

東アジア文献への TEI の適用をめぐる

王 一凡^{1,2} 永崎 研宣²

概要: 漢訳仏典を主体とする SAT テキストデータベースの TEI 移行をめざすにあたり、東アジア文献に多くみられる形式を現行の TEI 標準で記述する際に生じるいくつかの問題点に注目し、それらに対する可能な解決策を議論する。

Remarks on Application of TEI to East Asian Documents

WANG YIFAN^{1,2} NAGASAKI KIYONORI²

Abstract: This paper discusses some problems and possible solutions regarding TEI markup of conventional and pre-vailing styles in East Asian documents, which will be relevant to our TEI migration project of the SAT Text Database, mainly consists of Chinese and Japanese Buddhist scriptures.

1. はじめに

筆者らの取り組む SAT テキストデータベースプロジェクト^{*1}（代表：下田正弘）は大正新脩大藏經のデジタル化を推進するにあたり、従来のデータを国際的なガイドラインである TEI に準拠した XML 形式に改めることを検討している。なお、すでに中華電子仏典協会 (CBETA)^{*2} や叢ら [1] による取り組みをはじめ、多くの先行事例があるものの、大正新脩大藏經を表現するには未だ検討すべき事項が多く、TEI ガイドラインの側でも対応しきれない面がある。したがって、大正新脩大藏經の TEI エンコーディングには様々な観点からの探究が必要となっている。

TEI ガイドラインを策定する TEI コンソーシアムは、テキスト資料の電子的表現における世界的な標準を確立することを目的としているが、歴史的経緯から欧米の研究者が中心となって策定されており、人的交流や技術的な問題のため東アジアからの参加はこれまで限定的であった。そのため、東アジア文献にみられる構造上、形式上の様式を

TEI で表現するには課題が残る部分が散見される [2]。近年、永崎と A. Charles Muller 氏の主導により TEI に East Asian/Japanese SIG（東アジア・日本分科会）が設立され、東アジア文献のエンコーディング問題の解決に向けた動きが端緒に就いたばかりである。

本稿では大藏經の TEI エンコーディングに関連する若干の論点を取り上げ、初歩的な検討を加えてみたい。本内容が日本を含む東アジア文献の TEI 化に関する議論を深化させる契機や、前述の分科会への有益なリソース提供となれば幸いである。

2. 割注の表現

東アジアの伝統的な版式においては、注釈を対象となる本文の直後に、一行を二行に分割して細字で書き込む形式が一般的に行われている。このような注は付随的ないしは例外的な現象では決してなく、しばしば詳細であったり、また連綿と続く注釈の伝統のために、重層しながら数頁以上にわたるなど非常に長くなることがあり、当該書籍において主要なコンテンツを占めることもしばしばである。

割注を表現するにあたり、構造面では例えば `<gloss>` や `<note>` で囲むことによって取り扱うことができる。例として図 1 の一節は次のようにエンコーディングできる（以下 `type="double"` は割注を示す仮の属性値）。

¹ 東京大学大学院教育学研究科

Graduate School of Education, University of Tokyo

² 人文情報学研究所

International Institute for Digital Humanities

³ 東京大学大学院人文社会系研究科

Graduate School of Humanities and Sociology, University of Tokyo

^{*1} <http://21dzk.l.u-tokyo.ac.jp/SAT/>

^{*2} <http://www.cbeta.org/>



図1 割注を含む写本

<p>
...<phr xml:id="a">或用麼多
之文重增其麼多而音必兼之</phr>
<note target="#a" type="double">
如部林二合字從
<w xml:id="b"><g ref="#xxx" /></w>
<gloss target="#b" type="double">
菩侯反
</gloss>
<w xml:id="c">婁</w>
<gloss target="#c" type="double">
力鉤反
</gloss>
與第十一麼多也
</note>
</p>

しかし割注の形式的側面を再現しようとすると、割注の中の改行と親行の改行を区別する必要があるが、現行のTEIガイドライン (TEI P5[3]) で行区切りに使用できる要素は <lb> のみである。既存の自由度が高い語彙を使用して、

```
<lb rend="warichu" />
```

```
<lb type="warichu" />
```

などと表現できなくもないが、以下のような課題が生じる。

- 割注が非常によく使われる形式であるのに対し、共通化された語彙がないと交換可能性に支障が出る
- 割注は入れ子になることもあるため、拡張性が必要

したがって、割注内の改行を示す要素に行の階層を明示するしくみを追加する必要があるのではないかと考える。この時、

- (1) 割注のために新たな要素を追加する
- (2) <lb> に新たな属性を追加する
- (3) <lb> 要素の@corresp に割注要素を参照させるなどの方法が考えられる。

(1) の場合, 仮に `<wb>` という要素と割注の深さを示す `@stratum` を追加することで, 以下のようにエンコーディングできる.

```
<p>
...<phr xml:id="a">或用麼多
<lb />之文重增其麼多而音必兼之</phr>
<note target="#a" type="double">
  <wb />如
  <wb />部
  <lb />林二合字從
  <w xml:id="b">
    <g ref="#xxx" />
  </w>
  <gloss target="#b" type="double">
    <wb stratum="2" />菩侯
    <wb stratum="2" />反
  </gloss>
  <w xml:id="c">婁</w>
  <gloss target="#c" type="double">
    <wb stratum="2" />力鉤
    <wb stratum="2" />反
  </gloss>
  <wb />與第十一麼多也
</note>
</p>
```

