Vol.2017-CH-114 No.4
2017/5/13

南アジア古典文献の XML によるマークアップ手法に関する考察

鈴木洋平[1,a)]

概要：古典テクストの XML データベース作成，運用に有用であるとされるマークアップのガイドライン Text Encoding Initiative (TEI) はヨーロッパ諸語への対応は手厚いものの，アジア諸語での使用には十分であるとは言えない状況である．東アジア諸語に関しては近年議論されているが，南アジア諸語も特有の問題を抱えてり，TEI に準拠しようとすると未解決の課題が少なくない．そこで本稿ではサンスクリット語文献を事例としていくつかの課題と解決案を提示する．

Consideration on Markup Method by XML in Classical South Asian Literatures

Yohei Suzuki[1,a)]

Abstract: Text Encoding Initiative(TEI) Guidelines are usable for constructing a XML database of classical literatures. Although they are reliable about European languages, it is impossible to use them in Asian literatures sufficiently. There are debates on East Asian languages and letters but there are also many problems in South Asian languages to apply to TEI Guidelines. In this paper, some issues and solutions of Sanskrit texts are presented.

1. はじめに

古典文献の XML によるマークアップの有効な手法に Text Encoding Initiative(TEI[*1]) がある．TEI に準拠したサンスクリット語文献データベースとして，Search and Retrieval of Indic Texts(SARIT) が知られており，現在 50 弱の文献が公開されている．TEI では欧米諸語を念頭に置いた詳細なマークアップ基準が設定されているものの，アジア諸語への研究レベルでの対応が十分であるとは言いがたい．昨今ようやく東アジア諸語への対応が論じられはじめているが，南アジア諸語についても議論がなされるべきであろう [鈴木 2017]．勿論 SARIT では幾つかのサンスクリット語に特有の言語現象に対応しているものの，現状では完全とはいえない問題があり，またエディタや文献によってマークアップ手法の差異が見られる．そこで，本稿ではナーティヤシャーストラ（Nāṭyaśāstra）および，これにアビナヴァグプタ（Abhinavagupta）が施した注釈であるアビナヴァバーラティー（Abhinavabhāratī）を題材に，サンスクリット語が孕むマークアップ上の問題を挙げた上で，対応策を提案したい．

2. アビナヴァバーラティー の文献的特徴

インドの体系的な演劇論として最古の文献が Bharata に帰せられるナーティヤシャーストラ（BhNŚ）であり，成立年代は見解の一致をみないものの，8 世紀には現存の形で成立したとされる［辻 1977: 200］．その現存最古の注釈として，10 世後半から 11 世紀前半にカシミールで活躍したアビナヴァグプタによって記されたアビナヴァバーラティー（AbhiBhā）がある．一般にサンスクリット語の学術文献は師のテクストに注釈を施す形で発展しており，それは単に古い知を引き継ぐだけでなく，潮流の中において新たな動力を生み出している．AbhiBhā もかかる性格を

1 東京大学大学院
a) suzuki-yohei@g.ecc.u-tokyo.ac.jp

*1 欧米語における TEI の有用性とマークアップ手法に関しては小風尚樹氏が実例と共に提言し，文献の多層的な性格を把握する可能性を示唆している [小風 2015].

強く持っており，注釈という様式を取りながらも BhNŚ から大きく展開した独自の思想を有する．また，先行する注釈からの引用が多く，その種が多岐に渡る点も AbhiBhā の特徴といえる．

本稿の題材に Abhibhā を取り上げた理由として，上記に加え，韻文と散文を含んでいること，間接引用と直接引用の両者がなされており，その引用の長さも様々で，マークアップのベンチマークとして相応しいと考えられる点がある．

3. サンスクリット語文献のマークアップにおける諸課題と解決

サンスクリット語の表記に関する特徴として，連声（Sandhi）がある．すなわち，一定の条件下でアンシェヌマンを起こし，その結果が筆記に反映される．例えば，iti uktam（〜と言われた）[AbhBhā on BhNŚ: 28.1] は ityuktam と表記される．

