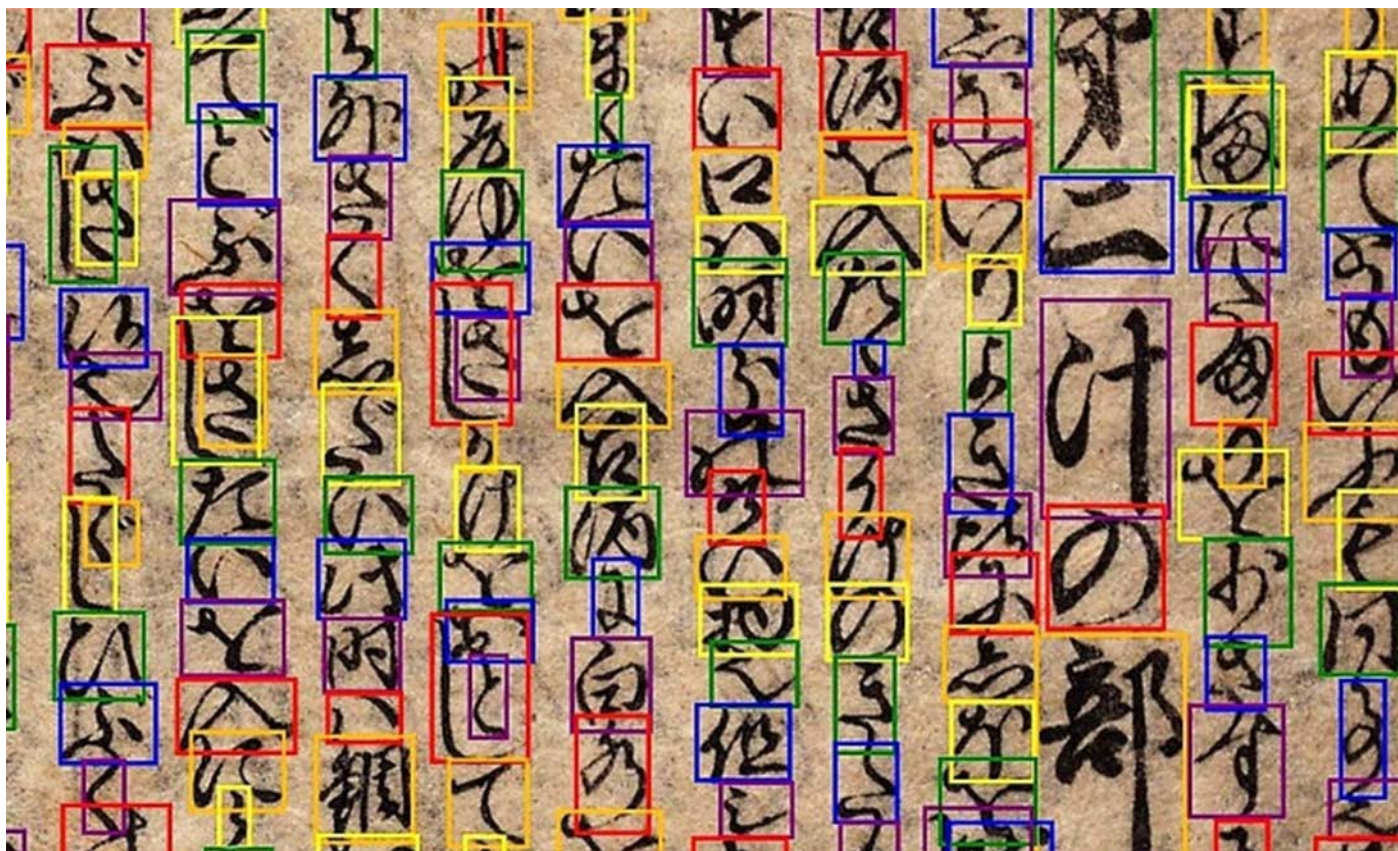


KogumaNet

画像は  
イメージ

KogumaNetがくずし字を習ったデータ

## くずし字データセット（国文学研究資料館+ CODH）



<http://codh.rois.ac.jp/char-shape/>  
<https://www.nijl.ac.jp/pages/cijproject/>

| Unicode | Image           | X   | Y    | Width | Height |
|---------|-----------------|-----|------|-------|--------|
| U+842C  | 200021853-00002 | 634 | 244  | 127   | 163    |
| U+58B6  | 200021853-00002 | 645 | 424  | 123   | 156    |
| U+65E5  | 200021853-00002 | 665 | 611  | 65    | 87     |
| U+7528  | 200021853-00002 | 650 | 727  | 97    | 123    |
| U+60E3  | 200021853-00002 | 644 | 883  | 121   | 140    |
| U+83DC  | 200021853-00002 | 640 | 1000 | 120   | 164    |
| U+4FCE  | 200021853-00002 | 638 | 1117 | 124   | 124    |
| U+4E0D  | 200021853-00002 | 638 | 1234 | 108   | 108    |
| U+6642  | 200021853-00002 | 638 | 1351 | 45    | 45     |
| U+73CD  | 200021853-00002 | 638 | 1468 | 29    | 29     |
| U+5BA2  | 200021853-00002 | 638 | 1585 | 141   | 141    |
| U+5373  | 200021853-00002 | 638 | 1702 | 124   | 124    |
| U+5E2D  | 200021853-00002 | 465 | 1819 | 145   | 145    |
| U+5E96  | 200021853-00002 | 469 | 1936 | 133   | 131    |
| U+4E01  | 200021853-00002 | 480 | 2053 | 121   | 100    |
| U+5408  | 200021853-00002 | 533 | 2170 | 179   | 127    |



