

書き下し文での訓点情報検索を可能とする 訓点資料データベースの試作

中村 海翔, 田島 孝治 (岐阜工業高等専門学校),

堤 智昭 (筑波大学), 高田 智和 (国立国語研究所), 小助川 貞次 (富山大学)

概要: 本稿では, 書き下し文を用いて訓点情報を検索することができる訓点資料データベースの試作について述べる. 今回は国立国語研究所蔵『尚書(古活字版第三種本)』を用いて検索フォームに書き下し文またはキーワードを入力することで該当箇所を表示できるデータベースを実装した. このデータベースは Web ブラウザからアクセス可能であり, 検索結果より IIIF Curation Viewer を用いて切り出した資料画像を参照することができるようになっている. 現在は巻 1 の冒頭部分の公開を行っており, 今後も改良を続けていく予定である.

キーワード: 訓点資料, フコト点, データベース

A prototyping database of Chinese text that enables kunten information search by transcription of Chinese classics into Japanese

Nakamura Kaito, Tajima Koji (Gifu National College of Technology),

Tsutsumi Tomoaki (University of Tsukuba),

Takada Tomokazu (National Institute for Japanese Language and Linguistics),

Kosukegawa Teiji (University of Toyama)

Abstract: This paper describes a proto-type database for kunten information search by transcription of Chinese classics into Japanese. We implemented a web-based database that can show the relevant sections of transcription and cropped materials using IIIF Curation Viewer by inputting the transcription or keywords. The current database can search into the first part of Volume 1 of the “Shangshu” (old type print version, Type3), which is owned by the National Institute for Japanese Language and Linguistics. the future, we plan to add functions and proofread data, and release all volumes.

Keywords: Kuntenmaterial, Wokototen, Database

1. まえがき

漢文に語順や送り仮名や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い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, 加減内容を理解することを中心に行われてきた. これまでにフコト点図や釈文を作り, 資料を解釈する方法が確立され, 著名な訓点資料の解説文が作成された. これらは, 研究資料として書籍等で公開されている[1-3].

一方で, 解説文は解説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などを調べるには適していない. これらの研究用途のために, 資料に付与されたフコト点などの加減情報を電子化する手法が検討されており, 漢字やフコト点の座標, 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の開発や分析を行ってきた[4-7].

これまでに作成してきたデータベース[8]は, フコト点や語順点などの訓点情報を直接指定して検索を行うことはできるが, 自然な文章での検索

が不可能で, フコト点に精通した利用者でなければ利用が難しかった. 今回は, 訓点資料の電子データを漢字+送り仮名といった現代語に近い形で検索できる可用性の高いデータベースを試作した. 訓点資料を利用したい他分野の研究者であっても, この形式での検索であれば訓点資料内の検索が容易となる. また, 切り出された資料画像を検索結果より参照できるため, 訓点情報を書き下し文と比較しながら確認を行うことも可能である.

本稿では試作したデータベースの実現方法と, 利用方法について述べ, デモンストレーションにより報告を行う.

2. 対象とした資料について

今回は国立国語研究所蔵「尚書(古活字版)」を対象として研究を行う. 尚書は書経とも呼ばれ政治史・政教を記した中国最古の歴史書で, 序文と 58 の通篇で構成される. 今回対象とするものは, 1596[慶長元]-1615[慶長 20]年刊のものであり, 巻 1 から巻 9 までの画像データが公開されている.

冊子本であるため、画像データは1丁に対し表裏が存在し、半丁あたり8行構成となっている。

今回の試作では、書き下し文作成のために行った精密解読が巻1の本文冒頭部分、具体的には1丁目と本文が連続する関係で2丁目の最初の割注終わりまでであるため、この範囲を対象とする。

3. データベースの必要条件

これまでに作成したデータベースにおける実用上の課題は、次の2点である。

- (1) フコト点に精通した利用者でなければ扱いが難しく、検索が容易でない
- (2) 検索結果は画像の全体のリンクであり、画像そのものが表示や切り出しが行われておらず原本の確認が難しい

(1)の課題は、検索時に訓点の体裁、訓点資料の特定の範囲の選択が必要で、フコト点など訓点に精通していなければ入力項目が分かりにくいことに依存する。これらの項目は文字や文と異なり、直感的に入力することが難しく、複数個の入力では手間がかかる。特に、訓点を専門としない日本語史、音声言語研究などの研究者や情報学研究、訓点資料を学び始めた人などにとってはフコト点を基軸とした検索は困難である。

