『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

池上　尚[†]・鴻野知暁・河瀬彰宏・片山久留美（国立国語研究所コーパス開発センター）

Morphological Analysis for the Konjaku-Monogatarishū Corpus Non-core data

Nao Ikegami　Tomoaki Kouno　Akihiro Kawase　Kurumi Katayama
(National Institute for Japanese Language and Linguistics)

要旨

『今昔物語集』のコーパス化における形態論情報の付与作業、特に非コアデータに対する精度向上作業の方針を示した。発表者らは、まず、コアデータとして 5 つの巻を選定し、これについては「中古和文 UniDic」による形態素解析の結果すべてに目を通し人手修正を加えた。残る非コアデータについては、はじめに、コアデータを学習用データとして作成した「和漢混淆文 UniDic」を用いて形態素解析を行い、約 94%の精度を得た。次に、非コアデータのサンプリングチェックによる誤解析結果から、コーパス公開までの短期間で精度を効果的に向上させる方針を打ち出した。すなわち、「漢字一字表記、かつ、活用語尾（一部）非明示の用言」、「助動詞の前接用言」、「欠字欠文・破損の前後」などのチェックである。上記の作業により精度は約 99%まで向上している。

１．はじめに

国立国語研究所コーパス開発センターでは、共同研究プロジェクト「通時コーパスの設計」と連携し、『日本語歴史コーパス』（Corpus of Historical Japanese, CHJ）[1]の開発を進めている。江戸時代以前の口語性の強い資料群から優先してコーパス化を進め、2014 年 3 月には中古和文 14 作品を収録した平安時代編、2015 年 3 月には『虎明本狂言集』を収録した室町時代編 I 狂言を公開してきた。

一方で、日本語史研究において重要な文語性の強い資料群のコーパス化にも着手しており、現在、和漢混淆文資料を中心に収録した鎌倉時代編 I （説話・随筆など）の構築を進めている。中でも、このコーパスに収録予定の『今昔物語集』[2]は規模が大きく、技術的な問題点を多くはらむため、形態素解析を施す研究に特に注力してきた（冨士池・田中 2012、冨士池ほか 2013 など）。本発表では、これまでの研究を踏まえた上で、『今昔物語集』のコーパス化の全体的な方針と作業の過程を示す。そして、形態論情報の付与作業、特に非コアデータに対する精度向上作業の方針と進捗について報告する。

２．『日本語歴史コーパス』の資料選定方針

２．１　代表性の担保

『日本語歴史コーパス』においてコーパス化の対象とする主な資料群は、日本語史研究において重要な位置を占めてきた文学作品である。『日本語歴史コーパス』の嚆矢となった

[†] nikegami@ninjal.ac.jp

[1] http://www.ninjal.ac.jp/corpus_center/chj/

[2] 平安時代末成立とされるが、『今昔物語集』から始まる説話の一群が鎌倉時代に集中するため、便宜的に鎌倉時代編に収録する。

平安時代編も、「日本語史研究の源流となった、藤原定家や本居宣長などに始まる古典学の主たる対象になってきた作品群がその中心をなしており、古典のコーパス化の対象として最初に取り組むのに妥当なもの」（田中 2014）として選定された中古和文 14 作品の全文がコーパス化されている。平安時代編収録の作品とその語数（短単位）[3]をまとめた表 1 から分かるように、ジャンルは歌集・作り物語・歌物語・日記・随筆にわたり、約 74 万語（短単位）規模のコーパスである[4]。

表 1　平安時代編の作品・語数

ジャンル	作品名	語数
歌集	古今和歌集	31,288
作り物語	竹取物語	10,317
歌物語	伊勢物語	13,824
歌物語	大和物語	23,090
歌物語	平中物語	12,403
日記	土佐日記	6,685
作り物語	落窪物語	54,583
作り物語	堤中納言物語	15,699
随筆	枕草子	66,044
作り物語	源氏物語	445,675
日記	和泉式部日記	10,891
日記	紫式部日記	17,440
日記	更級日記	14,659
日記	讃岐典侍日記	15,555
計		738,153

