

## くずし字のオープンデータとその活用

畑埜, 晃平  
九州大学基幹教育院

<https://hdl.handle.net/2324/2202982>

---

出版情報 : シンポジウム「オープンデータと大学」, 2019-01-30. Department of Library Science,  
Graduate School of Integrated Frontier Sciences, Kyushu University

バージョン :

権利関係 :

# くずし字のオープンデータと その活用



九州大学 基幹教育院

畑埜晃平

シンポジウム「オープンデータと大学」

2019.1.30@九大

※日本古典籍字形データセットより（国文研所蔵）

# 研究協力者

- **唐一平** (ライブラリーサイエンスM2)
- 石田栄美 (附属図書館/ライブラリーサイエンス)
- 中藤哲也 (情報基盤研究開発センター)
- 川平敏文 (人文科学研究院)

# 自己紹介

## □ 経歴 & 研究 :

九大システム情報研究院

(情報科学・機械学習)



→ 九大附属図書館研究開発室

統合新領域学府ライブラリー・サイエンス専攻

(+ 学術情報基盤)



→ 九大基幹教育院

(+ 学習データ分析)



※ 人文学に関しては全くの素人

# 概要

## 1. 背景

- デジタル・ヒューマニティーズ
- くずし字とその認識問題

## 2. くずし字のオープンデータ

## 3. くずし字のオープンデータの利用

# 背景： デジタル・ヒューマニティーズ

## □ デジタル・ヒューマニティーズ (DH)

### …情報技術を援用した人文学研究

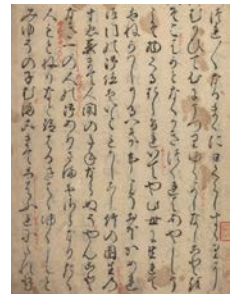
- 例：HarthiTrust (著作権付き人文学データを共有する枠組み)
- 国際会議 DH, EADH, JADH(日本発)など

## □ 古典籍を人間/機械可読な形で電子データ化することはDH発展に不可欠

## □ 日本でも日本古典籍のオープンデータが促進

- 歴史的典籍NW事業
- 人文学オープンデータ共同利用センター (CODH)

※ 古典籍におけるくずし字の自動認識がより重要に



徒然草

『日本古典籍データセット』  
(国文研所蔵)

# くずし字とは (\*)

- 文字資料のうち、楷書の点画を省略した手書き文字と、手書き文字をもとにした版本の文字
- 古典籍や古文書などの表記に用いられる



## 「あ」のくずし字

電子くずし字事典データベースより

<http://wwwap.hi.u-tokyo.ac.jp/ships/shipscontroller>

(\*)国文学研究資料館 平成27年度日本古典籍講習会テキスト

くずし字について「くずし字の見方・読み方」

[https://www.nijl.ac.jp/pages/event/seminar/2015/old\\_books\\_text.html](https://www.nijl.ac.jp/pages/event/seminar/2015/old_books_text.html)

# くずし字認識問題

テキスト（翻刻）

画像



認識機



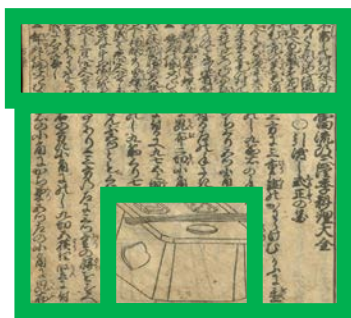
つれづれなるままに



# 文字認識の概要

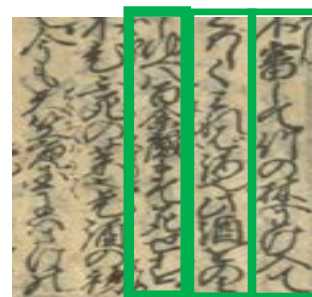
1. レイアウトの解析  
(文章・絵の切り分け)

簡単



2. 行の切り出し

簡単



4. 文字の認識



3. 文字の切り出し

難しい



参考 : <https://mediadrive.jp/technology/techocr05.html>

画像 : 節用料理大全 (国文研所蔵)