(2)の課題は資料画像のリンクが1ページ単位で定められおり、検索結果では画像を直接閲覧できない、また部分切り出しが行われていないことにより、該当箇所をページ内から探す必要があることが原因である。検索した文字や文章が複数回、同一ページに登場することもあり、検索時間に比べて、画像を開き、該当箇所を探し出す時間の方が長いという状況になることも多く、現在のデータベースではスムーズな資料画像の閲覧ができるとは言い難い。

そこで、今回新たに設計するデータベースでは書き下し文による検索を実装する。これにより自然な文章の形で検索できる。ただし、今回のデータベースにおいては現代語訳を掲載するわけではないため、検索においては現代語訳ではなく、本文中の文字を用いた書き下し文の一部を用いて検索することにする。また、資料画像の表示はあらかじめ該当箇所に対して切り出された状態の画像を用意する。書き下し文とともに検索結果として画像が表示されることで、利用者がページ内から探す必要はなくなり、書き下し文と資料の該当箇所を簡単に比較できるようにする。

4. データベースの設計

4.1 実装環境

実装するデータベースは、既存のデータベースとの将来的な連携を意識して、同一環境で動作させることにした。データベースの実装環境を表1に示す。一般的なLAMP環境を構築し、関係データベースを利用して検索を実装している。

表1 データベースの動作環境
Table 1 A Database Operating Environment.

ソフトウェア名	version
Ubuntu	20.04.2
MariaDB	10.4.26
Apache	2.4.41
PHP	8.0.10
PHPMyAdmin	4.9.5deb2

4.2 検索方法

実装するデータベースはWebブラウザを用いて検索、表示ができるものとし、検索キーワードなどは、検索フォームに任意のテキストを入力させる。テキストから書き下し文を検索する方法としては、LIKE検索とN-gramによる全文検索が行えるように実装を行う。

LIKE検索はあいまい検索であり、正規表現に対応したキーワードで、データベース中のすべての書き下し文を検索する。一方でN-gramによる全文検索では、あらかじめ書き下し文をN-gramに分割しておき、該当キーワードと一致する部分文字列を含む書き下し文を検索する。いずれの場合もスペース区切りで複数のキーワードを入力可能とし、検索されたすべての結果を表示することにする。

4.3 データベースのテーブル設計

今回のデータベースでは、整合性を考慮するとともに、N-gramによる全文検索を行うために、書き下し文と該当箇所の画像のURLが登録されたテーブルとn文字にトークナイズされた書き下し文が登録されたテーブルの2つをRDB (Relational Database) で保持する。今回のデータベース中のテーブルの構造を表2に、書き込んだデータの例を表3にまとめる。test_sentenceというテーブルには書き下し文と該当箇所の画像のURLを登録してあり、資料中のどこに記述されている文かを示すユニークなIDを付与した。このIDの構造は「1-1-F-1」のように、巻、ページ番号、表裏(表がF裏がB)、文番号をハイフンでつなぐこととした。文番号に関しては、文の長さがそれぞれ違うことや割注のように1行が2行に分かれる箇所があり、列や行など定まった物理量で定義すると分かりにくくなる恐れがあったため、最初の文を1として順番に番号を振る形をとった。

テーブルに登録する書き下し文の定義は以下のようにになっている。

- (A) 句読点がある箇所で区切る
- (B) 割注の始まり、終わりで区切る
- (C) 助字は含めない
- (D) 科段点は含めない
- (E) 音合符や訓合符は含めない

表 2 DB のテーブル構造

Table 2 The data structure of the DB Tables.

Table 名 : test_sentence	
要素名	型 (長さ)
id_sentence	varchar (8)
Sentence	varchar (30)
sentence_image_1	text
sentence_image_2	text

Table 名 : test_ngram	
要素名	型 (長さ)
id_ngram	varchar (10)
n_length	varchar (2)
text	varchar (30)
id_sentence	varchar (8)

表 3 DB の要素例

Table 3 An Example data of the DB Tables.

Table 名 : test_sentence	
要素名	値の例
id_sentence	1-1-F-6
Sentence	將に位を遜とし て虞舜に譲る
sentence_image_1	https://dglb01.ninja l.ac.jp/iiif/syousy o/001/tiff/sysy001-01 3.tiff/1975, 3359, 2 82, 416/full/0/default. jpg
sentence_image_2	https://dglb01.ninja l.ac.jp/iiif/syousy o/001/tiff/sysy001-01 3.tiff/1678, 450, 26 4, 1180/full/0/default. t.jpg

Table 名 : test_ngram		
要素名	値の例	値の例
id_ngram	N-1-1-F-6	N-1-1-F-6
n_length	2	5
text	將に	將に位を遜
id_sentences	1-1-F-6	1-1-F-6

