Vol.2018-CH-117 No.11
2018/5/12

日本古辞書マークアップ・ツール tagzuke の課題 –操作性・汎用性・維持性の改良–

劉冠偉[†1]

概要：発表者は部首分類体の日本古辞書における注文構造を効率的にマークアップするためのツール tagzuke を開発した．このツールは，翻刻テキストの CSV ファイルを読み込み，マークアップした要素に色付けて表示し，マークアップしていない要素にマウスのクリック操作によってタグ付けし，CSV ファイルとして書き出す，という一連の操作を効率的に行うことができる．これまで観智院本『類聚名義抄』の注文をサンプルとして，ツールの有用性を検証してきたが，改善すべき点をいくつか見出したので，本研究ではそれらについて，操作性，汎用性，維持性の三つの側面から述べていきたい．操作性では，キーボードのカーソル操作を加え，予備に仮タグを付与することなど，汎用性ではエクセルファイルの対応，フィールド・句切り符号の指定が可能とすることなど，維持性ではオープンソースのフレームワークを導入することである．

キーワード：マークアップ，JavaScript，類聚名義抄，言語資源

Challenges of Mark-up Tool tagzuke - Updatings of Manipulability, Applicability and Maintainability-

GUANWEI LIU[†1]

Keywords: Mark-up, JavaScript, Ruijumyogisho, Linguistic sources

1. はじめに

部首分類体の日本古辞書の項目構造は複雑である．例えば観智院本『類聚名義抄』の項目構造は掲出字と注文からなる．注文は大別して音注，意義注，字体注，和訓の 4 要素からなる．この 4 要素には多様な形式で注記が施され，それぞれの要素に数種のタイプが認められる．音注には，反切・類音注，仮名音注，声点などがあり，意義注は漢字 1 文字で注記されることが多いが，2 文字以上のこともある．字体注は，「正」「通」「俗」などが注記される．和訓は，万葉仮名や片仮名で注記され，声点が施されることも多い．

国語学の研究では，このような項目構造の多様性に対応した検索と表示が求められており，それを可能にするには，適切なマークアップが必要であるが，既存のマークアップ・ツールは効率性の点で満足できないところが少なくない．そこで，JavaScript と HTML でマルチデバイス対応のマークアップ・ツール tagzuke を独自に開発した．

このツールを使えば，翻刻テキストの CSV ファイルを読み込み，マークアップした要素に色付けて表示し，マークアップしていない要素にマウスのクリック操作によってタグ付けし，CSV ファイルとして書き出す，という一連の操作を効率的に行うことができる．これまで観智院本『類聚名義抄』の注文をサンプルとして，ツールの有用性を検証してきたが，改善すべき点をいくつか見出したので，本研究ではそれらについて，操作性，汎用性，維持性の三つの側面から述べていきたい．

操作性では，キーボードのカーソル操作を加え，予備に仮タグを付与することなど，ユーザの操作量を減らした．

汎用性ではエクセルファイルの読込み・書出しの対応，フィールド・句切り符号の指定が可能とすることなど，従来，マークアップに導入の前が必須な作業を軽減した．

維持性ではオープンソースのフレームワーク Vue.js[a]を導入することによって，インターフェースをコンポネントに分散して，開発が効率的に行い，オープンソースソフトウェアとして共有性が高めることができる．

順序として，第 2 章では tagzuke の必要性と現状，第 3 章 tagzuke の改善点について述べる．

2. マークアップツールの必要性と現状

2.1 文系研究資源の作成について

これまでの文系の研究では一次資料（原文データ）から二次資料（抽出・加工済みの紙カード），最後は三次資料としての研究成果という流れであった．最近では文系研究でもコンピュータを利用して，データベースやコーパスなどを作成して研究を進めている．つまり，二次資料はデジタルである場合が多くなっている．研究資料のデジタル化は選択的であり，デジタル化する要素があらかじめ標準化されている場合と標準化されていない場合の両方がある．

†1 北海道大学
Hokkaido University

a) https://vuejs.org/

古辞書は日本語史の資料として有益であり，例えば和訓アクセント，字音の声点，字体注記がある．しかし各研究者のやり方によって統一な記載方法はない. 前述のように，部首分類体の日本古辞書の項目構造は複雑である．それを研究するため，我々のグループでは蛍光ペンなどで標識して統計すること試みた．図 1 は観智院本『類聚名義抄』の注文に出現する要素を蛍光ペンで分類する例である. また，鉛筆とかで文献影写の隅に備考を追加したり，関係ある資料をカットしてまとめたりすることもある．これらの二次資料は原本文献の視点から見ると，原本に対して研究者の理解を含める一種な情報追加だともいえる．各研究者が工夫にしてエクセルなどでデータ処理を行なっており，現段階は共通の入力・加工方法をすることはなされていない．

