『今昔物語集』のテキスト整形

冨士池優美　（国立国語研究所　コーパス開発センター）†
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）
岩崎瑠莉恵　（国立国語研究所　コーパス開発センター）

The Text Formatting for *Konjaku-Monogatarishū*

Yumi Fujiike　(Center for Corpus Development, NINJAL)
Akihiro Kawase　(Center for Corpus Development, NINJAL)
Takahiro Noda　(Center for Corpus Development, NINJAL)
Rurie Iwasaki　(Center for Corpus Development, NINJAL)

１．はじめに

国立国語研究所では「通時コーパスの設計」プロジェクトの中で、古典テキストに対して形態素解析を施す研究を進めている。これまでは平安時代を中心とする和文についてその成果を発表してきたが（小木曽ほか 2010 など)、漢文の要素が交じる和漢混淆文の古典テキストに対して形態素解析を施すためには、和文の場合とは異なる研究が必要になる。

発表者らは、新編日本古典文学全集『今昔物語集』一～四[1]（小学館、以下新編全集『今昔物語集』とする）に基づくコーパス化を行っている。新編全集『今昔物語集』には読解の便宜のために様々な校訂がなされているが、漢文の要素が交じっているため、そのまま形態素解析をすると様々な問題が生じる。この問題をテキスト整形によって解消すべく検討を行った。その結果、テキスト整形が必要な主な要素として、①返読文字、②助詞・助動詞等の省略表記、③捨て仮名、④欠字欠文・破損、⑤字種（片仮名・万葉仮名)、⑥踊り字・くの字点・同の字点があった。本発表では、新編全集『今昔物語集』を例に、各要素の問題点を指摘した上で、問題を解消するためのテキスト整形の事例を示すとともに、テキスト整形後の形態素解析結果を紹介する。

２．問題の所在

和漢混淆文特有の問題点が三つある。

まず、漢文の要素が交じっていることに起因するものである。形態素解析を施すにあたり、語順の転換や形態素の重複・不足があると、文字と形態素との対応を上から順に取れないことが問題になる。『今昔物語集』の場合、①返読文字が語順の転換と形態素の重複、②助詞・助動詞等の省略表記が形態素の不足、③捨て仮名が形態素の重複に該当する。これらについては、語順及び形態素の重複・不足を解消すべく、前もって整備する必要があ

† yfujiike@ninjal.ac.jp

1 新編全集には『今昔物語集』の巻 11 以降（本朝部）が収録されている。なお巻 18、21 は欠巻である。

る。

次に、新編全集『今昔物語集』の本文校訂の問題がある。新編全集では、④欠字欠文・破損に関して、底本に存している空格が記号で表記されるほか、底本に存していないが、校訂者が推定して置いた空格も異なる記号で表記されている。空格が記号で残されると、形態素解析を施す際、その前後に影響を及ぼすため、可能であれば空格に文字列を補うことが望ましい。また、新編全集『今昔物語集』の本文は、「底本を忠実に活字化することを期した」とあり、⑥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。新編全集の中古和文資料の場合、繰り返し記号は用いられず、文字を繰り返して表記されるが、『今昔物語集』では繰り返し記号がそのまま表記された結果、文字とルビが対応しないという問題が生じた。この場合、中古和文資料と同様に、繰り返し記号を文字を繰り返す表記に置き換える必要がある。

さらに、形態素解析辞書に関する問題点がある。『今昔物語集』は漢字片仮名交じり文であり、万葉仮名の歌もあり、複数の⑤字種が含まれている。今回、形態素解析辞書としてUniDic[2]を用いており、片仮名の活用語尾や万葉仮名には対応していないため、字種を変更する必要が生じた。

３.『今昔物語集』のテキスト整形

３．１　概要

2 節で挙げた問題点について、前もってテキスト整形し問題を解消した上で、形態素解析を施すこととした。その際、テキスト整形前の状態を XML タグに記録し、元がどのようなものであったかの情報を取り出せるようにした。以下に、テキスト整形が必要であった要素ごとに、処理の詳細を述べる。

