古文コーパスから「かな表記の語彙」を抽出する方法とその評価

1M－1

北村啓子
国文学研究資料館研究情報部
keiko@nijl.ac.jp

1　はじめに

翻刻テキストの集まりである古文コーパスを分析し、古文のテキスト処理用辞書の構築を目指している。特に古文で重要な「かな表記の語彙」を抽出する方法を提案する。この方法を「源氏物語」のテキストに適用し、これまで研究者の人力で作られた語彙集との比較を行うことにより評価する。

2　源氏物語の分析

辞書作りの考え方、コーパス分析の処理手順と、初期辞書として使用した語彙表とコーパスの作品との組合せによって分析結果から読み取れる特徴の概略を [1] で報告した。本稿では特定の一作品源氏物語に絞り、どの位の語彙を拾えるか、そして何が拾えなかったのかを評価し、「かな表記の語彙」を抽出する方法を提案する。

語彙表との比較、分析の処理手順を以下に示す (図 1参照)。

[漢字表記の語彙抽出]

1. 漢字表記の文字列を抽出する 1,910
2. 最長一致法で初期辞書と照合を行い語彙と認定する
   マッチ 1,498 ／漢字語彙数 8,180 →残 6,649
3. 不照合の文字列は最長文字列のまま新しい語彙として抽出する 412
4. 2. の残り漢字表記からそのよみ (かな) で書かれていたものを抽出する 3,272/6,649 →残 3,377
5. 残りの漢字表記の分析→ a .

[かな表記の語彙抽出]

6. 漢字表記を抜いた残りのかな文字列を抽出する
7. 初期辞書の中の漢字表記語彙のよみとかな表記の語彙 (漢字表記を持たないものも含む) との照合を行う
   マッチ 5,389/ かな語彙数 10,828 →残 5,434
8. 7. の残りかな表記からその漢字表記で書かれていたものを抽出する 1,655/5,434 →残 3,757
9. 残りのかな表記の分析→ a .

[分析]

10. 残ったかな文字列の中から、一文字のかなを除く (助詞が多いという判断) 残 14,867
11. 残った 2 文字以上のかな文字列をリストし、最長一致文字列でグループ分けする 2,376/14,867
12. 残ったかな表記の候補を分析し、抽出のアルゴリズムを考案する→ b .

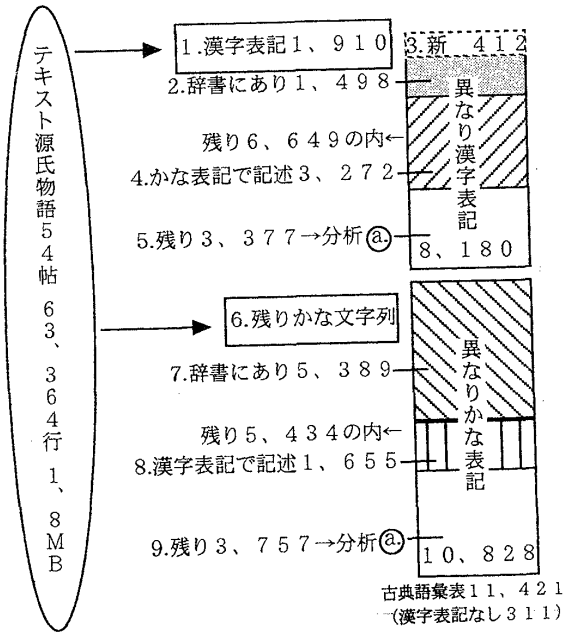

図 1: 語彙表との比較、分析の結果

源氏物語[1] ５４帖、６３、３６４行、１、８ＭＢのテキストと、古典対照語い表[2] １１、４２１語を利用する。

3　かな表記抽出の分析

a . 残りを分析する

[漢字表記]：抽出されなかった語彙表中の漢字表記残 6,649 の中の 3,272 語はよみ (かな表記) で記述されていた。ほんとうに出現しなかった漢字表記は、残 3,377 語である。

[かな表記]: 抽出されなかった語彙表中のかな表記残 5,434 の中の 1,655 は漢字表記で記述されていた。ほんとうに出現しなかったかな表記は、残 3,757 語である。

