Hanzi Dictionaries in Early Ages with Smartphone : A IDS Query System of Tenrei Banshō Meigi

スマホで古辞書
－『篆隷万象名義』の IDS 検索を例に－

劉　冠偉（北海道大学文学研究科博士課程）†
李　媛（北海道大学文学研究科博士課程）
池田　証壽（北海道大学文学研究科）

Hanzi Dictionaries in Early Ages with Smartphone
A IDS Query System of Tenrei Banshō Meigi

Guanwei Liu (Graduate School of Letters, Hokkaido University)
Yuan Li (Graduate School of Letters, Hokkaido University)
Shoju Ikeda (Graduate School of Letters, Hokkaido University)

要旨

近年,スマートフォンやタブレットのようなモバイル端末が普及し,日常生活を変えつつあり,日本語教育・日本語研究にも使えるようになると予想される。

しかしながら,構築･公開が盛んである古典籍･古文書のデータベースは PC 向けが多く,PC 以外の端末で利用する際は表示サイズのずれや機能障害がしばしば発生する。そこで,モバイル端末でデータベースを利用しているユーザを想定した利便性が高い言語資源データベースの Web インターフェイスを開発したい。漢字字形の構造情報を用いて古辞書のテキスト・画像を検索することによって文字の同定に利用できる Web アプリはまだないので,篆隷万象名義の掲出字について IDS 検索と画像表示を可能にするツールを試作した。本アプリによって,漢字のパーツで篆隷万象名義に掲載している文字の画像をスマートフォンなどの携帯端末で検索でき,写本の解読・翻刻する際に役立つと期待している。

１．はじめに

スマートフォンが急速に社会で普及している。インターネット上の言語資源もスマートフォンへの対応が求められている。一方,日本の古辞書は日本語の歴史的研究に有益であり,これまでの研究と教育において利用・活用されてきた。しかし,日本の古辞書をスマートフォンで利用しようとしたときに,解決しなければならない課題は多い。

(a)　利用に制限のない,デジタル化された翻刻本文と原文画像　［対象］
(b)　パソコンで利用できる古辞書関連サイトとスマートフォン対応　［構想］
(c)　古辞書に含まれる難字・異体字を入力・表示するシステムの開発　［設計］
(d)　サーバに実装する上での問題　［実装］

上記の課題を本文の第 2 節～第 5 節にわたってその詳細を論じていく。

まず第 2 節では,モバイル端末（スマートフォン・タブレットを含む）対応古辞書検索システムを構築するには,それらのデジタル化された翻刻本文と原文画像が必要となり,利用に制限のないことが必要であることを指摘する。本研究では,我々の HDIC プロジェクトで公開している篆隷万象名義データベースの翻刻本文と,利用・公開の許諾を得ている掲出字の原文画像を利用することでこの問題を解決しようとしたことを述べる。

次に第 3 節で,古辞書を検索・表示するスマートフォン対応のサイトを構築する上での課

† toyjack@gmail.com

題を検討し,字形は明白だが,部首・画数・音訓がわかりにくい漢字は,そもそも検索のための入力が困難となるので,入力メソッドの開発が必要であることを述べる。さらに,古辞書に含まれる難字・異体字を入力・表示するシステムの開発には,IDS（詳細後述）のデータを利用するのが有効であるので,IDS データを利用する上での問題と解決策を述べる。

第 4 節では,実際の Web アプリケーションの設計について述べる。そのあと第 5 節ではサーバに実装する上での課題を述べる。

２.『篆隷万象名義』の翻刻本文と原本画像

２.１『篆隷万象名義』

『篆隷万象名義』は,9 世紀前半,唐から日本に戻った弘法大師空海が,梁・顧野王撰述の原本『玉篇』(543)を抜粋した字書である。唯一の古伝本である高山寺本『篆隷万象名義』は研究資料としての価値が高いが,誤写・誤脱が多いことも早くから言われている。一方,中国南北朝以来の字体の古い情報を残すものもあって,字体研究においても重要な資料である。