３．２　処理の詳細

①　返読文字

『今昔物語集』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。返読文字とはこれらの表記における「不」のような文字を指す。返読を含む文を形態素解析するとき、語順の転換や形態素の重複があり、文字と形態素との対応を上から順に取れないため、「不知ズ（シラズ)」「不知リ（シラザリ)」「不知（シラヌ)」をそれぞれ「知ズ」「知ザリ」「知ヌ」といった形式にする必要がある[3]。この形式の変更にあたり、語順を転換すると同時に、返読文字を形態素解析対象から除外し、返読文字に対応する文字列を挿入した。助詞・助動詞及び「なし」[4]を挿入する際は仮名に改めた。返読であるかどうかの認定は新編全集のルビに従い、山田ほか（1959-1963）に基づき事前に洗い出した返読文字を検索して処理を行った。その際、「所謂」等、形態素解析辞書に 1 単位

---

2 小木曽ほか（2010)、小椋・須永（2012）参照。

3 『今昔物語集』における返読文字の詳細については、冨士池・田中（2012）を参照。

4 形容詞又は形容詞の一部。具体例は 4 節表 2 を参照。

として登録することが妥当と思われる語については、語順変更を要しないため、対象外とした。また、目録・説話タイトル・漢詩文の引用及びルビが付されていない部分も対象外とした。

元の返読文字のほか､返読の通し番号､返読タイプといった情報をXMLタグに記録した。返読タイプとは返読文字に該当する語の仮名表記に着目した分類で、次の4種類がある。

Ａ：返読文字＋□＋語の全体仮名表記
　例）不知ズ（シラズ）　令聞シム（キカシム）　不泣給ヒソ（ナキタマヒソ）
Ｂ：返読文字＋□＋語の一部の仮名表記
　例）不知リ（シラザリ）　令聞ム（キカシム）　難有タキ（アリガタキ）
Ｃ：返読文字＋□
　例）不知（シラヌ）　令聞（キカシム）　于今（イマニ）
Ｄ：上記Ａ～Ｃに当てはまらない変則的なもの
　例）不ジ見（ミセジ）　不被ラ止知ニケリ（シラデヤミニケリ）

返読文字は助動詞・助詞・接尾辞等（以下「助動詞等」とする）と意味が対応する漢文の助字に当たるものであり、□には主に動詞が入る。返読タイプとは助動詞等を助字を用いてどのように表記するかを示したものと言える。タイプＡは助動詞等を漢字で表しつつ仮名で併記したもの、タイプＢは助動詞等を漢字で表したものに送り仮名を付したもの、タイプＣは助動詞等を漢字のみで表したものである。

上に示した例は以下のように処理される。テキスト整形の面から見たとき、タイプＡでは返読文字に対応する文字列の挿入がなく、タイプＢは返読文字で表される語の一部が挿入され、タイプＣは返読文字で表される語全体が挿入されることになる。タイプＤは変則的な表記に合わせて、ルビに合うように対応する語を適宜挿入した。処理対象返読文字数は、タイプＡが最も多く3464箇所、タイプＢは2964箇所、タイプＣは2796箇所、タイプＤは30箇所であった。

Ａ：返読文字のみ除外される
　例）知ズ（シラズ）　聞シム（キカシム）　泣給ヒソ（ナキタマヒソ）
Ｂ：返読文字除外、対応する語の一部挿入
　例）知**ザ**リ（シラザリ）　聞**シ**ム（キカシム）　有**ガ**タキ（アリガタキ）
Ｃ：返読文字除外、対応する語の挿入
　例）知**ヌ**（シラヌ）　聞**シム**（キカシム）　今**ニ**（イマニ）
Ｄ：返読文字除外、ルビに合うように対応する語を挿入
　例）見**ジ**（ミセジ）　**知ラデ止**ニケリ（シラデヤミニケリ）
（**太文字**：挿入箇所）

また、「余日」「余歳」についてはルビに合わせ「日余」「歳余」に語順を変更した。

二十余日（はつかあまり）　→　二十日余　　二十余歳（はたちあまり）　→　二十歳余

②　助詞・助動詞等の省略表記

『今昔物語集』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。表記されていない文字は形態素解析対象とならないため、このような形態素の不足を補う必要がある。これを補読と呼ぶ。補読処理はルビに基づき行った。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している。なお、活用語尾の省略については形態素解析辞書 UniDic で対応可能であるため、補読処理の対象外とした。補読処理によって挿入された文字列であることは XML タグに記録した。補読処理の対象は 7334 箇所であった。以下に例を示す。

