テキストから「かな表記の語彙」を抽出する試み

—コーパスを利用して古典語彙を収集するために—

北　村　啓　子

要　旨　古文のテキスト処理をしようとすると、表記のゆらぎは切実な問題であり、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。古文のテキストデータ化が研究者個人で活発に行われるようになり十年を数え（国文学資料館でも二十年近く前から実験されていた）、大規模にテキストデータベースとして構築するプロジェクトもいくつか興っている。これらの活動で作られてきた古文テキストは、古文を対象にした一種の大規模コーパスを形成している。

この生データであるコーパスから直接古典語彙を抽出するというアプローチは、トップダウンに作られた辞書にはない古文のテキスト処理に実際に役立つ語彙集の抽出が期待できる。

特に古文を扱う上では、「もののあはれ」の例を出すまでもなく「かな表記の語彙」に重要な語彙が多く存在する。ここでは、この「かな表記の語彙」を抽出することに狙いを定め、現在利用できるテキストを分析することにより、コーパスから語彙を抽出する手法を検討し、いかに抽出できるかを試みる。

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く、翻刻したテキストの電子化、流通が活発に行われている。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。

しかし、十数世紀に渡り使われてきた文字数も語彙数も大きく、時代とともに変遷してきており、時代を遡った固有名詞の数は無限に近い数になる、などの古文固有の特徴から発する問題が大きな障害となっている。計算機で使える文字コードの不足は言うまでもなく、代替を行うためテキストから原本上の正確な表記は求められない。現段階では、計算機の大きな制約下での標準化の追求は学問上の価値を低下させることにもなる。

国文学研究資料館において1980-1981に取り組まれた研究では［1］［2］、語彙索引を作ることが第一の目的であった。そのため厳密な電子化の凡例を決め、テキストを人手で分かち書きし、品詞情報を付加するテキストデータ作成が行われた。研究者が個人でテキストを作成し、個人の研究に必要な付加情報を付け、研究目的にあったテキスト処理ができるようになってきた現在では、統一的な標準化は現実的ではない。また、人手をかけて加工したデータを作成するより、速くシンプルなテキストを作成し、より高度なテキスト処理機能を利用してゆるやかな標準化をカバーし、個別の研究上重要な情報を付加して利用できる自由度を確保していく方向が望まれるであろう。

一つのアプローチとして、厳しい制約下で電子化されたテキストの表記のゆらぎをカバーする辞書類の構築が考えられる。その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目指す。現在利用できるコーパスで実験を行い、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。

構文解析よりも軽い処理としてテキスト処理で使われる技法に、漢字表記の語彙だけを抜き出すという方法がある。対象と処理内容によっては有効である。だたし古文では、「もののあはれ」が代表するように、かな表記の重要な語彙が多く存在するため、「かな表記」の語彙の拾い方を考案することを重要テーマに据える。

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第一の目標とする。語彙数が集まった上で、極力人手を使わないで異表記の辞書化やシソーラス化の方法の検討に取り組む。

全体のフローとしては、コーパスと照合することにより新しい語彙を発見していく成長型の辞書である。まず、利用できる古語辞書、語彙表を集め、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

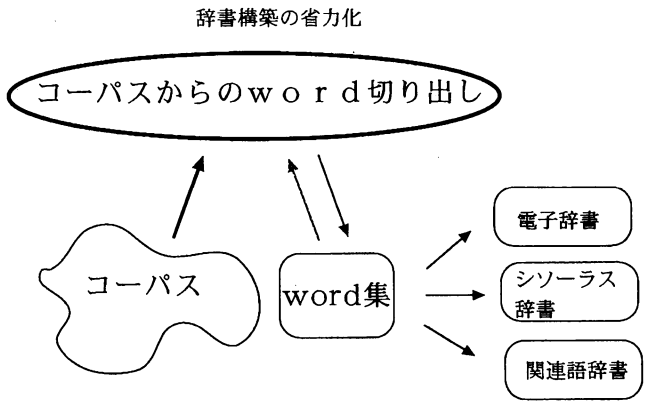

図１　全体のフロー

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

初期辞書として、万葉集、竹取物語、伊勢物語、古今和歌集、土佐日記、後撰和歌集、かげろふ日記、枕草紙、源氏物語、紫式部日記、更級日記、大鏡、方丈記、徒然草の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）を利用する。コーパスとしては、国文学研究資料館で構築されてきたテキストデータベースの中から利用する。**