# 単一文字の認識

□文字認識はすでに実用レベル

□背景：機械学習手法の発展

- サポートベクトルマシン

- ディープラーニング

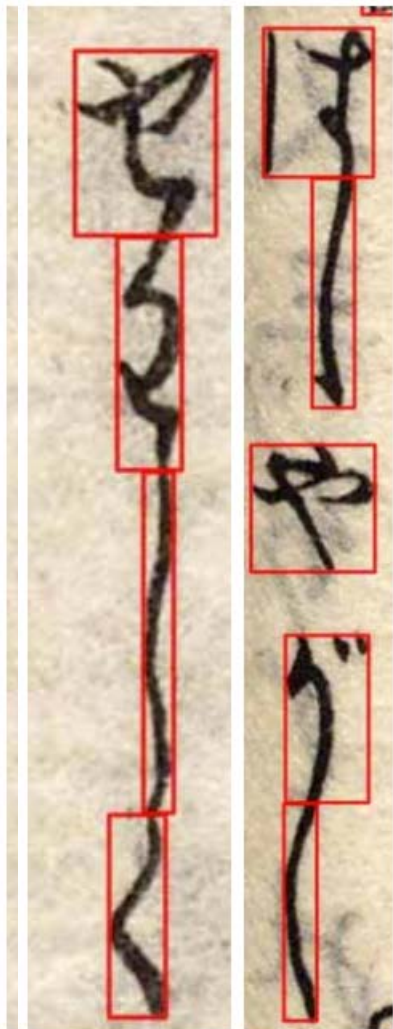
□アラビア数字(0-9)の手書き文字…99%(MNIST)

□くずし字：

- 変体仮名ひらがな48種類…70-80%[早坂ら 16]

- CODHデータセット頻出上位10文字…96-7% [北本 17]

# 問題点：文字の切り出し



□“長期的に欲しいのは、文字の認識の自動化に加えて、**文字の切り出しの自動化**である。”

–北本, “日本古典籍字形データセットの公開と活用への期待”  
第2回CODH セミナー くずし字チャレンジ  
～機械の認識と人間の翻刻の未来～

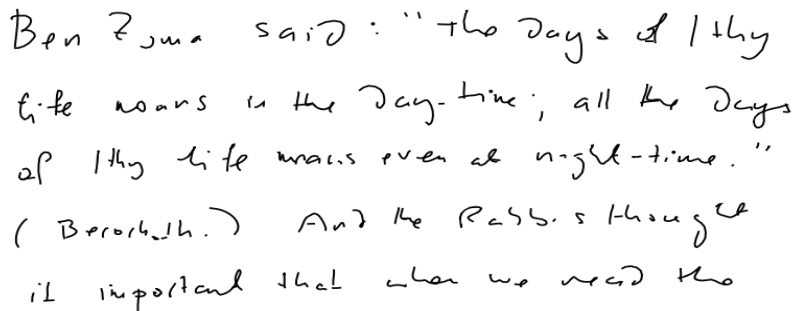
2017. 2

# 参考：非日本語の場合

## 英語：

"For the IAM Online Handwriting Dataset, our best result was a character error rate of 9.26% on the test set. The best previously published result is 11.5% character error rate by Graves using a different and much more extensive preprocessing."

— [Greff+15], LSTM: A Search Space Odyssey15



Ben Zoma said: "The days of lthy life means in the day-time; all the days of lthy life means even at night-time." (Berochoth.) And the Rabbis thought it important that when we read the

(a)

Ben Zoma said: "The days of lthy life means in the day-time; all the days of lthy life means even at night-time ." (Berochoth .) And the Rabbis thought it important that when we read the

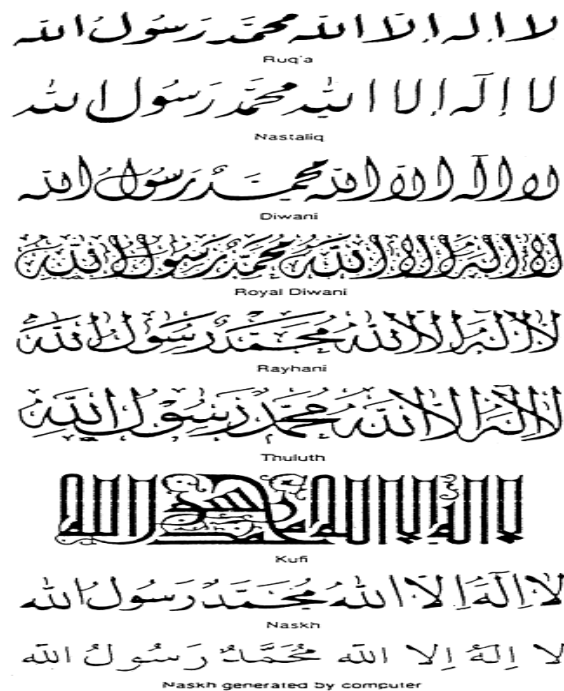
(b)

Fig. 2. (a) Example board (a08-551z, training set) from the IAM-OnDB dataset and (b) its transcription into character label sequences.

## アラビア語：

Arabic database resulting in an average character error rate of 1.9%.

