Vol.2020-CH-124 No.4
2020/9/5

Pali Text Society 版パーリ語文献を対象としたテキスト検索システムの構築

渡邉 要一郎[1,a)]　永崎 研宣[2]　大向 一輝[3]　下田 正弘[3]

概要：上座部仏教の聖典言語であるパーリ語の文献研究は，Vipassana Research Institute によって制作された電子テキストとその検索システムである Chattha Sangayana CD（CSCD）によるデジタル化の波を大きく受けた．しかしこの CSCD が依拠している電子テキストは，ビルマ第六結集版という研究者が標準的に用いるテキストでないものにもとづいたものであった．一般に研究者が用いている標準テキストは Pali Text Society（PTS）によって出版されたものであり，パーリ語の単語や文の位置している頁・行数は PTS 版のそれに従って記述されるのが通例である．そこで筆者は，研究者のニーズを踏まえ，PTS 版の電子テキストを用いて PTS 版の頁・行番号が簡単にとれる検索システムを作成した．

Developement of a Text Search System for Pali Text Society's Edition of Pali Text

Abstract: The study of the literature of Pali, the sacred language of Theravada Buddhism, has benefited greatly from the digitization of the Chattha Sangayana CD (CSCD), an electronic text and retrieval system produced by the Vipassana Research Institute. However, the electronic text relied on by the CSCD is based on the Sixth Buddhist Council Tipitaka Edition, which is not a standard text used by researchers. Generally, the standard text used by researchers is the one published by the Pali Text Society (PTS), and the number of pages and lines where Pali words and sentences are located is usually described according to the PTS version. Therefore, based on the needs of the researchers, the author has created a search system to easily find the page and line numbers of the PTS edition using the electronic text of the PTS edition.

1. はじめに

上座部仏教の聖典言語であるパーリ語の文献研究は，Vipassana Research Institute によって制作された電子テキスト検索システム Chaṭṭha Saṅgāyana CD（CSCD）[*1]によるデジタル化の波を大きく受け，従来は索引や手作業で行っていた作業が瞬時に実行可能となった．

しかしこの CSCD が依拠している電子テキストはビルマ第六結集版という研究者が標準的に用いるテキストでないものを基にしたものであった．一般に研究者が用いている標準テキストは 1881 年にイギリスに創設された Pali Text Society によって出版されたものである．例えばパーリ語の単語や文の位置している頁・行数は Pali Text Society 版のそれに従って記述されるのが通例となっている．

CSCD を使用する際の文献研究者にとっての最大の不便は，標準テキストである PTS 版への参照が不十分であることであろう．PTS 版との対応は，頁数だけが表示されるだけである．つまり，文献研究者は CSCD で検索した結果，あわせて表示された PTS 版の頁数をもとに，実際に PTS 版を開き，CSCD のテキストとの差異に気を配ったうえで，行数を数えて，それから PTS 版の頁数・行数を論文などに記すことになる．このような煩雑な手順がかかってしまう点に関しては改善の余地があるように思われた．

ここで付言しておかなければならないことであるが，筆者は PTS 版のテキストが最良のものであって，ビルマ第六結集版がそれに劣るものであるということを意図しているわけではない．テキスト出版の事情について詳細は省く

1 東京大学史料編纂所
Historiographical Institute, the University of Tokyo
2 一般財団法人人文情報学研究所
International Institute for Digital Humanity
3 東京大学大学院人文社会系研究科
Graduate School of Humanities and Sociology, the University of Tokyo
a) yo-watanabe@g.ecc.u-tokyo.ac.jp

