Vol.2016-CH-110 No.9
2016/5/14

『大字典』和訓データベース構築の現状と課題

劉冠偉[†1]

概要：『大字典』（上田万年ら，初版 1917 年）は，約 18,000 字の親字を収録する漢和辞典であり，全親字中 6,103 字にゴシック体の和訓が品詞名付きで 10,515 個付される．日本語史研究の観点から，これらの和訓や典拠を分析するために『大字典』和訓データベースを構築している．本報告では，オープンソースフレームワーク Laravel によるウェブインターフェースを開発することで当データベースの検索・照合・編集システムを実現する試みを紹介する．また，データの精度を上げるために，北海道大学池田証壽研究室で開発・整備している各種のデータベースと照合しながら点検を行う．

キーワード：漢和辞典，和訓，データベース

Construction of a Database of Japanese Readings in Daijiten: Its Progress and Problems

GUANWEI LIU[†1]

Abstract: Daijiten is a dictionary of Japanese readings of Chinese character which has collected over 18,000 characters which has 6,103 characters with Japanese reading. And the Database of Japanese readings in Daijiten is created in order to analyze these Japanese readings in view of Japanese language study's history. This report introduces the attempt to realize the search and matching-editing system of this database to develop a web interface with open source framework Laravel.

Keywords: Dictionary of Japanese readings of Chinese characters, Japanese readings, Database

1. はじめに

『大字典』は『康熙字典』と同じ部首画数順で配列する漢和辞典である．明治以降刊行された数多くの漢和辞典と比べると，国語学者による編纂であること，親字に通し番号を付けること，重要視する和訓にその品詞を示すことなどの特徴を持っている．『大字典』和訓データベースは『大字典』に掲載する 10,515 個ゴシック体で示される和訓とそれらの品詞・親字を全て収録している言語データベースである[1].

北海道大学池田証壽研究室は平安時代漢字字書総合データベース（Integrated Database of Hanzi Dictionaries in Early Japan, HDIC）[a]に収録する各種データベースを開発・整備している．HDIC は中国中世字書の『玉篇』残巻と逸文・『大広益会玉篇』・『切韻』残巻と逸文・『大宋重修広韻』などと，これらの中国字書を参照・利用している日本側の平安時代漢字字書『篆隷万象名義』・『新撰字鏡』・『類聚名義抄』より構成している総合漢字データベースである[2]．古典籍や古語の研究に用いられるデータベースとして，一定の精度が必要だと考えられる．HDIC では『大広益会玉篇』の全文テキスト[b]を公開しており，『大字典』和訓データベースはそれと照合しながら点検してデータの精度向上をはかっている．また作成したデータベースがより多くの人に利用されるため，インターネットに公開するウェブインターフェースを開発することが必要である．本報告では，『大字典』和訓データベースの点検とインターフェースの開発過程と現状，さらに残された解決すべき問題について述べる．

2. データベース開発開始まで

『大字典』和訓データベースを開発する当初に念頭に置いたのは，

（a）明治時代の国語の基礎資料として蓄積する
（b）HDIC に収録している和訓同定に役が立てる
（c）自己責任でゼロからデータベースを作成して，自由に実践できる場所がほしい
（d）できたデータベースをインターネットに公開して，さらに多くの人の利用に提供したい

などであった．

作業の手順は次のように進めている．

① データ入力
② 点検
③ CSV などによるファイルでの公開
④ 検索システムによる公開

『大字典』の親字に対して，高田（2001）・高田（2004）は一連の研究があり[3][4]，全親字を収録した「『大字典』データベース」を作成した．①と②の構築段階で，筆者はそれを利用して，入力時間を大幅に減少させることができ

a) http://hdic.jp

b) https://github.com/shikeda/HDIC

た．③の CSV ファイルなどの公開は，従来，人文系のデータベースによく使われているが，利用者の視点からやはり不便の場合もある．そこで④のオンライン検索システムの開発に至ったのである．④のシステムは，最初は筆者がゼロから PHP で開発した検索ページであり，専門知識がないため，開発のスピートと質とも低いものであった．また，開発に使える時間も少なかったので，断続的に進めたため，自分が書いているコメントアウトさえ分からなかったこともあった．