(A)と(B)に関しては、書き下し文の 1 文が長くないように考慮し、(C)～(E)に関しては、検索を実装するにあたり、文法の構成要素となる文字や記号を含めると自然な文での検索、容易な検索という目的から遠ざかると考えたため、これらのように一文を定義した。

画像 URL は、検索結果に直接資料画像を表示するためにデータとして登録を行った。今回対象としている『尚書（古活字版第三種本）』は単なる画像ファイルでの公開、専用ビューアでの公開に加えて、IIIF マニフェストが用意されている。そこで IIIF Curation Viewer の部分矩形領域指定機能を用いることによって書き下し文と対応している資料画像の範囲を切り出した上で、その座標が含まれた URL を取得することとした。この方法であれば、別の訓点資料のデータベースを作った際に、画像表示の枠組みを流用できるという利点がある。

次に、test_ngram というテーブルである。このテーブルはテーブル名の通り、N-gram を用いるために設計したテーブルであり、こちらもユニークな ID を id_ngram として振ってある。N-gram については、これは任意の文字数で文章を分割する手法である。test_ngram テーブルの text の箇所をそれぞれ見ると分かるように、2 文字と 5 文字で書き下し文が分割されている。分割した文字と一致するキーワードが入力された時、該当箇所の書き下し文を表示するための要素である。今回は分割数 n を要素名 n_length とし、2～5 の間で設定しテーブルを設計した。

分割数については、書き下し文の長さに合わせて最大値を変更することができるため、その上限値についてはデータ数を考慮したうえで今後決定していく予定である。特にデータ数が増えた場合に検索文字数 (n の値) による検索速度の比較し、適切な値に設定する必要がある。また、RDBMS には MariaDB を採用しているため、Mroonga プラグインを有効にすることで ngram parser による全文検索も可能であるが、実装上の問題と、パフォーマンスの検証を行うために、今回の実装では自身で新たにテーブルを作成する手法をとった。

新たなテーブルをデータベースに追加し同一ページ内で訓点情報も閲覧できると、書き下し文が作られた課程をより詳細に確認することができ、便利であると考えた。訓点情報の電子データは先行研究により得られているため、今回のデータベースに合わせて外部キーを登録するなど手直しを入れるのみで対応できる。現段階で想定している訓点情報のテーブル設計を表 4 に示す。これらテーブルは現在のデータベースに登録されていないが、今後実装予定である。テーブルの文字情報のみでは視覚的にわかりにくいいため、検索画面とは異なる新たなインターフェースを設け

表 4 訓点情報のテーブル
Table 4 Table of kunten information

Table 名:charac		
要素名	型(長さ)	値の例
id_charac	varchar(12)	1-1-F-3-19
Charac	varchar(1)	將
linename	varchar(8)	卷 1:1 オ 03
lineNumber	int(3)	3
columnNumber	int(2)	19
Warityu	int(1)	0
warityu_kaigyo	int(1)	0
id_sentence	varchar (8)	1-1-F-6

Table 名:elements		
要素名	型(長さ)	値の例
id_elements	int(5)	17
style	varchar(1)	朱
mark	varchar(1)	・
X	int(2)	-2
Y	int(2)	-2
id_charac	varchar(12)	1-1-F-03-19

Table 名:gojunelements		
要素名	型(長さ)	値の例
id_gojunelements	int(5)	5
style	varchar(1)	墨
mark	varchar(1)	レ
id_charac	varchar(12)	1-1-F-03-19

Table 名:kanaelements		
要素名	型(長さ)	値の例
id_kanaelements	int(5)	7
target_length	int(1)	1
position	int(1)	1
position_text	varchar(3)	左
style	varchar(2)	墨
text	varchar(12)	シ
id_charac	varchar (12)	1-1-F-03-19-

ることで利用者が理解しやすいような形を目指すとともに、検索条件の絞り方や細かな仕様を検字情報のみでは視覚的にわかりにくいいため、検索画面とは異なる新たなインターフェースを設けることで利用者が理解しやすいような形を目指すとともに、検索条件の絞り方や細かな仕様を検討する必要がある。これらについては、今後の展望として 6 章で述べる。

4.4 検索用の SQL 文

データベースに対する検索の種類は Like 検索によるあいまい検索とスペース区切りによる複数検索、N-gram を用いた一致検索とスペース区切りによる複数検索の計 4 種類が存在する。これらにおいて作られる SQL 文を順番に説明する。

LIKE 検索による検索は、あいまい検索、複数検索ともに同じ SQL 文の構造を持ち、

```
SELECT * FROM test_sentence
WHERE Sentence LIKE Keyword
```