*1 https://www.tipitaka.org/ 最終アクセス 2020 年 8 月 11 日（以下最終アクセス日に関しては同様）．

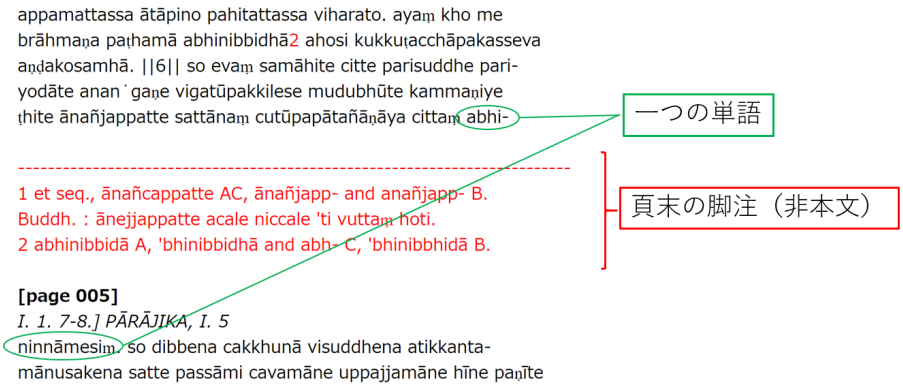

図 1 改頁時に単語が切られる例

が，19 世紀より校訂出版を行っていた PTS と並行する形で，東南アジア各国もその国の文字でのパーリテキストの校訂出版を行っていた．ある PTS のテキストはその現地版のテキストを参照し，また，現地版のテキストも PTS を実見したうえで読みの決定を行っている場合があり，両者の関係は複雑である．ともあれ，文献研究者は PTS 版だけではなく，シンハラ版，タイ版，ビルマ版といった現地テキストをも参照しながら研究を行う必要がある．その意味では，ビルマ第六結集版を検索できるシステムとしての CSCD の有用性と研究への貢献には疑問の余地はない．

ここで見られる問題をやや一般化して述べれば，情報技術は必ずしも研究者コミュニティの意向を反映してくれるわけではないが，技術が発明されてしまった以上，研究者はそれらを使用する必要があり，そこで研究の名目と研究の実情が必ずしも一致しない事態が発生する，ということである．

2. 問題の解決にあたって

2.1 使用する電子テキストの選定

上記のような問題を解決するためには，PTS 版の電子テキストに依拠した検索システムを文献研究の立場にある人間が開発すればよいと思われた．筆者は GRETIL（Göttingen Register of Electronic Texts in Indian Languages）[*2] にアップロードされている，Dhammakaya Foundation により入力された PTS 版の電子テキストを利用することとした．この電子テキストは PTS の判型がそのままに保たれている利点を持っており，頁数と行数が簡単に表示されるようにする目的を達成するために有用であった[*3]．

2.2 GRETIL 電子テキストを利用する際の諸問題

しかし，これらの電子テキストは単なる HTML 形式で PTS 版の外見をそのまま再現しただけのものであって，例えば一つの単語が行・頁をまたぐときにハイフンを付けるなどの，もとの PTS テキストの表記を継承している特徴がある．このような版型の継承は，テキスト検索を行う際には当然障害となってしまう．また，例えばその章のタイトルなど，後代の学者が挿入したテキスト外部の情報がタグ等の機械的に判読できる識別子なしで混ぜ込まれてしまっており，これを検索対象から除外する必要があった（図 1 参照）[*4]．

さらに研究者の慣例を踏まえる課題もあった[*5]．上記の例ではすべて頁・行数の問題ばかりを殊更取り上げてきたが，ある韻文のテキスト群では，一般に PTS 版の韻文番号で指定され，頁・行数は用いられないものがある．また，『スッタニパータ』という韻文・散文が混ざったテキストでは，散文部分は頁・行数で指定され，韻文部分は韻文番号で指定されている．このような文献研究者のローカルルールは，CSCD ではあまり顧みられている印象はなく，改善の余地があった．また，図 2 の例は『ジャータカ』というテキストであるが，刊本の段階でこのテキストは『ジャータカ・アッタカター』という注釈部分（ほぼ散文）と，インデントされてある聖典部分『ジャータカ』（韻文）とをまとめた形で出版されていた．本来，この二つは階層が違うものであるため，検索する際でも混合しないように注意が必要である．このように，個々のテキストの出版事情を考慮した上で，それぞれの電子テキストごとに個別の方針に従ったテキスト整形をする必要があった．

さらにもう一つの課題として，GRETIL で得られる電子テキストは，元となっている PTS 版の刊本が著作権切れとなっている場合であっても，PTS 自体が再配布に消極的であるという課題がある．PTS がパーリ文献研究に果たしてきた役割の重大性は言うまでもないことであり，出版社の意向は尊重されるべきである．

3. システムの実装

上記の課題を踏まえて，アプリケーション本体とデータ

*2 http://gretil.sub.uni-goettingen.de/gretil.html
*3 この電子テキストを利用した主要な成果として，西・逢坂 [2] が挙げられる．
*4 http://gretil.sub.uni-goettingen.de/gretil/2_pali/1_tipit/1_vin/vin3s1ou.htm
*5 一般に研究者は *A Critical Pali Dictionary* [1] の範例に従う．

