『今昔物語集（本朝部)』のデータについて

池上　尚

１．はじめに

『日本語歴史コーパス　鎌倉時代編 I 説話・随筆』収録作品の中で、データの性質について特に注意が必要なものとして『今昔物語集（本朝部)』がある。注意点は大きく分け二つある。第一に、コアデータ・非コアデータの区別があること、第二に、校訂されたテキストと原文文字列との関係である。本稿では、コーパス利用時に注意すべき『今昔物語集（本朝部)』のデータの特殊な点について述べる。

２．コアデータ・非コアデータの区別

文学作品の場合、一作品の全文をコアデータ（自動形態素解析の結果すべてに目を通し、人手修正を加えたデータ）としてコーパス化することが前提である。しかし、本コーパスでは、全体の約 70%を占める規模の大きな『今昔物語集（本朝部)』については、部分的に非コアデータ（自動形態素解析の結果のすべてには人手修正が加えられていないデータ）とし、これ以外の複数の説話作品を収めるコーパスへと拡張していくことで、代表性が担保されたコーパスの構築を目指した。

（１）本文テキスト

コーパス化の対象とする『今昔物語集』の本文は、コーパス構築のために小学館から国立国語研究所に提供された「新編日本古典文学全集」の『今昔物語集 1～4』（馬淵和夫・国東文麿・稲垣泰一校注）の電子テキストを利用している。これには巻 1～10 の天竺部・震旦部は収録されておらず、巻 11～31 の本朝部のみが収録されている。よって、コーパス化の対象もこの範囲となる。底本はそれぞれ、巻 12・17・27・29 が『今昔物語集』最古の写本である鈴鹿本（現在は京都大学図書館蔵)、巻 11・13～16・19・20・22・24 が実践女子大学本、巻 23・25・26・28・30・31 が東京大学国語研究室本となっている。

（２）コアデータ

『今昔物語集（本朝部)』のうち、コアデータとしてまず選定したのは、鈴鹿本を底本とする巻 12・17・27・29 である。『今昔物語集』の最初の方の巻は漢文訓読体の性格が強く、後ろの巻に進むにつれ和文体の性格が強まるという性質を有し、その境は巻 20 前後とされる。よって、上記 4 巻は、漢文訓読体的な 2 巻（巻 12・17)、和文体的な 2 巻（巻 27・29)となる。これに、文体から見た場合に中間的な性質を有するとされる巻 20 を加え、計 5 巻（本朝部の約 30.0%・約 15 万短単位）をコアデータとした。

（３）非コアデータ

コアデータである 5 巻を除いた残りの 14 巻（本朝部の約 70.0%・約 35 万短単位）を非コアデータとした。ただし、非コアデータと言っても、コアデータを学習用コーパスとして作成した「和漢混淆文 UniDic」により再解析し、効果的な精度向上作業を施したため、精度はコアデータに準ずる約 99%まで向上している。非コアデータに対する具体的な精度向上作業の内容については、池上ほか（2015）を参照されたい。

３．テキストの校訂と原文文字列の関係

「新編日本古典文学全集」の中でも『今昔物語集（本朝部)』は、写本を忠実に活字化する方針をとっている。加えて、漢文の要素が混在する和漢混淆文資料であることから、コーパス化する際にテキストを一部校訂する必要があった。3 節では、そうしたテキストの校訂の概要と、原文文字列との関係について述べる。

（１）仮名種別の変換

『今昔物語集（本朝部)』は漢字片仮名交じり表記の資料である。また、万葉仮名による和歌が含まれている。コーパス化するにあたり、片仮名・万葉仮名を平仮名に変換した上で、形態論情報を付与した。一部平仮名表記された本文は、コーパス上でも平仮名とした。変換前の片仮名・万葉仮名・平仮名は、中納言の「原文文字列」に表示される。