２．２　鎌倉時代編の構築

平安時代編に後続する鎌倉時代編の収録作品候補としては、和漢混淆文資料として重要な軍記・説話・随筆が挙げられる（田中 2014）。そこで、まずは鎌倉時代編 I として説話・随筆のコーパスの作成に着手し、2016 年 3 月の公開を目指して現在作業中である。このコーパスが鎌倉時代の説話・随筆の実態の縮図となり得るよう、収録作品は当代の代表的な説話・随筆 5 作品とした。すなわち、説話は『今昔物語集』（1120 頃か）本朝部[5]、『宇治拾遺物語』（1220）、『十訓抄』（1252）の 3 作品、随筆は『方丈記』（1212）、『徒然草』（1336）の 2 作品である。表 2 は、上記の作品の語数（短単位）[6]をまとめたものである。全体で約 71 万語（短単位）となり、規模としては平安時代編とほぼ同等となる。

ただし、表 2 の語数から明らかなように、『今昔物語集』（本朝部）が量的に大きな割合を占めている。文学作品の場合、一作品の全文をコーパス化することが前提であり[7]、『今昔

[3] 空白・記号・補助記号は含まない。語（短単位）の認定基準については小椋・須永(2012)を参照。

[4] 2016 年 3 月には『蜻蛉日記』『大鏡』の 2 作品を追加する予定である。

[5] 天竺部・震旦部を含まない理由については 3 節を参照。

[6] 空白・記号・補助記号は含まない。語（短単位）の認定基準については小椋・須永(2012)に従うが、鎌倉時代編収録の作品に適用するにあたり一部変更したところがある。

[7] 文学作品をコーパス化する場合、一ジャンルから一部の作品を収めるという意味でのサンプリングはあっても、作品の一部を収めるという意味でのサンプリングは望ましくなく、一作品の全文をコーパス化する必要がある（近藤 2014）。

物語集』（本朝部）のように規模の大きな作品であってもそれに変わりはない。しかしながら、限られた時間・人手の中にあっては、コーパス総語数の約 70%を占めるような一作品の全文をコーパス化することに専心するよりも、そ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方が、『日本語歴史コーパス』としての代表性は担保されよう。そこで、発表者らは、『今昔物語集』（本朝部）の全文コーパス化・公開を目標とした上で、巻ごとにコアデータ・非コアデータの区別（3 節）を設け、それぞれ異なる作業方法により形態論情報の付与を行うことにした（4 節）。

表 2　鎌倉時代編 I の作品・語数

ジャンル	作品名	語数
説話	今昔物語集（本朝部）	499,712
説話	宇治拾遺物語	101,250
説話	十訓抄	73,514
随筆	方丈記	4,605
随筆	徒然草	33,767
計		712,848

『今昔物語集』は全 31 巻（うち巻 8・18・21 は欠巻のため、現存するのは 28 巻）、1000 話あまりの説話から構成され、一つ一つの説話は基本的に「今昔」という書き出しに始まり「トナム語リ伝ヘタルトヤ」と結んで終わる形式をとる。つまり、一話完結の説話を集めた説話集である。一話一話、一巻一巻の繋がりが希薄である一話完結の説話集だからこそ、作品の一部分をコアデータとして選定することが可能になるという側面もある。

３.『今昔物語集』（本朝部）におけるコアデータ・非コアデータ

コーパス化の対象とする『今昔物語集』の本文は、小学館の「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）により、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電子テキストを利用している。『今昔物語集 1～4』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、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 は実践女子大学本、巻 23・25・26・28・30・31 は東京大学国語研究室本である。

このうち、まず、鈴鹿本を底本とする巻 12・17・27・29 をコアデータに選定した。『今昔物語集』は、最初の方の巻は漢文訓読体としての性格が強く、後ろの巻に進むにつれ和文体としての性格が強まるという性質を有し、その境は巻 20 前後と言われている[8]。よって、上記 4 巻は、漢文訓読体の性格が強い 2 巻（巻 12・17）、和文体の性格が強い 2 巻（巻 27・29）ということになる。この 4 巻に、文体から見た場合に中間的な巻となる巻 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。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）が非コアデータとなる。