3. 現状

3.1 ウェブインターフェースの開発

開発の効率を向上するため，フレームワークを利用して『大字典』和訓データベースのインターフェースを構築する．開発言語は筆者が馴染みある PHP を用い，フレームワークは近年，ネット上に流行している Laravel にする．Laravel は，RESTful[c]ルーティングのサポート，豊かなコミュニティなどの特徴があり，他の PHP フレームワークと比較すれば，プログラム初心者でもより容易に利用できる．かつ，MIT ライセンスのもとで配布されているので，利用上の制限が少ない．Google Trends によると，2016 年 15 週間(2016 年 1 月 3 日から 2016 年 4 月 16 日までの統計)PHP フレームワークのなかで一番人気だという．

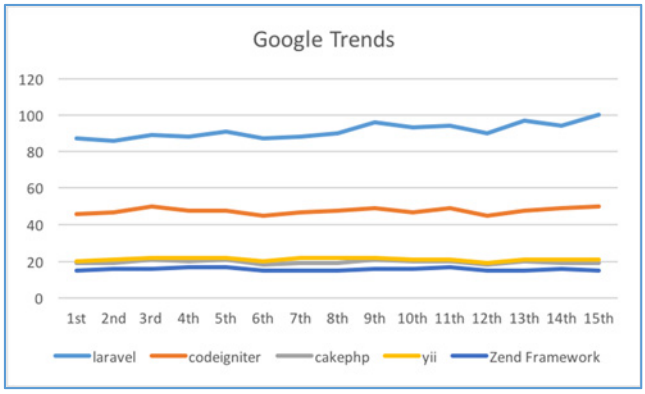

図 1　Google Trends による 2016/01/03-2016/04/16 人気度

筆者のようなプログラム専門家以外の人に対して，Laravel のような MVC 系フレームワークを利用して得られる一番のメリットはおそらくロジックのモデル（Model）およびコントローラ（Controller）を表示のビュー（View）より分離できることであろう．これによって，システムの機能デザインに専念でき，断続的な開発でも効率的に進めることが可能となる．

3.2 検索機能

当システムの基本機能は『大字典』和訓データベースに収録しているすべての和訓に対して検索できることにある．和訓を片仮名・平仮名，和訓の親字を漢字符号・ユニコード・大字典番号で検索する．検索の結果は検索画面の右側にリストとして表示する．リストにある結果をクリックすると，該当和訓や漢字の独自のページに入る．

図 2　検索画面

3.3 レスポンシブデザイン

近年，スマートフォンやタブレット端末のユーザが著しく増え[5]，研究資源の公開もそれらへの対応を考えなければならない時代を迎えている．当システムのフロントエンドは Twitter 社が公開されている CSS フレームワークの Bootstrap を利用することで，さまざまな端末に対応できるレウスポンシブデザインをサポートしている．

図 3　水平放置の iPhone 6 端末によるアクセス

3.4 URL の設計

『大字典』和訓データベースの URL はの設計は次の通りとする．

①検索ページを辞書名にする．
例：http://hdic2.let.hokudai.ac.jp/djt/
②各親字独自ページの URL は Unicode スカラ値にする．
例：http://hdic2.let.hokudai.ac.jp/djt/character/U+4E00
③各和訓独自ページの URL は/wakun/{カナ}にする
例：http://hdic2.let.hokudai.ac.jp/djt/wakun/ハジメ
④『大字典』番号で親字独自ページにアクセスできる．
例：http://hdic2.let.hokudai.ac.jp/djt/00001

3.5 データ精度の向上

データ精度の問題を解決するには，原文だけではなく，さらに別のデータベースと照合しなら点検するのは効率的だと考えられる．北海道大学池田証壽研究室では『大広益会玉篇』や『篆隷万象名義』など一連の古辞書総合データベース HDIC が公開されている．筆者は，そのなかの『大

c) Representational State Transfer

広益会玉篇』を利用して『大字典』和訓データベースを点検する．例えば，大辞典番号 1705 番の「坻」の和訓を入力する際，「シマ」を「レマ」に間違ってインプットしてしまったことがある．原文のページや画像を探すより，『大広益会玉篇』に掲載している同じ項目の注文[d]を参照すると，間違ったことを察知できる．図 4 は宋本玉篇データベースとの連携し、データ修正済みの画面である．