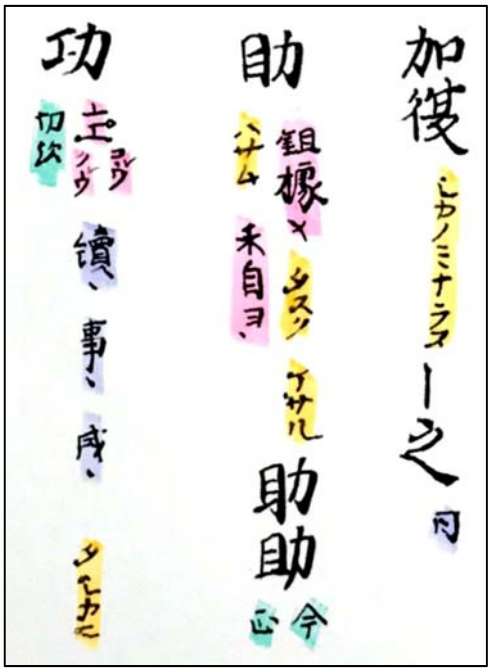

図 1　観智院本『類聚名義抄』の模写による注文マークアップ一例

デジタル二次資料を作成する際には，どのような方法で情報を追加して，どのような形式で追加した情報を記述したか、その過程を逐一記録しておくことは研究資料データベースを作成する際には重要な課題となる．注意したいのは，言語研究者（国語学者）は細部の言語事象へのこだわりが強いことである．一方，マークアップツールを開発するには全体像がなければ進むことができない．そこで大きな枠組の設定ができれば細部に入れる．そこでまず大きな枠組みを提案・改善し，それを踏まえて，さらに細部の検討を進めることにする．

2.2 データベースへのながれ

古辞書の研究の場合，文献に記載されている内容の符号化およびその後の ID 付与やスキーマ設計などによっての構造化ともに，データベースを構築すること自身はすでに情報追加を行った．画像データベースの構築では，文献を画像化にして，画像とテキストデータとの連携によってデータベースができる．図 2 はそれらのイメージである．

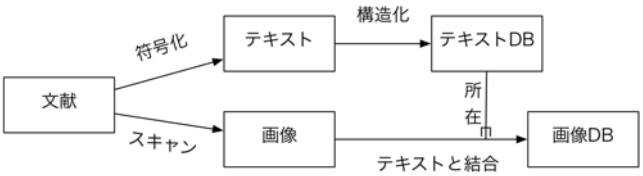

図 2　古辞書系データベース作成のながれ

以下では主としてテキストデータベースの構造化をする際に二次資料の情報追加方法として古辞書に記載されている注文要素の分類について述べていきたい．

テキストデータベースの実装では主にスキーマを事前に定義する RDB とそうではない NoSQL DB に分けられる．古辞書は資料としてすでに一定の構造を持っており，RDB[b]との相性が良いと言えるが，注文を扱う段階に入ると，スキーマの変動が多く生じるため，XML や JSON-LD などで実装するのが理想的であろう．現在では共有形式の統一まだされていないため，本研究で XML 風の簡易タグを用いて古辞書の注文要素にマークアップを行う．

古辞書の分野には，日本の平安時代に編纂された古辞書を中心としている HDIC[c]プロジェクトがある. HDIC は高山寺本『篆隷万象名義』，観智院本『類聚名義抄』などの辞書をデータベース化にして，現在一部分のデータは TSV 形式のテキストファイルとして公開している．本研究は HDIC で公開予定である観智院本『類聚名義抄』データベースを対象にして，古辞書マークアップツール tagzuke の有用性を実験する．図 3 は，図 1 に示した項目構造を注文要素に着目して分類したものである．