[8] 佐藤(1984)の序章に研究史が詳細にまとめられている。

４.『今昔物語集』（本朝部）のデータ整備

前述のコアデータ・非コアデータの区別を踏まえた上で、以下、『今昔物語集』（本朝部）のデータ整備の手順(1)～(7)について詳述する。はじめに概要を示し、次に詳細を述べる。

(1) テキスト整形　……　全データ
(2) 「中古和文 UniDic」による全文の形態素解析　……　〃
(3) コアデータの整備　……　コアデータ
(4) 「和漢混淆文 UniDic」による非コアデータの形態素解析　……非コアデータ
(5) サンプリングチェック　……　〃
(6) 非コアデータの精度向上作業　……　〃
(7) 現在の精度　……　〃

(1) テキスト整形

冨士池ほか(2013)で述べたように、漢字片仮名交じりの和漢混淆文である『今昔物語集』のテキストは、形態素解析を施す前処理としてテキストを整形する必要があった[9]。その理由として、第一に、和漢混淆文ゆえに語順の転換、形態素の重複、形態素の不足があり、上から順に文字と形態素との対応がとれないテキストであったこと、第二に、「中古和文 UniDic」では非対応であった片仮名活用語尾・万葉仮名を含んでいたことが挙げられる。以下、データ整備の手順(5)・(6)に関わるものを中心に具体例をいくつか紹介する。

まず、語順の転換、形態素の重複が問題となる①返読文字がある[10]。返読文字とは、「不」「令」といった助詞・助動詞・接尾辞等と意味が対応する漢文の助辞に当たるものを指す。代表的な処理例として、「不知ズ→知ズ」（シラズ）のように返読文字を除外するタイプ、「不知リ→知**ザ**リ」（シラザリ）、「不知→知**ヌ**」（シラヌ）のように返読文字を除外し対応する語（の一部）を挿入するタイプなどがあった（□は返読文字、**太文字**は挿入箇所）。

次に、形態素の不足が問題となる②助詞・助動詞等の省略表記がある。これについては、「今昔（いまはむかし）→今**ハ**昔」「此二（このふたり）→此**ノ**二**人**」のようにルビに基づき補読処理を施した（**太文字**は挿入箇所）。ただし、「畢（をはり）テ」のように活用語尾が非明示のものについては、語彙素「終わる」―語形「オワル」―書字形「畢る」の連用形として「畢」が登録されていれば UniDic でも対応が可能なため、補読処理の対象としなかった。

同じく形態素の不足が問題となるものに、空格で示される④欠字欠文・破損がある。これは、「破損による欠字」「意識的欠字」を指す。後者には、「綿厚ク␣タル」のように、漢字で表記することを意図しながらもその表記を保留した欠字や、「磐田ノ郡、⇦␣⇨ノ郡ニ」のように固有名などの具体表記を保留した欠字がある。

テキスト整形が必要だったもののうち、形態素の不足については平安時代編を構築していた段階では特に問題とならず、『今昔物語集』のコーパス化に着手して初めて直面した課題であった。平安時代編のコーパス化の対象となった「新編日本古典文学全集」所収の中

[9] テキスト整形前の原文の状態は XML タグに記録してある。

[10] 『今昔物語集』の返読文字の詳細は冨士池・田中(2012)を参照。なお、本文中の丸数字①・②・④は冨士池ほか(2013)をそのまま引用する。

古和文 14 作品においては、読解の便をはかり、送り仮名などを適宜補入するという校訂方針がとられていたためである[11]。

(2) 「中古和文 UniDic」による全文の形態素解析

(1)の整形を経たテキストに対し「中古和文 UniDic」を用いて自動形態素解析を施した（解析器：MeCab 0.993）。

(3) コアデータの整備

(2)の解析結果のうち、コアデータとして選定した 5 巻について目視で確認し、誤解析の修正や揺れの統一、未知語の辞書登録を手作業で行い、短単位データを整備した。

(4) 「和漢混淆文 UniDic」による非コアデータの形態素解析

(3)の人手修正が完了したコアデータを学習用コーパスとして利用し、和漢混淆文を対象とした辞書「和漢混淆文 UniDic」を作成した[12]。さらに、この「和漢混淆文 UniDic」を用いて、人手修正の入っていない非コアデータ 14 巻の再解析を行った（解析器：MeCab 0.993）。結果は次の表 3 に示す通りである[13]。