今昔（いまはむかし）　→　今**ハ**昔　　此二（このふたり）　→　此**ノ**二**人**　　況（いはむや）　→　況**ムヤ**　（**太文字**：補読箇所）

③　捨て仮名

『今昔物語集』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ことで漢字の読みを明確にする捨て仮名が多用されている。「汝ヂ」の「ヂ」のように最後の一字が主に捨て仮名となるが、「候フウ」のように「さぶらう」の「ぶ」を表記したと思われるものもある。頭注の記述を参考に、捨て仮名を洗い出し[5]、捨て仮名部分は形態素解析対象から除外した。捨て仮名として除外された文字列であること、元の表記を XML タグに記録した。処理対象となった捨て仮名は 1312 箇所であった。以下に捨て仮名の処理例を示す。

汝**ヂ**　→　汝　　此**カク**　→　此　　候**フ**ウ　→　候ウ
努々**メ**　→　努々　　今夜**ヒ**　→　今夜　　（**太文字**：捨て仮名）

表 1 に各巻の捨て仮名出現状況を示した。頻度 1～2 は省略した。「此カク」のような全訓捨て仮名については、「全訓」列に○を付した。捨て仮名は表記にして 240 種類超あるが、およそ半数は頻度 1 となっており、臨時的に付されることが多い様子がうかがえる。

④　欠字欠文・破損

2 節で示したように、新編全集『今昔物語集』では空格が記号で示されている。これを、頭注の記述に基づき[6]、3 種類に分けてテキストの整形を行った。

---

[5] 頭注の記述は一様ではないため、捨て仮名に関する頭注を洗い出し、誤写の可能性があるもの、衍字か捨て仮名かが明確でないもの等については、テキスト処理対象から除外した。

[6] 頭注の記述は一様ではないため、欠字欠文・破損に関する頭注を洗い出し、3 種類の分類、「推定文字列の候補が一つ示されている」とするかどうか、「何の表記が保留されたのか」を判定する基準を定めた上で、欠字欠文・破損に関するテキスト整形の作業を行った。