また，サンスクリット語文献では，文藝作品のみならず，学術文献においても韻文が多用される伝統がある．内容が韻律に収まらないとき，例えば以下の例 [BhNŚ: 6.27–28] におけるように，偈の句切れと内容のまとまりが異なる場合がある．

śārīrāś caiva vaiṇāś ca
sapta ṣaḍjādayaḥ svarāḥ /
tataṃ caivāvanaddhaṃ ca
ghanaṃ suṣiram eva ca // 27 //
caturvidhaṃ ca vijñeyam
ātodyaṃ lakṣaṇānvitam /
tataṃ tantrīgataṃ jñeyam
avanaddhaṃ tu pauṣkaram // 28 //
身体に属するものと竪琴に属するものの
七種のシャッジャ音に始まる音がある．
弦楽器，膜鳴楽器
鳴り物，管楽器という，（27）
四種と知られるべきである，
定義を具えた楽器は．
弦楽器は弦に拠り，
膜鳴楽器は太鼓と知られるべきである．（28）

この場合，27cd[*2] と 28ab（太字部分）が意味上の区切れとなり，韻律上の区分とは一致しない．

以下本稿では，この二点を中心に考察を進める．

3.1 連声の処理

連声の処理についてはマカリスターによって論じられている [鈴木 2017]．本稿でも，マカリスターの方法に従って，上記の例では<q><w lemma=”iti”>ity</w></q><q>uktam</q>
とマークアップする．連声は母音間のものと子音間のものがあり，母音間の場合には kanyā iva（乙女の如く）を kanyeva と表記するなど，文字数の変化が発生する上に，連声の後分の語頭母音までもが音声変化を起こす．従ってこの場合には連声前の形を記すことによって，コーパス検索などにおいてより有利になる．子音間の連声に関しては，通常ローマ字表記では単語間での分かち書きを行うため，SARIT でも処理を行わないケースが目立つ．しかし，デーヴァナーガリー文字の刊本では単語末が子音である場合には分かち書きをしないため，ローマ字への転写を行う際に分かち書きをしてしまった場合, 分け方に関して多かれ少なかれ編集者による解釈が介在することになる．このような恣意性を排除するために，分かち書きの基準は刊本に従うことにした [永崎 2006]．従って，子音連声に関しても母音と同様の処理を施した [*3]．

これに準じて，連声が起きていないときでも，単語が子音で終わり次の単語が母音で始まる際は連声の表記に従ってマークアップした．というのもデーヴァナーガリーの分かち書きの基準は当然発音上の都合に基づいており，発音に際しては分かたれていない単語間ではアンシェヌマンを起こしていたと考えられる．このことは分かち書きの法則と連声規則が本質的には同問題であることを示唆する．

3.2 韻文のマークアップ

韻文のマークアップ関しては，次の二点を考慮して決定する必要がある．第一に，詩論上，韻律の分析は重要であり [*4]，伝統的な韻律解釈に堪えなければならない．第二に，前節のような読解上の問題にも対応すべきである

このことは，伝統学問と近代文献学の両立を迫る課題に他ならない．そこで，前者については区切りのマークアップで，そして後者については複層的なメタデータの付与で対応する．韻律の区切りについてはマカリスターが極めて有意義な提案をしている [鈴木 2017]．すなわち，a パーダと b パーダ間，c パーダと d パーダ間は<caesura/>タグで区切り，パーダ間で単語を跨ぐ場合や連声にも応じる [*5]．更に一偈を<l>タグで，一行を<lg>でマークアップ

[*2] サンスクリット語の韻文は二行からなり，それぞれの行が半分ずつで別れ，それぞれ（四分の一偈）がパーダ（pāda）と呼ばれる．上記の BhNŚ: 6.27 では, śārīrāś caiva vaiṇāś ca が a パーダ，sapta ṣaḍjādayaḥ svarāḥ が b パーダで，ここまでが一行分である．

