精緻な表記情報を有する「延喜式祝詞」コーパスの構築
Construction of the Corpus of “Engishiki Norito” with Detailed Annotation of Notation

間淵 洋子
MABUCHI, Yoko
国立国語研究所，東京都立川市緑町 10-2
National Institute for Japanese Language and Linguistics,
10-2 Midori-cho, Tachikawa City, Tokyo

概要：本論文では、現在開発を進めている「延喜式祝詞」のコーパスについて、その仕様と整備状況について報告する。延喜式祝詞は，上代から中古にかけての日本語を研究するための言語資源として期待される資料で，万葉仮名の使用や，宣命書きといった特徴的な表記形態が見られるほか，現存する写本には，ヲコト点や傍訓などの漢文訓読用の書き入れが施されている。本研究では，これらの表記特徴をできるだけ精緻に写した XML による構造化テキストを整備する。加えて、訓読文に対する形態素解析についても検討を進める。

Abstract： In this paper, we report on the specification and developmental status of the corpus of "Engishiki Norito" which is currently under construction. "Engishiki Norito" is considered a valuable material as a language resource for studying Japanese of Nara-Heian period because it has characteristic notation styles such as "manyo-gana", "senmyo-gaki", "wokoto-ten", "kana-ten", etc. We are developing XML structured texts with detailed annotation of these notational features. In addition, we consider morphological analysis of Sino-Japanese gloss reading sentences for the corpus with morphological information.

キーワード：コーパス，延喜式祝詞，表記情報，XML
Keywords：corpus, “Engishiki Norito”, annotation of notation, XML

1．はじめに

現在，国立国語研究所（以下「国語研」）では，上代から中古にかけての日本語を研究するための言語資源として，「延喜式祝詞」（『延喜式』巻八所収の祝詞）のコーパス構築を進めている。

「延喜式祝詞」は，内容語や活用語の活用語幹を通常の文字大で，助詞・助動詞などの付属語や活用語の活用語尾を小書きの万葉仮名で表記する，上代日本語の独特な表記法「宣命書」に表記上の特徴がある。また，現存する『延喜式』の写本には，本文筆写の後に，ヲコト点や仮名点（付訓）等，漢文訓読用の書き入れ（訓点）が施されている点においても，日本語研究資料としての価値が高い。

昨今，典型的な和文とは性質を異にする宣命や訓点資料といった資料を電子化し，有用なアノテーションを加えることで，計量的にこれらの資料における語彙・語法や表記の実態を捉えるための試みが進められているが[1][2]，両資料の性質を併せ持つ「延喜式祝詞」のコーパス化が実現すれば，万葉仮名や宣命書の表記傾向，訓点の使用実態等を比較することにより，資料間の差異や個々の資料の特徴を浮き彫りにすることができよう。

そこで，本研究では，表記史研究に資する言語資源としてのコーパス形態のプロトタイプを策定することを目指し，精緻な表記情報を付与した「延喜式祝詞」のコーパス構築を試みる。本稿では，主に，(1)延喜式祝詞の資料特性，(2)表記情報を精緻に写すためのデータ仕様，(3)データ構築方法，の 3 点について報告する。

2. 延喜式祝詞

『延喜式』は，平安時代中期に律令の施行細則をまとめた法典（「格式」）で，延喜 5（905）年 8 月，醍醐天皇より命を受けた藤原時平により編纂が開始された。延喜に撰修を始めたことに，『延喜式』の名前の由来があるが，その完成は 22 年後の延長 5（927）年，施行は康保 4（967）年である。全 50 巻中，巻一から巻十までが神祇祭祀に関する規定を収めたもので，そのうちの巻八が「祝詞」の巻となっている[3]。

「祝詞」は，神前に奏上する際の独特の文体を持つ言葉で，『延喜式』巻八所収の祝詞（以下「延喜式祝詞」）は，現存する最古の祝詞とされる。「延喜式祝詞」は 27 の祝詞を所載する。それぞれの祝詞の成立は，祭祀の開始時期により奈良朝以前から平安初期までとされているが，「延喜式祝詞」自体の作成時期はその時期と一致しないという指摘もあり[4]，安易に上代日本語資料として扱うことには問題があるものの，残存する資料の少ない上代・中古初期において貴重な資料の一つである。