表１　各巻の捨て仮名出現状況

種類	読み	全訓	11	12	13	14	15	16	17	19	20	22	23	24	25	26	27	28	29	30	31	総計
汝ヂ	ナンジ		24	20	34	44	11	18	51	19	16	1	3	7	3	1	3	0	0	1	5	261
形チ	カタチ		7	1	0	2	1	11	26	12	4	4	4	9	2	4	5	1	5	12	9	119
夜ル	ヨル		3	2	12	4	1	5	3	5	2	0	0	2	3	0	2	2	4	4	2	56
何ニ	ナニ		0	0	1	1	0	3	2	8	1	1	0	6	0	2	6	4	7	4	4	50
智リ	サトリ		13	3	2	0	5	1	3	5	7	0	0	0	0	0	0	0	3	0	3	45
昔シ	ムカシ		8	5	1	3	3	1	1	5	1	0	2	1	0	0	2	0	0	0	0	33
公ケ	オオヤケ		0	0	0	0	0	4	4	9	3	1	2	0	4	0	0	0	1	0	2	30
験シ	シルシ		1	1	1	6	0	4	0	1	8	0	0	1	0	0	0	1	0	1	1	26
者ノ	モノ		0	1	0	1	0	2	1	3	1	0	0	0	4	0	3	4	1	0	1	22
奉ツ	タテマツル		1	2	3	2	1	5	4	0	0	0	0	0	0	0	0	1	0	0	0	19
許リ	バカリ		0	0	1	5	0	0	2	3	1	1	1	2	0	1	0	1	0	0	0	18
後チ	ノチ		0	1	1	0	1	2	3	3	0	0	0	1	0	0	0	3	0	0	1	16
昼ル	ヒル		0	1	1	2	1	0	0	5	0	0	0	0	2	1	1	0	1	0	1	16
上へ	ウエ		0	0	0	0	0	1	1	12	0	0	0	0	0	0	0	0	0	0	1	15
万ヅ	ヨロズ		0	0	0	1	0	0	1	0	1	1	0	0	0	1	1	1	4	2	2	15
為メ	タメ		0	1	2	1	0	0	7	2	0	0	0	0	0	0	0	0	0	0	0	13
現ハ	アラワ		0	0	0	0	3	1	1	2	1	0	0	0	1	1	1	1	0	0	1	13
間ダ	アイダ		0	1	1	1	0	0	1	4	2	0	0	0	0	0	1	1	0	0	0	12
抑モ	ソモソモ		0	1	0	0	0	1	0	0	1	0	0	2	1	4	0	1	1	0	0	12
皆ナ	ミナ		0	0	1	1	1	0	3	0	0	1	0	0	1	0	1	1	1	0	0	11
古へ	イニシエ		0	0	0	0	0	0	2	1	0	0	0	2	0	1	2	2	0	1	0	11
程ド	ホド		0	0	1	0	0	1	1	2	1	0	0	0	0	0	0	2	1	1	1	11
長ケ	タケ		0	0	0	1	0	1	0	1	1	0	0	3	1	0	0	1	0	0	1	10
各ノ	オノオノ		0	0	1	0	0	0	1	1	1	1	0	0	2	0	0	1	0	0	1	9
人ト	ヒト		0	1	1	0	0	1	2	0	1	0	0	0	1	0	0	0	0	0	2	9
適マ	タマタマ		0	1	0	2	0	3	0	0	1	0	0	0	0	0	0	2	0	0	0	9
夢メ	ユメ		0	1	1	3	0	1	2	1	0	0	0	0	0	0	0	0	0	0	0	9
志シ	ココロザシ		3	0	1	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	8
事ト	コト		0	0	0	3	0	0	2	1	0	0	0	0	0	0	1	0	1	0	0	8
前へ	マエ		0	0	1	0	0	3	2	2	0	0	0	0	0	0	0	0	0	0	0	8
君ミ	キミ・ギミ		0	1	0	3	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	7
罪ミ	ツミ		0	0	1	1	0	0	2	1	1	0	0	0	0	0	0	0	1	0	0	7
其コ	ソコ		0	0	0	0	0	1	0	0	0	0	1	1	0	2	0	1	0	1	0	7
怠タ	オコタル		0	0	2	2	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	7
男コ	オトコ		0	0	1	0	0	3	0	1	2	0	0	0	0	0	0	0	0	0	0	7
物ノ	モノ		0	0	0	1	0	3	0	0	0	0	0	0	0	0	1	0	0	1	1	7
去キ	ノク		0	0	0	0	0	1	0	2	1	0	0	0	0	0	2	0	0	0	0	6
後ロ	ウシロ		0	0	0	2	0	3	0	0	0	0	0	0	0	1	0	0	0	0	0	6
尚ヲ	ナオ		0	0	1	0	0	2	1	0	0	0	0	0	0	0	0	2	0	0	0	6
数タ	アマタ		0	0	0	0	0	1	2	0	1	0	0	0	0	0	0	1	0	0	1	6
注シ	シルシ		2	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	6
度ビ	タビ		2	2	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	6
仏ケ	ホトケ		0	0	0	0	0	2	0	4	0	0	0	0	0	0	0	0	0	0	0	6
豈ニ	アニ		1	0	2	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	6
下ダ	クダル		0	0	0	1	2	0	0	0	0	0	0	0	1	0	1	0	0	0	0	5
共モ	トモ・ドモ		0	1	0	0	0	0	0	0	0	1	0	0	0	2	0	0	0	0	1	5
今夜ヒ	コヨイ		0	0	0	2	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	5
時キ	トキ		0	0	0	0	1	0	2	1	1	0	0	0	0	0	0	0	0	0	0	5
前キ	サキ		0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	1	1	0	0	5
渡タ	ワタル		0	0	0	1	0	1	0	2	0	1	0	0	0	0	0	0	0	0	0	5
年シ	トシ		0	0	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	5
偏へ	ヒトエ		0	2	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	5
赤タ	マタ		0	3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	5
為スル	スル	○	0	0	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	0	4
給マ	タマワル		0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	1	0	4
経へ	ヘル	○	0	0	0	1	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	4
高カ	タカイ		0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	4
最モ	モットモ		0	0	0	0	0	0	0	0	1	0	0	0	3	0	0	0	0	0	0	4
聖リ	ヒジリ		0	1	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	4
傍ラ	カタワラ		0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	4
忝ナ	カタジケナイ		0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	1	0	0	0	4
為セ	スル	○	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	3
員ズ	カズ		0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
肝モ	キモ		0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	3
強ヨ	ツヨイ		0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	0	0	0	3
空ラ	ソラ		0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	3
此カク	カク	○	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	3
取ト	トル	○	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	3
情ケ	ナサケ		0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	3
心ロ	ココロ		0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	3
祖ヤ	オヤ		0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	3
男コ	オノコ		0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	3
殿ノ	トノ・ドノ		0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	3
童ハ	ワラワ		0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	3
如ト	ゴトシ		0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	3
非ラ	アラズ		0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	3
辺リ	ホトリ		0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	3
頻度2　計	37種類																					74
頻度1　計	129種類																					129
総計			68	64	88	122	42	119	192	166	83	16	20	43	36	30	45	53	37	33	55	1312