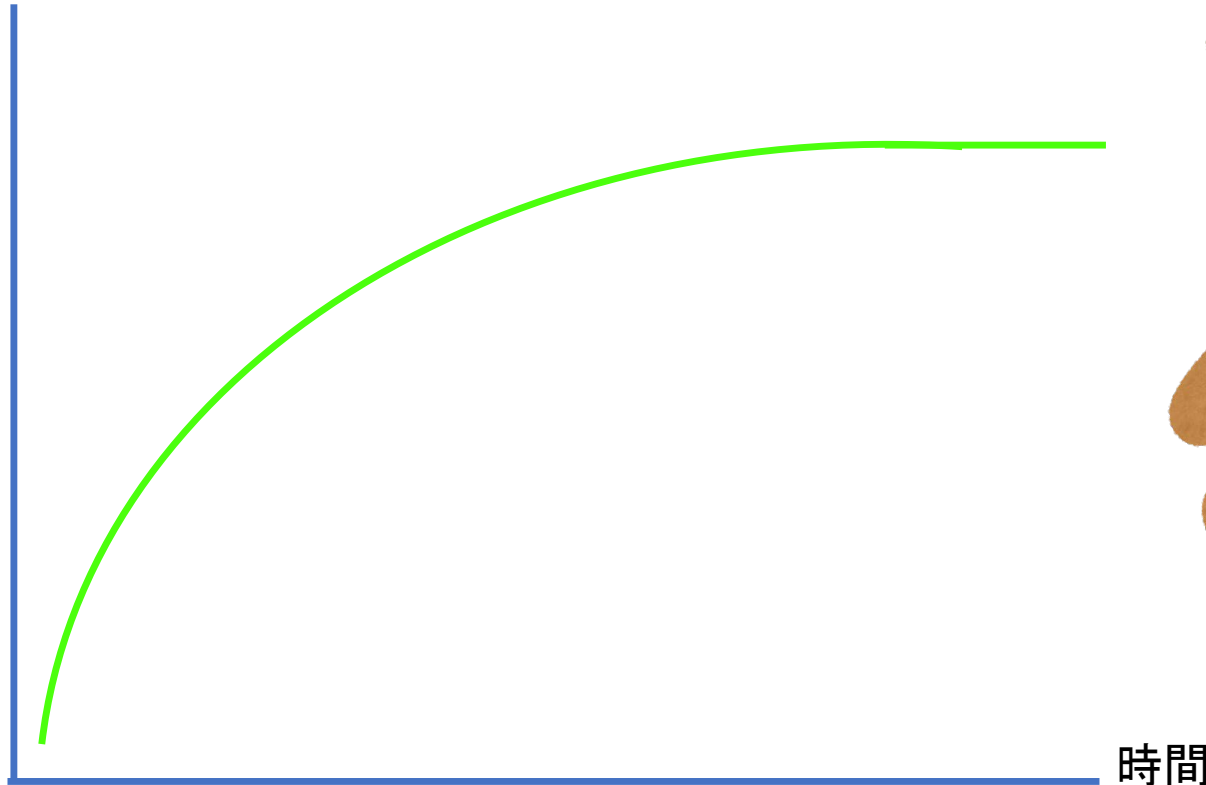
100万  
文字！

KogumaNet 画像はイメージ。

KogumaNetがくずし字を習ったデータ

## くずし字データセット（国文研 + CODH）

精度



KogumaNet

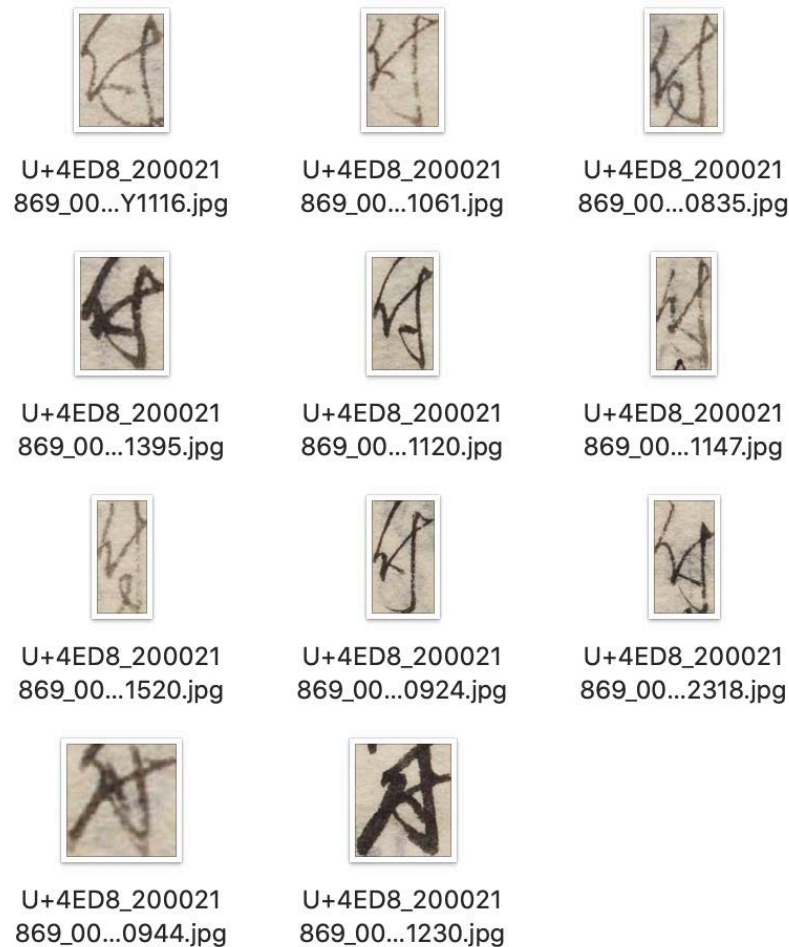
これ以上学習しても精度が上がらない。

画像はイメージ。

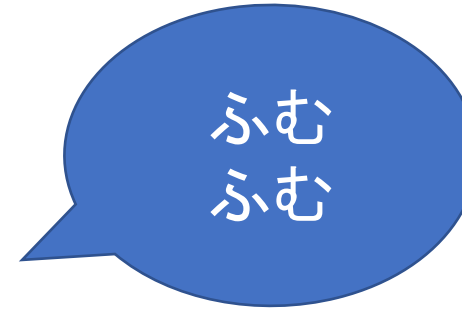


KogumaNetがくずし字を習ったデータ

# くずし字データセット（国文研 + CODH）



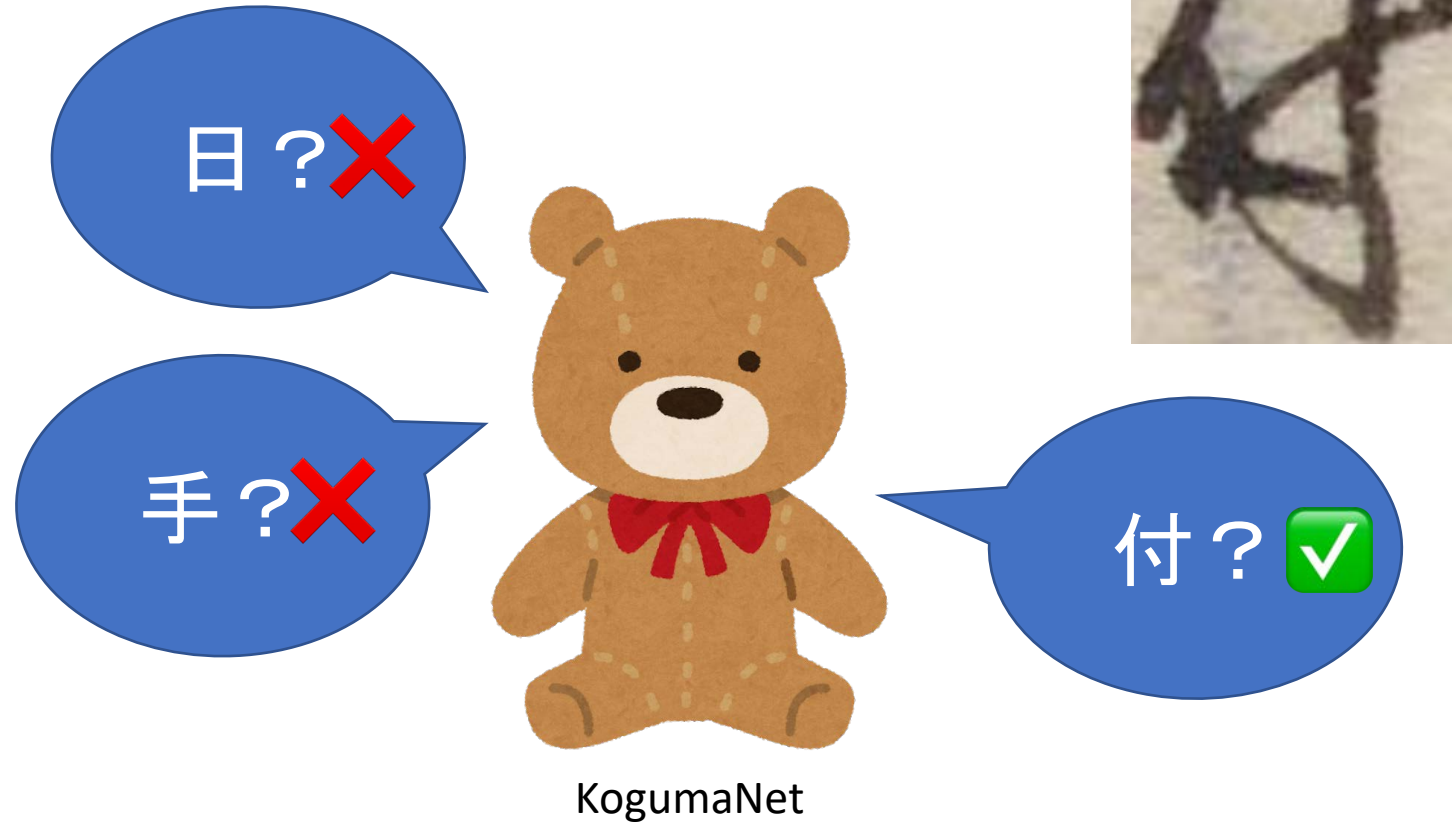
KogumaNet



画像はイメージ。