『篆隷万象名義』は,約 16,000 字の掲出字に対して,字音・字義・字体の記述を収録する。漢字字体研究において,HNG に収録された標準文献に比べて,次の二つの特徴が指摘できる。

(1) 掲出字は古辞書の説明対象としての骨組みであり,少数の重複字以外,ユニークな存在である。一方で,個々の掲出字そのもののバリエーションが僅少であるが,掲出字を網羅的に収録し,異体字も併記するため,漢字字体の多様性を備える。

(2) 異なる掲出字の間に,同一漢字部品が持つものが多く存在する。漢字部品レベルでは,字体の同一性（単一パタン）と多様性（複数パタン）が観察できる。

また,掲出字画像は,字体研究資料であると同時に,掲出字の字形の細部を確認することを可能にするもので,古写本のデータベース構築に不可欠である。さらに,データ化する過程に,Unicode テキストの不足を補う機能している。

２.２　翻刻本文

『篆隷万象名義』の全文テキスト［http://github.com/shikeda/HDIC］は TSV データで公開済みである。その詳細を李・池田（2016）では報告した。Unicode で扱える漢字の『篆隷万象名義』全掲出字に占める割合は,99.2%となる。掲出字「語」の TSV データは次の表 1 の通りである。01～10 の番号は,説明の便宜ため付けたものである。

表 1「語」の TSV データ

01	TBID	3_007_B62
02	TB_vol_radical	v9#91
03	TB_radical	言
04	Entry	語
05	Entry_type	Regular
06	Entry_diff	無
07	TB_def	魚擧反。説也,言也,喜也。
08	SYID	a082b061
09	YYID	無
10	TB_remarks	無

解　説

01　第 3 帖 7 丁裏 6 列の 2 字目（所在）
02　巻 9・部首 91 番目（巻数・部首番号）
03　言部（部首）
04　語（掲出字）
05　隷書掲出字（掲出字タイプ）
06　諸家認定に異同なし（先行研究照合）
07　魚擧反。説也,言也,喜也。
08　対応する宋本玉篇の所在は上篇 82 丁裏 6 列 1 字目（関連字書所在）
09　原本玉篇残巻に存せず（関連字書所在）
10　なし（校勘意見）

２.３　原本画像

『篆隷万象名義』掲出字の原本画像は HDIC のプロジェクトで作成したものを利用している。詳細は池田（2014）・池田他（2016）で述べた。図 1 に「語」（第 3 帖 7 丁裏）と「諒」（第 3 帖 8 丁表）の高山寺本・崇文叢書[1]の項目画像,ならびに掲出字画像を示す。

