仏教文献のための構造的なデジタルテクストの記述と活用

永崎研宣

仏教文献に限らず，歴史的文化的に意義のあるテクストをデジタル化しようとする営みは今や世界中に広がっている．テクストを効率的に共有し活用することはデジタル化の意義をより深めることになる．そのためにはテクストの構造を記述しつつ，その構造を共有していくことが今のところ一番の近道である．そして西洋古典から近代に至るテクストについてはすでに様々な構造的記述手法が開発され，TEI Guidelines（Text Encoding Initiative P5 Guidelines）として広く共有されている．仏典に関してはCBETAがこの手法を漢文大蔵経に採用しており，SARITプロジェクトではサンスクリット文献に採用している．しかし一方で，現在の所，TEI Guidelinesは西洋文献の研究者が中心となって策定されてきたルールであり，西洋以外の文化圏におけるテクストの状況を十分に反映しているとは言いがたい面がある．テクストの構造は言語・文化によって様々な特性があり，それを十分に活かした形でなければ，構造化することで逆に多様性の大切な芽を摘み取ってしまう状態に陥る可能性もある．本稿では，TEI Guidelinesの状況を踏まえつつ，仏教文献に焦点をあてた構造的な記述手法について提示するとともに，その活用方法・メリットについて検討したい．

構造的なデジタルテクストの記述

一次資料にせよ二次資料にせよ，テクストには様々な仕方で構造が埋め込まれている．これには意識して埋め込まれたとは限らないものもあり，意識的なものかどうかの境界線を引くことは困難だが，それでも，記述した人が意識的に埋め込んだ構造は少なからず存在するはずである．それは，読者として想定するコミュニティにおいて理解されるであろう仕方で構成され記述されている．一次資料，校訂テクスト，校訂テクストに基づく発展的な研究，そしてそれぞれに関する研究論文，といった形で，それぞれに構造があり，それらが棲み分けられつつ，必要に応じていくつかの構造が同時に採用されることもあるというのが現状だろ

う．

一方，人文学に関わる資料を対象として，こういった様々な構造を共有できる範囲でなるべく効率的に共有しようとするのがTEI（Text Encoding Initiative）協会であり，指針としてこのTEI協会から公表されているTEI Guidelinesである．効率的に共有するためには，コンピュータで読み取りやすくすることも必要であり，そのためには，なるべく広く普及しているコンピュータ上の規格に依拠しつつ，コミュニティにおいて受容可能なルールを作っていくことが比較的妥当な方向性だろう．TEIコンソーシアムでは，そのような方針に基づいてTEI Guidelinesを作製・公開し，現在も改良を続けている．もちろん，人文学資料においてあり得るすべての構造を網羅できているわけではなく，汎用性の向上と個別への配慮とのバランスの中で，現在も模索が続けられている．たとえば，ごく近年では，2011年末に，「ドキュメント志向」を全面的に採り入れてTEI Guidelinesを改訂したことがあった．「ドキュメント志向」とは，それまでのTEI P5 Guidelinesの方針であった「テクスト志向」に対置される，テクストの構造の記述の仕方についての考え方である．「テクスト志向」は，当初からのTEIの基本方針であった，内容的な論理構造を持つテクストを前提とし，その論理構造を記述することを目指すという考え方である．頁や行を記述するルールも用意されてはいたものの，あくまでも従属的な情報という位置づけであった．これについて，写本や自筆原稿等の手書き資料を扱う研究者からは不十分であるとの声が多かった．そういった資料について内容的な構造を見いだしてから記述していこうとしたなら，共有できるようになるまでに相当の時間を費やしてしまうことにもなりかねない．むしろ，文書に見えているテクストの状況をなるべくそのままに，見えている構造を記述し，共有していくことが最初の一歩としては重要である．「ドキュメント志向」が台頭した背景にはそういった問題意識と必要性があったようである．そこで，相当の時間をかけた検討の結果，「ドキュメント志向」のテクスト構造化の手法がTEI GuidelinesのP5 version 2.0から採り入れられたのである．ここで確認しておきたいのは，TEI Guidelinesがそのようにして人文学の方法論的な要請に着実に対応しながら改訂を進め，人文学資料全体にとって有益な構造的記述手法たらんとしてきているという点である．