表 3　「和漢混淆文 UniDic」による『今昔物語集』（本朝部）非コアデータの解析精度

評価レベル	Level 1 単語境界	Level 2 品詞認定	Level 3 語彙素認定	Level 4 発音形認定
解析精度（F値）	0.9889	0.9585	0.9479	0.9449

(5) サンプリングチェック

35 万短単位の規模になる(4)の解析結果から、2000 語を無作為に抽出するサンプリングチェックを行い、誤解析の傾向を確認した。

(6) 非コアデータの精度向上作業

(5)で確認した誤解析の結果からその要因を検討し、コーパス公開までの短期間で精度を効果的に向上させる方針を打ち出した。以下、特に重点的に行った作業の内容を述べる。

a. 漢字一字表記、かつ、活用語尾（一部）非明示の用言

誤解析の中でも特に目立ったのが、漢字一字で表記され、活用語尾が（一部）明示されない用言の語彙素・発音形の誤りである。テキストにルビが振られていればそれを参考に語彙素・発音形を決定する[14]が、機械解析ではテキストのルビを参照しないため、正しい語彙素・発音形を認定できない可能性が高くなる。「新編日本古典文学全集」の『今昔物語集』

[11] 作品ごとの校訂方針については「新編日本古典文学全集」当該巻の「凡例」を参照。

[12] 今後公開する予定である。なお、コアデータ 5 巻は約 15 万短単位あり、学習用コーパスに必要な 5 万～10 万語という目安（小木曽 2014）をクリアしている。

[13] 解析精度は 4 つのレベルで評価される。すなわち、「単語境界」（単語の境界の正しさ）、「品詞認定」（「単語境界」＋単語の品詞・活用型・活用形の正しさ）、「語彙素認定」（「品詞認定」＋UniDic の見出し語である語彙素認定の正しさ）、「発音形認定」（「語彙素認定」＋読み方の正しさ）の 4 つである。

[14] 小椋・須永(2012)に従い、ルビよりも「中古基本読み」を優先する場合は、ルビと発音形は一致しない。

は校注者によって漢字表記語ほぼ全てにルビが振られており[15]、このルビを尊重しつつ語彙素・発音形を決定しようとすると、機械解析の結果とずれが生じやすい（表 4）。

表 4 “漢字一字表記、かつ、活用語尾（一部）非明示の用言”誤解析例

No.	ファイル名	前文脈	キー	後文脈	ルビ	出現発音形	語彙素読み	語彙素	品詞	解析活用型	活用形
1	35_今昔物語集 01_14c_S037_令誦方広経知父成牛語第三十七	家の主悲で、牛の辺に寄て、藁の座を敷て云く、「牛、実の我が父に	在さ	ば、此の座に登り給へ」と。	ましま	オワサ	オワス	おわす	動詞-一般	文語四段-サ行	未然形-一般
2	38_今昔物語集 04_30c_S003_近江守娘通浄蔵大徳語第三	持来べき便も思ず。奇異き事かな」と思て、「今は此の事	止め	て、偏に行ひをせむ」と思けれども、尚愛欲の思ひに勝ずして、	とど	ヤメ	ヤメル	止める	動詞-一般	文語下二段-マ行	連用形-一般
3	35_今昔物語集 01_13c_S042_六波羅僧講仙聞説法花得益語第四十二	愛執の過に依て、小蛇の身を受て、彼の木の下に住す。	願く	は、我が為に法花経を書写供養じて、此の苦を抜て	ねがは	ネガワシク	ネガワシイ	願わしい	形容詞-一般	文語形容詞-シク	連用形-一般
4	35_今昔物語集 01_11c_S015_聖武天皇始造元興寺語第十五	「東西二町に外閣を廻す事は、菩提涅槃の二果を證ずる相を	表す	。南北四町なる事は、生老病死の四苦を離れむ事を表す。	あらは	ヒョース	ヒョウスル	表する	動詞-一般	文語サ行変格	終止形-一般
5	37_今昔物語集 03_26c_S008_飛弾国猿神止生贄語第八	衣は思に随て着す、食物は	無	物無く食すれば、有しにも似ず、引替たる様に太りたり。	なき	ム	ム	無	名詞-普通名詞-一般		