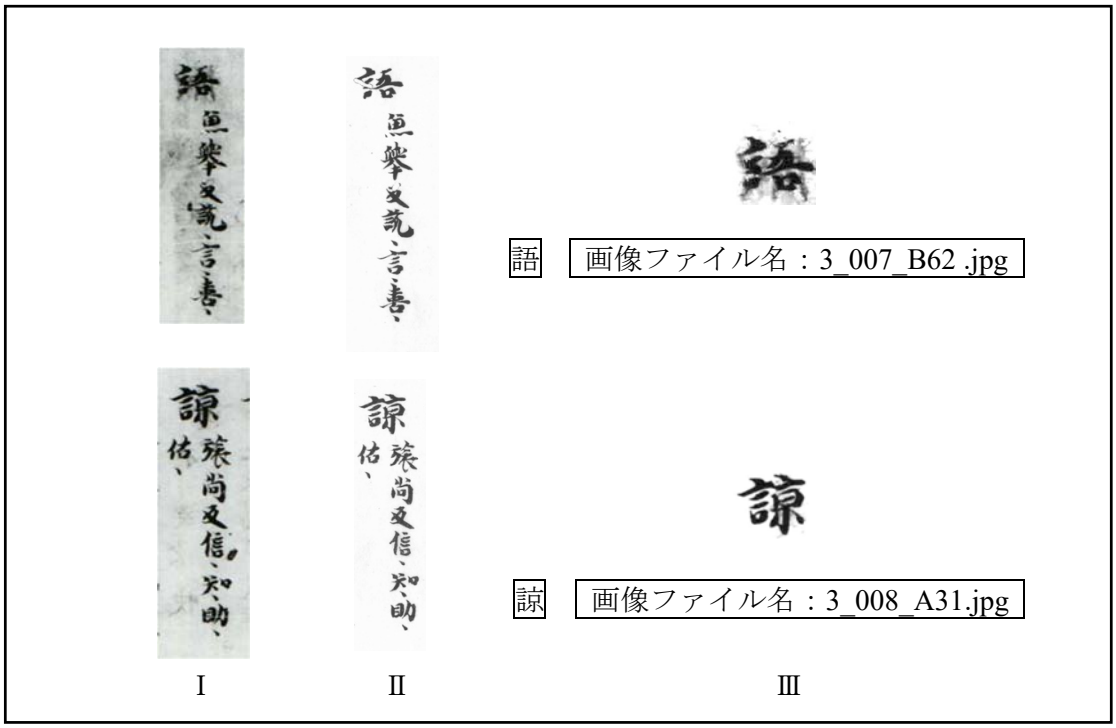

図 1 『篆隷万象名義』高山寺本（Ⅰ）と崇文叢書（Ⅱ）の原文画像・掲出字画像（Ⅲ）

掲出字のテキスト化の際に,画像データベースを構築して,掲出字のテキストの効率化をはかる。また,「諒」のように,旁の「京」の部分について,翻刻本文「京」と原本字形「亰」と相異があるが,テキスト化のとき「亰」を「京」に統一して翻字する。

[1] 図 1 に示した「語」・「諒」の崇文叢書画像は著者の個人蔵書によったが,『篆隷万象名義』崇文叢書のテキストの一部（第 1 輯の第 32 至 43）は,国立国会図書館デジタルコレクションにて公開されている。

３．IDS によっての漢字検索・入力

３.１　漢字の IDS 検索

古辞書に収録される漢字の中には,直ちには音訓がわからないような難字があり,それらの漢字を効率的に検索・入力する方法も問題である。すなわち,字形は明白だが,部首・画数・音訓がわかりにくい漢字は,そもそも検索のための入力が困難となるので,入力メソッドの開発が必要なのである。

古版本・古写本を研究するに際して,翻刻は必須な作業として研究者が多くの時間をかけている。近年,辞書・典籍の電子データベース化と公開がなされており,それらの利用によって,作業の手間が格段に軽減されているが,これらの電子データを検索・編集するために,漢字の入力が常に必要となる。その際,読み方が不明であることや,入力メソッドに未収であることが原因で,漢字を簡単に入力できないケースも少なくない。このような漢字の形しか知らずに漢字を入力したい場合は,まさに紙の字書を引く時と似ている。紙の字書のように,部首と画数を用いて漢字を検索できるデータベースでは Unihan データベースが権威的である。しかし実際に利用する際,次の二つの難点がある。

(1) 同部首同画数の字数が多い場合,欲しい漢字を探すのは難しい。

(2) 所属する部首が分からない場合,利用できない。

部首より小さい漢字構造上の要素によって検索するシステムを作ることでこの二つの問題は解決できる。そのようなシステムを実現するための漢字記述の方法として,「漢字構成記述文字列（IDS)」がある。IDS とは,漢字の構成を文字列で記述したものである。IDS は IDC[2]と漢字の部品からなる。符号化されていない漢字を表すことのできる漢字記述言語の一種である。IDS をすでに符号化した漢字に用いて,漢字の検索方法とすることもできる。このような漢字検索システムはいくつか開発されており,もっとも代表的なものは CHISE/ids-find[3]である。