（１）破損による欠字

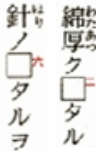
頭注に「破損による欠字」とあるものについては、推定文字列の候補が一つ示されている場合のみ、その読みを検討後、該当文字列を補った。推定文字列が特定できない場合は空格を示す記号を挿入した。

(1)破損

（２）意識的欠字

ａ．漢字表記保留

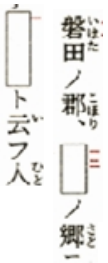
頭注に「漢字表記を期した意識的欠字」のようにあるものについては、推定文字列の候補が一つ示されている場合のみ、その読みを検討後、該当文字列を補った。推定文字列が特定できない場合は空格を示す記号を挿入した。



(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

上記 3 種類とも、空格を示す記号を挿入する際には、1 字は“␣”、2 字以上は“⇦␣⇨”というように文字列長によって空格を示す記号を分けた。3 種類の別は XML タグに記録した。新編全集『今昔物語集』では底本に存する空格と校訂者の判断により補った空格とで記号を分けていたが、この違いについても XML タグに記録した。また、（２）ｂ．具体表記保留については、何の表記が保留されたのか（UniDic の品詞相当の情報[7]）をあわせて XML タグに記録した。

以下に、前ページに示した欠字欠文・破損の処理例を示す。

（１）		鳩**ゾ**現ニ来テ	聴聞⇦␣⇨疑ヒ	
（２）	ａ．	針ノ**サビ**タルヲ	綿厚ク␣タル	
	ｂ．	⇦␣⇨⇦␣⇨ト云フ人	磐田ノ郡、⇦␣⇨ノ郷ニ	（**太文字**：補入箇所）

⑤　字種（片仮名・万葉仮名）

『今昔物語集』は漢字片仮名交じり文である。この片仮名については、全て平仮名に置き換えた。

『今昔物語集』中に出てくる万葉仮名の歌についても、全て平仮名に置き換えた。万葉仮名から平仮名への置き換えについてはルビを参照し、元の文字が万葉仮名であったことと、元の万葉仮名がどの文字であったのかを確認できるよう XML タグに記録した。

---

7 具体表記を保留された内容については、頭注の記述に基づき「人名-一般」「人名-姓」「人名-名」「地名」「数詞」「一般」の 6 種類に分類した。例えば「国司の姓名の明記を期した意識的欠字」とある場合、「人名-姓」「人名-名」に該当する二つのタグを補入した。「一般」は、寺院を含む建物名・官職名・方位・食物名など、UniDic において概ね「名詞-普通名詞-一般」が適用されるもの、及び、具体表記を保留された内容に複数の品詞の可能性がある（例えば「□ノ比（ころ）」の場合、年号＋年代、年号のみ、天皇名、干支等、品詞が特定できない）ものである。