とした。単一キーワードの検索においては、Keyword に単一の単語を指定する。ただし、SQL インジェクションや、不正な文字による DB 全体の検索ができないようにした。具体的には PHP の preg_replace 関数により、入力されたキーワードに対して、正規表現で(?=[! %])の文字を取り除き、%を前後に加えて、文字列がどこかに含まれてキーワードとして検索する。

スペース区切りによる複数検索の場合は、OR 演算子を使って検索式自体を複数結合した SQL 文を作り、検索を行う。

いずれの場合も、データベースへの接続は PHP のライブラリである PDO を用いて行い、SQL 文のチェックを行ったうえで検索変数を代入することで、不正な SQL 文が実行されないようにした。

検索結果は、主キー以外は表示用のデータであり書き下し文、資料画像の URL として各要素は利用する。

次に N-gram 検索の SQL に関しても、一致検索とスペース区切りによる複数検索で同じ構造で、

```
SELECT test_sentence. id_sentence,
       test_sentence . Sentence ,
       test_sentence. sentence_image_1,
       test_sentence. sentence_image_2
FROM test_sentence, test_ngram
WHERE
  test_sentence. id_sentence = test_ngram. id_sentence
AND (test_ngram. text = keyword)
```

とした。キーワードのチェックと OR でつなぐ部分に関しては LIKE での検索と同じ処理を行い不正な検索文字列は入力できないようにした。

SELECT 句で抜き出したいデータは*と同様に test_sentence テーブルの全要素であるが、この SQL の場合はテーブル名を選択する FROM 句に test_ngram テーブルも含まれているので、こちらの全要素も抜き出してしまうため個別に指定した。WHERE 句では test_sentence テーブルの主キーと test_ngram テーブルの外部キーが一致しているかつ、入力したキーワードが test_ngram テーブルの text と一致している場所を抜き出すようにした。

5. データベースの検索画面

今回試作したデータベースは Web ページとして実装した。実装には、マークアップ言語として HTML を用いてページ内要素を構成した。また、CSS フレームワークである Bootstrap (ver.5.0.2) を用いることにより、文章のより詳細な配置、装飾を施した。そこに、スクリプト言語である PHP を埋め込む形で、動的ページを作成した。

本データベースの検索フォームの画面を図 1 に示す。検索フォームには、各検索方法が選べるようにあいまい検索と書かれたチェックボックスと複数検索 (スペース区切り) と書かれたチェックボックスの二つを配置した。また、検索ワードに関してシンプルなテキストボックスを配置し、ここに入力できるようにした。あいまい検索のチェックボックスによって、LIKE 検索と N-gram による一致検索を切り替えることができる。標準は N-gram による一致検索とし、チェックを入れることで LIKE 検索に切り替わる。また、複数検索 (スペース区切り) にチェック入れると、それぞれの検索において、スペース区切りにより複数の検索ワードを指定できるようになる。

テキストボックスに任意の書き下し文を入れ、検索ボタンを押すことで、検索を行い、検索結果の表示画面に遷移する。検索結果の画面を図 2 に示す。検索結果画面では、該当箇所の書き下し文



図 1 検索画面
Figure1 Search page.

が縦書きで右側に表示され、該当箇所の資料画像が左側に表示される。複数検索 (スペース区切り) の場合、該当箇所が縦に複数表示されるためスクロールバーを配置することで閲覧しやすいようにしている。また、test_sentence テーブルの設計で資料画像を二つ含むように設計してある理由は、資料上の物理的なレイアウトにより、一文であるはずの本文が複数の行にわたって書かれている場合に対応するためである。上下に分かれた一文を 2 枚の画像として並べて表示することで、関係性を分かりやすくしている。また、割注においても同様に 2 行に分かれるケースがある。こちらは行中の位置などによって画像の切り出し方法が異なるが、基本的に本文の表示方法と同じになるように切り出し範囲を調整している。検索結果においても、元の資料と同じ配列になるように sentence_image_1 の画像を右に、sentence_image_2 の画像を左に配置している。

6. 今後の展望

今後の展望として、まず LIKE 検索と N-gram での検索速度を比較する実証実験と行う予定である。また Mroonga を有効にした場合のパフォーマンスの変化も確認する必要があると考えている。特に、現段階ではまだ各テーブルに登録されている本文の量が少なく、情報量が増えたときの挙動が確認できていない。今回実装した部分では、精密にテキストを解読したうえで書き下し文を生成してデータ化しているが、まずは機械的に作った書き下し文をデータベースに追加し、データ量と検索速度の関係を調べていきたい。