祝詞には，体言や用言の語幹を通常の文字大で，助詞・助動詞や活用語尾などを万葉仮名の小書き右寄せ（または2行の割書）で表記する独特の表記法が用いられている。この表記法は，祝詞と同様，このスタイルで記された「宣命」（天皇の勅命を和文で記した文書）にちなみ「宣命書」と呼ばれている。

『延喜式』の原本は現存しないが，伝本は少なくない。そのうち，巻八を存するもので代表的な古写本には，「九條家本」「卜部兼永自筆本」「卜部兼右自筆本」などがあるが，中でも「九條家本」（東京国立博物館蔵）は，全 50 巻中 27 巻を存し，平安中期の書写と見られる現存最古の写本として貴重である。現在，国立文化財機構が提供する Web コンテンツ「e 国宝」（http: //www.emuseum.jp/）により高精細画像を閲覧することができる。

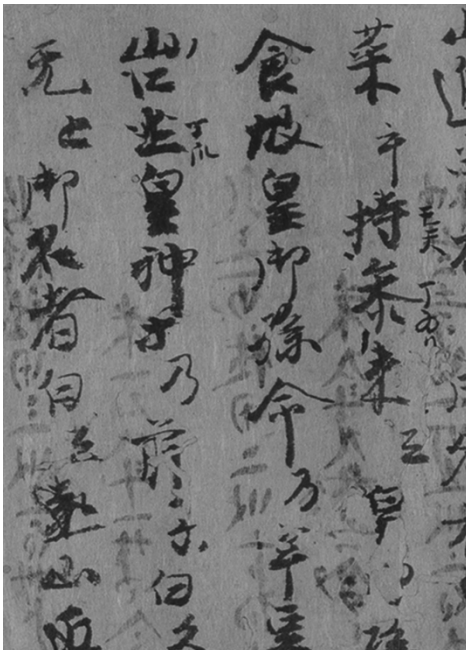

図 1 「延喜式 巻八（九條家本）」1巻
（東京国立博物館蔵，e 国宝）[5]

九條家本巻八には，墨点による仮名点（付訓）やヲコト点が施されており，平安中後期加点資料として貴重なものである。

そこで，本研究では，最古の写本，九条家本を底本とし，緻密な翻刻とこれに基づく訓読文を提示した『東京国立博物館蔵本 延喜式祝詞総索引』[3]に基づき，九條家本「延喜式祝詞」のコーパスを開発することとした。その際，万葉仮名，宣命書，訓点などの特徴的な表記情報を精緻に写し取り，日本語表記史研究に資するコーパスとすることを目指した。

3. 電子化フォーマットの設計

コーパス構築にあたっては，まず，資料を電子化するための仕様を定める必要がある。本研究において

は,「延喜式祝詞」の資料性および研究目的に鑑み,設計方針を以下の通り設定した。

(1) 記された文字をできるだけ忠実に写す。
(2) 宣命書や訓点等の表記情報を明確に表現する。
(3) 文書要素・言語単位と表記との対応関係を明確に示す。
(4) 原文(翻刻本文)と訓読文との対応関係を明確に示す。
(5) 原文と訓読文をそれぞれ抽出できるようにする。
(6) 分析・解析用ツールで扱いやすい形とする。
(7) 情報の追加を想定し,拡張しやすい形とする。
(8) 底本と容易に対照できるようにする。

この設計方針に基づき,データの電子化フォーマットを以下の通り定めた。

- 文字コードは UTF-8,文字セットは JIS X0208 に準じる。
- XML 形式による構造化テキストとする。以下のタグにより情報を格納する。

【文書構造情報を表すタグ】

text テキスト全体(延喜式における1巻)。属性に書誌情報を含む。
@title="延喜式"
@volume="巻八"
@source="九條家本"
title 文書または特定の文章範囲の見出し。
div 文章範囲(各祝詞の範囲)。属性に各祝詞の情報を含む。
@title="祈念祭(例)"
@type="(奏上|宣命)"
s 文
lb 原典における行頭位置。属性に行番号(@n)を含む。『東京国立博物館蔵本 延喜式祝詞総索引』において認定・付与された行番号に一致する。

