和文体および漢文体をもつ資料の構造化 —法華百座聞書抄の事例研究—

河瀬 彰宏[†]（国立国語研究所コーパス開発センター）
野田 高広 （国立国語研究所コーパス開発センター）

Structuring Documents Having both Japanese and Chinese Style of Writing: Encoding *Hokke hyakuza kikigaki sho* as a Case Study

Akihiro Kawase (National Institute for Japanese Language and Linguistics)
Takahiro Noda (National Institute for Japanese Language and Linguistics)

要旨

本研究では、国立国語研究所の「通時コーパス設計」プロジェクトの一環として『法華百座聞書抄』について、和漢混淆文の文書—『今昔物語集』のテキスト—と比較しながら、文書構造化を検討し、具体的事例を示す。『法華百座聞書抄』とは、天仁 3 年（1110 年）2 月 28 日から 300 日間にわたって講じられた法華経・阿弥陀経・般若心経の説教の聞き書きである。唯一の伝本である法隆寺蔵本には、20 日分の講説（説教・説話）が片仮名主体の漢字仮名交じり体で筆録されている。本研究では、この翻刻テキストを底本として構造化を進める。このテキストには、傍書に加えて、校注者による振り仮名・振り漢字が付されている。この振り漢字は、『今昔物語集』の文書構造化には見られなかった特徴であり、構造化上の問題となる。本研究では、文書全体の構成から文字レベル（#PCDATA）に至るまでのタグセットを精緻に考慮し、構造化における問題点を整理する。

１．はじめに

国立国語研究所（以下、国語研）では、「通時コーパス設計」プロジェクトの一環として古典資料の形態素解析を実施している。形態素解析を行うためには、基礎資料となる古典テキストの電子化および構造化が必要となる。

筆者らは、これまでに平安時代の和漢混淆文の構造化（冨士池ほか 2013）、近世口語テキストの構造化（河瀬ほか 2013；Kawase et al. 2014）、狂言台本の形態素解析（河瀬ほか 2014）などを進めている。また、国語研では他にも『太陽コーパス』（田中・小木曽 2000）や『明六雑誌コーパス』（近藤・田中 2012）、BCCWJ（前川 2008、山口ほか 2009）などの様々なテキストコーパスを電子化し、公開している。

上記のテキストコーパスは、国語研が独自に考案したタグセットに基づく XML（eXtensible Markup Language）を用いて文書構造のマークアップを行っている。しかし、各々のコーパスを規定する要素には、共通のタグが使用される場合が少なからずあるものの、基本的には共通のタグセットを使用していない。そのため、同一コーパス内での文書構造の比較や文字列の抽出は可能である一方で、複数のコーパス間の構造比較や計量分析を機械的に実施することが現状では難しいという問題を抱えている。したがって、複数のコーパスの構造を高次の視点から統一的に記述することが求められている。

本研究では、この問題を解決するために、和漢混淆文の重要資料である『法華百座聞書抄』の翻刻テキスト（小林 1975）を事例に、タグセットを考案し、文書構造化を試みる。そして、過去に構造化を実施した和漢混淆文の資料である『今昔物語集』のタグセットと

[†] a_kawase@ninjal.ac.jp

の相違点を整理し、和漢混淆文の資料における構造化の問題点を明確化する。

２．『法華百座聞書抄』の特徴と電子化の意義

『法華百座聞書抄』とは、天仁３年（1110 年）２月 28 日から 300 日間にわたって講じられた法華経・阿弥陀経・般若心経の講説の聞き書きである。唯一の伝本である法隆寺蔵本には、計 35 の説話を含む 20 日分の講説が片仮名主体の漢字仮名交じり体で筆録されている。『日本語学研究事典』（飛田良文ほか（編）2007）によれば、おもに次の３点の特徴をもつことから、重要な言語資料であると考えられている：（１）説法に因縁話や比喩談などの説話を加える形式をもつため、説話集の成立と説法の関係を知る手掛かりとなること。（２）典拠を仏典に求めているため、漢文訓読語の影響が著しく見受けられること。（３）中世語の萌芽と目される新語が出現していること。

一方で、片仮名主体の漢字仮名交じり体で書かれた本文は、説教・説話のどの部分に和文脈・漢文脈の要素が現れるのか、新語がどのように現れるのか、文体の指向が鎌倉時代の和漢混交文とどのような関連をもつのか、などが未解明である（飛田良文ほか（編）2007）。したがって、『法華百座聞書抄』を機械可読な形式に整備することは、これらの問題を計量的観点から分析することを実現させ、日本語史や書誌学などの人文学研究を促進する意義がある。また、同時代には、漢文脈傾向の強いテキストが数多く存在しているため、これらの文献資料をアーカイヴ化するためのフォーマットを新たに統一的観点から提供する意義がある。