KogumaNetがくずし字を習ったデータ

## くずし字データセット（国文研 + CODH）



画像はイメージ。



KogumaNetがくずし字を習ったデータ

## くずし字データセット（国文研 + CODH）

十分学習  
したと  
思う！



KogumaNet

Quantization  
量子化

小さく  
なったよ！



KogumaNet

ロボット  
になった



画像はイメージ。

# くずし字の難しいところ 1 : 変体仮名

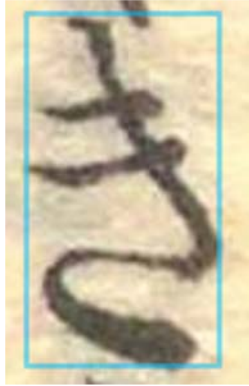
か  
Ka

香 閑 加 我  
家 歟 荷 駕  
嘉 可 佳 賀

字母

| か |   |   |   |
|---|---|---|---|
| 佳 | 香 | 閑 | 加 |
| 佳 | 香 | 閑 | 加 |
| 嘉 | 家 | 閑 | か |
| 嘉 | 家 | 歟 | か |
| 荷 | 我 | 閑 | 可 |
| 荷 | 我 | 閑 | 可 |
|   | 賀 | 駕 |   |
|   | 賀 | 駕 |   |

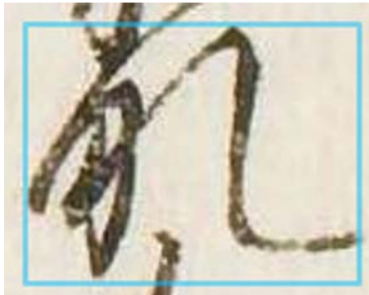
## くずし字の難しいところ 2 : 似たような文字が多い！



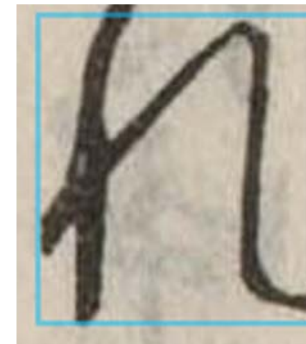
き



れ



な



れ



# くずし字の難しいところ 3 : 同じ文字なのに似てない！



前



前



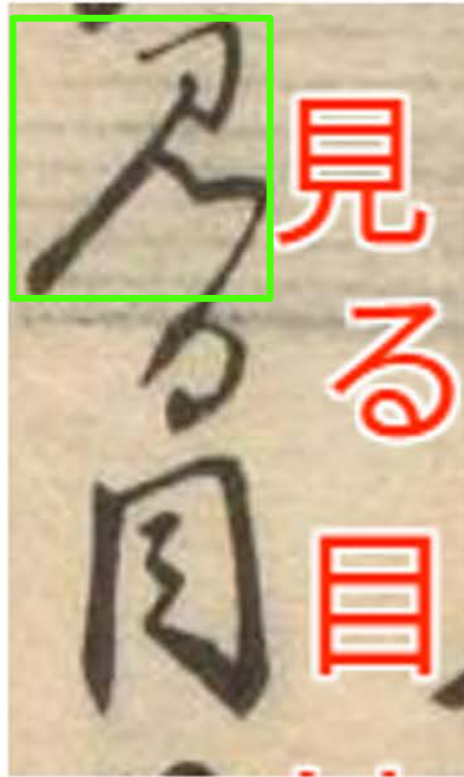
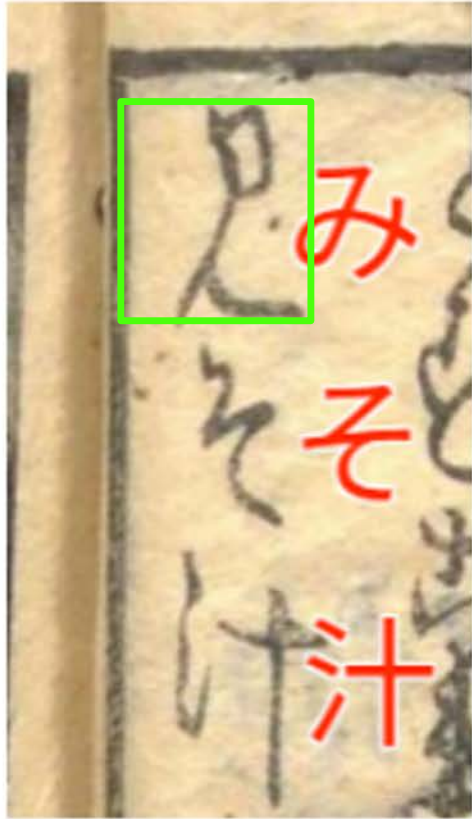
花