CHISE は漢字符号をコード制限なしの環境で処理するためのプロジェクトである。CHISE IDS はそれのサブプロジェクトとして,漢字の IDS 情報を整備している。IDS-FIND はそれらの IDS 情報を検索するためのウェブアプリである。

図 2　CHISE/ids-find の PC 画面

図 3　CHISE/ids-find のスマートフォン画面

図 2 に示すように,CHISE の IDS-FIND 機能は PC 向けで開発されている。図 3 に示すように,PC 以外の端末でアクセスすると画面の表示が PC とほとんど変わらず,携帯端末によって操作が難しい場合が生じる。

---

[2] Ideographic Description Character 構造を表す符号であり,「⿰⿱⿲⿳⿴⿵⿶⿷⿸⿹⿺⿻」12 個からなる。

[3] http://www.chise.org/ids-find

CHISE/IDS-FIND における検索結果の数が多い場合,一回の検索結果の表示に数十秒間かかることがある。同様の問題が我々のシステムにも生じるため,解決策が必要となる。表示スピードの問題は 5.2 で検討する。

３.２　IDS データの利用

Unicode 委員会が公開している Unihan データベースは漢字データベースとして最も広く知られているものであるが,IDS に関する情報は現時点まで公開されていない[4]。

現在公開中の漢字 IDS データの中で,整備状況が一番良好なのは CHISE IDS である。CHISE IDS の ReadMe ファイルによると,「<CODEPOINT><CHARACTER><IDS>」三つのフィールドをタブで区切ったデータ構造をとっており,つまり TSV（Tab-separated Values）で示している。CODEPOINT は UCS のコードポイントである。拡張漢字 A までは「U+hhhh」のような「U+」と 4 桁の 16 進数で示す。それ以降は「U-hhhhhhhh」のような「U-」と 8 桁の 16 進数で示す。CHARACTER では CODEPOINT が対応する漢字の符号化字形を示す。IDS はその漢字の構成記述情報を示す。次の表 2 に例を示す。

表 2　CHISE IDS データの一例

CODEPOINT	CHARACTER	IDS
U+5B9A	定	⿸宀&CDP-8BCE;
U-0002A76B	儼	⿰亻⿸厂亚取
U-0002B7AA	甚	⿸甘⿱&AJ1-04307;匹

また,漢字データベースプロジェクトの「漢字構成データベース」に「字形 IDS データ」があり,現在はオープンソース共有プラットホームの GitHub［https://github.com］を用いて「CJKVI-IDS」という名称で公開している。CJKVI-IDS の構造は主に CHISE IDS と同様で,拡張漢字 B まで CHISE IDS のデータがそのまま利用されているようである。拡張漢字 C・D・E のところでは独自の IDS データを採用している。ただし,CHISE IDS と比べると,CJKVI-IDS は CDP 漢字[5]など表外漢字をそれらの画数である「①②③…」のような丸数字に変換している。Unicode の表示方法も少し異なる。IDS の問題点は川幡（2009）に詳しい。次の表 3 に例を示す。

表 3　CJKVI-IDS データの一例

CODEPOINT	CHARACTER	IDS
U+5B9A	定	⿸宀正
U+2A76B	儼	⿰亻⿸厂一④取
U+2B7AA	甚	⿸甘⑤区

今回の試作における IDS データは,高速な検索を実現するため,部品から漢字を合成するデータベースと,漢字から部品に分解するデータベースの二つに分けて作成する。合成データ

[4] http://www.unicode.org/L2/L2015/15065-ids-links.pdf

[5] Chinese Document Processing 台湾の中央研究院が開発した漢字処理システムである。2011 年の最終更新まで約 16 万 5 千の字形が収録されている。