漢字／かなそれぞれのアプローチで処理をしたが、この残りは、異なりの漢字／かな表記のカウントの仕方が違うだけで、実態は同じ語彙が残っている。この中のサンプリング調査により抽出されなかった理由を分析する。抽出されなかった原因は次の種類に分類される。

- 活用形 (辞書は終止形)(c [3]) ６３％
  愛揺, めでゆする (cめでゆす - り)、逢難, あひがたし (cあひがた - き)
- 歴史的かな使い、かなの同音異表記 (～) ２４％
  愛執, あいしふ (～あいしう)、威言, おどしいふ (～をどしい

A Method and Evaluation of Building Dictionaries for Classical Text from Corpus
Keiko Kitamura
National Institute of Japanese Literature
142-8585, Tokyo, JAPAN

[1] テキストは国文学研究資料館中村康夫による底本「国文学研究資料館蔵承応版絵入源氏物語」の翻刻テキストを利用させて頂いた。

[2] 宮島達夫、中野洋、鈴木泰、石井久雄編、笠間書院版

[3] 原因を分類しコード化した 複数の原因が複合したものあり

ふ)、委伏, しをれふす (～ｃしほれふし)、一故, ひとつゆゑ (～ひとつゆへ)、一香, ひとつかをり (～ひとつかほり)、一昨日, をととひ (～”おとゞひ)、一類, ひとるゐ (～ひとるひ) 烏帽子, えぼうし (～ゑぼうし)、駅, うまや (～むまや)、遠遠, とほどほし (～とを＋＋し)[4]、押, おす (～ｃをし)、王家裔, わかんどほり (～”わかんとをり)、家損, けそん (～けそむ)、花文綾, けもんりよう (～けもんれう)

- 漢字、かな混じり (m) １８％
  逢坂山, あふさかやま (m あふ坂山)、葵草, あふひぐさ (m あふひ草)、梓弓, あづさゆみ (あづさ弓) 粟田山, あはたやま (あはた山)、伊勢人, いせびと (m いせ人)、異人人, ことひとびと (m こと人々)
- 異体字、新字旧字 (=) ３％
  一涙, ひとつなみだ (=m ひとつ泪)、阿弥陀経 (= 經)、壱越調, いちこちてう・いちこつてう (～ =m 一越でう)、卯月早月, うづきさつき (sub= 四月)、栄華, えいぐわ (= 栄花)、伽陵頻伽, かりようびんが (= 迦陵■伽)[5]、歌枕, うたまくら (m= 哥まくら)
- 濁点のありなし (読み不明)(”) ３％
  あいぎやうづく (あいぎやうつき)、向心, ひたぶるごころ (～”ひたふるごゝろ)、一腹, ひとつはら (”ひとつばら)、雨注, あまそそぎ (m”雨ぞゝき)、横様雨, よこざまあめ (”よこさまあめ)
- 複合語の間に助詞が入るケース (＾) ２％
  哀知顔, あはれしりがほ (＾あはれをしりがほ)、亥子餅, ゐのこもち (＾ゐのこのもち)、宇治川, うぢがは (＾＝宇治の河)、加持僧, かぢそう (＾加持の僧)、暇日, いとまび (＾ｍいとまの日)
- 複合語 (sub) １％
  沖玉藻, おきつたまも (sub 玉藻) 王家裔腹, わかんどほりばら (～”sub わかんとをり)
- 人名に付く呼称、固有名詞 (k)
  惟光様, これみつやう (惟光朝臣)、沖玉藻, おきつたまも (sub 玉藻) 王家裔腹, わかんどほりばら (～”sub わかんとをり)
- (残りは) 古典語彙表の底本とテキストの底本の記述の差 (異本の差) の可能性が高いのでそれぞれの原本に当らないと分析できない (今回は未調査)

b．残ったかな文字列の分析

残ったか文字列を頭から最長一致する文字列でグループ分けをする。活用形、複合語はここで吸収できる。14,867 語→5,376 のグループに分類できた。まだ一部のサンプル分析しかできていないが、4 割近くのかな表記語彙を抽出できている (リスト中の○)。また、語彙表にない新規の語彙として 5 割近くのかな表記語彙を抽出できる (リスト中の→)。グループのサンプルリストを掲載しておく。