図 4　宋本玉篇データベースとの連携

4. 課題

4.1 端末における字体表示の問題

『大字典』和訓データベースはレスポンシブデザインをサポートして，スマートフォンなどのデバイスにも正常に操作できるが，現時点において，iOS 9.3.1 を実装している iPhone や iPad では Unicode の追加漢字面にある符号のサポートは非常に不十分である．図 5 はテストページ本来の表示，図 6 は iPhone 端末に表示される様子である．テストページは各拡張漢字集合の最初の四文字で作成したものである．図 6 を参照すると，iPhone は拡張漢字をほぼ表示できない状態といえる．これは『大字典』和訓データベースだけにとどまることではなく，データベースに収録の文字をありのままで表示するのはウェブインターフェースとして越えなければならない問題であろう．

図 5　OS X El Capitan 10.11.4 Safari 拡張漢字表示

図 6　iOS 9.3.1 Safari 拡張漢字表示

表示ページの CSS に@font-face，つまり Web Font を使うと，これを解決できるようである．しかし，拡張漢字を全て収録しているフォントファイルのサイズが大きく[e]，サーバよりダウンロードする時間が長すぎてページに応用するのは難しい．そこで，各ページに表示される符号だけを抽出して，何十字や何百字の小さいフォントファイルを作れば，符号の表示とページのダウンロード時間が両立できるようである．こういう機能が実現できるオープンソフトを調査すると，現在では fontSPIDER[f]と Fontmin[g]などがある．かつ，Fontmin より作った server-fontmin[h]は動的な内容でも常にフォントファイルとスタイルファイルを生成できる．しかし，筆者のパソコンの環境では両方とも花園フォントの SIP[i]面にある符号が認識されない．二つとも node.js[j]で開発されたので，これはおそらく node.js が 4 バイト符号に対するバグであろう．

4.2 画像データベースへの拡張

データを点検する際には『大字典』原本との照合作業は，その原本画像をデータベースに追加することによって行いたいが，さまざまな著作権の問題も考えなければならない．一つの手としては二つのバージョンの『大字典』画像を公開している近代デジタルライブラリー（以下は近デジと呼ぶ）へのリンクを付けることである．近デジは特定のページをアクセスすることが可能であり，近デジの「このデータベースについて」[k]では次のように説明している．

d）直飢切．水中可居曰坻．《方言》云：坻場也，梁宋之間，蚍蜉犁鼠之場謂之坻．又音底．《埤蒼》云：坂也．俗作坵．
e）花園フォント 2016 年 02 月 01 日版，拡張漢字 B～E を収録している HanaMinB.ttf ファイルは約 27.2MB である．
f）http://font-spder.org
g）http://ecomfe.github.io/fontmin
h）https://github.com/junmer/serve-fontmin
i）Supplementary Ideographic Plane
j）https://nodejs.org/en/
k）http://kindai.ndl.go.jp/ja/aboutKDL.html#aboutKDL2_6

個々の資料の，特定のコマにリンクを張る場合は，閲覧画面の「URL」ボタンで表示される「http://kindai.ndl.go.jp/info:ndljp/pid/{数字}/{コマ番号}」の形の URL をご利用ください．

方法は,『大字典』和訓データベースに収録されている親字にそれらのページなどの所在情報を追加して，これらの所在情報を利用して各親字に所在している近デジの『大字典』画像ページへのリンクを作成することである．作業には難点がないが，所在情報の追加は時間の余裕を見て行いたいと考えている．

5. おわりに

以上,『大字典』和訓データベースを開発する過程と現状,および当面の解決が求められる課題を述べた．これからの『大字典』和訓データベースをさらに多くのデータベースと連携させて拡張する．機能面では，各データ項目に正規表現を使える検索を実現したい．

また，データ精度の向上の際に使われる対象データベースは, やはり別の和訓データベースと照合したい. しかし，管見の限りで，利用制限が少ないかつ信用できる公開中の和訓データベースはまだ目に入らなかった．

参考文献

[1] 劉冠偉，李媛，池田証壽．平安時代漢字字書総合データベースの拡張と和訓対応．情報処理学会研究報告．人文科学とコンピュータ研究会報告，2015，vol. 2015，issue 5，p.1-8．
[2] 池田証壽．平安時代漢字字書総合データベースの構築．北海道大学文学研究科紀要．2014，vol. 142，p.79-90．
[3] 高田智和.『大字典』データベースをつくる．じんもんこん 2001 論文集．2001，vol. 109，p. 107-99．
[4] 高田智和.『大字典』データベースをつかう．情報処理学会研究報告．人文科学とコンピュータ研究会報告．2004，vol. 2004，issue 58，p. 45-52．
[5] “Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020 White Paper”. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html（参照　2016-04-17）