校訂情報の記述手法と現状

本稿では，構造的なデジタルテクストの記述と活用の事例として，脚注におけ

る校訂情報に焦点を当てる．校訂情報のデジタル化の手法としては，TEI Guidelines には（1）Location-Referenced Method,（2）Double End-Point Attachment Method,（3）Parallel Segmentation Method の三種が紹介されている．技術的には一長一短あるが，いずれも，どの一次資料に基づく異読かということを明示できるようになっている．これに基づいて作成されたデータは，任意の書式に整形したり，必要な情報だけを抜き出して統計処理を行ったり，といった活用が可能となる．SAT 大蔵経テキストデータベース研究会（代表：下田正弘東京大学教授）において作成・公開されている大正新脩大蔵経（以下，大正蔵）のテクストデータ（以下，SAT テクスト）の場合には，XML が登場する以前の 90 年代半ばに計画されたものであり，校訂情報の記述方法については XML を念頭に置かずに設計されたものであり，また，TEI Guidelines についても，当時はまだ東洋の文献に適用することはかなりの困難さを伴うものであったことから，独自の形式での構造化が行われている．この方式は，TEI Guidelines における Location-Referenced Method とほぼ同じ構造となっていることから，必要であれば，そのルールに則って TEI/XML 形式文書へと変換することはある程度自動的に可能である．

大正蔵の校訂情報には，いくつかの独特な規則があり，プログラムで処理するにはやや曖昧で難しいものも散見される．データ量として，約 600 万行，約 1 億字，脚注数が約 75 万という膨大な数となっているため，当初の枠組みを変更することは難しいことから，当面は，現状のものに依拠する形で Web サービスに供しており，現在の Web サービスでは，本文中の脚注番号のボタンをクリックすると脚注の内容がポップアップ表示されるようになっている．

また，校訂情報を分析するには，それがどのような出自の資料を用いてどのように遂行されたのかという状況についての確認が欠かせない．大正蔵の校訂の歴史的経緯と問題点についてはすでにいくつかの研究があるので詳しくはそちらをご参照いただくとして，以下に簡潔に触れておきたい．

大正蔵の校訂について

我国では黄檗版大蔵経開版以降，これが大蔵経としては特に流布していた．明代の嘉興蔵に基づくものであったが，その後，明治時代に入り，活版を用いた約 400 巻に上る縮刷版の和装本，『大日本校訂縮刷大蔵経』（縮刷蔵）が刊行された．ここで校合に用いられたのは，増上寺に所蔵されていた木版大蔵経である，高麗大蔵経再雕本，宋本（思渓蔵），元本（普寧蔵）であり，これに加えて，明代の大

蔵経として，上述の黄檗版が用いられたようである．高麗大蔵経再雕本は，10世紀宋代に蜀の成都にて開版された『開宝蔵』の復刻版として11世紀に高麗にて作られた初雕本をただ再雕しただけでなく，唐の長安の中原写経を淵源とするとされる『契丹蔵』をも照合して編纂されたものであり，伝統的にはもっとも良い大蔵経であるとされてきたものである．縮刷蔵では，これを底本としつつ他の三版を照合して校合を行っていった．

その後，仏教学の発展とともに聖語蔵や敦煌写本の内実が明らかになるにつれ，新しい知見を反映させた新たな校訂大蔵経の出版の機運が高まり，着手されるに至ったのが大正蔵である．高楠順次郎・渡辺海旭を都監に，仏教学者の力を結集して行われたこの出版事業は，まずは，縮刷蔵を少し大きな活字で印刷し直した頻伽精舎刊行のものをベースとして作業が進められていった．このため，縮刷蔵の校訂情報がそのまま引き継がれることも少なくなかったようである．全体として，膨大な仕事を短期間に行ったことから，誤植の混入は無理からぬ事であり，近年においても誤植の指摘が方々でなされている．しかしながら，大正蔵の校訂情報の正確性について全般的に検証するには，諸般の事情により，まだかなりの時間を要するものと思われる．