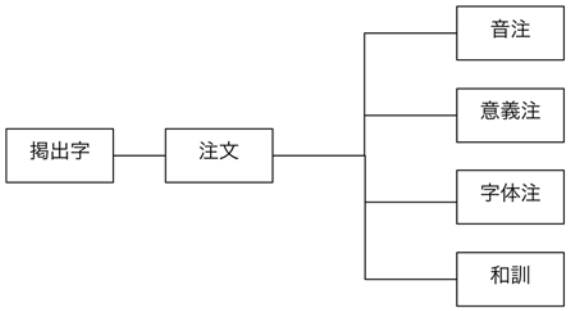

図 3　観智院本『類聚名義抄』の注文要素

2.3 tagzuke の必要性

観智院本『類聚名義抄』の掲出字は約 42,000 字，注文は約 200,000 字と膨大である．その注文の要素をマークアップには効率的な手法が強く求められた．

旧バージョンでは CSV ファイルを読み込み，マークアップした要素に色付けて表示し，マークアップしていない要素にマウスのクリック操作によってタグ付けし，CSV ファイルとして書き出す，という操作のながれであった．

b) 関係データベース
c) http://hdic.jp

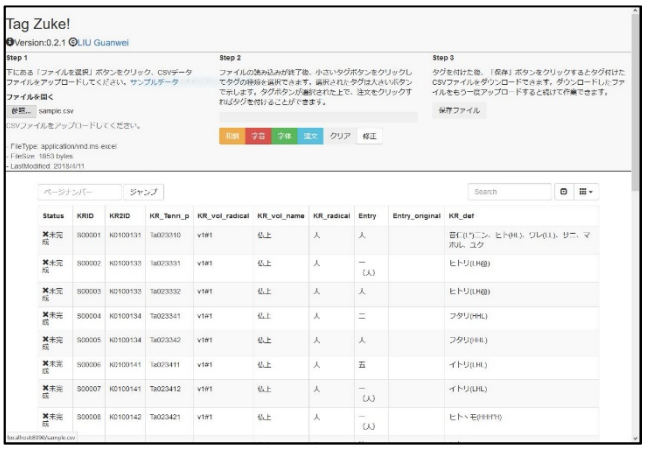

図 4　旧バージョン tagzuke の操作画面

この旧バージョンを用いて，3 名の作業者が約 200 項目をマークアップする実験を行った. 実験の結果として, 1 項目の注文をマークアップするには平均約 6 秒かかった．また，原本のコピーに蛍光ペンで分類する作業では 1 項目に 4.5 秒かかった. 旧バージョンによる作業は蛍光ペンの作業より遅いが，デジタルデータを作成することによって，数量を統計する作業効率の向上に一定の成果があった.

3. tagzuke の改善

3.1 新しい tagzuke の全体像

今回は，旧バージョンのコードを利用せず，一から作り直したものであった。可能な限り，タグ付与の自動化を目指した. 新バージョンの操作の流れはつぎのとおりである。

1 ファイルを開く（図 5）
2 フィールド・区切り文字などの設定（図 6）
3 タグ付け作業（図 7）
4 確認と保存　（図 8）

部首分類体の日本古辞書における注文要素は一定のパターンがある．そのパターンを用いて注文要素の自動マークアップを実現した．詳細は後述する.

tagzuke は左のメニューパネルと右の操作パネルからなる．操作パネルの上部分はタイトル・ツールバー，真ん中部分はテキストデータ，下部分はタグ操作区域である．マルチデバイスに対応するため，画面のサイズによってメニューパネルを自動に隠すなどのレスポンシブデザインを応用している．旧バージョンと比べた相違点を次の表 1 に示す.

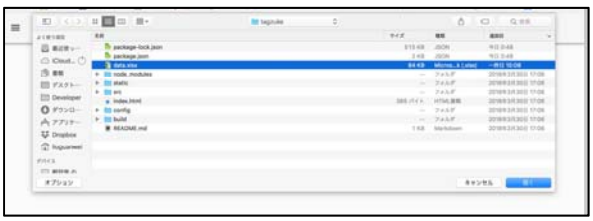

図 5　ファイルを開く画面

図 6　読取りの設定画面

図 7　タグ付けの操作画面

図 8　確認画面

表 1　新旧 tagzuke の機能比較

機能	旧	新
仮タグ	なし	あり
操作	マウス	キーボード・マウス
対応ファイル	CSV	CSV・エクセル

3.2 操作性に関する改善

操作性では，予備に仮タグの付与とキーボードのカーソル操作機能の追加となる.

