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

1V－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

古文の世界でもフルテキストデータベースの可能性への期待が大きい。情報検索、テキスト処理などでの表記のゆらぎは現代語より切実な問題であり、これをカバーするシソーラスや異表記辞書、読み辞書、固有名詞辞書などの語彙に関する電子辞書の構築が待望されている。一つのアプローチとして、辞書構築に最初のステップとして、極力人手を介さずコーパスから大量な語彙を抽出することを狙う。現在利用できる材料で実験を行い、その分析結果からこの手法での具体的な戦略を検討し、提案したい。

2　辞書作りの考え方

ここでの「辞書」は、表記のゆれをカバーすることが目的で、語彙の「表記」と「よみ」のみで文法情報は持たない。人手を使わず自動的に語彙を粗々に集めることをポリシーとする。語彙数が集まって上で、異表記の辞書化やシソーラス化を極力人手を使わない方法の検討に取り組む。

テキスト処理で使われる漢字表記の語彙だけを抜き出すという方法がある。構文解析よりも処理が軽く、対象と処理内容によっては有効である。だたし、古文ではかなで表記される重要な語彙が多く存在するため、「かな表記」の語彙の拾い方を重要テーマに据える。

全体のフローとしては、コーパスと照合することによる新しい語彙を発見していく成長型の辞書である。利用できる古語辞書、語彙表の類を集め初期辞書を構築する。初期辞書を使ってコーパスを分析し、新たな漢字／かな表記語彙の候補を抽出し、語彙の認定と読みを確定して新しく辞書に登録する。この手順を踏んで、コーパスが順次多溜まって行くに従って辞書も成長していく。

3　コーパスの分析

著名な１４作品について既に手作業で作られた総索引の電子化された「フロッピー版古典対照語い表および使用法」(古典語彙表[2])を利用する。

1. 同じ作品の語彙表とテキストを比較することにより分析する
2. １４作品分の語彙表を初期辞書として使用し、他の作品のテキストとの比較を行う

処理手順:

[漢字表記の語彙抽出]

1. 漢字表記の文字列を抽出する
2. 最長一致法で初期辞書と照合を行い語彙と認定する
3. 不照合の文字列は最長文字列のまま新しい語彙候補とする

[かな表記の語彙抽出]

4. 漢字表記を抜いた残りのかな文字列を抽出する
5. 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)との照合を行う
6. 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)
7. 残った２文字以上のかな文字列をリストし、かな表記語彙として抽出すべき文字列の条件を分析する

テキストはそれぞれの入力形式(凡例)に従っている。実験では、タグや付加情報は除き本文のみを使用している。ただし本文中の文字で、躍り字など凡例によりコード化されているものはできる限り元の字を復元した。「行」は物語はテキストに書かれた「。」の単位、和歌は意味的な識別タグの単位にあわせた。分析結果を表1.2.にまとめる。

手順中の3.4.6.のサンプルリストを掲載しておく。

```
----- ミスマッチの漢字文字列(新しい語彙候補)- ------""
葦邊 羽 翁 俄
梓 雨 河 我見
一 雲 河内 海人
右大將藤原 延 花 :
----- 4.-6.のプロセス ---------------- ""
残かな文字列: むかし、おにしるよしして、、とこ、かいまみてけり。
2> かな表記 : むかし おに しる よし いま しる とこ しる かい にし よし よし
ほんとの残り:,,,, して,, か, みてけり
残かな文字列: おもほえずふるさとに、いとはしたなくてありければ、、まどひにけり。
2> かな表記 : あり はし ふるさと はし さと なく なく おも
ほんとの残り:, ほえず, にいと, た, て, ければ, まどひにけり
残かな文字列: おきてやる。
2> かな表記 : おき
ほんとの残り:, てやる
残かな文字列: そのおたりける。
2> かな表記 : その
ほんとの残り:, おたりける
残かな文字列: かすが, られずとなむをいつきていひやりける。
2> かな表記 : いつ つき かす いつき かすが
ほんとの残り:, が, られずとなむを, きていひやりける
残かな文字列: ついでおもしろきこととともや, けん。
2> かな表記 : こと とも とも とも とも ついで ついで おも ついで
ほんとの残り:,, しろき,, や, けん
----- のサンプル(かなコード順)---------- ""
あかねども あひ あらぬ ...
あけ あらざりけり あらねば
あけてかへるに あらざりければいるかづけんとす あらねばいと
あはじ あらず いきけり
---------- ""
```

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

[2]宮島達夫、中野洋、鈴木泰、石井久雄編、笠間書院版。収納作品と元データの総索引のリストはフロッピー同梱の使用方を参照。凡例についてはそれぞれの総索引を参照。フロッピー版ならびに偉大なる元データの総索引の作成者の方々に深謝。

