Vol.2013-CH-98 No.6
2013/5/11

論理構造と物理構造が混在するテキストの XML によるマークアップに関する考察

高橋晃一[†1]

XML は文献学的なテキスト分析にも応用が期待されるマークアップ言語である．特に TEI によってガイドラインが策定されてから，その利用価値は一層高まっている．本研究では，『中辺分別論疏』という仏教文献を取上げ，複雑な構造を持つテキストを TEI P5 によって分析する具体的事例について報告する．『中辺分別論疏』は 6 世紀頃にインドで作成されたサンスクリット語文献である．写本が全体的に 3 分の 1 程度欠損しているため，校訂テキストは欠損部を諸資料から再現し，イタリック表記で補完している．すなわち「欠損した写本」という物理的な事情を反映したテキストであり，通常の論理構造と物理構造が混在した状態になっている．こうした複雑なテキストを TEI P5 に準拠してマークアップする方法について考察する．

Consideration about Marking up a Complicated Critical Edition Influenced by the Substantial Form of Manuscript by Using XML

KOICHI TAKAHASHI[†1]

Today XML (Extensible Markup Language) is expected to be applied to the philological analysis. Especially the TEI P5, the useful guidelines for encoding texts, provides many tools to mark up various kinds of documents. The present paper aims to consider about the potentiality of TEI P5 through the examination of marking up a text with a complex structure. The manuscript of the *Madhyāntavibhāgaṭīkā*, a Buddhist text composed in Sanskrit about 6th century, is partially damaged. The third part of each folio is completely lost. Then, the modern critical edition of this text attempts to reconstruct the lost part by means of philological investigation. The reconstructed parts, which are shown in italics in that edition, are irregularly inserted without any association with the logical context. In this sense, this critical edition reflects the complicated material form of the manuscript as well as the logical structure. This paper reports how to mark up such a complex structure according to the TEI P5.

1. はじめに

XML（Extensible Markup Language）は自由にタグを決められるということから，文献学上のテキスト分析にも応用が期待されている．しかし，一方で標準的なタグセットがないということは，使用者ごとに異なるタグを恣意的に設定することになり，情報の共有，データの可読性に問題が生じる可能性がある．TEI P5 はこうした課題に対する一つの解決である．TEI P5 とは，Text Encoding Initiative という団体によって策定されたガイドラインであり，文献資料を XML によってマークアップする際に用いるタグを予め標準化することで，利用者ごとに使用するタグが異なるという状況を解消しようとしている[a].

現行のガイドライン P5 はかなり豊富なタグセットを提供しており，それらを組み合わせて用いることで，一般的に扱われる文献はほとんど処理することができるように思われる．しかし，このガイドラインは基本的に欧米の文献を念頭に置いて定められているため，アジア・アフリカなど，より広い範囲の文献研究に対して十分に対応できているか否かは未知数である．今後は各分野の文献研究者が個々の関心にしたがって TEI P5 を実際に利用することにより，その有効性を検証しながら，課題を提起することが求められている．言い換えれば，ガイドラインの汎用性を高め，情報の共有化を進めるために，人文学者の貢献が不可欠になりつつあるのが現状と言える．少なくとも XML による文献分析の領域では，文献学者はコンピュータ技術とは無縁であると決め込み，情報学側から提供される成果のみを受動的に利用するということはあり得ない．分析対象となる文献は，それを扱う専門家でなければ，的確にマークアップすることができないからである．

こうした視点から，今回は『中辺分別論疏』という文献を XML による分析する際の課題について検討する．

2. 『中辺分別論疏』について

『中辺分別論疏』とは，『中辺分別論』というインド仏教の哲学文献に対する注釈である。原題は *Madhyāntavibhāgaṭīkā* という．*Madhyāntavibhāga* が注釈対象である『中辺分別論』の原題であり，*ṭīkā* は「複註」という意味である．この文献の由来は 4 世紀頃まで遡る．その頃，インドの大乗仏教に新たな一学派が登場した．ヨーガの実践を重視したことから瑜伽行派（ヨーガーチャーラ）と称され，また「一切唯識」を標榜したことから，唯識学派とも呼ばれた．

†1 東京大学大学院
Graduate School, the University of Tokyo

a) 詳しくは参考文献[1]を参照.