花

くずし字の難しいところ4：

同じ文字であっても、現代日本語だと違う文字になる。

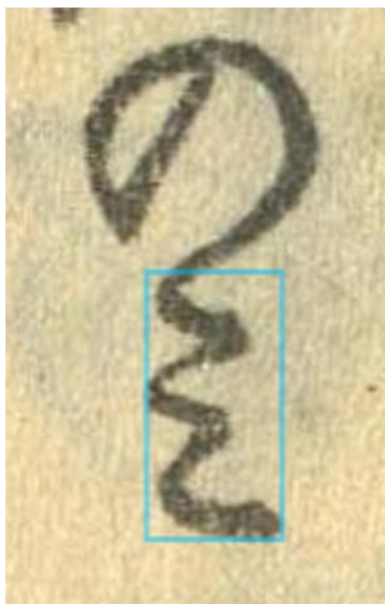


一文字ずつしか見てないから区別できないよ

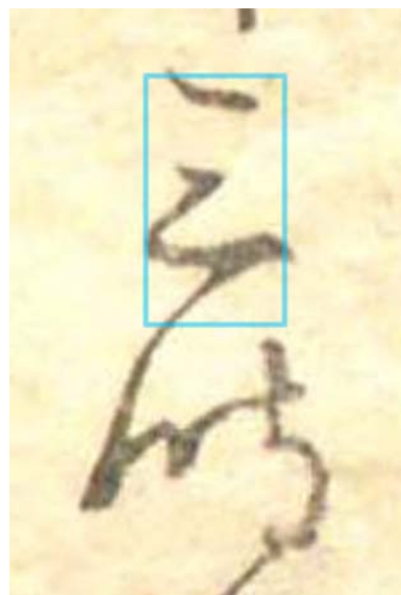


KogumaNet

# 区別しにくい文字



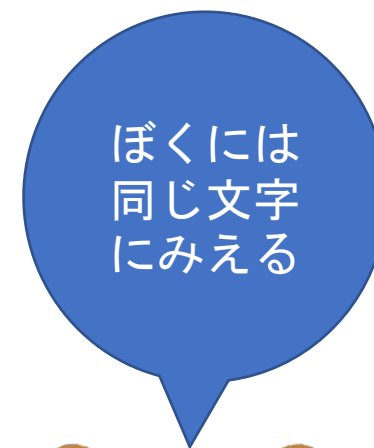
のみ



三時



キミ



KogumaNet

# AIを用いてくずし字を読む: くずし字一文字認識 (KogumaNet)

- Web Browser上でくずし字認識を行う。
  - 認識するのにデータ送信が不要。
  - モデルはとても小さい (4 MB)。
  - モデルのロード時間がかかる。
  - モデルを小さくするため精度が落ちる。
- 
- 他の研究にも使われている。