【表記情報を表すタグ】

span 特殊な表記様式を持つ範囲。属性に,種別(type)を持つ。主に割注(割書)箇所を示す。
ruby 傍訓とその対象本文の対。
rt 本文に傍書されるテキスト(傍訓)。属性に,原文文字列(@originalText;任意)を持つ。
rb 傍訓の対象となる本文行のテキスト。熟字訓を除き1文字を単位とする。
corr 原文(翻刻本文)からの修正(補読等)。属性に校訂情報を含む。
@type="(omission|excess|erratum|hendokuB|hendokuA)" 脱字 or 衍字 or 誤字 or 返読前 or 返読後
@resp="(editor|annotator)" 校注者 or 作業者
@id(任意,返読前後を対応付ける固有番号)
kunten ヲコト点・仮名点等の訓点による訓読。属性に文字情報を含む。
@type="(wokoto|kana)" ヲコト点 or 仮名点
@originalText(任意,仮名点の場合のみ)
kana 大書きの万葉仮名,宣命書による文字。
@type="(manyo|senmyo)" 万葉仮名 or 宣命書
@originalText(万葉仮名字体)

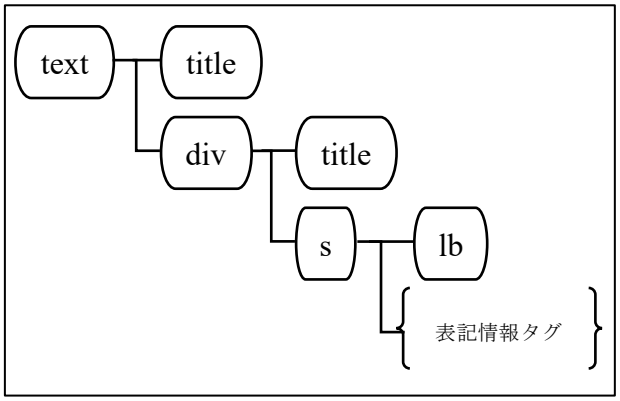

図 2 文書構造情報に関する要素のスキーマ

4. データ構築作業

データ構築にあたっては,コーパスの底本を『東京国立博物館蔵本 延喜式祝詞総索引』所収の訓読文と定め,これを文字入力した。次に,入力本文と「翻刻本文」「影印」(原典画像)とを対照しながら,訓読文との差異を 3 節に示したタグを用いて写し取った。作業の効率化を図るため,簡易的なタグを使用しタグ付けしたものを一括で XML タグへ変換する方法で実施し

た。「延喜式祝詞」に出現する特徴的な表現と，その電子化形式について，以下に例示する（用例画像は，いずれも『東京国立博物館蔵本 延喜式祝詞総索引』より引用した）。

【宣命書と万葉仮名】

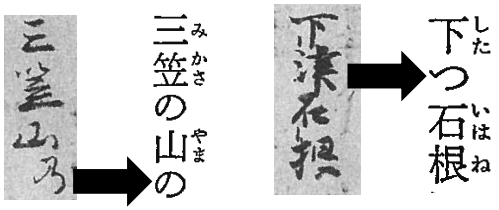

図 3 原文・訓読文の対応（左：宣命書，右：万葉仮名）

```
<ruby><rb>三</rb><rt>み</rt></ruby>
<ruby><rb>笠</rb><rt>かさ</rt></ruby>
<corr type="omission" resp="editor">の</corr>
<ruby><rb>山</rb><rt>やま</rt></ruby>
<kana type="senmyo" originalText="乃">の</kana>
```

図 4 宣命書の形式化

```
<ruby><rb>下</rb><rt>した</rt></ruby>
<kana type="manyo" originalText="津">つ</kana>
<ruby><rb>石</rb><rt>いは</rt></ruby>
<ruby><rb>根</rb><rt>ね</rt></ruby>
```

図 5 万葉仮名の形式化

【訓点】

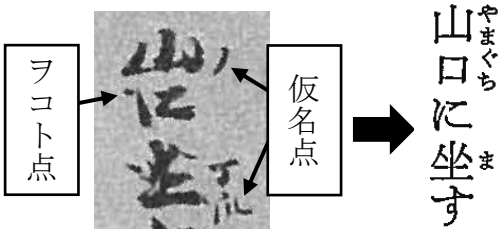

図 6 原文・訓読文の対応（ヲコト点と仮名点）

```
<ruby><rb>山</rb><rt>やま</rt></ruby>
<corr type="excess" resp="editor"><kunten
type=”kanaten” originalText="ノ"/></corr>
<ruby><rb>口</rb><rt>ぐち</rt></ruby>
<kunten type=”wokoto”>に</kunten>
<ruby><rb>坐</rb><rt originalText="マ">ま
</rt></ruby>
<kunten type=”kanaten” originalText="ス">す
</kunten>
```