い．模擬実験

「源氏物語」のテキストについて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、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、評価する。

ろ．大量コーパスの処理による特徴分析

使用する語彙表の作品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からコーパスから語彙を効率的に抽出する方法を検討する。

コーパス分析の処理手順を図2に示す。

以下それぞれの処理結果とその分析を報告する。

---

*宮島達夫、中野洋、鈴木泰、石井久雄編、笠間書院版。元データの総索引のリストはフロッピー同梱の使用法を参照。凡例についてはそれぞれの総索引を参照。

**研究情報部データベース室で構築中の原本テキストデータベースならびに中村康夫助教授、安永尚志教授により構築されたデータベース（データベース科研による）の中から利用させて頂いた。[3] [4] [5]

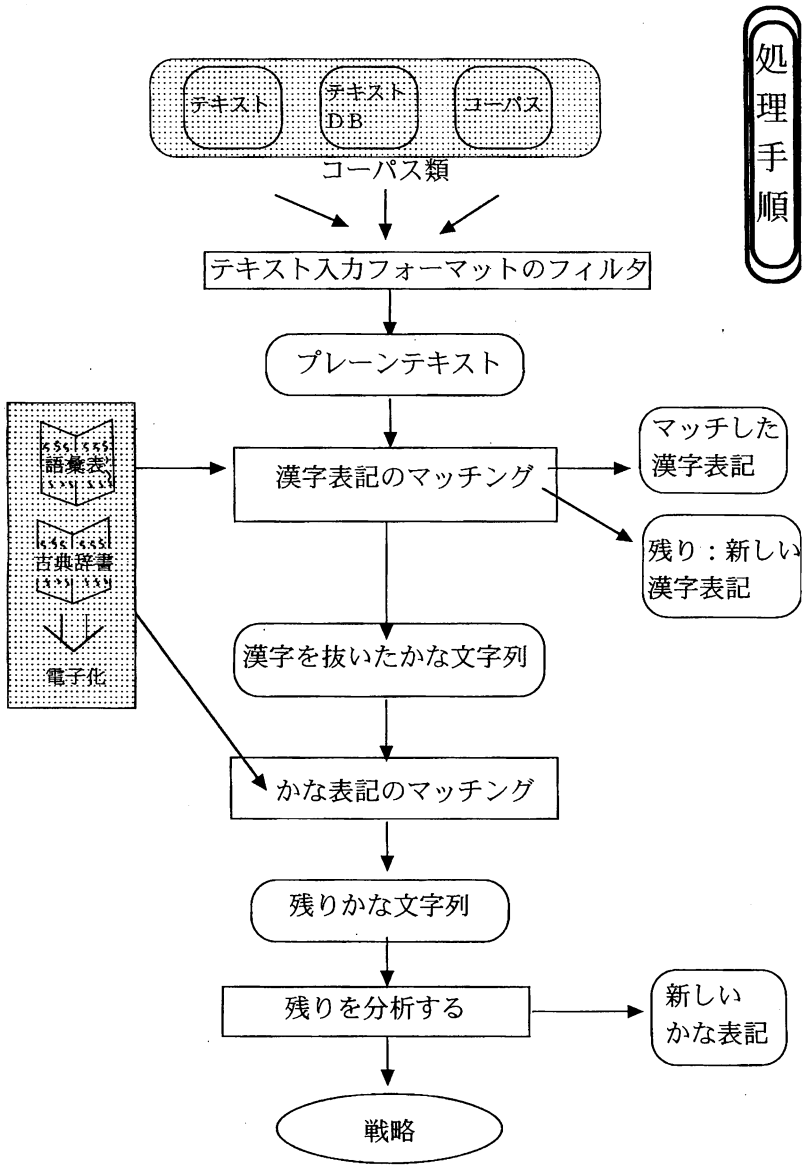
処理手順
テキスト
テキストDB
コーパス
コーパス類
テキスト入力フォーマットのフィルタ
プレーンテキスト
語彙表
古典辞書
電子化
漢字表記のマッチング
マッチした漢字表記
残り：新しい漢字表記
漢字を抜いたかな文字列
かな表記のマッチング
残りかな文字列
残りを分析する
新しいかな表記
戦略