ベースは CJKVI-IDS の ids.txt と ids-ext-cde.txt ファイルをベースとして,IDC と符号化されない部品,すなわち入力困難なものと作者の備考などを削除して,さらにシンプルなデータベースを作成する。次の表 4 に例を示す。

表 4　簡略 IDS データの一例

CODEPOINT	CHARACTER	IDS
U+5B9A	定	宀正
U+2A76B	儼	亻一取
U+2B7AA	甚	区

４．ウェブアプリケーションの設計

４.１　設計上の問題点

スマートフォンやタブレットなどのモバイル端末は PC と比べると大きな相違がある。古辞書データベースの場合では,主として次の三つの問題が生じる。

(1) 画面サイズが小さく,同時に表示できる情報が少ない。

(2) 入力メソッドが軟弱であり,難字の入力に弱い。

(3) システムフォントが不足している。また新しいフォントをインストールできない。

４.２　レスポンシブデザインの採用

(1)に対しては,レスポンシブデザインの応用によって解決できる。レスポンシブデザインとは,端末の画面サイズにより,アプリが相応な仕組みへ変えて表示するデザインである。図 4 は PC の画面,図 5 は携帯端末の画面である。同一のアプリが自由に切り替えることが可能となった。

図 4　PC での画面

図 5　携帯端末での画面

４.３　難字入力方法の開発

(2)については,Unihan データベースで公開している漢字総画数データを使用することによって,漢字の部品と残りの画数で検索できるようにした。例えば,「謐」を入力したい場合に,CHISE/ids-find などの漢字検索システムであれば,「言」と「必」との組み合わせで検索すると「謐」が出てくる。本システムは「言」と残り部分「必」の画数「10」,つまり「言 10」

の組み合わせで検索できる。画数で検索できるようにすることで,難字に対応しない入力メソッドでも負担が少なくなった。

４.４　対応するフォント

(3)については,ウェブフォント技術などを用すれば解決できるが,Unicode 漢字を全すべて含めるとフォントファイルが大きくなる,通信の制限で実現するためにさらに努力が必要と考える。ウェブフォントの圧縮・区分をさらに検討することが必要であるが,これは今後の課題とする。

５．サーバへの実装

５.１　利用したフレームワーク

ウェブアプリに用いるフロントエンドのフレームワークは多くある。データを処理する JavaScript フレームワークは jQuery［https://jquery.com/］,React［https://facebook.github.io/react/］などがよく利用され,表示の仕方を定める CSS では Bootstrap［http://getbootstrap.com/］が定番となっている。だが,今回の開発は JavaScript のフレームワーク Vue.JS［https://vuejs.org/］と CSS のフレームワーク bulma［http://bulma.io/］を用いた。

React や Bootstrap は機能性と汎用性が強く,各分野の開発によく見られるが,学習の労力を考えると,よりシンプルなフレームワークを利用して開発したいと考えた。

Vue.JS は軽量なインターフェイス利用を中心とした JavaScript フレームワークであり,学習しやすいながら強い性能を持っていることで評価されている。bulma も軽量ながら,レスポンシブデザインをサポートする CSS フレームワークである。より広範囲で使われている React と Bootstrap の代わりに,この二つのフレームワークを選択するのは,専門のプログラマーではない筆者（劉）にとって,学習が比較的容易であるというメリットがある。

５.２　検索速度の向上

検索速度を向上するため,IDS データをローカルに保存する。検索プログラムを JavaScript にして,検索の計算をクライアント側に負担させる。より効率的に IDS データベースを利用するため,オープンソース漢字検索システム刹那字引[6]［https://github.com/g0v/z0y］の部分コードを利用した。「刹那字引」は拡張漢字 E までサポートする[7]漢字検索システムである。特に検索スピードに優れている。ただし,現在では開発が止まっているようである[8]。「刹那字引」の検索がはやい理由は,検索用 IDS データベースの再構築である。「刹那字引」では,表 3 のようなデータを逆引きにして利用する。再構築したデータは次のようである。