[*3] サンスクリット語では文末に立てる子音が制限されており，他の子音が文末に来ると，規則的に文末に立てる音に変化する．この法則による変化にはビューワーでの対処を見込んだため,XML ファイルで変化前の音を明示することは避けた．

[*4] 韻律はヴェーダの系譜で学問として成立し，その後言語哲学的思索と文藝技術の一致でもある詩論の中で高度に発展した．複雑かつ多岐に渡る韻律が生み出され，また審美の的となった．

[*5] 一方で行や偈末は文法上文末とみなされるので，原則として単語がこれらを跨いだり，行や偈を超えて連声や綴りが生じることは

し，パーダ，行，偈という三様の区分を反映している．

韻律のメタデータについて一偈を ID でタグ付けし，他の文献へのリンクを統一した．これは SARIT でも少なくない文献に施されているが，現状では徹底されているとは言いがたい．また，筆者は行毎にラベリングを行い，より詳細な参照を可能にした．

上記の規則に従って，ここで先述の BhNŚ: 6.27 のマークアップ案を提示する．

```
<lg xmlìd=”BhN6.27”>
 <l><label>6.27ab </label>
  <q>śārīrāś</q><q><w lemma=”ca”>cai</w></q>
  <w lemma=”eva”>va</w> <q>vaiṇāś</q>ca<caerura/>
  sapta ṣaḍjādayaḥ svarāḥ /
 </l>
 <l><label>6.27cd </label>
  tataṃ <q><w lemma=”ca”>cai</w></q>
  <q><w lemma=”eva”>vā</w></q>
  <w lemma=”avanaddham”>vanaddhaṃ</w>
  ca<caerura/> ghanaṃ <q>suṣiram</q>eva ca // 27 //
 </l>
<lg>
```

4. むすび

以上，連声と韻文という二点から精確なマークアップへの提示を行った．前者はサンスクリット語の言語的特徴を照らし出して反省する営みであり，文法学や言語学の知見を拠所としながら，表記と音声の関係を再考することとなった．一方，後者において，伝承の流れとそれを客観的に観察する近代インド学を包含する新しい視座を設定する必要が生じた．本稿で取り扱いきれていない問題は多い．例えば，複合語 *6 や文法的な分析，辞書機能とのリンク，そしてビューワーへの高度な対応など，様々な積み残した課題がある．しかし，多層的な性格を持つサンスクリット語文献に対して TEI に準拠しつつサンスクリット語の環境に応用した XML によるマークアップが構造上極めて有用であることが一層明らかになった．

参考文献

AbhiBhā　Vyas, K. ed. *Nāṭyaśāstra of Bharatamuni, with the Commentary Abhinavabhāratī by Abhinavagupta.* 4th edition. 4vols. Baroda, 1992–2006
BhNŚ　Ibid.
小風尚樹：19 世紀イギリス政府文書における財政・統計関連史料のマークアップ例提示, 情報処理学会研究報告, Vol. 2015-CH-106, No. 7, pp. 1–5, 2015
鈴木洋平：イベントレポート　講演会”Encoding Sanskrit Śāstra: The TEI for Indic Scientific Treatises”, 人文情報学月報, No. 068, 第 68 号［後編］, 2017
辻直四郎：『サンスクリット文法』岩波書店, 1974
――：『シャクンタラー姫』岩波書店，1977
永崎研宣：シラブルを最小単位とする仏教哲学文献データベースについて, 情報処理学会研究報告, 2006-CH-071, pp.33–40 (2006)
――：インド学仏教学分野におけるデジタル媒体の活用と課題, 印度学仏教学研究, 60-2, 2012
Search and Retrieval of Indic Texts(SARIT), http://sarit.indology.info/exist/apps/sarit/works/, 2017 年 4 月 18 日参照
Text Encoding Initiative(TEI), http://www.tei-c.org/index.xml, 2017 年 4 月 18 日参照

ない．

*6 SARIT には複合語を”.” で区切っている文献もあるが，複合語の前分と後分の間で連声が起きている場合や，複数の解釈が存在するケースに対応しきれているとは言いがたい．