[page 311]
9. Visavantajātaka. (69) 311
ti. "Tayā daṭṭhaṭṭhānato tvaṃ yeva mukhena visaṃ ākaḍḍhāhīti". "Mayā ekavāraṃ jahitavisakaṃ puna na[1] gahitapubbaṃ, nāhaṃ mayā jahitavisaṃ kaḍḍhissāmīti". So dārūni āharāpetvā aggiṃ katvā āha: "sace attano visaṃ nākaḍḍhasi imaṃ aggiṃ pavisā" 'ti. Sappo "api aggiṃ pavisissāmi na c'; attanā ekavāraṃ jahitavisaṃ paccāvamissāmīti" vatvā imaṃ gātham āha:

Ja_I,7.9(=69).1: Dhi-r-atthu taṃ visaṃ vantaṃ yam ahaṃ[2] jīvitakāraṇā vantaṃ paccāvamissāmi, matam me jīvitā varan ti. || Ja_I:68 ||

図 2 ジャータカの例

GRETIL Pali Text Searcher

Input word(s) to be searched
Use KH-transcription system Input Unicode characters by yourself
Show line-changes Neglect line-changes

Please input maximum number of results shown at ones 100

Please select texts all texts

Vin
DN MN SN AN all
Khp Dhp Ud It Sn Pv Vv Th Thi J Nidd I Nidd II Paṭis Ap Bv Cp
all
Dhs Vibh Dhātuk Pugg Kv Yam all
Mil Vism Sp Ja all

送信

図 3 システムのインターフェイス

を一緒に配布せずに，GRETIL の電子テキストに手を加えた検索用テキストをこちら側から再配布せずに，あくまでも GRETIL の電子テキストをユーザー側が自動でダウンロードし，それをアプリケーション側で自動的に編集し，検索用テキストを生成するスクリプトを作成した．

解決すべき課題としては，第一に検索用テキストの生成，第二に実際の検索システムの作成の二つに分けられた．

3.1 検索用テキストの生成

検索用テキストの生成に際しては，それぞれのテキストがすべて同じ形態をとったり，意味に基づいた画一的なマークアップが行われていない以上，原則すべての電子テキストを実見し，それぞれのテキストごとに固有の検索置換を繰り返し，パーリ語テキストのみを抽出し，これを一つの文字列と見做してテキストファイルに整形したうえで，それに合わせて行数・頁数の区切りが生じる箇所のインデックスの配列を記したファイルでを生成するスクリプトを作成した．韻文テキストに関しては，韻文の区切りと判断できる部分が比較的明確であるので，これを韻文ひとつずつに分割し，合わせてそれぞれの韻文が固有の韻文番号と紐づけられるような CSV ファイルを作成するようになっている．

3.2 検索システムの作成

3.2.1 転写方式への配慮

上記のプロセスを経て生成されたテキストを利用し，単語検索を行う．図 3 のようなシンプルなインターフェイスを作成した．検索対象とするテキストがラジオボタンで選択できるのはもちろんのこと，検索対象とする単語の入力に関しては Kyoto-Harverd 転写方式（KH 方式）が使用できるようにした．というのも，パーリ語やサンスクリット語はもとよりローマ字で書かれていたわけではなく，研究にあたって文字をローマ字に転写することが行われていたわけである．ところが，アルファベットの数だけでパーリ語やサンスクリットの音素全てを表現できるはずはないので，補助的な記号をアルファベットに付けてその音素を表現する必要が生じる．例えば，ā などの a の長母音を記す必要が生ずる．ところで，一般のキーボードではすぐさま ā を書き込むことはできず，手間がかかる．そこで，一般のキーボードで容易にこれらの音素を表現できるような拡張的な転写方式がいくつか生まれた．KH 方式はその一つであり，例えば ā に対して A を当て，ṣ（反舌音の s）に対して S を当てはめるなどである．その転写された A や S に対して，上からフォントをかぶせることで，外見上 ā や ṣ を表現することが可能となる．筆者の作成したシステムでは，この KH 転写方式を用いて単語を入力しても検索可能であるようにした[*6]．これは実装の上では単純な作業であるが，CSCD では配慮されていなかった点であった．