こうした誤解析は、テキストの校訂方針、和漢混淆文である『今昔物語集』本来の表記の在り方に加え、出来る限り原文を尊重するという(1)テキスト整形の方針も影響している。

(1)テキスト整形における①返読文字の処理では、返読文字を除外（し意味の対応する助動詞（の一部）を挿入）しても、動詞の活用語尾を送り仮名として補入しなかった（「不知ズ→知ズ」、「不知リ→知**ザ**リ」など）。その結果、動詞の活用語尾が正しく解析されず、誤解析に繋がりやすくなった。

これと同様のことが、(1)テキスト整形における②助詞・助動詞等の省略表記に対する処理についても指摘できる。用言の活用語尾が非明示の場合は、UniDic に登録された活用形によって対応可能であると考え、ルビに基づく補読処理を施さなかった（「畢（をはり）テ」など）。しかし、実際には、非コアデータを扱う中で初めて出現したもの（新たに活用形として登録すべきもの）も多く、それらが結果として誤解析に繋がった。

発表者らは、まず、誤解析の大きな割合を占める“漢字一字表記、かつ、活用語尾（一部）非明示の用言”について、集中的に修正作業を行うことにした。そのためには、誤解析の可能性をもつ“漢字一字表記、かつ、活用語尾（一部）非明示の用言”の全例を洗い出す必要がある。そこで、非コアデータ中、ルビと発音形が不一致となっているキーに着目し、【ルビ 1 文字目と発音形 1 文字目が一致しないもの】、【ルビ 1 文字目と発音形 1 文字目は一致するが、ルビ 2 文字目と発音形 2 文字目が一致しないもの】の 2 パターンのリスト[16]を作成した上で、特に頻度の高いものから修正を施していった。表 5 には、活用語尾が明示されない漢字一字表記のもの[17]の中で、頻度・修正率ともに高かったものを示す。

別語彙素でありながら同一表記となりうるものが誤解析を起こしやすいのは、容易に想像がつく。表 5 で言えば、6「焼（ヤケル）」—9「焼（タク）」、17「行（オコナウ）」—22「行（アリク）」などである。このタイプには、7「畢（オエル）」—19「畢（オワル）」、29

[15] ルビは、「もし当時、仮名で書くとしたならばこう書いたであろうと校訂者が再構した仮名づかいで付してある（ただし、これには「平安仮名づかい」［発表者注：いわゆる「古典仮名づかい」とは違う、平安時代に行われた仮名づかい］は採用しなかった）。いわば校訂者の試論ともいうべきものである。」「新編日本古典文学全集」『今昔物語集 1』凡例

[16] ルビが歴史的仮名遣い、発音形が現代仮名遣いであることからリストに挙がってくるキーも多く（「可咲（をかし）」など）、目視での確認が必要であった。また、このリストは全ての品詞を対象とし作成したため、これを基に用言以外の修正も行っている。

[17] 活用語尾が（一部）明示される場合もあるため、語彙素自体の頻度とは必ずしも一致しない。

「下（クダス）」—30「下（クダル）」のように、動詞の自他で別語彙素となるものも含まれる。また、28「来（キタル）」のような漢文訓読体に特徴的な語が頻出する一方で、和文体に特徴的な「来（クル）」も使用されるため、類義語で文体差のある語彙素の対にも注意して修正作業を進める必要がある。