例

コーパス本文	原文文字列	備考
したてるや	志太弖留耶	万葉仮名
首尾を取て	首尾を取テ	平仮名

（２）外字の処理

本文テキストは JISX0213 に準拠している。外字となった 193 字（異なり）については、原則として JISX0213 内の別字で代用しているが、Unicode 内の別字で代用したものもある。漢字 1 字に対しては漢字 1 字で代用できるものを優先しているが、文字数の変更が入ったものもある。なお、別字での代用表示が困難な場合に限り、仮名に開いた。処理方針の詳細については、須永・堤（2013）を参照されたい。

例

新編全集　本文	コーパス本文	ルビ	原文文字列	処理内容
此ノ事ヲ［左「弟」＋右「令」］デ	此の事を矜で	あはれむ	此ノ事ヲ矜デ	代用
夜モ皆［左「日」＋右「差」］畢ル	夜も皆晞畢る	あけ	夜モ皆晞畢ル	代用
三［左「日」＋右「关」］ノ門	三咲の門	せう	三咲ノ門	代用

［上「日」＋下「下」］部ノ	日下部の	くさかべ	日下部ノ	合字
［左「巾」＋右「蔓」］	帯獲	おびとり	帯獲	文字数変更
［左「巾」＋右「百」］袼	抹額	もこう	抹額	2字に対する読み
潀［左「氵」＋右上「孑」＋右下「日」］	潀ちゅう	しふちう	潀チュウ	代用字候補なし
靁	ひひめき	ひひめき	ヒヒメキ	代用字候補なし

（3）踊り字・くの字点・同の字点

『今昔物語集（本朝部)』には踊り字・くの字点・同の字点といった繰り返し記号が用いられている。コーパス化にあたっては、踊り字・くの字点は、原則としてすべて文字を繰り返す表記に改めた。また、同の字点は、複数の文字を繰り返しているもののうち読みが確定しているもの、文節を越えるもの、動詞の終止形が二つ重なる形式、動詞・形容詞の連用形が二つ重なる形式のいずれかに当てはまる場合は文字を繰り返す表記に改めた。同の字点は繰り返す文字数が一定ではないが、ルビを参照して、原則として文字数が同じになるように、文字を挿入した。同の字点で繰り返す文字のルビが短単位を越えている場合や同の字点と挿入する文字の字数が一致しない場合は、補読処理をあわせて行った。なお、読みが確定できないもの（ルビがなく、どこから繰り返しているかが不明確なもの）については、繰り返し記号を処理せずに残した。変換前の表記は「原文文字列」として表示される。

例

コーパス本文	原文文字列
つつ	ツヽ
給はば	給ハゞ
をいをい	ヲイヽヽ
ほろほろ	ホロ／＼
さめざめ	サメ〴〵
今や今や	今ヤ／＼
返返す	返々ス
穴怖し穴怖	穴怖シ々々
一つ一つ	一（ひとつひとつ）々
参ぬや参ぬや	参（まゐり）ヌヤ々（まゐりぬや）々
君達や有君達や有	君達ヤ有（きむだちやある）々

（４）欠字欠文・破損

『今昔物語集（本朝部)』では、欠字欠文・破損箇所が記号で示されている。コーパス化にあたっては、頭注の記述に基づき、文字列ないし空格を示す記号の挿入を行った。

［１］破損

頭注に「破損による欠字」とあるものについては、推定文字列の候補が一つ示されている場合のみ、該当文字列を補った。推定文字列が特定できない場合は空格を示す記号を挿入した。

(1)破損

［２］意識的欠字

ａ．漢字表記保留

頭注に「漢字表記を期した意識的欠字」のようにあるものについては、推定文字列の候補が一つ示されている場合のみ、その読みを検討後、該当文字列を補った。推定文字列が特定できない場合は空格を示す記号を挿入した。