平仮名に置き換えた万葉仮名は 94 箇所であった。例えば、右に示した歌は以下のように置き換えられる。

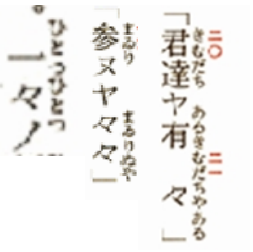
みづはさす　やそぢあまりの　おひのなみ　くらげのほねに　あふぞうれしき

⑥　踊り字・くの字点・同の字点

2 節に示したように、新編全集『今昔物語集』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。このうち踊り字・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。元が踊り字・くの字点・同の字点であったこと、元の表記を XML タグに記録した。処理対象となった繰り返し記号は 1495 箇所であった。以下に踊り字・くの字点・同の字点の処理例を示す。

ツヽ　→　ツ**ツ**	給ハゞ　→　給ハ**バ**	ヲイヽヽ　→　ヲイ**ヲイ**
ホロ/＼　→　ホロ**ホロ**	サメ/〃＼　→　サメ**ザメ**	今ヤ/＼　→　今ヤ**今ヤ**
返々ス　→　返**返**ス	穴怖シ々々　→　穴怖シ**穴怖**	

『今昔物語集』の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[8]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。



一々（ひとつひとつ）　→　一ツ一ツ
参ヌヤ々々（まゐりぬや）　→　参ヌヤ参ヌヤ
君達ヤ有々（きむだちやある）　→　君達ヤ有君達ヤ有

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。

３．３　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＞

8 短単位については小椋・須永（2012）参照。

最後に、これは和漢混淆文だけではなく和文と共通の問題点となるが、漢字の字体に関する問題がある。新編全集『今昔物語集』では、原漢字はすべて漢字で転写されているが、漢字のうち、常用漢字字体のあてられるものはおおむねその字体としたほか、「常用漢字に該当するか否かの判定に迷う」場合は「底本の字体に従った」とある。これに対し、『今昔物語集』のコーパス化にあたっては JISX0213 に依拠して文字処理を行っており、JISX0213 外となる文字については別字代用を行うといった前処理を上記テキスト整形前に行っている。漢字の処理方針の詳細については、須永・堤（2012）を参照されたい。

４．形態素解析例

『今昔物語集』巻第十二「遠江国丹生茅上起塔語第二」をテキスト整形後、形態素解析し、人手修正を加えたものの一部を表 2 に示した。キー（本文）とルビは新編全集『今昔物語集』の情報である。これを短単位に分割し、代表形・代表表記に当たる語彙素読み・語彙素、品詞、活用型、活用形等を付与している。表 2 の「テキスト整形」列には 3.2 節①～⑥のうちどの要素に当たるのかを、「紙面」列には新編全集『今昔物語集』における表記を示した。

テキスト整形の結果、漢字片仮名交じり文だった本文は漢字平仮名交じり文となり、返読文字は語順を変換した形、助詞・助動詞等の表記省略箇所は挿入した文字が形態素解析対象となっている。空格は記号に置き換え、空格がどのような欠字欠文・破損であるかについては、XML タグの情報に基づき、人手で情報を付与した。表 2 の空格「⇦␣⇨ノ郷」は具体表記保留によるものである。紙面列の頭注からもわかるように、ここは地名の具体表記を保留した空格であるため、品詞情報として「意識的欠字（地名）」を付与した。「可産（さんすべ）キ」の場合、「可」が返読文字であるため形態素解析対象から除外し、「産す」の「す」がないため「す」を挿入し、返読文字「可」にあたる「べし」の「べ」を挿入した結果、「産すべき」が形態素解析対象の文字列となっている。

５．おわりに

本発表では、和漢混淆文の形態素解析における問題点を、新編全集『今昔物語集』のテキストを例に指摘した。その問題点に対して、テキストをどのように整形することで解決してきたかを具体的に示すとともに、整形後のテキストを形態素解析した結果、どのような情報を付与するのかをあわせて紹介した。各処理の説明で触れたが、テキスト整形の理由と整形前の状態は XML タグに記録し、必要に応じて参照できるようになっている。