３．『法華百座聞書抄』の構造

コーパス言語学の観点から『法華百座聞書抄』を分析するためには、文書の外形的情報だけでなく、文書構造および言語構造を精緻にマークアップすることが求められる。

図 1 は、『法華百座聞書抄』の翻刻テキスト（小林 1975）をワープロ印字し直したものである。

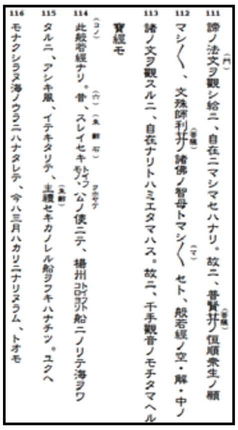


図 1　『法華百座聞書抄』のワープロ印字（抜粋）

『法華百座聞書抄』は、（ａ）前付け部分と（ｂ）講説部分の順に構成される。（ａ）前付け部分には、内題「天仁三年二月廿八日令始修法一百座」、書写者「大安寺僧都永」、序文が記されている。（ｂ）講説部分には、計 35 の説話が収録されており、その基本構成は聞き書きされた日付、経の品名、講師名などのコンテンツ情報と、片仮名主体の漢字仮名交じり体で筆録された本文を含む（図 2）。以下では、このような構造をもつ『法華百座

聞書抄』のテキストを精緻にマークアップしていく。なお、片仮名で書かれた本文はすべて平仮名に置き換えた上で作業を実施している。

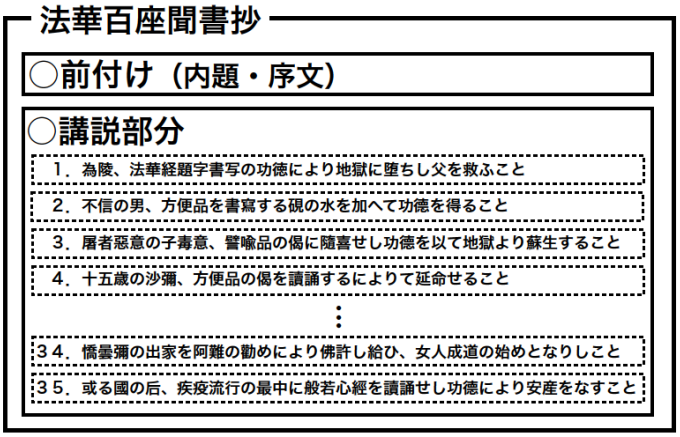


図 2　『法華百座聞書抄』の基本構成

４．文書全体の構成に関わる構造化

上述のように『法華百座聞書抄』のテキストは、（ａ）前付け部分（ｂ）講説部分をもつ。この構成は、一般的な欧文の写本と概ね一致するため、TEI P5:Guidelines（Burnard and Bauman 2007）準拠の要素を用いて構造化する場合、テキスト全体は<text>、（ａ）前付け部分は<front>、（ｂ）講説部分は<body>、をそれぞれ対応させることができる。そして、<front>内部に置かれる内題、序文、<body>内部に置かれるコンテンツ情報と講説については （text division）によって規定することができる。あるいは<article>を用いることもできる。<article>以下は基本的にコンテンツ情報と本文の塊であるパラグラフによって構成されるため、それぞれ<head>と<p>（paragraph）を対応させて明確に区別する。

以上のテキストの大局的構造を支える６つの要素を階層構造に留意して一覧表にまとめると表 1 のようになる。

表 1 <text>から<p>までの要素の一覧

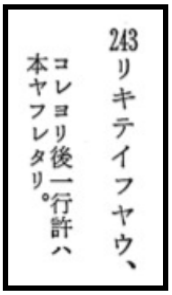
要素	説明
<text>	テキスト全体
<front>	前付け部分
<body>	講説部分
<article>	序文、説話・説教
<head>	内題、コンテンツ情報（日付・経の品名・講師名）
<p>	パラグラフ

５．パラグラフ以下の要素の構造化

次に『法華百座聞書抄』の<p>（paragraph）以下の階層について述べる。

（ａ）前付け部分<front>および（ｂ）講説部分<body>に含まれる<p>以下の内容は、講説（本文）と経典からの引用文（経文）が含まれる。本文の体裁をとるものは、欧文の電子化と同様に TEI の<s>（sentence unit）を用い、経文は<q>（quoted）と属性@type に値“経