『西遊記』の三蔵法師のモデルで知られる玄奘がインドで学んだのはこの学派の思想であり,彼の創始した法相宗は,南都六宗の一つとして日本でもなじみ深い.

インドの瑜伽行派の開祖はマイトレーヤとされる.いわゆる弥勒菩薩のことであり,半ば伝説と考えるべきであろう.しかし,瑜伽行派の草創期の思想家達はこのマイトレーヤから新たな教えを受けたと信じていた.その教えの中の一つが『中辺分別論』であった.マイトレーヤの教示は韻文で著されていたが,それに対して,唯識思想の大成者の一人であるヴァスバンドゥ(400-480)が散文で注釈を付したと言われている.したがって,厳密には韻文の部分が『中辺分別論』本論ということになるが,実際には散文注釈の中に韻文の本文が織り込まれる形で伝承されている.韻文だけの単立のテキストがあったわけではないらしい.そのため,韻文と散文をあわせて『中辺分別論』と呼び習わしている.

『中辺分別論疏』とは,この『中辺分別論』に対する注釈である.『疏』とは複註を意味する.ヴァスバンドゥの注釈に対する,さらなる註ということだが,実際には韻文箇所も解説している.注釈者はインドの 6 世紀の仏教学者スティラマティである.瑜伽行派の思想に精通していたほか,仏教学一般に詳しく,多くの注釈文献を残している.ただし,独自の著作はない.彼の注釈家としての態度は,極めて客観的であり,文章や術語について複数の解釈があり得る場合には,それらをすべて併記している.今日,瑜伽行派のみならず,サンスクリット仏教文献の研究にとって,非常に貴重な資料を提供している.

3. 写本と校訂テキスト

『中辺分別論疏』にはサンスクリット語原典の写本が現存している.全体は 85 葉よりなり,20 世紀の初頭にネパールで発見された.ただし,状態は非常に悪く,各葉とも 3 分の 1 程度が欠損している.そのため,全体像は古典期のチベット語の翻訳に依らなければならない.現在は,この写本はドイツのハンブルク大学の Nepalese-German Manuscript Cataloging Project により管理されている[b].

このサンスクリット写本に基づき,1934 年に山口益教授が校訂テキストを公表している[c].このテキストでは欠損部分の原文をチベット語訳から想定し,補完している.

専門的な話になるが,サンスクリット語仏教文献の多くは,8 世紀後半からチベットにもたらされ,チベット語に翻訳されている.訳経事業に当たって,欽定訳語という形で翻訳に用いる語彙を統制し,またある程度定式的な翻訳をしているため,サンスクリット語の原文を想定しやすいという事情がある.そのため,かつてはチベット語訳から,サンスクリット原典の復元を試みる研究がなされた時期があった.山口校訂本もそうした試みの一つと言える.

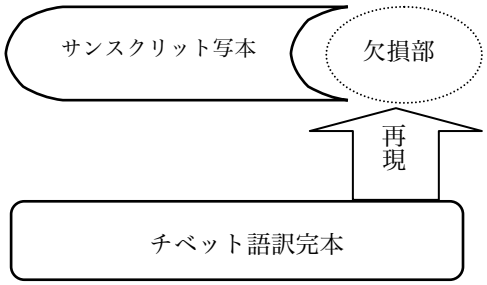

図 1 サンスクリット語写本欠損部の再現イメージ

山口校訂本の体裁では,想定された原文は次のようにイタリックで表記されている.

atha vā praṇetṛipraṇeya[vaktṛivākyasamādāna]-pravacanāt sūtra*praṇetṛivaktṛivṛttiṣu gāurvotpādan-ārtham āha/*