語順の転換や形態素の重複・不足は和漢混淆文をはじめとした漢文の要素が交じる資料に形態素解析を施す際の大きな課題である。新編全集『今昔物語集』は漢文の要素が交じる資料にある様々な問題を含んだテキストであった。今回の検討によって、漢文の要素が交じる資料を形態素解析する際の問題点のいくつかに対して、解決がつく見込みが立ったものと考える。

表 2　形態素解析例

キー	ルビ	語彙素読み	語彙素	出現発音形	品詞	解析活用型	活用形	語義	テキスト整形	紙面
遠江国丹生茅上起塔語第二	とをたうみのくにのにふのちがみたふをたつることだいに				題					
					空白					
今	いま	イマ	今	イマ	名詞-普通名詞-副詞可能					
は	は	ハ	は	ワ	助詞-係助詞				助詞・助動詞等の省略表記	
昔	むかし	ムカシ	昔	ムカシ	名詞-普通名詞-副詞可能					
、			、		補助記号-読点					
聖武	しやうむ	ショウム	ショウム	ショーム	名詞-固有名詞-人名-一般					
天皇	てんわう	テンノウ	天皇	テンノー	名詞-普通名詞-一般					
の		ノ	の	ノ	助詞-格助詞					
御代	みよ	ミヨ	御代	ミヨ	名詞-普通名詞-一般					
に		ニ	に	ニ	助詞-格助詞					
、			、		補助記号-読点					
遠江	とをたうみ	トオトウミ	トオトウミ	トートウミ	名詞-固有名詞-地名-一般					
の		ノ	の	ノ	助詞-格助詞					
国	くに	クニ	国	クニ	名詞-普通名詞-一般					
、			、		補助記号-読点					
磐田	いはた	イワタ	イワタ	イワタ	名詞-固有名詞-地名-一般					
の		ノ	の	ノ	助詞-格助詞					
郡	こほり	コオリ	郡	コーリ	名詞-普通名詞-一般					
、			、		補助記号-読点					
⇦⎵⇨					意識的欠字（地名）				欠字欠文・破損：2字以上の意識的欠字（地名）	
の		ノ	の	ノ	助詞-格助詞					
郷	さと	サト	里	サト	名詞-普通名詞-一般					
に		ニ	に	ニ	助詞-格助詞					
、			、		補助記号-読点					
丹生	にふ	ニュウ	ニュウ	ニュー	名詞-固有名詞-人名-姓					
の		ノ	の	ノ	助詞-格助詞					
直	あたひ	アタイ	直	アタイ	名詞-普通名詞-一般					
茅上	ちがみ	チガミ	チガミ	チガミ	名詞-固有名詞-人名-名					
と		ト	と	ト	助詞-格助詞					
云ふ	い	イウ	言う	イウ	動詞-一般	文語四段-ハ行	連体形-一般			
人	ひと	ヒト	人	ヒト	名詞-普通名詞-一般					
有	あり	アル	有る	アリ	動詞-非自立可能	文語ラ行変格	連用形-一般			
けり		ケリ	けり	ケリ	助動詞	文語助動詞-ケリ	終止形-一般			
。			。		補助記号-句点					
（中略）										
父	ちち	チチ	父	チチ	名詞-普通名詞-一般					
有	あり	アル	有る	アリ	動詞-非自立可能	文語ラ行変格	連用形-一般			
て		テ	て	テ	助詞-接続助詞					
母	はは	ハハ	母	ハハ	名詞-普通名詞-一般					
に		ニ	に	ニ	助詞-格助詞					
云く	いは	イウ	言う	イワク	動詞-一般	文語四段-ハ行	ク語法			
、			、		補助記号-読点					
「			「		補助記号-括弧開					
汝	なむ	ナンジ	汝	ナンジ	代名詞				捨て仮名：汝ヂ→汝	
、			、		補助記号-読点					
齢	よは	ヨワイ	齢	ヨワイ	名詞-普通名詞-一般				捨て仮名：齢ヒ→齢	
産す	さんす	サンスル	産する	サンス	動詞-一般	文語サ行変格	終止形-一般		返読文字、助詞・助動詞等の省略表記：可産キ→産す	
べき		ベシ	べし	ベキ	助動詞	文語助動詞-ベシ	連体形-一般		べき	
齢	よはひ	ヨワイ	齢	ヨワイ	名詞-普通名詞-一般					
に		ナリ	なり	ニ	助動詞	文語助動詞-ナリ-断定	連用形-ニ	断定		
非	あら	アル	有る	アラ	動詞-非自立可能	文語ラ行変格	未然形-一般			
ず		ズ	ず	ズ	助動詞	文語助動詞-ズ	連用形-一般			
し		スル	為る	シ	動詞-非自立可能	文語サ行変格	連用形-一般			
て		テ	て	テ	助詞-接続助詞					
産せ	さん	サンスル	産する	サンセ	動詞-一般	文語サ行変格	未然形-一般			
り		リ	り	リ	助動詞	文語助動詞-リ	終止形-一般			
。			。		補助記号-句点					
（中略）										
其	そ	ソ	其	ソ	代名詞					
の		ノ	の	ノ	助詞-格助詞					
塔	たふ	トウ	塔	トー	名詞-普通名詞-一般					
今	いま	イマ	今	イマ	名詞-普通名詞-副詞可能				返読文字：于今→今に	
に		ニ	に	ニ	助詞-格助詞					
有り	あ	アル	有る	アリ	動詞-非自立可能	文語ラ行変格	終止形-一般			
。			。		補助記号-句点					
磐田	いはた	イワタ	イワタ	イワタ	名詞-固有名詞-地名-一般					
寺	でら	テラ	寺	デラ	名詞-普通名詞-一般					
の		ノ	の	ノ	助詞-格助詞					
内	うち	ウチ	内	ウチ	名詞-普通名詞-副詞可能					
の		ノ	の	ノ	助詞-格助詞					
塔	たふ	トウ	塔	トー	名詞-普通名詞-一般					
、			、		補助記号-読点					
此	これ	コレ	此れ	コレ	代名詞					
も	なり	ナリ	なり	ナリ	助動詞	文語助動詞-ナリ-断定	終止形-一般	断定		
と		ト	と	ト	助詞-格助詞					
なむ		ナム	なむ	ナン	助詞-係助詞					
語り	かた	カタル	語る	カタリ	動詞-一般	文語四段-ラ行	連用形-一般			
伝へ	つた	ツタエル	伝える	ツタエ	動詞-一般	文語下二段-ハ行	連用形-一般			
たる		タリ	たり	タル	助動詞	文語助動詞-タリ-完了	連体形-一般	完了		
と		ト	と	ト	助詞-格助詞					
や		ヤ	や	ヤ	助詞-係助詞					
。			。		補助記号-句点					