KogumaNet



# AIを用いてくずし字を読む: くずし字一文字認識 (KogumaNet)

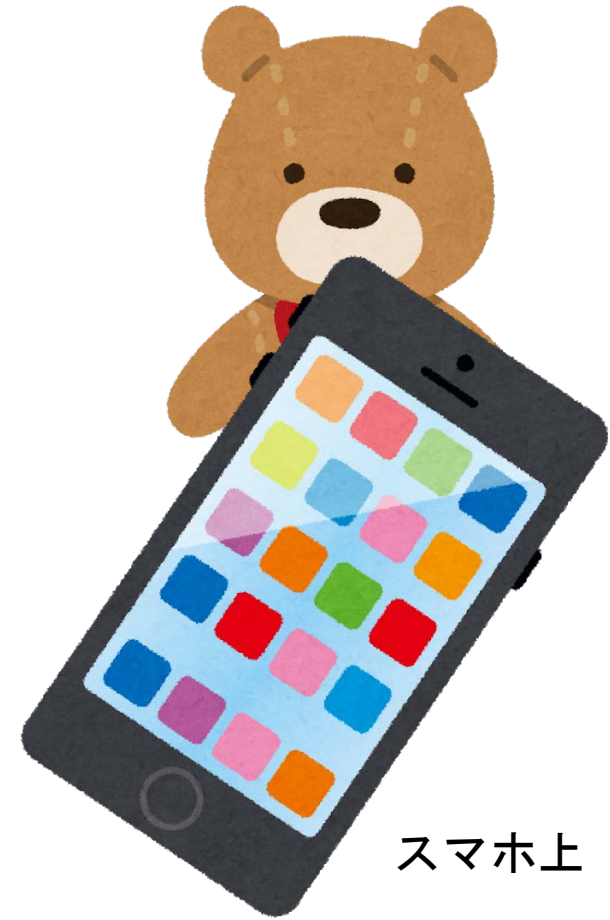
いろんなところ  
で活用できる!



KogumaNet



ウェブブラウザ上  
(サーバーなし)



スマホ上

# KuroNet



画像はイメージ。



KuroNet

- Tarin Clanuwat, Alex Lamb, Asanobu Kitamoto, KuroNet: Pre-Modern Japanese Kuzushiji Character Recognition with Deep Learning (ICDAR, Sep 2019) <https://arxiv.org/abs/1910.09433>
- Alex Lamb, Tarin Clanuwat, Asanobu Kitamoto, KuroNet: Regularized Residual U-Nets for End-to-End Kuzushiji Character Recognition (Springer Nature Computer Science Special Issue on Document Analysis and Recognition, May 2020) <https://link.springer.com/article/10.1007/s42979-020-00186-z>



# KuroNet



KuroNet Kuzushiji Recognition Viewer

伊勢物語 (Character List)

« Selected 1 / 1 » Selected Thumbnails



サーバーで認識するスマホアプリ  
(開発中)