śāstrasyāsya [Tib.19,a] *praṇetāram/*

iti sarvam/

*tatra praṇetrā vaktum upadiṣṭāt sūtre gāurava*m utpadyate/ yasmād asya kārikāśāstrasyāryaMaitreyaḥ praṇetā/ sa caikajā*tipratibaddhāt sarvabodhisattvā-bhijñādhāraṇīprtisaṃvitsamādhīndriyakṣāntivimokṣāḥ paramampāraṃgataḥ sarvāsu bodhisattvabhūmiṣu niḥśeṣam api prahīṇāvaraṇaḥ/ vaktṛisamādānadvāreṇa vṛittyāṃ* gāuravam utpadyate/

(*Madhyantāvibhāgaṭīkā* ed. by S. Yamaguchi, 1,11-2,5)

図 2 山口校訂本の表記例

その後も,類似の研究がいくつか発表されたが,学術的に山口校訂本が学術的に最も評価されている[d].

4. テキストの論理構造と物理構造

『中辺分別論疏』の刊行テキストは上記の図のようにイタリック表記を含んでいる.この箇所は写本が欠損しているために,校訂者によって復元されたサンスクリット語であり,注意を促すために,このように表記されている.出版形態としては当然の配慮といえる.しかし,イタリック表記の箇所は,写本の損傷という偶発的な事故によって生じたものである.したがって,テキストの内容とは全く無関係にイタリック表記が現れることになる.その点で,山口校訂本の形態は「欠損した写本」という物理的な事情に少なからず影響を受けていると言える.

また,一方で,山口校訂本は段落分けがなされている.

b) 文献[2]参照.
c) 文献[3]参照.
d) 参考文献[4]参照

通常サンスクリット語の写本では，文章構造に従った段落分けはなされず，行の左端から右端まで文字を隙間なく書き込んである．現代の校訂では，これを論理構造に従って段落分けすることが常識となっている．山口校訂本もこの形式をとっている．すなわち，現代の文章表現としては常識的な，論理構造を明確にしたテキストでもある．

さらにもう一つこのテキストが内包している情報がある．『中辺分別論疏』は先にも述べたように「複註」であり，これ自体が単立の著作ではない．そのため，注釈対象である『中辺分別論』から常に術語や文章が引用され，それに対する解説がなされるという構造になっている．テキストの性格上，必然的に内在する情報である．これは厳密には論理構造ではないが，読み手が文献の性質を理解し，それにしたがって読解しなければ，誤解を生じかねないという意味で看過できない．なお，山口校訂本では，注釈対象として『中辺分別論』本論から引用された章句に下線を付す場合もあるが，必ずしも統一されてはいない．

このように，山口校訂本『中辺分別論疏』は「物理構造」と「論理構造」を内包したテキストと言える．どのような古典文献の校訂テキストでも，基本的には「物理構造」を含まざるを得ない．例えば，写本の改行位置やファリオあるいはページの変わり目など，論理構造とは無関係の情報も，通例，校訂テキスト内に何らかの形で明示されている．しかし，山口校訂本の場合は単なる改行・改頁の表記では収まらない．写本の欠損の状態と本来そこにあるべきだったテキストの再現がなされているからである．架空のテキストなのだから，そもそも無視するという判断もあり得るが，学術的に評価されている校訂本であり，また原文の再現も文献学上の手続きを踏んでいる「研究成果」であるため，それらすべてを含めて一つのテキストの形態として扱うのが望ましいように思われる．

5. 基本的構造のマークアップ

次にこのような複雑な構造を持つテキストのマークアップについて考察する．今回は XML でタグ付けを行うにあたって，TEI P5 に準拠する．先に引用した例文を用いて，まず一般的にテキストの構造を表すために段落構造を表す (paragraph)タグと，改頁を表す<pb/>(page break)，改行を表す<lb/>(line break)を付けると次のようになる．なお，下図でイタリック表記されている個所は実際のテキストファイルではローマン体になるが，本論文では，便宜上イタリックで表記する．

ちなみに，改頁・改行を表すタグは開始・終了が一つのタグで完結している[e].