付　記

本発表は、国立国語研究所共同研究プロジェクト「通時コーパスの設計」（プロジェクトリーダー：近藤泰弘）、及び、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、研究代表者：田中牧郎）の成果の一部である。

文　献

小木曽智信・小椋秀樹・田中牧郎・近藤明日子・伝康晴（2010）「中古和文を対象とした形態素解析辞書の開発」情報処理学会研究報告 人文科学とコンピュータ vol.2010-CH85,No.4

小椋秀樹・須永哲矢（2012）「中古和文 UniDic 短単位規定集」、平成 21（2009）－平成 23（2011）年度科学研究費補助金基盤研究（C）「和文系資料を対象とした形態素解析辞書の開発」研究成果報告書 2 （http://dl.dropbox.com/u/73297026/report/unidic-EMJ_rulebook2012.pdf よりダウンロード可能）

須永哲矢・堤智昭（2012）「小学館新全集『今昔物語集』での漢字活字－コーパス化のための調査と処理方針の検討－」、通時コーパスプロジェクト・オックスフォード大 VSARPJ プロジェクト合同シンポジウム「通時コーパスと日本語史研究」予稿集、pp.15-22

冨士池優美・田中牧郎（2012）「今昔物語集の返読文字について―形態素解析の前処理を通して―」、日本語学会 2012 年度春季大会予稿集、pp.223-228

馬淵和夫・国東文麿・稲垣泰一（1999-2002）『新編日本古典文学全集 今昔物語集』1～4（小学館）

山田孝雄・山田忠雄・山田英雄・山田俊雄（1959-1963）『日本古典文学大系 今昔物語集』1～5（岩波書店）