図 7 訓点の形式化

【漢文式語順】

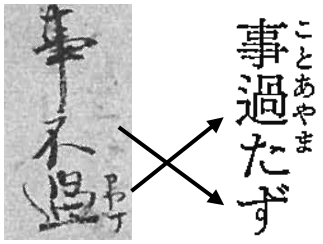

図 7 原文・訓読文の対応（漢文式語順）

```
<ruby><rb>事</rb><rt>こと</rt></ruby>
<corr type="hendokuB" resp="editor"
id="07401">不</corr>
<ruby><rb>過</rb><rt originalText="アヤマ">
あやま</rt></ruby>
<corr type="omission" resp="editor">た</corr>
<corr type="hendokuA" resp="editor"
id="07401" originalText="不">ず</corr>
```

図 9 漢文式語順の形式化

上記 XML による構造化が終了した後，原文（翻刻本文），訓読文の各テキストを抽出するテストと，タグ付けのチェックとを兼ねて，XSL により，訓読文・原文を HTML 文書へと変換，出力し確認した。その際，要素や属性によって表示を工夫することで，タグ付けの実

態を容易に判別でき，底本となる訓読文，あるいは原文との照合がしやすい環境を整えた。

図 10　訓読文 XHTML の Web ブラウザ表示

図 3　翻刻本文 XHTML の Web ブラウザ表示

これにより，原文との関係性を明確にしつつ，原典の表記情報を精緻に写し取った「延喜式祝詞」の訓読文ベース XML データが完成した。

5. 今後の課題と展開

本研究では，日本語表記史研究に資する言語資源としてのコーパス構築を目指している。ここまでの整備状況として，前節までで述べてきた通り，表記情報を精緻に写し取った訓読文ベースの本文を XML 形式により作成したが，今後，研究利用に供するためには，少なくとも以下の二つのフェーズで開発・検討を進める必要があると考える。

まず一つは，形態論情報の付与である。

日本語表記史研究においては，どのような言語要素，どのような語に対して，どのような表記を用いているかといった，表記と言語単位との相関の解明が大きな課題の一つであるが，計量的手法によりこれを可能にするためには，適切な言語単位に分割しておくことが不可欠となる。すなわち，形態素解析を施す必要がある。

日本語の古代語・近代語に対する形態素解析手法については，国語研が開発を進める「日本語歴史コーパス」（https://pj.ninjal.ac.jp/corpus_center/chj/）の構築過程において，これまで多くの試みがなされており，各時代の日本語資料に即した解析用辞書の開発が行われてきた[6][7][8]。しかし，それらはあくまでも漢字仮名で表記された和文を対象としたものであり，漢字文に適用できるものではない。そのため，本研究でも，ベースとなる本文を原文ではなく漢字ひらがな表記による訓読文として整備を進めてきた。これにより，既存の形態素解析システムに適用可能な電子データの準備は整ったが，「延喜式祝詞」は表記のみでなく，語彙の面でも既存の上代・中古の作品とは，大きく異なっている。現在，上代（万葉集用）と中古（和文）の 2 種の解析辞書を用いて形態素解析試行を実施しているが，いずれも解析精度は高くない。今後，3 月に国語研から「日本語歴史コーパス　奈良時代編Ⅱ宣命」として公開予定の五国史宣命を資料としたコーパスなども用いて，「延喜式祝詞」の本文解析に適した辞書を開発することも，残された課題の一つである。

なお，形態論情報アノテーションの結果は，国語研の「日本語歴史コーパス」との連携を確保し，国語研のコーパス検索用オンラインシステム「中納言」での公開を視野に開発を進める予定であり，言語単位として，現在国語研で公開されているコーパスが統一的に採用している「短単位」[9][10]を使用する。その際，上代語，あるいは祝詞や宣命に特有の問題点として，語形（語の読み方とそれに伴う単語の区切り方）と表記単位の不整合（「王臣等」の文字列を「おほきみたちまへつきみたち（＝大君／達／公卿／達）」と読む類）が生じるため，これらの対応についても検討が必要である。