文”を入力して表現する。また、書写者による注記（体裁注記）がある箇所については、属性に値@type=“体裁注記”を入力して区別する。図 3 は、その例である。




```
<s>
    <lb n="u243" />りきていふやう、
</s>
<q type="体裁注記">
    <s>これより後一行許は本や
        <vMark>ぶ</vMark>れたり。
    </s>
</q>
```

図 3 書写者による注記（体裁注記）の XML 表現

以上、パラグラフ以下で用いた要素を階層構造に留意してまとめた一覧を表 2 に示す。

表 2 <p>以下の要素の一覧

要素	説明
<q>	経文、書写者による注記（体裁注記）
<s>	講説（本文）、文単位

６．センテンス以下の要素の構造化

次に『法華百座聞書抄』の<s>（sentence unit）以下の階層について述べる。

（ａ）前付け部分<front>および（ｂ）講説部分<body>に含まれる<s>以下では、ルビ付き文字が頻出する。

著者らはこれまでに古典作品におけるルビ文字に対して、マークアップ上の問題点や言語構造に留意した構造化を提案してきた（e.g. 河瀬ほか 2013、Kawase et al. 2014）。ここでは先行研究—『今昔物語集』や BCCWJ のマークアップ—の方針を踏襲し、<ruby>を用いて、ルビとして振られた文字列は属性@rubyText に記述する。とくに、語（後述する短単位）を越える文字列に付与されるルビは、属性@rubyBase の値にルビが振られている文字列全体を入力して表現する（図 4）。


```
<ruby resp="校注者" rubyText="コノ" rubyBase="此の">
此
</ruby>
<corr type="hodoku">の</corr>
僧のかくのたまふ
```

図 4 @rubyBase を用いた XML 表現

通常縦書きの文書では、ルビは文字の右側に置かれる。しかし、『法華百座聞書抄』では文字の左側にも同様にルビを付与することがある。近世口語資料—洒落本—の場合、右側ルビには文字の読みや宛て語を、左側ルビには右側ルビ以外の読みや語の意味を記す傾

向があった。ここでは左側ルビについて<lRuby>を規定して表現する。

ところで、本文<s>の内容について、形態論情報（品詞・活用形・読みなど）を付与することにより、言語資源として質の高いコーパスを設計することが適う。形態論情報の付与や形態素解析の精度を向上させるためには、文字レベル（#PCDATA）の変換・修正を実施する必要がある。『法華百座聞書抄』では、具体的には（イ）本来は濁点を付けるべき文字、（ロ）踊字（ヽ・ゝ・/ \）の多用、（ハ）文字種の混在、（ニ）JISX0213 外の文字の使用といった問題がある。ここでは先行研究—『明六雑誌コーパス』と洒落本のマークアップ—の方針を踏襲し、次のようにタグを規定し、本文を整形する：（イ）<vMark>、（ロ）<odoriji>、（ハ）<char>、（ニ）<g>（gaiji）。これらの具体例を図 5 に示す。

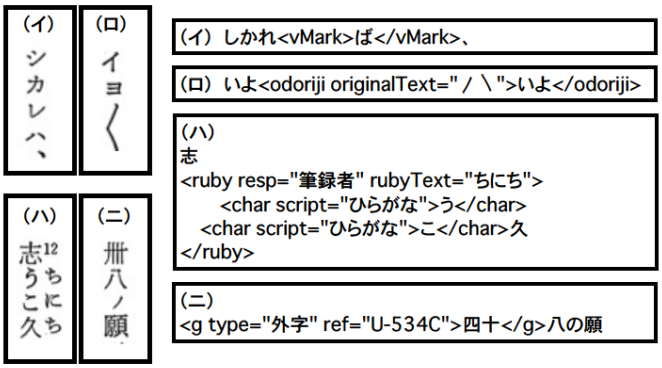


図 5 文字レベル（#PCDATA）の変換・修正
（イ）<vMark>、（ロ）<odoriji>、（ハ）<char>、（ニ）<g>

また、タグの階層構造は前後するが、<s>とは国語研が規定する短単位の集合体である。したがって、<s>直下に短単位<SUW>（Short-Unit Word）を定義し、その属性に形態論情報—語彙素、語形、書字形、品詞、活用型、活用形、発音形、語種—を付与する。この作業は<s>に該当する文の形態素解析結果に対して人手で修正しつつ付与していくものである。