この方法の利点は、非対応のパーサに影響を与えず処理が可能であることだが、新しい要素という比較的大きな変更を加えることとなる。

次に, (2) を用いれば, 例えば

<lb stratum="1" />

を（１段階目の）割注として解釈することとなる。

```
<p>
...<phr xml:id="a">或用麼多
```

```
<lb />之文重増其麼多而音必兼之</phr>
<note target="#a" type="double">
  <lb stratum="1" />如
  <lb stratum="1" />部
  <lb />林二合字從
  <w xml:id="b">
    <g ref="#xxx" />
  </w>
  <gloss target="#b" type="double">
    <lb stratum="2" />菩侯
    <lb stratum="2" />反
  </gloss>
  <w xml:id="c">婁</w>
  <gloss target="#c" type="double">
    <lb stratum="2" />力鉤
    <lb stratum="2" />反
  </gloss>
  <lb stratum="1" />與第十一麼多也
</note>
</p>
```

この場合、新属性を解釈しないパーサには区別なく改行として扱われることとなるが、要素を変更せず、また属性値を限定することで、割注標示を標準化することができる。

(3) を採用した場合、

```
<p>
  ...<phr xml:id="a">或用麼多
  <lb />之文重増其麼多而音必兼之</phr>
  <note xml:id="xyz" target="#a"
    type="double">
    <lb corresp="#xyz" />如
    <lb corresp="#xyz" />部
    <lb />林二合字從
    <w xml:id="b">
      <g ref="#xxx" />
    </w>
    <gloss xml:id="zxy" target="#b"
      type="double">
      <lb corresp="#zxy" />菩侯
      <lb corresp="#zxy" />反
    </gloss>
    <w xml:id="c">婁</w>
    <gloss xml:id="yzx" target="#c"
      type="double">
      <lb corresp="#yzx" />力鉤
      <lb corresp="#yzx" />反
    </gloss>
    <lb corresp="#xyz" />與第十一麼多也
```

```
</note>
</p>
```

このように既存の語彙のみで正規化可能であるが、割注は非常に一般的に使われるため煩雑さが否めないだろう。

また、いずれの方式でも、<gloss>などの意味的な囲み要素に割注範囲の規定というレイアウト上の処理を求めることとなる。テキストの構造を表現する目的においては問題ないとはいえ、改行位置を適切に共有しようとする場合には、専用の処理系を実装する必要があるため、なお検討を要する。

3. 尾題の表現

東アジアの古典籍では書籍または巻の題を先頭と末尾に示す慣習がある。これは題ではあるが内容と連続しているわけではないため<explicit>には該当しない。また奥付でもないため<colophon>でもない [3]。この尾題を表現するには多くの方法が考えられる。

(1) <finalRubric> 内の <title> として <msItem> 内に記載

```
<msDesc>
  <msItem>
    <finalRubric>
      <title>尾題</title>
    </finalRubric>
  </msItem>
</msDesc>
```

(2) 特別な標識つきの <title> として <body> の下に記載

```
<body>
  <p>
    <title type="final">尾題</title>
  </p>
</body>
```

(3) <closer> 内の <title> として <body> 内に記載

```
<body>
  <closer>
    <title>尾題</title>
  </closer>
</body>
```

(4) <title> として <back> の下に記載

```
<back>
  <p>
    <title>尾題</title>
  </p>
</back>
```

いずれも一長一短と思われ、意味的には (2) が最も近いのではないかと考えられるが、形式上 (1) のような処理が望ましいとも考えられる。これらは取り扱い方により記述箇所が大きく異なってしまうことから、どのような選択肢を採用するか、よく検討する必要があるだろう。

4. 外字の取り扱い

漢籍をはじめとして、東アジア文献をエンコーディングする際に避けて通れないのが外字である。TEI は UTF-8 に準拠しているため、Unicode の未登録字が外字として扱われることとなる。漢字外字に関していえば、依然として古典籍をエンコーディングするには多くの未登録字に直面する一方、Unicode 側でも学術用の漢字追加が急ピッチで行われているため、現時点での外字がかなりの割合、近い将来に符号化文字で代替できる可能性が高い。

TEI には現在 `<g>` `<glyph>` などの外字を取り扱うしくみがある [3] が、個別の文書内で外字を定義すると、将来外字の情勢が変化した際に追従する負担が大きい。また、標準化の過程においては諸般の事由によりグリフの統合や分離を行う必要があり、当該文献内の字との対応が不明瞭になる可能性が常にある。そのため、特に大規模な TEI コレクションを維持していくためには、スキーマなどを通して外字のデータベース化を進め、かつ Unicode の議論と同期できるようなメカニズムが必要と考えられる。

この点については、王が Unicode の関連会議である IRG での漢字符号化の審議を追跡するシステムを開発している [4] ため、それとの連携を視野に入れて取り組んでいきたい。

謝辞 本稿の基となった議論の一部は Marjorie Burghart 先生、Dot Porter 先生のご教示を得ました。感謝申し上げます。

参考文献

- [1] 叢 艶, 高久雅生: 唐詩作品の本文フルテキストに対する TEI マークアップ手法の提案, 情報知識学会誌, Vol. 28, No. 2, pp. 174–185 (2018).
- [2] 永崎研宣: デジタル文化資料の国際化に向けて: IIIF と TEI, 情報の科学と技術, Vol. 67, No. 2, pp. 61–66 (2017).
- [3] TEI Consortium: TEI P5: Guidelines for Electronic Text Encoding and Interchange (online), 3.4.0, 入手先 <http://www.tei-c.org/Guidelines/P5/> (参照 2018-07-25).
- [4] 王 一凡, 永崎研宣, 下田正弘: グラフデータベースによる文書リポジトリ統合管理システムの設計, 研究報告人文科学とコンピュータ (CH), Vol. 018-CH-117, No. 8, pp. 1–6 (2018).