(2)a. 漢字表記保留

ｂ．具体表記保留

頭注に「(地名・僧名…）の明記を期しての欠字」のように具体表記を保留した欠字であることが示されているものについては、推定文字列が一意に決まる場合でも、欠字の補入は行わず、空格を示す記号を挿入した。

(2)b. 具体表記保留

空格を示す記号を挿入する際には､1 字は“␣”､2 字以上は “⇦␣⇨”と、文字列長によって空格を示す記号を分けた。推定文字列挿入前の表記は「原文文字列」として表示される。また､「原文文字列」は空格を示す記号が補われた形式とした。

例

コーパス本文	原文文字列	欠字欠文・破損種別	推定文字列
鳩**ぞ**現に来て	鳩␣現ニ来テ	破損	あり
聴聞⇦␣⇨疑ひ	聴聞⇦␣⇨疑ヒ	破損	なし
針のさびたるを	針ノ␣␣タルヲ	漢字表記保留	あり
綿厚く␣たる	綿厚ク␣タル	漢字表記保留	なし
⇦␣⇨⇦␣⇨と云ふ人	⇦␣⇨⇦␣⇨ト云フ人	具体表記保留	あり
⇦␣⇨の郷に	⇦␣⇨ノ郷ニ	具体表記保留	なし

（５）漢文の要素に関する変換

『今昔物語集（本朝部)』には表記に漢文の要素が交じっているため、コーパス化にあたっては以下の変換を行った。

［１］補読

『今昔物語集（本朝部)』では、「今昔」を「いまはむかし」と読むように、助詞・助動詞等の表記 が省略されることが多い。コーパス化にあたってはこれを補った。助詞・助動詞のほか、「二」に対して「フタリ」のようなルビがある場合に「人」を補う、漢語サ変動詞の「ス」に該当する箇所が省略されている場合に「ス」を補うといった処理もあわせて行った。なお、活用語尾の省略については、補っていない。変換前の表記は「原文文字列」として表示される。

例

コーパス本文	原文文字列	ルビ
今は昔	今昔	いまはむかし
此の二人	此二	このふたり
況むや	況	いはむや

［２］捨て仮名

『今昔物語集（本朝部)』には、他の読み方をされる可能性のある漢字の読みの一部を補う捨て仮名が多用されている。コーパス化にあたってはこれを本文から除いた。変換前の表記は「原文文字列」として表示される。

例

コーパス本文	原文文字列	備考
度	度ビ	
此	此カク	
候う	候フウ	「さぶらう」の「ぶ」を表記したと思われる

［３］返読文字

『今昔物語集（本朝部)』には、「不知ズ（シラズ)」「不知リ（シラザリ)」「不知（シラヌ)」のよ うな表記がある。コーパス化にあたってはこれをそれぞれ「知ズ」「知ザリ」「知ヌ」と変換した。変換の際、助詞・助動詞および「なし」を仮名に改めた。また、「余日」「余歳」については「日余」「歳余」に変換した。変換前の表記は「原文文字列」として表示される。

例

コーパス本文	原文文字列
聞かしむべからず	不可令聞カ
今に	于今
知らで止にけり	不被ラ止知ニケリ
二十日余（はつかあまり）	二十余日

（６） その他

会話中の心中思惟については紙面で「」（「」の太字）と表されていたが、これを＜＞に置き換えた。以下の 1 箇所である。

＜悩スラム所ノ悪鬼ヲ揮へ＞　巻十四　極楽寺僧誦仁王経施霊験語第三十五

参考文献

池上　尚・鴻野知暁・河瀬彰宏・片山久留美（2015）「『今昔物語集』のコーパス化における非コアデータの精度向上作業」『第 8 回コーパス日本語学ワークショップ予稿集』pp.65-74

須永哲也・堤智昭（2013）「『日本語歴史コーパス』のための書籍活字の電子化 : 小学館新全集『今昔物語集』を事例として」『国立国語研究所論集』6, pp.163-181, 2013-11、国立国語研究所

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