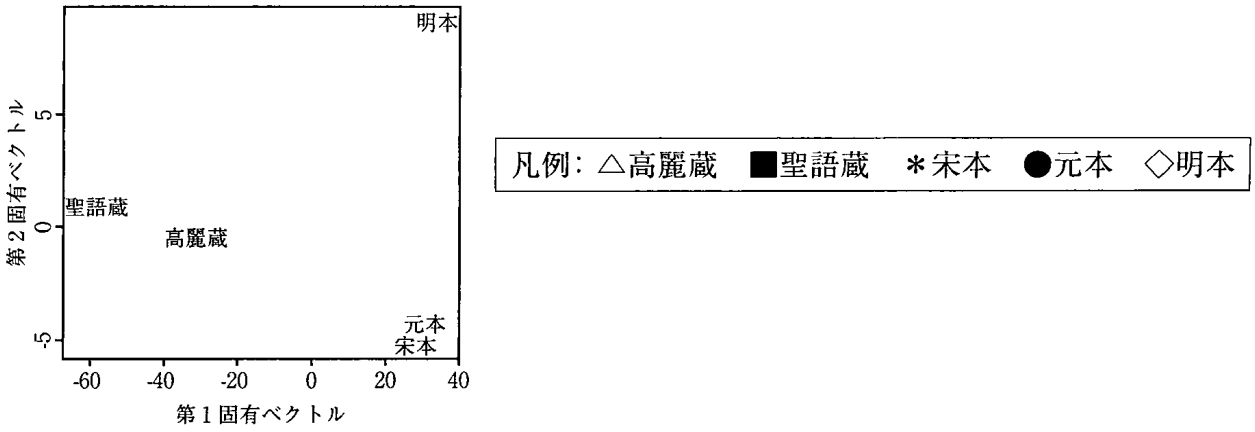
このようなことから，新たな研究成果に基づく校訂情報の追加だけでなく，既存の校訂情報の検証という仕事もまた，より良い研究の基盤としての大蔵経データベースを共有していくためには避けては通れない．そして，そこでは適切な計画と着実な実践が必要となる．この計画にあたっては，既存の校訂情報について事前に分析しておくことが有益であると考えられることから，筆者は共同研究者らとともに，デジタル化された大正蔵の校訂情報の分析を試みた．分析結果自体はすでに人文情報学関連のシンポジウムにおいて公表済みだが，ここでは，それを紹介するとともに，仏教学の立場からその意義についてさらに検討したい．

校訂情報の解析

SATテクストの脚注における校訂情報はすでに上述のような形式でデジタル化されているため，「現在の大正蔵において校訂情報として記載されている各版同士の相違が何文字ずつあるか」というデータ（以下，これを編集距離と呼ぶ）を取り出すことができるように思われる．しかし，実際のところ，上述のとおり，大正蔵の記述方式が機械処理という観点からはあいまいな面があることや誤植があること，それに加えて，デジタル翻刻に際しての誤植があり得ることを考慮に入

れると，大正蔵で校合された資料間の編集距離を正確に導き出すことは現時点では困難であると言わざるを得ないが，それに近い数字を取り出すことはできるだろう．

また，大正蔵に収録されたテクストには様々なものがあり，すべてが一様に同じ一次資料群に基づくものではない．宋・元・明の各木版大蔵経を参照しているものは多いが，それらに加えて正倉院聖語蔵を校合対象としたテクストとなるとかなり限られてくる．現在の脚注データを検索して聖語蔵と校合したテクストを抽出したところ，全体で190点であった．そこで，まずはこの190点全体での編集距離を計測し，多次元尺度法を用いてプロットしてみたのが以下の図1である．編集距離の計測にあたっては，1文字あたりの相違・増加・欠如をすべて1とした上で，図では最大値を100とした相対値を用いている．

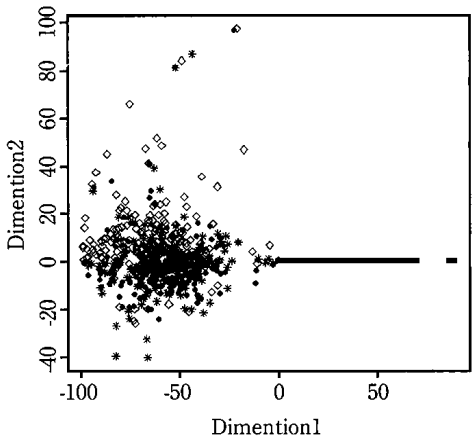


（図1：多次元尺度法を用いて計測した各版の関係）

全体としては，高麗蔵はどれからも距離があり，聖語蔵は他の諸版から離れつつ，高麗蔵が最も近くなっていることがわかる．さらに，縦軸は横軸に比べて目盛の数値差が小さいことから，宋本と元本との違いは極めて小さく，明本も図表上は離れているが聖語蔵に比較するとそれほど離れていないということがわかる．この図からは，木版大蔵経に関してはその成り立ちを大体反映していると言ってよいのではないかと思われる．