さらに，今一つの課題として，校異情報の付与を挙げておきたい。

古体を保持する祝詞の性質上，「延喜式祝詞」は本研究で底本とした九条家本に代表される古写本の他にも，長く後世に伝わっており，伝本が多数存在する。本研究プロジェクトの連携機関でもある歴史民族博物館（以下「歴博」）が所蔵する，江戸初期の写本「土御

門家旧蔵本」もその一つであり，金子(2012)により巻八祝詞の翻刻も試みられている[11]。史料画像の共有によって加点情報などを含めた異同を確認することができれば，筆写時期による表記・形態の差異を明らかにすることができる。その他，「延喜式祝詞」諸本の詳細な校異を有する金子(2014) [12]など，「延喜式祝詞」の異文研究の成果を統合的に格納したコーパスというのも，本研究で目指すべき一つの形態であると考える。

なお，校異情報の付与は，TEI（Text Encoding Initiative; https://tei-c.org）による電子化フォーマットが確立されており，これに準拠した形式でのデータ構築が有用であると思われる。それと同様に，祝詞特有の表記体系をアノテーションするために本研究で策定した枠組みも，より標準的な枠組みとの連携を視野に検討を続ける必要があるだろう。

6．おわりに

本稿では，現在構築を進めている「延喜式祝詞」のコーパスについて，以下を報告した。

（1）日本語の表記史研究に資するコーパスとして，祝詞の特徴的な文体・表記（宣命書，万葉仮名）や，古写本に見られる訓点の情報を，精緻に付与した XML 文書としてデータ構築を進めている。

（2）形態素解析処理を前提に，訓読文をベースとしたコーパス本文を整備した。訓読文と原文との対応を保持し，表記情報と単語情報の相互参照が可能な形式での公開を検討している。

（3）更なるアノテーションの拡張として，異本と底本（九条家本）との校異情報の付与を計画している。

付記

本研究は，人間文化研究機構広領域連携型基幹研究プロジェクト「異分野融合による「総合書物学」の構築」国語研ユニット「表記情報と書誌形態情報を加えた日本語歴史コーパスの精緻化」及び国立国語研究所プロジェクト「通時コーパスの構築と日本語史研究の新展開」による成果の一部である。

参考文献

[1] 池田幸恵・須永哲矢，「五国史」宣命コーパスの設計とその利用，訓点語と訓点資料, 2015, no. 134, pp.98-80.

[2] 柳原恵津子，『金光明最勝王経』平安初期点の形態素解析用本文作成: その方法と問題点，「東洋学へのコンピュータ利用」第 31 回研究セミナー発表論文集, 2019.

[3] 沖森卓也編，東京国立博物館蔵本 延喜式祝詞総索引,古典研究会, 1995.

[4] 沖森卓也，日本古代の表記と文体，吉川弘文館, 2000.

[5] http://www.emuseum.jp/detail/100162

[6] 小木曽智信，旧仮名遣いの口語文を対象とした形態素解析辞書，じんもんこん 2012 論文集, 2012, pp.25-32.

[7] 小木曽智信・小町守・松本 裕治，歴史的日本語資料を対象とした形態素解析，自然言語処理, 2013, Vol.20, No.5, pp.727-748.

[8] 小木曽智信・市村太郎・鴻野知暁，近世口語資料の形態素解析の試み，第 4 回コーパス日本語学ワークショップ予稿集, 2013, pp.145-150.

[9] 伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵，コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用，日本語科学, 2007, Vol.22, pp.101-123.

[10] 小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕，『現代日本語書き言葉均衡コーパス』形態論情報規程集 第 4 版（下），文部科学省科学研究費特定領域研究「日本語コーパス」データ班, 2011.

[11] 金子善光，翻刻・歴史民俗博物館所蔵「延喜式 巻八祝詞」，神社と実務, 2012, no.11, pp.95-74.

[12] 金子善光，翻刻・京都大学図書館蔵「陽明文庫本 延喜式・巻八・祝詞，『文化史史料考證』刊行委員改編，嵐義人先生古稀記念論集文化史史料考證, 2014, pp.23-60.