** **atha vā praṇetṛipraṇeya[vaktṛivākyasamādāna]-pravacanāt **<lb n="12"/>**sūtra*praṇetṛivaktṛivṛttiṣu gāurvotpādanārtham āha/*

<pb n="2"/><lb n="1"/>*śāstrasyāsya* [Tib.19,a] *praṇetāram/*

<lb n="2"/>*iti sarvam/*** **

** <lb n="3"/>***tatra praṇetrā vaktum upadiṣṭāt sūtre gāurava*m utpadyate/ **<lb n="4"/>**yasmād asya kārikā- śāstrasyāryaMaitreyaḥ praṇetā/ sa caika**<lb n="5"/>***jāti- pratibaddhāt sarvabodhisattvā bhijñā-dhāraṇīprtisaṃvitsamādhī***<lb n="6"/>***ndriyakṣānti-vimokṣāḥ paramampāraṃgataḥ sarvāsu bodhi-sattva***<lb n="7"/>***bhūmiṣu niḥśeṣam api prahīṇā-varaṇaḥ/ vaktṛisamādānadvāreṇa* **<lb n="8"/>***vṛittyāṃ* gāuravam utpadyate/** **

図 3　一般的なタグ付けの例

一般的なテキストであれば，これで十分に構造を表すことができる．しかしながら，山口校訂本はすでに述べたような複雑な構造を持っている．以下では，その最大の特徴である「写本欠損部の再現箇所」を表記する方法を考える．なお，<pb/><lb/>は煩雑になるのを避けるためにこれ以降省略する．

校訂テキストの表現形式自体に着目すれば，イタリックにより強調されていると見なすことができるので，<emph rend="italic"> (emphasized/ rendered in italics)を用いることも考えられる．あるいは<hi>(hilighted)でも構わない．しかし，<emph>あるいは<hi>は単独では強調表現となっている「理由」を示すことができない．山口校訂本は，独自の想定原文であることを示すためにイタリック表記を使用しており，そこにはこの校訂者の特殊な意図が込められている．したがって，単に表現形式によるだけでは，十分に校訂テキストの内実を反映できないことになる[f].

一方，<damage>や<supplied>というタグもある．特に後者は，校訂者による補完も念頭に置いている[g]．今回の例には適していると言える．なお，<emph>, <hi>, <damage>, <supplied>は，テキスト構造を表すための<floatingText>タグを子要素として取ることができる[h]．この<floatingText>は何らかの挿入的な文章をマークアップするものである．厳密には「本文を一旦妨げる挿入文であり，またその挿入文の終了後，本文が再開するような場合」を想定している．“floating text”という概念は XML のある種の幾何学的な構造と文献資料の持つ「割り切れなさ」を媒介するために，TEI P5 で考え出された概念といえる．XML では，タグ同

e) 文献[1]Appendix C Elements の各項目を参照.

f) 文献[1] Appendix C Elements の各項目を参照.
g) 文献[1] Appendix C Elements の各項目を参照.
h) 文献[1]. Appendix C, Elements <emph> May contain *textstructure* floatingText

士は「ネスト」と呼ばれる入れ子構造が保たれなければならない．これは単にタグをまたいではいけないというだけでなく，そもそも XML が，樹木が一つの根から次第に枝分かれしていくようなイメージを持っていること，数学的な幾何学模様を描くように，分析対象を階層的に要素ごとに分解可能であると見ていることに関わっている．しかし，これはすべての文献に当てはまるわけではない．文献資料には時として，不意に入り込んだ文章というものがある．そのような文章は，文脈上，必然性のある引用でもなく，そのため，引用一般を表す<quote>タグでは対応できない．そのために考え出されタグが，<floatingText>である[i].