活用形ごとに見てみると、未然形・連用形の修正件数が多い。これには、その活用形自体の頻度が高いことに加え、未然形・連用形接続の助動詞の頻度が高い（後述）ことも関係していよう。漢字一字表記用言の発音形と関連する活用形については、次に述べる「助動詞の前接用言」の処理によって正しく修正されたものも多いことを補足しておく。

表 5　“漢字一字表記、かつ、活用語尾非明示の用言”修正例

№	表記	語彙素読み	頻度	誤解析	修正率	活用形別修正件数					
						未然形	連用形	終止形	連体形	已然形	命令形
1	開	ヒラク	84	84	100.0	3	79	2	0	0	0
2	咲	ワラウ	66	66	100.0	11	51	1	3	0	0
3	寄	ヨセル	41	41	100.0	8	32	1	0	0	0
4	合	アワセル	38	38	100.0	6	30	2	0	0	0
5	生	ウマレル	31	31	100.0	1	19	11	0	0	0
6	焼	ヤケル	22	22	100.0	1	21	0	0	0	0
7	畢	オエル	14	14	100.0	3	11	0	0	0	0
8	遣	オコセル	13	13	100.0	4	9	0	0	0	0
9	焼	タク	11	11	100.0	1	10	0	0	0	0
10	聞	キコエル	10	10	100.0	3	6	1	0	0	0
11	勝	スグレル	10	10	100.0	0	10	0	0	0	0
12	小	チイサイ	31	30	96.8	0	0	0	30	0	0
13	通	カヨウ	21	20	95.2	2	15	2	1	0	0
14	下	オロス	14	13	92.9	8	4	1	0	0	0
15	上	アガル	41	37	90.2	0	35	2	0	0	0
16	御	オワシマス	17	15	88.2	1	14	0	0	0	0
17	行	オコナウ	67	58	86.6	20	29	4	5	0	0
18	生	イキル	88	76	86.4	0	67	9	0	0	0
19	畢	オワル	42	36	85.7	2	34	0	0	0	0
20	遣	ツカワス	30	25	83.3	3	19	2	1	0	0
21	出	イダス	51	38	74.5	15	22	0	1	0	0
22	行	アリク	21	15	71.4	1	12	0	2	0	0
23	替	カワル	26	18	69.2	7	11	0	0	0	0
24	悪	アシイ	27	18	66.7	2	0	0	16	0	0
25	見	ミエル	82	53	64.6	34	18	1	0	0	0
26	入	イレル	157	100	63.7	15	84	1	0	0	0
27	立	タテル	138	80	58.0	9	69	2	0	0	0
28	来	キタル	466	265	56.9	33	215	2	9	1	5
29	下	クダス	22	12	54.5	9	2	1	0	0	0
30	下	クダル	103	54	52.4	4	47	2	1	0	0

b. 助動詞の前接用言

非コアデータに出現する助動詞のうち、用言を前接するものを抽出し、前接語の活用形や発音形について確認した。対象となったのは以下の助動詞である（語彙素で示す）。併せて、接続する活用形ごとのおよその頻度、括弧内には前接用言の修正件数を示した。

未然形接続：れる・られる・せる・させる・しむ・ず・じ・む・むず・まし・まほし
……約 8500(1730)
連用形接続：き・けり・つ・ぬ・たり（完了）・たし・けむ　……約 17000(1692)
終止形接続：べし・まじ・らむ・めり・なり　……約 1500(425)
連体形接続：なり（断定）　……約 8000(216)
命令形接続：り　……約 800(57)

また、助動詞として抽出されたキーそれ自体が正しい語彙素・活用形であるかについても確認している。特に、次のような、全体で 1 短単位とすべき他動詞「輝かす」「動かす」が「輝か｜す」「動か｜す」のように分割されていないか確認した（表 6）。