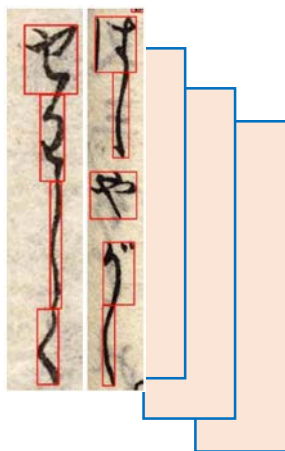
—[AdnanAmin98] Off-line Arabic character recognition: the state of the art,1998



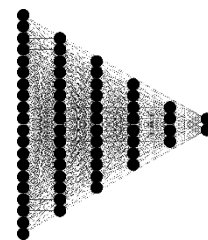
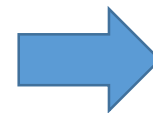
# 機械学習に基づく くずし字認識問題へのアプローチ

□機械学習：“一を聞いて十を知る”ための情報技術

学習



切り出し情報・テキスト  
付きくずし字画像



文字区切り/認識  
ルール

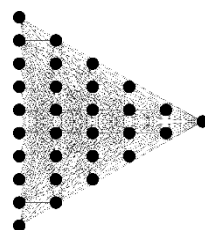
(一般に)  
多量に必要

文字区切り, テキスト

予測



未知のくずし字  
画像



け  
した

# 概要

## 1. 背景

- デジタル・ヒューマニティーズ
- くずし字とその認識問題

## 2. くずし字のオープンデータ

## 3. くずし字のオープンデータの利用

# くずし字オープンデータ

## 人文学オープンデータ共同利用センター(CODH)

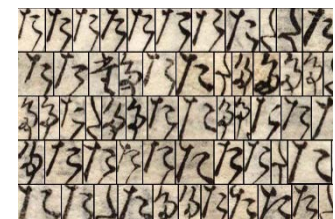
### □日本古典籍データセット

- 日本の古典籍3126点（2019.1現在）
- 画像データ（約60万）、書誌データ、テキストデータ（一部）
- オープンデータ（CC-BY-SA）



### □日本古典籍字形データセット

- 日本古典籍データのセットの28点から得たくずし字4645文字種の字形データ約68万文字（2019.1現在）
- オープンデータ（CC-BY-SA）



# くずし字オープンデータ (2)

- KMNISTデータセット (new!)
  - 機械学習用くずし字データセット
  - Kuzushiji-MNIST: 10種類のくずし字 約70万個
  - Kuzushiji-49: 49種類のくずし字 約27万個
  - Kuzushiji-kanji: 3832種類の漢字 約14万個
  - オープンデータ (CC-BY-SA)

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, David Ha, "Deep Learning for Classical Japanese Literature", arXiv:1812.01718.



# CC-BY-SA 4.0 (簡略版)



- あなたは以下の条件に従う限り、**自由に**：
  - **共有** — **複製**したり、**再配布OK**
  - **翻案** — 資料を**リミックス**したり、**改変可能**
  
- あなたの従うべき条件は以下の通りです。
  - **表示** — あなたは**適切なクレジット**を表示し、ライセンスへのリンクを提供し、変更があったらその旨を示さなければなりません。
  - **継承** — もしあなたがこの資料をリミックスしたり、改変したり、加工した場合には、あなたはあなたの貢献部分を元の作品と**同じライセンスの下**に頒布しなければなりません。
  - 追加的な制約は課せません