Vin I 10.22-25: imesaṃ hi
Sāriputta pañcannaṃ bhikkhusatānaṃ yo pacchimako bhikkhu so sotāpanno avinipātadhammo niyato sambodhiparāyano 'ti.
Vin I 11.3-4: api ca yo deyyadhammo
so na dinno.
Vin I 19.28-29: purāṇadutiyikāya
methuno dhammo paṭisevito,
Vin I 19.34-35: nanu āvuso bhagavatā anekapariyāyena
virāgāya dhammo desito no sarāgāya,
Vin I 19.35-36: visaṃyogāya dhammo
desito no saṃyogāya,
Vin I 19.36-37: anupādānāya dhammo desito no saupādānāya.

図 4 検索結果の例

Vin I 10.22-25: imesaṃ hi Sāriputta pañcannaṃ bhikkhusatānaṃ yo pacchimako bhikkhu so sotāp
Vin I 11.3-4: api ca yo deyyadhammo so na dinno.
Vin I 19.28-29: purāṇadutiyikāya methuno dhammo paṭisevito,
Vin I 19.34-35: nanu āvuso bhagavatā anekapariyāyena virāgāya dhammo desito no sarāgāya,
Vin I 19.35-36: visaṃyogāya dhammo desito no saṃyogāya,

図 5 改行を反映しない場合

3.2.2 もとの文脈の確認

検索結果は図 4 のようにもとの PTS 版の改行が反映された状態で表示される．前後の文脈も踏まえられるように文章単位で検索結果が表示され，Vin I 10.22-25 などのテキストの頁・行数を示す箇所をクリックすると元の電子テキストの対応箇所にジャンプできる仕組みとなっている．

*6 http://list.indology.info/pipermail/indology_list.indology.info/2009-December/033855.html を参照．

また，図５のように改行を反映することなしに，一行で表示するオプションも用意した．例えばすぐにテキストをノートに引用したいという時などはこちらの方が便利であろう．

3.2.3 利用の簡便性

検索システムの構築にはプログラミング言語として Python，アプリケーションのフレームワークは Flask を利用しており，ウェブブラウザを使用する形とした．また，初回のテキストダウンロードと検索用テキスト作成時にのみオンライン環境が必要で，それ以後にはオフラインで使用可能となっている．また，このシステムを一般に公開するにあたり[*7]，インストーラーを作成し，情報技術に詳しくない一般的な文献研究者であっても容易に使用できるものとした．

4. 意義と課題

前述したような研究者の需要に合わせた検索システムが作成されたということ以外に，電子テキストの再利用という観点からも意義があるように思われる．GRETIL に掲載されている電子テキストは，GRETIL 自身が組織的に電子テキストを作成したというより，既に各地のサイトで公開されている電子テキストを網羅的に収集し，再掲載することを目指しているように見受けられる．そのため，パーリ語のものも含め，あまり統一的な入力方針は存在しておらず，その結果として全体としての利便性が損なわれている面がある．したがって，折角入力された電子テキストが利用され難く，そのポテンシャルがほとんど発揮できてない状況にある．本アプリケーションの作成によって，そのような電子テキストの利用価値を提示できたのではないかと思われる．

今回使用したものは GRETIL に掲載された限りのパーリ語テキストであるが，それ以外の聖典注釈文献などは今回のシステムには組み込めなかった．これは単純に PTS 版の著作権がまだ切れておらず，電子テキストも公開されていないためである．CSCD はビルマ第六結集版の聖典注釈文献等をコンテンツに含めており，使用されているテキストの充実度においてはまだ CSCD の方が優勢である．CSCD が利用していない，サンスクリット語等の他言語の文献との横断検索などを可能にすることによって，CSCD との差別化が更に図られるであろう．

また，現状では使用できる電子テキストが筆者のみによって選択されている状況であるが，行と頁数が明示されているのであれば，利用者自信が作成した他の電子テキストも併せて検索できるようにする等の改良の余地があるように思われる．

[*7] https://github.com/wyoichiro1125/Pali_searcher

参考文献

[1] Ed. V. Trenckner et al.: *Critical Pali Dictionary*, Copenhagen 1924–.
[2] 西 康友・逢坂雄美: 「全パーリ聖典総語彙索引作成の研究—パーリ文献協会編纂文献に基づいて—」, 『中央学術研究所紀要』47, pp. 137–149, 2018.