表 6　1 短単位とする他動詞例

No.	ファイル名	前文脈	キー	後文脈	ルビ	出現発音形	語彙素読み	語彙素	品詞	解析活用型	活用形
1	35_今昔物語集 01_11c_S004_道照和尚亘唐伝法相還来語第四	|其の後夜に至て、其の光房より出て寺の庭の樹を	曜かす	。|久く有て、光西を指て飛び行ぬ。	かかや	カカヤカス	カガヤカス	輝かす	動詞-一般	文語四段-サ行	終止形-一般
2	35_今昔物語集 01_14c_S009_美作国鐵堀入穴依法花力出穴語第九	|底の人此れを引て	動す	。|然れば、|「人の有る也けり」|と知て、忽に葛を以て籠を造て、	うごか	ウゴカス	ウゴカス	動かす	動詞-一般	文語四段-サ行	終止形-一般

c. 欠字欠文・破損の前後

(1)テキスト整形で述べたように、『今昔物語集』に見られる欠字欠文・破損は空格を示す記号「␣」「⇦␣⇨」で置き換えている。これらの前後の文字列は誤解析が生じやすい（表 7）。

表 7 欠字欠文・破損前後の誤解析例

No.	ファイル名	前文脈	キー	後文脈	ルビ	出現発音形	語彙素読み	語彙素	品詞	解析活用型	活用形
1	35_今昔物語集 01_13c_S038_盗人誦法花四要品免難語第三十八	二つの手をば、上に大なる木を渡して、其れを␣	か	せて縛り付けつ。		カ	カ	か	助詞-係助詞		
2	36_今昔物語集 02_19c_S018_三条大皇大后宮出家語第十八	簾の内の女房␣て泣事糸⇦␣⇨	し	。|挟み畢奉て、聖人居去かむと為る時に、聖人音を高くして云く、		シ	スル	為る	動詞-非自立可能	文語サ行変格	連用形-一般
3	35_今昔物語集 01_13c_S015_東大寺僧仁鏡誘誦法花語第十五	|或時には、夢の中に白象来て随ひ␣	ふ	。|「此れ定て普賢文殊の護り給ふ也」|と知ぬ。		フ	フ	符	名詞-普通名詞-一般		
4	36_今昔物語集 02_16c_S038_紀伊国人邪見不信蒙現罰語第三十八	て、大きに嘖て、即ち、往きて妻を喚ぶ彼の導師此れを見て、慈の心を発して教へて	導	␣す。|而るに、夫此れを⇦␣⇨|「汝は此れ我が妻を婚むと為る盗人の法師也。|速に、	だう	ドー	ドウ	ドウ	名詞-固有名詞-人名-一般		

例 1 は「␣か」で 1 語の動詞・未然形、例 2 は「⇦␣⇨し」で 1 語の形容詞・終止形、例 3 は「␣ふ」で 1 語の動詞・終止形とそれぞれ推測される。例 4 は「導␣す」のどこで短単位が切れるのか不明である。例 1・2 は意識的欠字（漢字表記保留）に後続する文字列、例 3･4 は破損の前後に位置する文字列であったために誤解析となった例である。このように、語の一部が「␣」「⇦␣⇨」となっているとほぼ誤解析になる。もちろん、語がそのまま欠字欠文・破損である場合も、その前後では誤解析の生じる場合がある。

欠字欠文・破損は計 705 箇所（欠字・欠文：479 箇所、破損 226 箇所）あり、これらについては空格を表す記号「␣」「⇦␣⇨」を抽出した上で、その前後の修正を行った。例えば、例 1「␣か」・例 2「⇦␣⇨し」・例 3「␣ふ」であれば、空格直後の「か」「し」「ふ」にそれぞれ「解釈不明」という品詞を付与した。例 4「導␣す」であれば、空格前後の「導」「す」にそれぞれ「解釈不明」という品詞を付与した。

d. 題

一つ一つの説話冒頭には、その説話の題と当該巻中で第何話にあたるかが示されている。コアデータではこの「題＋第〇」のまとまりに対して、人手で「題」という品詞を付与していった。そのため、「和漢混淆文 UniDic」を用いたとしても、非コアデータの「題＋第〇」部分は本文同様に解析されてしまい、誤解析となっていた（表 8）。計 477 箇所あるこれらは、コアデータと同様に人手で品詞を付与した。