ただし，山口校訂本の場合は，「本文の文脈を遮る別な文章」のではなく，むしろ「本文が遮られないようにするための本来あるべき文章」なので，TEI P5 の想定する floating text には厳密には当てはまらない．しかし，仏教文献研究者の立場から，あえて言えば，山口校訂本のイタリックの個所は，まさに「読み手の思考を一旦妨げる」文章でもある．再現された原文の妥当性を無意識のうちに検証しようとしてしまうのである．その意味では floating text と呼んで差し支えがないように思われる．いずれにせよ<floatingText>が分析対象の割り切れなさを埋めるための概念であるなら，その適用の可能性を広げることも，今後の課題として提唱すべきであろう．

以上の検討を踏まえて，ここでは，<floatingText>タグを用いてタグ付けを行う[j].

atha vā praṇetṛipraṇeya[vaktṛivākyasamādāna]-pravacanāt sūtra**<supplied rend="italic"><floatingText type="reconstructed">** *praṇetṛivaktṛivṛttiṣu gāurvotpādanārtham āha/ śāstrasyāsya* [Tib.19,a] *praṇetāram/ iti sarvam/***</floatingText></supplied>**

<supplied rend="italic"><floatingText type="reconstructed">*tatra praṇetrā vaktum upadiṣṭāt sūtre gāurava***</floatingText></supplied>**m utpadyate/ yasmād asya kārikāśāstrasyāryaMaitreyaḥ praṇetā/ sa caikajā**<supplied rend="italic"><floatingText type="reconstructed">***tipratibaddhāt sarvabodhisattvābhijñādhāraṇīprtisaṃvitsamādhīndriya kṣāntivimokṣāḥ paramampāraṃgataḥ sarvāsu bodhisattvabhūmiṣu niḥśeṣam api prahīṇāvaraṇaḥ/ vaktṛisamādānadvāreṇa vṛittyāṃ***</floatingText></supplied>** gāuravam utpadyate/

図 4　再現箇所のタグ付案

6. 注釈文献としての論理構造をマークアップ

これまでの要領で基本構造を記述することができるようになったが，これだけではテキストデータを XML で電子化するメリットはあまりない．そもそも『中辺分別論疏』は注釈文献であり，語句の解釈を提示している点が重要なのである．実際の文献学研究の場面においても，語句の解釈を確認するために利用されることが多い．したがって，将来，語句説明の検索を行うことなどを考慮すると，注釈対象となる語句とそれに対する解説文をそれぞれマークアップし，関連付けておくことが重要であろう．

語釈のマークアップは<gloss>タグを用いる．注釈対象となる章句に<term>タグを付し，ID を与えておいて，それに対して@target を用いて<gloss>タグと関連付ければよい[k].

<term xml:id="vklp"><supplied rend="italics"><floatingText type="reconstructed">grāhyagrāhaka**</floatingText></supplied>**vikalpaḥ**</term>/ <gloss target="#abhtprklp" type="etym">**hastyādyākāra śūnyamāyāyām iva hastyākārādayaḥ/ abhūtam asmin dvayaṃ parikalpyate 'nena ve**</gloss>**ty **<term xml:id="abhtprklp">**abhūtaparikalpaḥ**</term>**/ **<gloss target="#abhtprklp" type="gloss" n="1">**abhūtavacanena ca yathāyaṃ parikalpyate grā**<supplied rend="italics"><floatingText type="reconstruction">**hyagrāhakatvena tathā nāstīti prdarśayati/ parikalpavacanena tv artho yathā prikalpyate tathārtho na vidyata iti pradarśa**</floatingText></supplied>**yati/ evam asya grāhyagrāhakavinirmuktaṃ lakṣaṇaṃ paridīpitaṃ bhavati/**</gloss>** kaḥ punar asau/ **<gloss target="#abhtprklp" type="gloss" n="2">**atītānāgata vartamānā hetuphalabhūtās traidhātukā anādikālikā nirvāṇaparyavasānāḥ **<supplied rend="italics"><floating Text type="reconstruction">**saṃsārānurūpāś citta caittā aviśeṣeṇābhūtaparikalpaḥ/ viśeṣatas tu grāhyagrāhakavikalpah/ tatra grā**</floatingText></supplied>**hakavikalpaḥ arthasattvapratibhāsaṃ vijñāna m/ grāhakavikalpa ātmavijñaptipratibhāsam/**</gloss>**