次に，テクスト毎に確認してみよう．ここで前提として留意しておきたいのは，まず，聖語蔵収録経典には欠巻が少なくないにも関わらず，「聖語蔵が校合されているテクスト」という単位で確認してしまっていること，宋元明本との校合については上述のような事情を反映していること，といった点である．それを

踏まえた上で，図2に注目されたい．



（図2: 多次元尺度法を用いて計測したテクスト毎の各版の関係1）

図2では，190点のテクストのそれぞれについて編集距離を測って相対値をとった上で，高麗蔵を中心点としつつ高麗蔵と聖語蔵を正の方向で同一線上に揃えつつプロットした．この図が意味するところは，聖語蔵に関しては，どのテクストにおいても，高麗蔵が最も近く，宋・元・明本はそれよりも遠いということである．

ただし，ここでまず注意しておかねばならないのは，「校訂情報の脚注がない箇所は同じとみなされる」という点である．機械処理上の制約により，今回はテクスト単位で比較してしまったことから，聖語蔵の場合には，欠巻となっている箇所は，この方式では高麗蔵と同じであり異文が存在しない，とみなされることになる．そのようにして，「聖語蔵が最も近い」と判定されたテクストも少なくないであろうことは念頭に置いておいていただきたい．

しかし，それでもなお，すべての聖語蔵テクストから見て最も近いのが高麗蔵であるというこの結果には，木版の4大蔵経のうち高麗蔵だけが契丹蔵の系統を継いでいるという事実を反映しているのかもしれないという期待を抱かせるものがある．

また，個別にテクストを見ていった時，特に編集距離が極端な形で現れるものに，『出生無辺門陀羅尼経』（大正No. 1018），『過去荘厳劫千仏名経』（大正No. 446），『開元釈教録』（大正No. 2154），『顕揚聖教論頌』（大正No. 1603），『東方最勝燈王如来経』（大正No. 1354）があった．このうち，『開元釈教録』については，聖語蔵の目録には見当たらず，さらに，縮刷蔵において明本となっている脚注が大正蔵

では「聖」と記載されていたため，大正蔵編纂時の誤植であったと思われる．より正確な編集距離の測定にあたっては，校訂情報の再チェックが欠かせないことが改めて確認されたとも言えるが，このように，まずは極端なものを丁寧にチェックしていくことで，留意すべき特徴が明らかになっていく可能性があることから，今後は，上記のテクストを中心に，校訂情報の確認を行っていきたい．また，今回は取り組めなかった巻毎の傾向についても，確認してみたい．

終わりに

ここまでの検討から，構造化されたテクストが計算処理を通じてテクスト上の気づきにくい事実に気がつくための手がかりを提供し得ることについて，おぼろげながら明らかにできたと言ってよいだろう．今後は，この結果を踏まえつつSATテクストの校訂情報の状況について調査を進めていくと同時に，TEI Guidelinesのより深い適用可能性についても検討していきたい．

〈謝辞〉本稿は，SATテキストデータベース研究会（代表：下田正弘東京大学教授）によって構築されたSATテクストに依拠して執筆されたものであり，SATテクスト構築にご尽力いただいた皆様に深く感謝する．

〈参考文献〉

1. Nagasaki, Kiyonori, Toru Tomabechi, and Masahiro Shimoda. 2013. "Towards a Digital Research Environment for Buddhist Studies." *Literary and Linguistic Computing* 28 (2): 296–300. doi:10.1093/llc/fqs076.
2. Burnard, Lou, and Syd Bauman. 2007. *TEI: P5 Guidelines*. http://www.tei-c.org/Guidelines/P5/.
3. 島田蕃根．1908.『島田蕃根翁』島田蕃根翁延寿会．http://kindai.ndl.go.jp/info:ndljp/pid/781562.
4. 山崎清華．1928.「異字の撰擇に就いて」『現代仏教』11月号，103–15.
5. 大蔵会編．1964.『大蔵経――成立と変遷』百華苑.
6. 船山徹．2008.「漢語仏典――その初期の成立状況をめぐって」『漢籍はおもしろい』，71–118．研文出版.
7. 永崎研宣，三宅真紀，苫米地等流，A. Charles Muller, 下田正弘．2013.「人文学資料としてのテクスト構造化の意義を再考する」『人文科学とコンピュータシンポジウム論文集』2013 (4), 239–46.

〈キーワード〉 TEI，大正蔵，聖語蔵，校訂情報

（一般財団法人人文情報学研究所主席研究員）