以上、センテンス以下で用いた要素を階層構造に留意してまとめた一覧を表 3 に示す。

表 3 <s>以下の要素の一覧

要素	説明		要素	説明
<SUW>	短単位（語）		<ruby>	ルビ付き文字
<lRuby>	左側ルビ		<vMark>	濁点表記の文字
<corr>	本文の修正		<char>	原文表記の文字種
<odoriji>	踊字表記の文字		<g>	JISX0213 外の文字

7．本文の修正

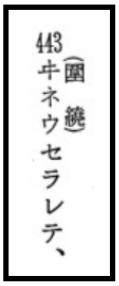
ここでは本文の修正を行う場合に用いる<corr>（correction）の用途について整理する。『法華百座聞書抄』の場合、その目的を（1）原文の誤りを正すための修正、（2）形態素解析の精度向上のための修正、に大きく分類することができる。

（1）には、原文の（a）誤字、（b）脱字の修正が存在する。（a）誤字は、属性@type に値“erratum”を、（b）脱字は、属性@type に値“omission”を入力して区別する。

そして必要に応じて属性@originalText に値として本文修正前のテキストを入力する。さらに、修正がどの段階で行われたものかを明示するために、任意に属性@resp を準備し、その値として「筆録者」、「校注者」、「作業者」を入力する。

（2）には、（c）補読（d）捨て仮名（e）振り漢字（f）返読の4種類の修正が存在する。（c）補読は、属性@type に値“hodoku”を、（d）捨て仮名は、属性@type に値“sutegana”を、（e）振り漢字は、属性@type に値“furikanji”を、（f）返読は、属性@type に値“返読前”もしくは“返読後”を入力して区別する。そして、（2）についても（1）と同様に、必要に応じて属性@originalText に値として本文修正前のテキストを入力し、修正が行われた段階を明示するために、属性@resp を準備し、その値として「筆録者」、「校注者」、「作業者」を入力する。

とくに、（f）返読については、和漢混淆文の『今昔物語集』の構造化（冨士池ほか 2013）と共通する特徴であるが、（e）振り漢字については、今回の文書構造化において新たに追加した内容である。図6に（e）振り漢字の XML 表現を示す。




```
<lb n="o443" />
<corr type="furikanji" resp="校注者" originalText="キネウ">
圍繞
</corr>
せられて、
```

図6　（e）振り漢字の XML 表現

８．位置情報および本文以外の情報

上述のタグに加え、文の改行位置には空要素の<lb/>（line break）を割り当てる。ただし、図 1（右）にも表れているように、『法華百座聞書抄』の翻刻テキストでは行番号を紙面の表裏にあわせて「オ」「ウ」とともに示している。ここでは<lb/>の属性@n に“o 行番号”や@n に“u 行番号”（o/u は表/裏に対応）を入力して表現する。

また、本文以外の情報については空要素の<info/>（information）を準備し、記述できるようにする。例えば、翻刻テキストには、書写者による傍書（注記）が本文中に一箇所だけ存在する。これを構造化する上で<info/>とその属性@text に値として傍書の記述内容を入力して表現する。

最後に、本文の欠損箇所については、<missing>を用いて表現する。とくに、文字単位の欠損と２文字以上にわたる欠損をそれぞれ図7のように表す。


```
<s>
    天仁三年二月廿八日
    令始修法一百座
    <missing>␣</missing>
    <missing>⇦␣⇨</missing>
</s>
```

図7 本文の欠損箇所の XML 表現

以上、位置情報と本文以外の情報について用いた要素を階層構造に留意してまとめた一覧表を表 4 に示す。

表 4 位置情報と本文以外の情報に関する要素の一覧

要素	説明
<missing>	欠損箇所
<lb/>	行番号（体裁注記）
<info/>	本文以外の情報（書写者による傍書）
#PCDATA	文字（character data）

９．まとめと今後の課題

本研究では、国立国語研究所の「通時コーパス設計」プロジェクトの一環として『法華百座聞書抄』について、和漢混淆文の文書―『今昔物語集』のテキスト―と比較しながら、文書構造化を検討し、具体的事例を示した。

『法華百座聞書抄』と『今昔物語集』の文書構造化における最大の違いは、とりわけ振り漢字の扱いにあった。本研究では、本文の修正処理について整理し、それらを<corr>の属性@type の値に応じて区別する方針を提案した。また同時に、修正作業の段階を明示するための工夫として<corr>の属性@resp の値として「筆録者」、「校注者」、「作業者」を割り当てる方針も提案した。これにより、原文の状態を保持しつつ、特定の修正段階における本文を抽出することが可能となった。