(*Madhyantāvibhāgaṭīkā* ed. by S. Yamaguchi, 13,17-14,3)

図 5　注釈文献としての論理構造

しかし，ここでもやはり物理構造との衝突は避けられない．下線を施した箇所は，以下に示すように，詳細には二つの語釈で構成されている．

i) 文献[1]4.3.2 Floating Texts 参照
j) このほかに，
k) 文献[1] Appendix C Element 各項目参照

（1）
atītānāgata vartamānā hetuphalabhūtās traidhātukā anādikālikā nirvāṇaparyavasānāḥ *saṃsārānurūpāś citta caittā aviśeṣeṇābhūtaparikalpaḥ/*
（2）
*viśeṣatas tu grā*hyagrāhakavikalpaḥ/ tatra grāhakavikalpaḥ arthasattvapratibhāsaṃ vijñānam/ grāhakavikalpa ātmavijñaptipratibhāsam/

図 6 下線部の詳細な構造

（1）の後半から（2）の冒頭にかけて再現テキスト（イタリック表記）となっている．これにタグを付す場合，やはり論理構造を優先させ，その後，物理構造に支配されている再現テキストをタグ付することになる．すなわち，まず<gloss>によって（1）（2）それぞれをマークアップし，それぞれのイタリック表記の部分を，先と同様に<supplied> + <floatingText>で表現する．結果として，次のようになる．

（1）
<gloss target="#abhtprklp" type="gloss" n="2_1"> atītānāgata vartamānā hetuphalabhūtās traidhātukā anādikālikā nirvāṇaparyavasānāḥ **<supplied rend="italics"><floatingText type="reconstruction">**saṃsārānurūpāś cittacaittā aviśeṣeṇābhūtaparikalpaḥ/ **</floatingText></supplied> </gloss>**
（2）
<gloss target="#abhtprklp" type="gloss" n="2_2"> <supplied rend="italics"><floatingText type="reconstruction">viśeṣatas tu grāhyagrāhakavikalpaḥ/ tatra grā**</floatingText></supplied>**hakavikalpaḥ arthasattva- pratibhāsaṃ vijñānam/ grāhakavikalpa ātmavijñapti- pratibhāsam/**</gloss>**

図 7 「図 6」に対するタグ付けの例

7. 今後の課題

今回のマークアップは人文学者の視点が色濃く反映している．そのため，タグ付されたファイルはかなり複雑な様相を呈している．ただし，これらはテキストを分析する上では不可欠な要素であるため，仮にその結果としてソースファイルの可読性が下がるとしても，やむを得ない面もあろう．

問題は，複雑化したソースファイルを XSLT で扱うには限界があるということである．このような複雑な XML ソースファイルを扱うために必要な技術を，各人文学者が修得すべきか，議論が分かれるところだろう．しかし，XML による文献分析が人文学の専門家の責任でなされるべきであるのと同様に，その利用法まで視野に入れて，研究がなされるべきであろう．

参考文献

1) TEI Consortium: *TEI P5: Guidelines for Electronic Text Encoding and Interchange 2.3.0*, originally edited by C.M. Sperberg-McQueen and Lou Burnard for the ACH-ALL-ACL Tex Encoding Initiative Now entirely revised and expanded under the supervision of the Technical Council of the TEI Consortium (Last updated on 17th January 2013). http://www.tei-c.org/Guidelines/P5/(2013/04/15)
2) Nepalese-German Manuscript Cataloging Project, http://catalogue.ngmcp.uni-hamburg.de/(2013/04/15)
3) Yamaguchi, S.(ed.): *Madhyāntavibhāgaṭīkā Exposition Systématique du Yogācāravijñaptivāda*, Nagoya, Librairie Hajinkaku (1934)
4) 塚本啓祥他編, 梵語仏典の研究 IV 論書編, pp.334-338(1990).