表1．源氏物語は源氏物語語彙表、古今和歌集は古今和歌集語彙表を使用

テキスト	辞書中のかな表記／語彙数	テキスト行数／文字数*2	マッチした(漢字表記)辞書語彙	マッチしなかった(漢字表記)新しい語彙)	マッチしたかな表記＋よみ辞書語彙	マッチしなかった2文字以上かな／残りかな表記
絵入り源氏物語	278/11421	63366/1891960	62658	340	3738	32673
校訂源氏物語	278/11421	10762/1059124	52465	530	1885	20660
古今和歌集	29/1994	2474/110776	6160	223	659	3057
校訂古今和歌集	29/1994	2674/112758	5191	789	1038	3757

表2．１４作品総合彙表 (23877 words) を使用

テキスト	行数／文字数	辞書にある語彙(漢字)	新しい語彙(漢字)	辞書にある語彙(かな)
絵入源氏	63366/1891960	12686	63	2828
拾遺和	3328/141290	8836	96	1354
後拾遺和	3500/158312	12089	133	1255
金葉和	3199/124515	10524	134	1163
詞花和	1222/51520	4170	83	836
千載和	3525/142079	16033	178	1057
新古今和	5261/194349	20133	188	1238
新勅撰和	3602/132966	14559	187	1149
続後撰和	3721/128219	15244	185	995
続古今和	5368/183130	21056	177	1211
新後撰和	4350/146662	18034	178	990
玉葉和	7746/280360	32567	260	1367
続千載和	5783/194141	24826	243	1078
続後拾遺和	3762/126154	15748	177	969
風雅和	6009/204365	25694	238	1133
新千載和	6601/238329	30397	272	1150
新拾遺和	5368/184566	22447	212	1139
新後拾遺和	4206/135682	17123	192	958
新続古今和	6175/212069	28057	256	1173
方丈記	159/7653	516	80	243
伊勢物語	463/26337	1474	229	243
蜻蛉日記	251/26929	1893	298	422
枕草子	2366/170332	8315	694	1187
紫式部日記	755/65471	3666	433	714
大鏡	2418/316788	27163	1651	1336
更級日記	251/26929	1893	298	422
竹取物語	232/14456	890	139	242
土佐日記	390/21648	73	36	520
徒然草	1315/69191	5612	1315	724

4　考察

1. 残りのかな文字列の辞書照合では、漢字語彙に付く助詞が頭に出てくることが多いため、最長一致法はあまり適していない。残りのかな文字列に対して任意の組み合わせのパターンマッチングで辞書照合を行った。このため一文字かな表記が多く出現した。

2. 随時利用可能な小さな辞書を使って軽い処理、かつ極力人手をかけないで、コーパスから語彙を抽出することを目的とした。したがって、構文的や意味的に正しいかどうかには触れず、既に辞書に存在する文字列(かなも含む)は既存語彙であるという大雑把な判断を採用した。また以下の点でも厳密性に欠けている。

   - 最長一致法で後ろから短くして辞書照合を行っているため、複合語の後ろの語彙は拾えてない
   - ミスマッチの文字列は最長文字列を新しい語彙候補としているため、複合語の分割はできていない
   - かな表記の抽出で一文字のかなを除いた(助詞が多いという判断)が、実際は一文字のよみを持った漢字表記の語彙は結構ある

3. 分析結果の数値で明らかなように、かな表記語彙の占める割合は多い。またテキストによる差が大きいこともわかる。(語彙表も凡例を決めて人手で分析したという意味では一つのテキストを作ったのと同値である。)これは底本表記の実際の差もあるが、電子化の際のルール(凡例)に依存する部分が大きい。

5　課題

本稿では詳細な分析報告までは行えなかった。以下の点について引続き調査分析を行い、論文発表または別の機会に報告を行いたい。

1. 辞書照合には成功したが、文脈エラーは発生する、現在の大雑把な照合は辞書全体で見て許容範囲か否か
2. 作品ごとまたは作品のジャンルごとに特徴がないか
3. 一定本から翻刻したものと校訂したもの(異本比較の上、判断の入ったもの)との差がないか
4. 異本を使うことで相互補間できるか
5. 電子化の凡例と辞書化に敵不敵の関連があるか、凡例を利用できないか

[3] テキストは国文学研究資料館中村康夫代表[本稿中のボルド体作品名]、同館安永尚志代表[本稿中の明朝体作品名]のDB科研により構築された中から利用させて頂いた。前者は一つの底本から翻刻がなされ、後者は校訂本に依る。