国立国語研究所学術情報リポジトリ

〈受賞紹介〉　近世口語資料を構造化するときに生じる問題点

著者	河瀬 彰宏
雑誌名	国語研プロジェクトレビュー
巻	5
号	3
ページ	135-138
発行年	2015-02
URL	http://doi.org/10.15084/00000782

国語研プロジェクトレビュー
NINJAL Project Review　Vol.5 No.3 pp.135–138（February 2015）

〈受賞紹介〉

The Alliance of Digital Humanities Organization では，Digital Humanities の年次国際大会における若手研究者の口頭発表・ポスター発表の中から特に優れていると認められる発表に対して，Bursary Award（若手研究奨励賞）を授与しており，河瀬氏は Digital Humanities 2014 における発表により受賞しました。

受賞対象　Akihiro Kawase “Problems in Encoding Documents of Early Modern Japanese”
（Digital Humanities 2014; University of Lausanne, Switzerland）
（※発表は Akihiro Kawase, Taro Ichimura, and Toshinobu Ogiso の連名）

近世口語資料を構造化するときに生じる問題点

河瀬 彰宏

1. はじめに

本稿は，2014 年 7 月，ローザンヌで発表した研究内容 “Problems in Encoding Documents of Early Modern Japanese”（Kawase et al. 2014）のエッセンスを紹介するものである。

国立国語研究所コーパス開発センター（以下，国語研）では，「通時コーパスの設計」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。形態素解析を正しく行うためには，基礎資料となる古典テキストの電子化が必須である。国語研では，これまでに様々な時代のテキストコーパスを電子化し，Web サイトや刊行物を通して公開している（河瀬 2014）。

しかし，これら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[1] を用いて文書をマークアップしているため，各コーパスを規定するタグ（要素）は，基本的に統一されていない。そのため，複数のコーパス間の構造比較や計量的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。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できれば，この問題は解決される。

河瀬ほか（2013）では，洒落本の一冊『傾城買二筋道』の版本を事例に，TEI（Text Encoding Initiative）（Burnard and Bauman 2007）に従う XML 形式による文書構造化を提案し，構造化の過程で生じる問題点を整理した。TEI は，人文学で用いられる様々な文書を電子化する共通のガイドラインを網羅的に定めることを究極の目標としており，欧文資料を中心に XML 化のガイドラインを設けてきたが，縦書きの文書についてはほとんど方針が未整備で

[1] XML（Extensible Markup Language）とは，文書に情報を付与するためのタグを自由に作成できるコンピュータ言語（約束事）のことである。XML の基本単位は，開始タグ “<・>” と終了タグ “</・>” を使って内容を囲んだ「要素」と呼ばれる塊である。また，要素には「属性」とその値（属性値）をもたせて情報の性質を表現できる。「通時コーパスの設計」プロジェクトにおける文書の XML 化の指針については，河瀬（2014）を参照されたい。

あるのが実情である。本発表では，洒落本の構造化をめぐる問題点に対して具体的な解決策を示した。

2.　構造化の問題点と解決策

一般に洒落本は，（a）前付け部分，（b）会話と地の文を混ぜた物語本文，（c）後付け部分の順に構成される。この（a）（b）（c）の構成は，一般的な欧文の写本と一致するため，TEI 準拠の要素を用いておおむね構造化できる。しかし，例えば，以下に取り上げる（A）「表記の正規化」や（B）「縦書きルビ」は，欧文に見られない特徴である。

（A）言語資源として質の高いコーパスを設計するためには，本文に形態論情報（品詞，活用，読み，語種など）を付与する必要がある。しかし，洒落本の本文には，次の 3 点の問題があるため，正しく形態素解析を実施することができない：（A-1）本来濁点を付与すべき文字に濁点がない；（A-2）平仮名と片仮名の混在が解消されていない；（A-3）踊り字が読み通りに修正されていない。これらは表記の正規化に関わる問題として共通項で括れることから，TEI 準拠の要素を用いて統一的に構造化の方針を提案できる（図 1）。

```
<seg type="vMark">
 <choice>
  <orig>か</orig>
  <reg>が</reg>
 </choice>
</seg>
```
(A-1)

```
<seg type="kana">
 <choice>
  <orig>か</orig>
  <reg>カ</reg>
 </choice>
</seg>
```
(A-2)

```
<seg type="odoriji">
 <choice>
  <orig>ゝ</orig>
  <reg>か</reg>
 </choice>
</seg>
```
(A-3)

図 1　表記の正規化 XML 表現例

図 1 の例では，TEI 準拠の <seg> タグに type 属性をもたせ，その値として（A-1）濁点の変換を示す vMark，（A-2）仮名の変換を示す kana，（A-3）踊り字の修正を示す odoriji を取らせることで，正規化の方針を区別している。さらに，<seg> に <choice> タグを組み合わせることで，必要に応じて正規化前 <orig> と正規化後 <reg> の表記を参照できるようにしている。