IIIFでのKuroNetサービス

# モデルアンサンブル

前？

則？

前？



KuroNet



他のモデル



KogumaNet



「前」で  
いいかも



人間



# AIをもっと賢くするには...



KuroNet

データが  
欲しい



他のモデル

データが欲  
しい



KogumaNet

データが  
欲しい

# しかし...

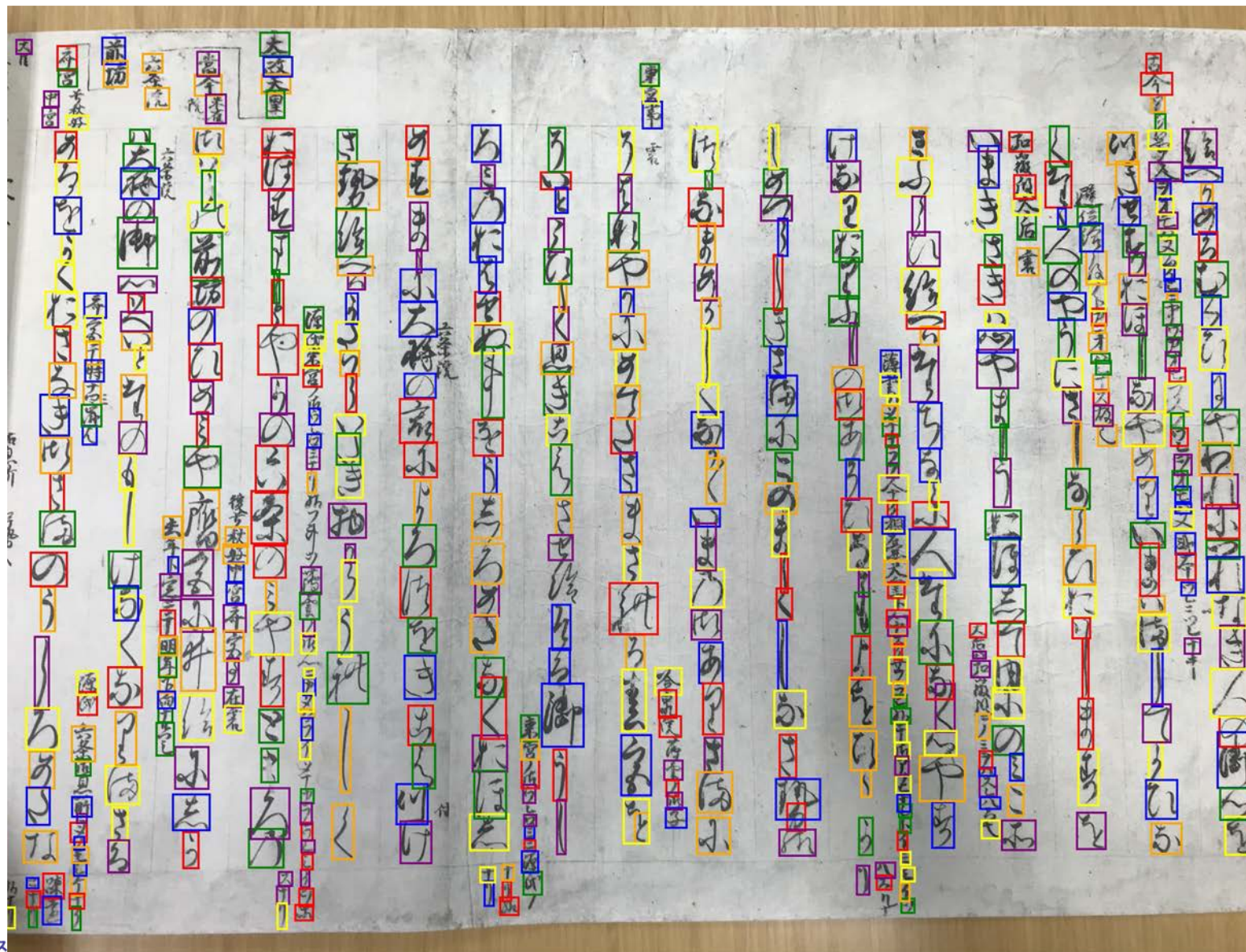
| Unicode | Image      | X    | Y    | Block ID | Char ID | Width | Height |
|---------|------------|------|------|----------|---------|-------|--------|
| U+6885  | umgy001_00 | 1402 | 398  | B0001    | C0001   | 192   | 155    |
| U+3053  | umgy001_00 | 1453 | 564  | B0001    | C0002   | 102   | 13     |
| U+3088  | umgy001_00 | 1464 | 689  | B0001    | C0003   | 55    | 124    |
| U+7F8E  | umgy001_00 | 1421 | 797  | B0001    | C0004   | 133   | 221    |
| U+306E  | umgy001_00 | 1448 | 1050 | B0001    | C0005   | 92    | 109    |
| U+5E8F  | umgy001_00 | 1418 | 1157 | B0001    | C0006   | 143   | 266    |
| U+5357  | umgy001_00 | 1169 | 397  | B0001    | C0007   | 152   | 101    |
| U+679D  | umgy001_00 | 1158 | 601  | B0001    | C0008   | 193   |        |
| U+306B  | umgy001_00 | 1214 | 751  | B0001    | C0009   | 58    |        |
| U+96EA  | umgy001_00 | 1166 | 973  | B0001    | C0010   | 153   | 14     |
| U+306E  | umgy001_00 | 1187 | 1126 | B0001    | C0011   | 101   | 92     |
| U+7A4D  | umgy001_00 | 1148 | 1228 | B0001    | C0012   | 170   | 166    |
| U+9803  | umgy001_00 | 1171 | 1390 | B0001    | C0013   | 134   | 157    |
| U+3088  | umgy001_00 | 1217 | 1539 | B0001    | C0014   | 44    | 2      |
| U+308A  | umgy001_00 | 1208 | 1633 | B0001    | C0015   | 57    | 102    |
| U+4E00  | umgy001_00 | 1177 | 1752 | B0001    | C0016   | 123   | 38     |
| U+8F2A  | umgy001_00 | 1161 | 1809 | B0001    | C0017   | 157   | 158    |