## 2次利用可能なライセンスの代表例

<http://creativecommons.org/licenses/by/4.0/deed.ja>

## 第21回 PRMUアルゴリズムコンテスト

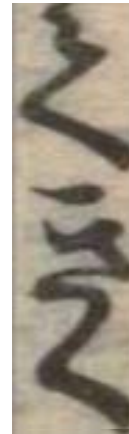
この文字読めますか？～くずし字認識にチャレンジ！～

- 2017年 電子情報通信学会 パターン認識・メディア理解研究会 (PRMU) がくずし字の認識コンテストを開催
- CODHのくずし字データ・セットを利用



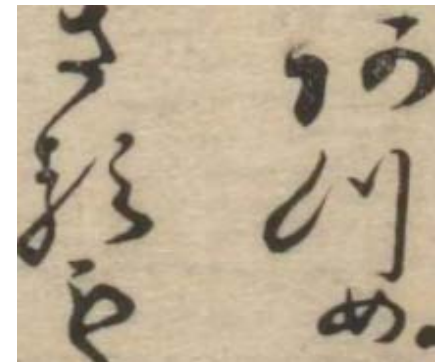
LV1 : 1文字の認識

認識率 **97.2%** (1位),



LV2 : 3文字認識

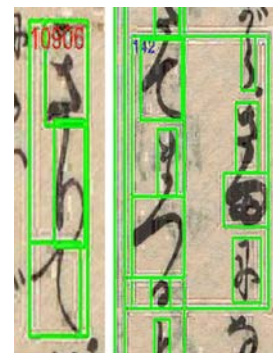
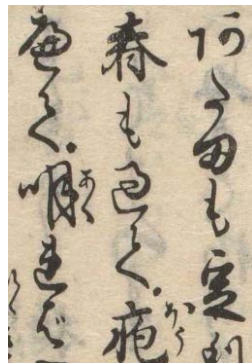
認識率 **87.6%** (1位)



LV3 : 4文字以上の認識

認識率 **39.1%** (1位)

# 我々の取り組み (1): 文字切り出し情報つきデータの作成



CODH日本古典籍字形  
データセット  
文字切り出し情報あり,  
学習用データなし

PRMUコンテスト  
データセット  
文字切り出し情報なし  
学習用データあり  
(LV1,2,3)

**2次加工データ**  
**文字切り出し情報あり**  
**学習用データあり**  
**(LV1,2,3)**

- 77953 枚の3文字画像(Lv2)と12582枚の多文字画像(L3)を作成
- 人手によるダブルチェックにより判別困難なデータを除去
- オープンデータとして公開予定

**Construction of Japanese Historical Hand-Written Characters Segmentation Data from the CODH Data Sets**

T. Yiping, K. Hatano, E. Ishita, T. Nakatoh, T. Kawahira, JADH 2018.

# 概要

## 1. 背景

- デジタル・ヒューマニティーズ
- くずし字とその認識問題

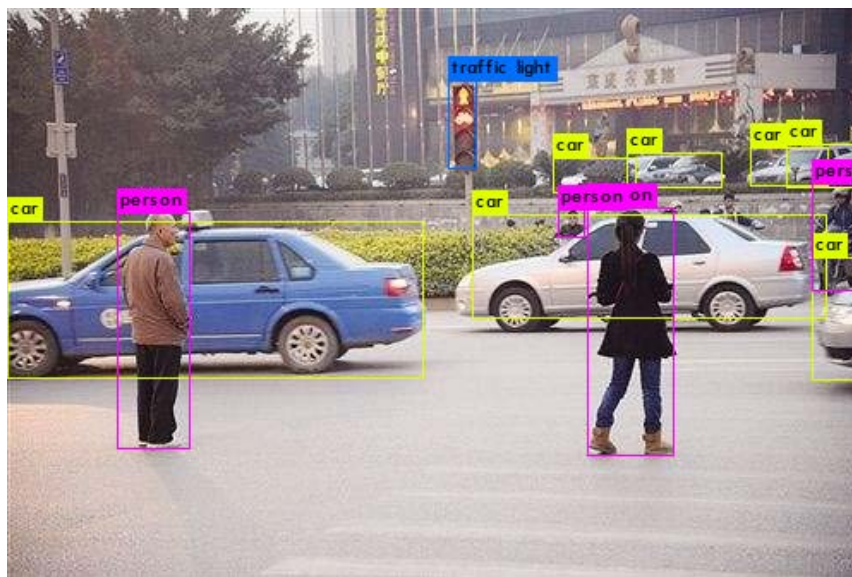
## 2. くずし字のオープンデータ

## 3. くずし字のオープンデータの利用

# 我々の取り組み (2): 文字切り出し情報つきくずし字データの利用

## □ アイデア :

画像認識分野における物体認識の手法を応用



## □ くずし字画像に対して

切り出しと認識を同時にできるのでは？

# 予備実験

- 7万枚のLV2文字画像(3文字)を学習させた認識ルールを7000枚のテスト画像データに適用した結果

手法	誤認識率(LER)
物体認識手法 (YOLO)に基づく文字区切り/認識手法	4.29 %
物体認識手法 (YOLO) <b>+<math>\alpha</math></b>	<b>0.7%</b>

参考：くずし字認識コンテスト優勝チームの誤認識率…12.4%

**文字切り出し情報つきデータが認識率の向上に寄与**

# まとめ

- 九大におけるくずし字オープンデータの利用事例
  - 解きたい問題に合わせてオープンデータを2次加工
    - ※ CC-BY-SAライセンスのもとでは加工データもオープンデータ化可能
  
- デジタル・ヒューマニティーズの発展に向けて
  - オープンデータは異分野研究者の参入を容易に
  - 「問題」の共有（オープンプロブレム）も重要