一方、データベースのテーブル設計に関しては、書き下し文に対し読み仮名が不足していると考えている。読み仮名を検索結果に表示することで、

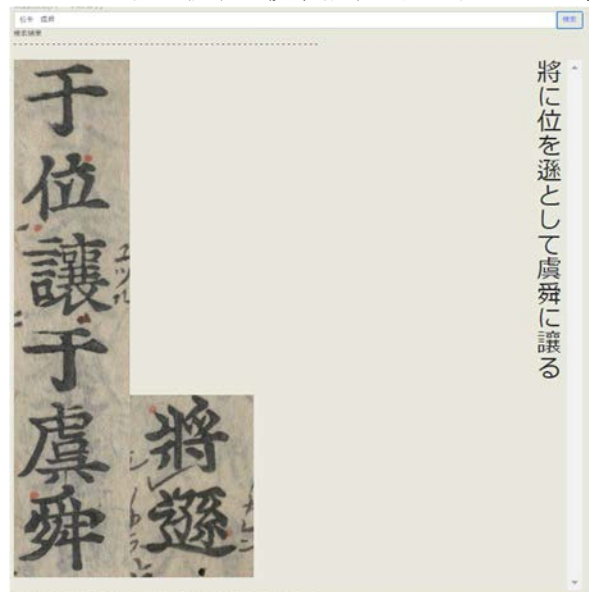


図 2 検索結果
Figure2 Search result.

書き下し文における漢字の読み方が一目で分かるという利点に加え、検索そのものもひらがなで行えるようになりなるため、より検索しやすいデータベースになると考えている。

また、訓点情報の電子データを検索結果に合わせて表示すると、1 ページあたりの情報量が多くなりすぎて、逆に書き下し文を認識しにくくなる可能性がある。そのため、検索結果の書き下し文の漢字に訓点情報が閲覧できる画面へ遷移するリンクを紐付ける、新たにチェックボックスやセレクトボックス等を追加することによって検索の段階で任意の情報の選択ができるなど、GUIを更新する必要がある。

7. まとめ

本稿では、書き下し文での訓点情報検索を可能とするデータベースの試作について述べた。このデータベースでは漢字+送り仮名といった現代語訳文に近い形で検索することができ、該当箇所の資料画像を閲覧できるため、ヲコト点や語順点なども同時に確認することができる。これにより、訓点の専門家以外であっても訓点資料を広く研究に利用することが可能となる。一方で、実証実験が行えておらず検索方法の妥当性が評価できていないため、早急にこれを解決し最終的な検索方法の決定をする必要がある。またテーブル設計やGUI等、課題はまだ存在するためアンケートを採ることを始めとしてデータベース全体の改善を図っていく。

謝辞

本研究は JSPS 科研費 20K00654 の助成を受けたものです。また、本研究は、人間文化研究機構広領域連携基幹研究プロジェクト「異分野融合による総合書物学」の国語研ユニット「表記情報と書誌形態情報を加えた日本語歴史コーパスの精緻化」による成果の一部である。

参考文献

- [1]春日政治:西大寺本金光明最勝王經古点の国語学的研究, 斯道文庫 (1942).
- [2]築島裕:興福寺本大慈恩寺三蔵法師伝古点の国語学的研究(訳文篇・索引篇・研究篇), 東京大学出版会 (1965-1967).
- [3]築島裕:訓点語彙集成<第1巻>, ヲコト点概要, 汲古書院 (2007).
- [4]高田智和:ヲコト点の座標表現, 国立歴史民俗博物館研究報告, Vol. 192, pp. 171-181 (2014).
- [5]堤智昭, 田島孝治, 高田智和:点図情報入力支援ツールによるヲコト点図の電子化, じんもんこん 2015 論文集, Vol. 2015, pp. 185-190 (2015).
- [6]堤智昭, 田島孝治, 小助川貞次, 高田智和:訓点資料の構造化記述方式と計算機を用いた基礎計量, 情報処理学会論文誌, Vol. 59, No. 2, pp. 278-287 (2018).
- [7]林昌哉, 田島孝治, 堤智昭, 高田智和, 小助川貞次:訓点資料の加点情報計量のためのデータ構造—国立国語研究所蔵「尚書(古活字版)」を対象として—. じんもんこん 2017 論文集, Vol. 2017, pp. 45-52 (2017).
- [8]田島孝治, Baptiste Jannequin, 堤智昭, 高田智和:IIIF Viewer と連携可能な訓点資料の加点情報データベースの試作, Vol. 2019, No. 1, pp. 109-114 (2019).