表 8　題の誤解析例

No.	ファイル名	前文脈	キー	後文脈	ルビ	出現発音形	語彙素読み	語彙素	品詞	解析活用型	活用形
1	38_今昔物語集 04_31c_S029_蔵人式部拯貞高於殿上俄死語第二十九	蔵人式部	拯	貞高於殿上俄死語第二十九｜ 今は昔、円融院の天皇の御時に、	くらうどしきぶのじやうさだたかてんじやうにしてにはかにしぬることだいにじふく	スクイ	スクウ	救う	動詞-一般	文語四段-ハ行	連用形-一般
2	37_今昔物語集 03_24c_S056_播磨国郡司家女読和歌語第五十六	播磨国郡司家女読和歌語第	五十	六｜ 今は昔、高階の為家の朝臣の幡磨の守にて有ける時、指せる事無き侍有けり。	はりまのくにのぐんじのいへのをむなわかをよむことだいごじふろく	ゴジュー	ゴジュウ	五十	名詞-数詞		

(7) 現在の精度

(6)の精度向上作業を経て、2000 語のサンプリングチェックを再度行った。非コアデータの現在の精度は Level 4（発音形認定）で 99.1%まで上昇している。

５．おわりに

『今昔物語集』のコーパス化は、テキスト整形、コアデータ整備と「和漢混淆文 UniDic」の作成、非コアデータの精度向上作業の 3 つの柱からなる。本発表では、その 3 つ目の柱について、作業方針・作業内容を明らかにし、精度が約 94%から約 99%まで向上したという結果をもってその方針の妥当性を示した。『日本語歴史コーパス』鎌倉時代編 I には、コアデータに準ずる精度となった非コアデータも含め、『今昔物語集』（本朝部）全文の収録を予定している。

また、『今昔物語集』非コアデータの精度向上作業によって、今後のコーパス開発、『今昔物語集』研究に次のような展開が期待されよう。まず、コーパス開発においては、今回、特に注力した(6)a「漢字一字表記、かつ、活用語尾（一部）非明示の用言」の誤解析処理によって新たに辞書登録した活用形も多く、他の和漢混淆文資料のコーパス化におけるコスト軽減に繋がると期待される。研究面においては、(6)a で散見された“同一漢字表記でありながら別語彙素の語”に着目することで、語から表記、表記から語へと往還しながらの網羅的な調査が可能になる。これまでの先行研究では『今昔物語集』の用字法が一語一表記で安定しているとされてきたが、語によって表記の安定性が異なる点については慎重に検討する必要がある（田中 1988）。表記の安定性を考察するにあたっては、語から表記、表記から語へといった双方向の検索が瞬時に可能な『今昔物語集』コーパスにより、示唆的なデータが提供されるのではなかろうか。

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／田中牧郎）の成果の一部である。

参考文献

小木曽智信(2014)「歴史コーパスにおける形態素解析と辞書整備」『日本語学』33:14, pp.83-95

小椋秀樹・須永哲矢(2012)『中古和文 UniDic 短単位規程集』科研費　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」（課題番号 21520492）研究成果報告書 2（中古和文 UniDic HP からダウンロード可）

近藤泰弘(2014)「歴史コーパスとは何か」『日本語学』33:14, pp.6-15

佐藤武義(1984)『今昔物語集の語彙と語法』明治書院

田中牧郎(1988)「仮名交じり文 3『今昔物語集』」『漢字講座 5 古代の漢字とことば』明治書院

田中牧郎(2014)「『日本語歴史コーパス』の構築」『日本語学』33:14, pp.56-67

冨士池優美・岩崎瑠莉恵(2014)「『今昔物語集』の捨て仮名」『第 5 回コーパス日本語学ワークショップ予稿集』pp.261-270

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

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

関連 URL

「通時コーパスの設計」プロジェクト　http://historicalcorpus.jp/

『日本語歴史コーパス　平安時代編』　http://www.ninjal.ac.jp/corpus_center/chj/

「中古和文 UniDic」http://www2.ninjal.ac.jp/lrc/index.php?UniDic

「MeCab: Yet Another Part-of-Speech and Morphological Analyzer」http://code.google.com/p/mecab/