観智院本『類聚名義抄』の注文要素は漢字のみ，仮名のみ, 漢字と仮名の混在の三つのパターンしかない. さらに，さらに各要素は独自な特徴がある．例えば，字音注記の要素は「音」字で始まるか「反」字で終わるかのものがほとんどである．これらの特徴を利用してファイルを読み取りとともに注文要素の区別を文字列で判定し，特徴に合致し

たタグを事前につけるというアイディアである． 表 2 は注文要素の特徴を整理したものである．

表 2　観智院本『類聚名義抄』の注文要素の特徴

要素	特徴
字音注	①「音」で始まる漢字のみの要素 ②「反」で終わる漢字のみの要素
和訓注	仮名のみの要素
字体注	「正」「俗」「籀」などの字体注記がある漢字のみの要素
漢文注	上記した字音注・字体注以外の漢字のみ要素

また，キーボードのカーソル操作を加えて，仮タグを確認しながらタグを選択する．Tab キーと Shift キー+Tab キーで前後の項目を切り替える．カーソル移動キーの左方向キーと右方向キーで注文の要素を切り替える．上方向キーと下方向キーで要素のタグを選択する．

図 9　tagzuke のタグ操作区域

3.3 汎用性に関する改善

汎用性ではエクセルファイルの読込み・書出しの対応，従来エクセルから CSV テキストファイルまでの変換は不要となる．文字化けや変換のミスで符号ずれなどを避けることができる．

フィールド・句切り符号の指定が可能とする．以前の tagzuke では ID，見出し，注文は必ず既定のフィールドに格納しないとデータを認識してくれなかった．要素の間の区切り符号もユーザが自分で入力できるようになった．

また，観智院本『類聚名義抄』だけではなく，ほかの古辞書データセットにも対応できるようになった．例えば，HDIC ではすでに公開されている高山寺本『篆隷万象名義』データベースのテキストファイルを読み込むとd, 次の図 X で示したように，マークアップすることができる．高山寺本『篆隷万象名義』には和訓がないため，字音注・字体注・意義注をマークアップする．

図 10 高山寺本『篆隷万象名義』をマークアップする画面

3.4 維持性に関する改善

維持性ではオープンソースのフレームワーク Vue.js を導入することによって，インターフェースをコンポネントに分散して，開発が効率的に行い，コードの難読性を減らした．オープンソースソフトウェアとして共有性が高めることができる．ソースコードはオープンソースコード共有プラトホーム GitHub の https://github.com/toyjack/tagzuke に格納して公開している．

3.5 改善後の効率

新しい tagzuke を利用して，二人の実験者が観智院本『類聚名義抄』の約 200 項目を各自にマークアップして，かかった時間を記録した．仮タグによって作用の速度はだいぶ上がって，項目数と時間を加算して，合計 408 項目のマークアップ作業は 19 分で完成した．原本コピーに蛍光ペンおよび旧バージョンの tagzuke と比べると，1 項目をマークアップする平均時間を次の表 3 に示す．

表 3　1 項目あたりマークアップする平均時間の比較

方法	時間
蛍光ペン	4.5 秒
旧 tagzuke	6 秒
新 tagzuke	2.8 秒

4. おわりに

本研究では自動タグ付与，キーボード操作などの機能追加によってマークアップツール tagzuke を改善して，作業の効率向上を行なった．

課題として，仮タグを付与するための判別ルールを追加して仮タグの正確率の向上，エクセル以外の形式の読取りと書出し，electron などによって SPA からデスクトップアプリケーション化の実装などがある．

謝辞　本研究にあたり直接の御指導を戴いた北海道大学・池田証壽先生に深謝する．本研究第 2 章の図 1 で模写画像を提供して戴くとともに，tagzuke の効率検証を戴いた北海道大学・鄭門鎬氏に感謝の意を表する．また，tagzuke の効率検証を戴いた北海道大学・張馨方氏に感謝の意を表する．

参考文献

[1] 池田証壽, 李媛, 申雄哲, 賈智, 斎木正直．平安時代漢字字書のリレーションシップ．日本語の研究. 2016, vol. 12, No. 2, pp. 68-75

[2] 劉冠偉, 李媛, 鄭門鎬, 張馨方, 池田証壽．部首分類体日本古辞書の項目構造の多様性に対応したマークアップ・ツールの開発．じんもんこん 2017 論文集. 2017, vol. 2017, pp. 97-102.

d) ファイルの初頭にあるコメント情報を削除する上に