（B）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，洒落本の本文には，次の 3 点の問題があるため，紙面の外形情報と言語構造を同時に包摂する構造化が実施できない：（B-1）ルビに外字，破損，誤字を含むことがある；（B-2）ルビ付き文字とルビの語が一対一対応でない場合がある；（B-3）文字に対して右左同時にルビを置くことがある。ここでは，日本語の文書における紙面の外形情報と言語構造の構造化を融合的に進めることの難しさを整理し，ルビの構造化について方針を提案した（図 2）。

本文中の各文字に対して，TEI 準拠の <c> タグと id 属性を使ってそれぞれ区別する。そして，外部ファイル中に，<note> タグに target 属性を使って各文字がどのようなルビをもつかを記述していく。こうすることで，（B-1）（B-2）（B-3）の各場合に対処することができる。

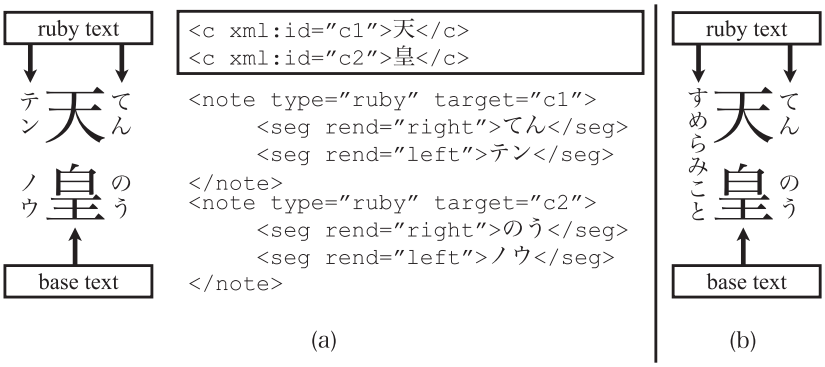

図 2　縦書きルビの XML 表現例

図 2（a）の例では，本文中の「天」に id 番号 c1，「皇」に id 番号 c2 を割り当て，外部ファイルには，<note> タグを使って id 番号が c1 および c2 の文字，すなわち「天」と「皇」のルビの様子を記述している。ただし，ここでは <seg> タグも組み合わせて「天」の文字の右側ルビに「てん」，左側ルビに「テン」を，「皇」の文字の右側ルビに「のう」，左側ルビに「ノウ」が付く場合を表現している。

3.　おわりに

本稿では，洒落本の一冊『傾城買二筋道』の版本を事例に，TEI に従う XML 形式による文書構造化の問題点とその解決策について端的に紹介した。とくにルビは，古典の写本から現代の漫画に至るまで幅広く利用されており，日本語資料にとって必要不可欠な表現方法であるものの，これまで建設的な指針は提案されてこなかった。本研究で提案した指針は，今後汎用的に日本語以外の縦書きの文書にも展開でき，これまで未着手だった漢字文化圏の文書について人文学的研究を促進することが期待できる。

●参照文献●

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/（2007 年発表。2014 年 11 月 11 日参照）.

河瀬彰宏（2014）「『日本語歴史コーパス』の設計を支えるマークアップとはなにか」『日本語学』33（14）: 68−82. 東京：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」『情報処理学会・人文科学とコンピュータシンポジウム IPSJ SIG-CH/PNC/ECAI/CIAS Joint Symposium 予稿集』7−12.

Kawase, Akihiro, Taro Ichimura, and Toshinobu Ogiso（2014） Problems in encoding documents of Early Modern Japanese, *Proceedings of Digital Humanities Conference 2014(DH2014)*, 225−227.

河瀬 彰宏（かわせ・あきひろ）

国立国語研究所コーパス開発センター プロジェクト非常勤研究員。博士(工学)(東京工業大学)。2011 年 5 月より現職。2014 年 4 月より京都大学地域研究統合情報センター共同研究員。

主な著書・論文：「第 4 章　日本民謡の計量分析」『量から質に迫る――人間の複雑な感性をいかに「計る」か』(新曜社, 2014)，「日本民謡の大規模音楽コーパスを用いた旋律の構造抽出」(『国立国語研究所論集』7, 2014)，Construction and verification of the scale detection method for traditional Japanese music（*Affective Engineering* 12, 2013)．

受賞：手島精一記念研究賞（東京工業大学，2012)，優秀発表賞（日本感性工学会第 13 回大会，2012)，最優秀ポスター発表賞 Gold Prize（IPSJ SIG-CH/PNC/ECAI/CIAS Joint Symposium，2013)．

社会活動：日本民俗音楽学会理事.