

訓点資料内の片仮名訓点データのデータベースへの追加 Add Katakana-Character of Classical Chinese Text to Kunten Database

2023Y37 森 駿 (Mori Shun)

担当教員 出口 利憲 (Deguchi Toshinori)・安田 真 (Yasuda Makoto)

1. 序論

漢文に語順や送り仮名、読み方などの注釈を、文字や記号で付与し、日本語として理解しやすくなった資料を訓点資料と呼ぶ。訓点資料の分析は、記述内容の正確な理解を目指し、加点内容を理解することを中心に行われてきた。これまでにヲコト点図や積文を作り、資料を解釈する方法が確立され、著名な訓点資料の現代語訳が作成されている。これらは、書籍や Web などで公開されている。一方で、現代語訳は訳者の意図や他の文献の記述が混在しているため、訓点資料そのものの研究や、資料に記された日本語の読み方の変化を調べるには適していない。これらの研究用途のために資料に付与されたヲコト点などの加点情報を電子化する手法が検討されており、漢字やヲコト点の座標、形状から資料に付与された訓点を検索するデータベースが開発されている^[1]。

引継ぎ元の研究では、訓点資料の本文内にある文字を検出し、その位置情報をデータベースにまとめることで Figure1 のように個別の文字へのアクセスを実現し、文字毎の比較等を可能とした。しかし、依然としてその周囲にある片仮名などを含む訓点に関しては未検討のままであり、訓点資料の情報を網羅するには至っていない^[2]。

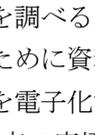
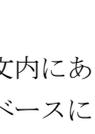
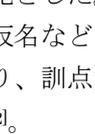
No	Page	Place	Character	Link	Image
巻1	2才	01行18文字	平	Link	
巻1	2才	02行24文字	平	Link	
巻1	2才	07行08文字	平	Link	
					

Figure 1 Image for the character.

Table 1 New table

Elements	Type
kana_charc_id	int(11)
id_kanaelements	int(11)
character_index	tinyint(4)
charc	varchar(1)
x1	double
x2	double
y1	double
y2	double

2. 目的

本研究では、検出された訓点資料内の文字の周囲にある訓点のうち、「片仮名」に関してその正確な座標をデータベースに登録し、訓点資料データベースの可用性を高めることを目標とする。同時に、漢字・平仮名と比較して一次資料に基づくデータが欠けている片仮名に関するデータベースの構築も目標とする。

3. 現状の課題

訓点資料内のどの文字の周囲にどんな片仮名が存在しているかは、事前作業によってデータベースに登録済みである。そのため、本研究では片仮名に関する情報を保存するテーブルをデータベースに作成し、そこに個別の情報を登録し、活用する方法を実装する必要がある。

4. 機能の構成

4.1 データベースの設計

現在のデータベースには、一連の片仮名訓点の内容と、その対象となっている文字列の所在が要素として保存されている。そのため、追加するテーブルは Table 1 に示す要素構成で設計し、片仮名の 1 文字ごとの正確な座標を保存しておき、「id_kanaelements」から上にたどれるようにした。また、それぞれに対して注釈をつけられるように、「id_kanaelements」ごとに「note」をつけるテーブルも追加した。

4.2 APIの実装

データベースとのデータのやり取りを行うために、Ubuntu22.04+Apache+PHP 8.3 を用いてAPIサーバーを実装した。すべてのパラメータはJSON形式でやり取りを行い、認証関連のAPI以外の利用にはアクセストークンが要求される。

片仮名訓点の登録時「kana_charc_id」を用いてクライアントから片仮名の指定を受け取ると、「id_kanaelements」をたどって紐づいている文字列の画像を返す。その後クライアントから文字を切り出した結果の座標を受け取り、データベースに保存することでデータの登録を行う。

この時に「id_kanaelements」をたどって得られる画像は、訓点資料の本文の文字を対象に切り出したものなので、訓点は画像から見切れていることも多い。そのため「img_size」というパラメータを設定し、そのままの画像を0として、数字を1増やすことで1周り外側の画像までを渡せるようにし、見切れている片仮名も切り出せるようにした。

また、画像を切り取った際に送られてくる座標はクライアントに渡した画像内の座標となっているため、渡した画像のサイズをもとに補正を行い、ページ全体の画像における座標として保存されるようになっている。

4.3 アプリケーションの実装

実際に画像から片仮名を切り出すため、Next.js + Typescript で Web アプリケーションを作成した。Figure 2 は、実際にアプリを用いて片仮名を切り出している様子である。このアプリでは、左のタブで「kana_charc_id」を指定することで片仮名を指定し、文字列の画像を呼び出す。呼び出された画像は右側に表示され、この画像内の2点を

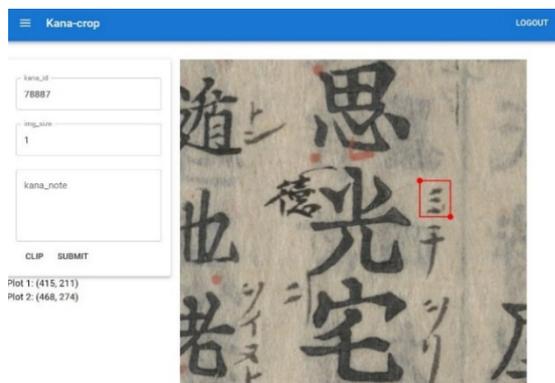


Figure 2 Screenshot of application 1.

をクリックすることで領域の指定が可能である。これを用いて、図に示すように片仮名を切り出し、その座標を保存する。また、「img_size」や「kana_note」では前述した画像サイズの指定や、注釈の入力が可能である。

5. まとめ・今後の展望

本研究では、訓点資料内の文字に振られた片仮名訓点を切り出しデータベースに登録するアプリケーションを実装し、データベースを構築した。しかし、作成したデータベースを扱うアプリケーションの実装が不十分で、Figure 3のような無骨な実装となってしまう、訓点資料データベースの可用性に関して改善点が残った。今後はアプリケーションの未実装機能の開発を進め、構築したデータベースの可用性を高めていく予定である。

また、今回実装した範囲の機能に関しても、片仮名や領域の指定方法などでUXに改善点が残っているため、リファクタリングを行い、アプリケーションの完成度をより高めることも予定している。

参考文献

- [1] 田島 孝治, 堤 智昭, 高田智和. “訓点資料の書き下し文自動生成を目的としたマコト点を中心とする訓点の計量分析”, 情報処理学会論文誌, Vol.61, No.2, PP.162–170, (2020-02-15)
- [2] 苫米地 康太, 田島 孝治. “訓点資料画像の文字位置検出とデータベースへの追加”, 情報処理学会研究報告, Vol.2023-CH-132, No.14, PP.1-4, (2023-05-13)

文字一覧

参照キー: 1-F-01

- 巻 1 : 1 才01
 - 書
 - 巻
 - 第
 - 一
- 巻 1 : 1 才02
 - 堯
 - 典
 - 第
 - 一
 - 〇
 - 虞
 - 書
 - 〇
 - 孔
 - 氏
 - 傳
- 巻 1 : 1 才03
 - 昔
 - ((右) ムカシ, [X王(あり)])
 - △{画像}(2221, 556) -> (2270, 600);
 - 九{画像}(2213, 599) -> (2273, 662);
 - ≧{画像}(2213, 665) -> (2283, 731);
 - 在

Figure 3 Screenshot of application 2.