くずし字データセットを作成するのはとても大変な手作業。



# 文字の位置をAIで認識する。





みんなで翻刻データはAIを賢くする。

に？



KuroNet

て？



他のモデル

二？

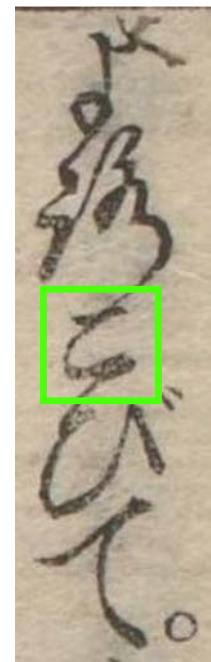


KogumaNet

これは  
「こ」だよ

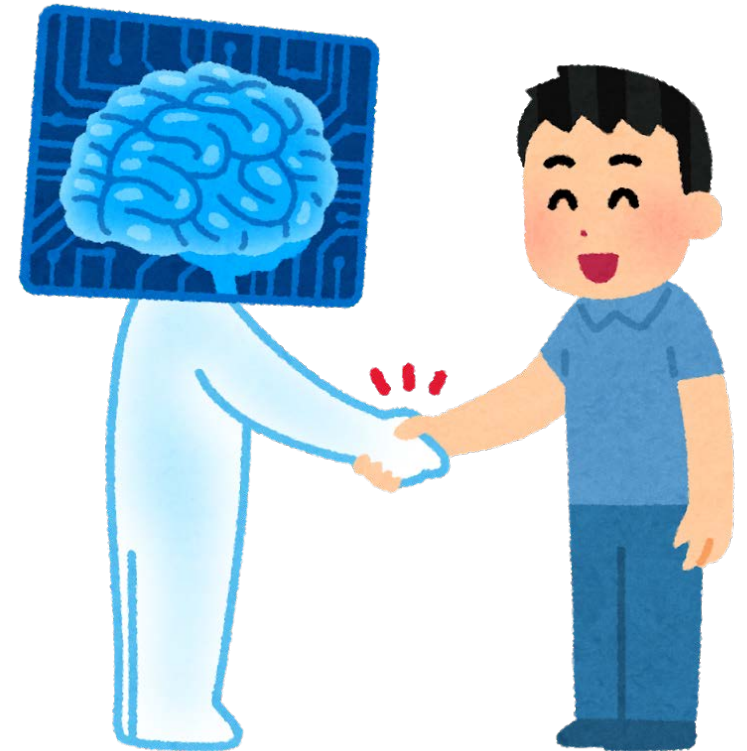
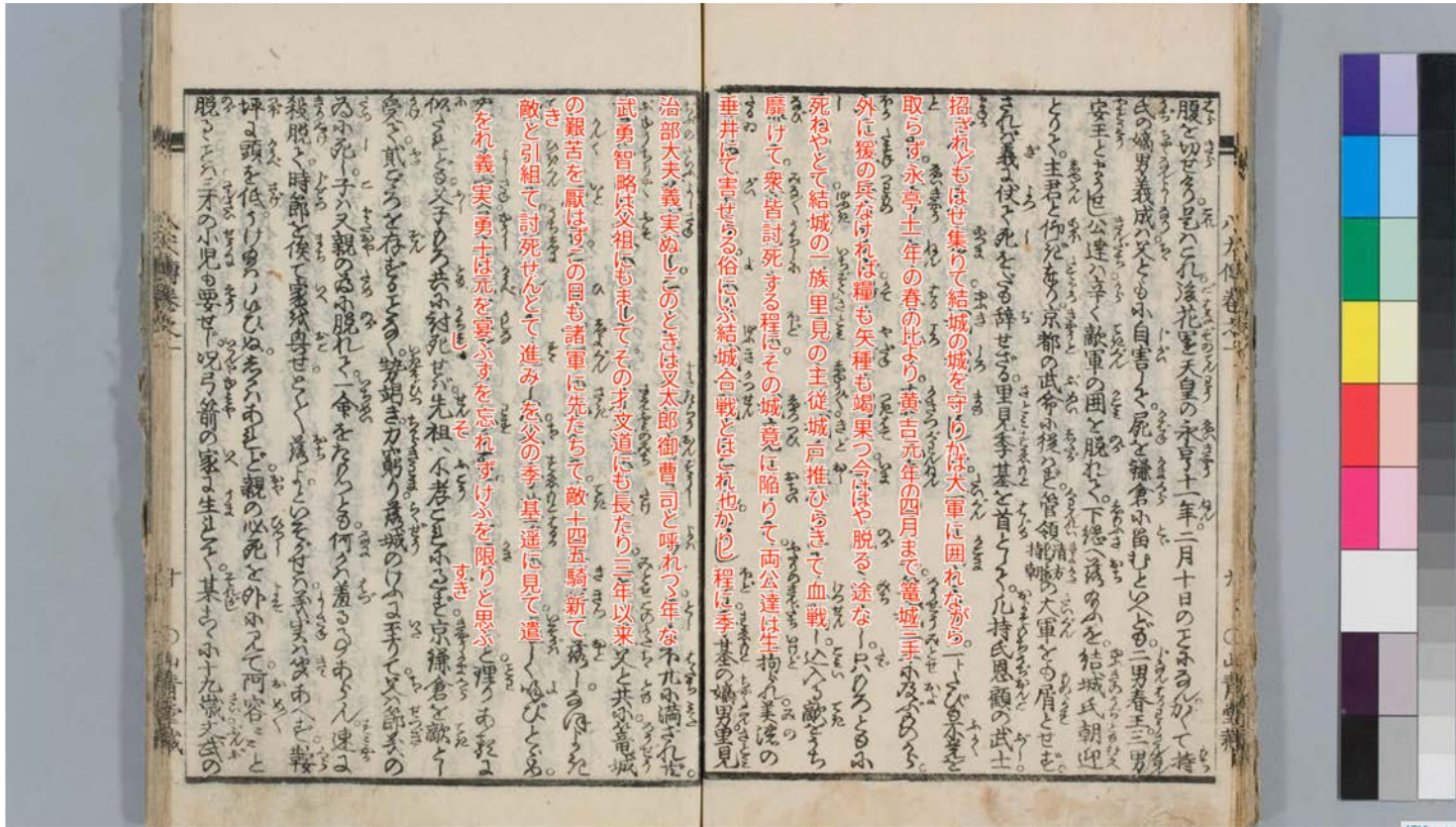


人間



# みんなで翻刻

AIを賢くするためにとても大事なデータです。



ありがとうございました。