"宀": "…宖宑宒宓宔宔宕宖宗官宙定宛宜宝实…",

"疋": "定従是珽埏媞揕疋碇蝭蝷　",

また,CHISE IDS/Find とは異なり,毎回検索で画面を更新する必要がない。ただし,画像ファイルの集合のデータ量が大きいので,サーバ側に保存する。アプリの初回利用と画像を請求する時のみサーバと通信する。

---

[6] http://www.ksana.tw/kzy/

[7] バージョン 1.0 までは拡張漢字 B までをサポードしており,拡張漢字 E までの検索はその以降の再開発バージョンである。コードがだいぶ変わったので,別のシステムであるともいえる。

[8] 最後のアップデートは 2016 年 3 月であり,筆者が提出したバッグ修復の「Pull Request」となった。その前の最後の更新は 2015 年 10 月であった。

作成したウェブアプリを KTB Image IDS Database を名付けて,https://hdic2.let.hokudai.ac.jp/ids で公開する予定である。

６．おわりに

本稿では HDIC プロジェクトの篆隷万象名義全文テキストデータベース・掲出字画像データと CJKVI-IDS データベースを利用して,篆隷万象名義の掲出字画像をスマートフォンやタブレットなどのモバイル端末での IDS および画数によっての漢字検索を実現した。

本研究は言語資源研究のツールの開発を目的に行ったものである。モバイル端末の利用法について,いろいろな意見をいただいて,さらに改良していきたい。

謝　辞

本研究は JSPS 科研費 16H03422 による成果の一部である。篆隷万象名義全文翻刻テキストと掲出字の画像公開については,高山寺当局ならびに石塚晴通教授（高山寺典籍文書総合調査団団長）のご許可・ご指導のもとに行われている。記して感謝の意を表す。

文　献

池田証壽(2014).「平安時代漢字字書総合データベースー現状と課題 2014 夏ー」『漢デジ 2014 : デジタル翻刻の未来』, 京都大学人文科学研究所附属東アジア人文情報学研究センター編.

池田証壽・李媛・申雄哲・**賈**智・斎木正直(2016).「平安時代漢字字書のリレーションシップ」『日本語の研究』12:2, pp. 68-75.

上地宏一(2005).「CHISE IDS FIND (ソフトウェア レビュー 多言語情報処理)」『漢字文献情報処理研究』2005-10:6, pp.163-165.

川幡太一(2009).「IDS による情報処理」2009 年漢字文献情報処理研究会年次大会.

守岡知彦・師茂樹(2004).「文字素性に基づく文字処理」『人文科学とコンピュータ研究会報告』2004:58(2004-CH-062), pp.53-60.

李媛(2016).「IDS データと HDIC 原本画像・翻刻テキストとを利用した古辞書の漢字字体研究について −『大広益会玉篇』を中心に−」『人文科学とコンピュータ研究会報告』,2016-CH-110:6, pp. 1-6.

李媛・池田証壽(2016). 「篆隷万象名義の全文テキストと公開システムについて」『じんもんこん 2016 論文集』, pp. 95-102.

劉冠偉・李媛・池田証壽(2015).「平安時代漢字字書総合データベースの拡張と和訓対応」『人文科学とコンピュータ研究会報告』2015-CH-106:4, pp.1-8.

The Unicode Consortium(2016). *The Unicode Standard*, Version 9.0.0, Unicode Consortium.

関連 URL

平安時代漢字字書総合データベース(HDIC)	http://hdic.jp/
KTB Image IDS Database	https://hdic2.let.hokudai.ac.jp/ids
CHISE project	http://www.chise.org
CHISE IDS	http://www.chise.org/ids/
漢字データベース	http://kanji-database.sourceforge.net/index.html
UniHan Database	http://www.unicode.org/charts/unihan.html