本研究で使用したタグセットの階層関係をダイヤグラムで表現すると図 8 のようになる。

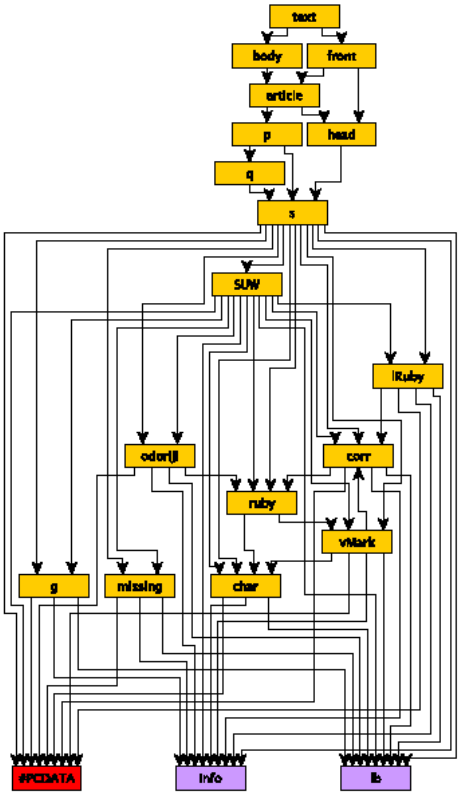


図 8 『法華百座聞書抄』のタグセットのダイヤグラム

今後の課題として、本研究において提案した文書構造化の方針を和文体および漢文体をもつ『法華百座聞書抄』以外の言語資料に適用しながら、漢文脈傾向の強いテキストを網羅的・統一的観点からアーカイヴ化するためのフォーマットを構築していく。

謝　辞

本研究は、日本学術振興会科学研究費基盤研究（B）「和漢の両系統を統合する平安・鎌倉時代語コーパス構築のための語彙論的研究」（24320086、代表者：田中牧郎）および国立国語研究所の共同研究プロジェクト「通時コーパスの設計」に基づく成果の一部である。

文　献

Burnard, Lou and Syd Bauman（2007）TEI P5: Guidelines for electronic text encoding and interchange, *Text Encoding Initiative.Arlington*, MA: TEI Consortium.
（http://www.tei-c.org/Guidelines/P5/）　（参照 2014-08-01）

冨士池優美・河瀬彰宏・野田高広・岩崎瑠莉恵（2013）「『今昔物語集』のテキスト整形」、第 4 回コーパス日本語学ワークショップ予稿集、pp.125-134

飛田良文ほか（編）（2007）『日本語学研究事典』、明治書院

河瀬彰宏・市村太郎・小木曽智信（2013）「TEI：P5 に基づく近世口語資料の構造化とその問題点」、じんもんこん 2013 論文集、Vol.2013、4、pp.7-12

河瀬彰宏・市村太郎・小木曽智信（2014）「『虎明本狂言集』における会話文の計量分析」、言語処理学会第 20 回年次大会発表論文集、pp.662-665

Kawase, Akihiro, Taro Ichimura, Toshinobu Ogiso（2014）Problems in encoding documents of early modern Japanese. *Proceedings of the Digital Humanities 2014*
（http://dharchive.org/paper/DH2014/Paper-934.xml）（参照 2014-08-01）

小林芳規(編)（1975）『法華百座聞書抄総索引』、武蔵野書院

近藤明日子・田中牧郎（2012）「『明六雑誌コーパス』 の仕様」、国立国語研究所共同研究報告 12-03 近代語コーパス設計のための文献言語研究成果報告書、pp.118-143

前川喜久雄（2008）「KOTONOHA『現代日本語書き言葉均衡コーパス』の開発」、日本語の研究、Vol. 4、No.1、pp.82-95

田中牧郎・小木曽智信（2000）「総合雑誌『太陽』 の本文の様態と電子化テキスト」、日本語科学、Vol. 8、pp.141-152

山岸徳平（開題）（1976）『法華修法一百座聞書抄』、勉誠社文庫 4、勉誠社

山口昌也・高田智和・北村雅則・間淵洋子・大島一・小林正行・西部みちる「『現代日本語書き言葉均衡コーパス』における電子化フォーマット」、Ver2.2、LR-CCG-10-04
（http://www.ni njal.ac.jp/corpus_center/bccwj/doc.html#02）（参照 2014-08-01）

関連 URL

国立国語研究所の言語コーパス整備計画 KOTONOHA　http://www.ninjal.ac.jp/kotonoha/