1 あい: あいの
2 あいだ: あいだち○
○あいだれ: あいだれたり あいだれて
あえ: あえなうおぼ
4 あけ: あけざりければ あけず あけながらおりにけるを あけに あけぬ あけよ→sub
○5 あけた: あけたてば あけたり
6 あけて: あけてみた あけてみんよ あけてゐたり
7 あげ: あげずは あげも あげらるるを あげを
あげさ: あげさせ あげさせて →あげさす
8 あげた あげたり →あげつ
9 あげて: あげてみ あげてみたて →あげつ
10 あげの: あげのでうどめく
11 あざ: あざわらひて →あざわらふ
12 あざむ:
○13 あざむき: あざむきて あざむきゐてたて
○14 あざや: あざやぎ あざやぎて →あざやぐ
○15 あざれ: あざれか あざれたり あざれて あざればまんも あざればみ
16 あそはせど: →-
17 あそべば: →sub
18 あぢき: →sub
19 あぢきな: あぢきなの あぢきなふぞ →sub
20 あぢきなう: あぢきなうおぼ あぢきなうも →v あぢきなふ
21 あぢきなき: あぢきなきこ →v あぢきなし

[4] ＋＋は濁点躍り字 「とをどを」 に戻す処理をしている
[5] 異体字が JIS 内ににあため外字 (■) にしたためマッチしなかった

4 考察

1. 当初文法を使わないでどこまで可能かを見極めることを目標とした。活用語については最長文字列一致を押えることで、かなりの確率で抽出可能ではある。しかし、活用変化程度は辞書照合の際に活用形展開した方が計算コストが小さく確実である。改善したい。
2. 古文特有の問題である異体字、新字旧字、かなの同音異表記、歴史的かな使い、漢字－かな混じり、複合語の間に入る助詞、濁点のありなし (読み不明) など表記上のシソーラスの整備が必要である。
3. 今回はＪＩＳ第２水準までででエンコードしたテキストを使用した。文字コード不足は言うまでもなく、文字の代替や外字化による弊害が見られた。語彙を抽出する方向の処理においては、新語彙が多く抽出されることになるので問題ないが、後で辞書化する際に同定作業の負担が増える。
4. テキストエンコーディング時の凡例を吸収するフィルターをテキストの凡例の種類ごとに用意している (例えば躍り字)。元の字を復元できる範疇であれば問題ないが、必ずしもそうでないものもある。コーパスとしてテキストを分析する立場から、テキストエンコーディング時の凡例を決めるステップへフィードバックをしたい。
5. 今回は、語と語の照合による評価までで、原本に戻っての確認までは分析できなかった。異本による記述の差は大きく、原本の記述に当らないと正確には判断できない。
6. 辞書照合には成功したが、文脈エラーは発生する。辞書項目の語を抽出することが目的なので、使用する辞書にない語を発見することを重要視し、現在の大雑把な照合は辞書全体で見て許容範囲と考える。

5 課題

1. 異体字、新字旧字、かなの同音異表記、歴史的かな使いは、一意に決まるので、表記上のシソーラスとして蓄積し、検索時のフィルターとして使えるようにする。
2. 予想より漢字かな混じりで表記した例が多く見られた。これは原本の表記の特徴やエンコードの際の凡例に依存はするが、一般的に出現する可能性は高い。「かな表記」のみではなく、「漢字かな混じり表記」についても取り組む必要がある。
3. 原本での記述の仕方に特徴があり、またエンコーディングの際にもエンコーディングする人が記述のルールを決める。このルールを記述できるようにして、テキストの特徴をフィルタリングすることにより、語彙抽出の際の処理の効率化をはかる。
4. 異本を使うことでどの位相互補間できるか評価し、抽出の手法はシンプルで異本を使うことでカバーすることを目指したい。

参考文献

[1] 北村啓子：コーパスを利用した古文テキスト処理用辞書構築のための一考察, 情処全大 57 回,Vol3, pp.201-202,(1998)