図2　処理手順

4．源氏物語の分析

源氏物語*54帖、63,364行、1.8MBのテキストと、源氏物語の古典対照語彙表11,421語を利用する。古典語彙表では、表記上同じでも異なる語は別の語として認識してあるが、ここでは表記しか認識しないため異なり表記のみをカウントする。語彙表に納められた語彙の内、異なりの漢字表記8,180語，異なりのかな表記10,828語、その内漢字表記のないもの311語である。

テキストは凡例に従って作られているが、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元の字を復元した。「行」はテキストに書かれた「。」の単位にあわせた。

語彙表との比較、分析の処理手順を以下に示す。処理結果は、図3に示す。（結果の数値は手順にも書き込んである。）

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　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

---

*原本デキストデータベースで作成された底本「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。

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　最長一致で語彙と認定する

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．

手順中の3.6.-9.のサンプルリストを掲載しておく。

--------- 3.不照合の漢字文字列（新しい語彙候補）----------

固有名詞：　按察 伊与 尉 衛 大液 王經 …

旧字：　大將 哥 戀 爰 當 兒 齋 螢 …

一般的：　逢 逢瀬 逢夜 衣 卯 浦 悦 河 介 学 給 …

数詞がついた：　九尺 四五人 十月中十日 …

部分的に照合した：　官 儀 吉 宮 月廿 …

接続した： 悦給（～悦-給ふ） 限四五人（～限り-四五人） …

……

--------- 6.-9.のプロセス ----------

テキスト：いづれの御時にか。

残かな文字列：いづれの,にか。

2>かな表記　：3いづれ 2いづ

ほんとの残り：„,にか,

テキスト：女御更衣あまたさふらひ給けるなかに。

残かな文字列：,あまたさふらひ,けるなかに。

2>かな表記　：3あまた 2また 2なか 2あま

ほんとの残り：„さふらひ,ける„,

テキスト：いとやむごとなききはにはあらぬが。

残かな文字列：いとやむごとなききはにはあらぬが。

2>かな表記　：2やむ 2むご 2には 2きは 2いと

ほんとの残り：„ごとなき„あらぬが,

テキスト：すぐれてときめき給ふありけり。

残かな文字列：すぐれてときめき,ふありけり。

2>かな表記　：2とき 2すぐ 2あり

ほんとの残り：,れて,めき„,けり,

テキスト：はじめよりわれはと思ひあがり給へる御かた＋＋$。

残かな文字列：はじめよりわれはと,ひあがり,へる,かたがた。

2>かな表記　：4かたがた 3はじめ 2われ 2たが 2かた 2あが

ほんとの残り：,より,はと,,,,,へる,,,

テキスト：めざましきものにおとしめそねみ給。

残かな文字列：めざましきものにおとしめそねみ,。

2>かな表記：4めざまし 4しきもの 3そねみ 2もの 2まし 2とし 2しめ 2しき 2きも 2おと

ほんとの残り：,,,に,,,,,お,,,,,

テキスト：おなしほど。

残かな文字列：おなしほど。

2>かな表記　：2ほど 2なし 2しほ

ほんとの残り：,,,,,

テキスト：それより下らうの更衣たちは。

残かな文字列：それより,らうの,たちは。

2>かな表記　：2らう 2たち 2それ

ほんとの残り：,より,,,,,,,

テキスト：ましてやすからず。

残かな文字列：ましてやすからず。

2>かな表記　：3まして 2やす 2まし 2から

ほんとの残り：,,,,,

,,,,,,

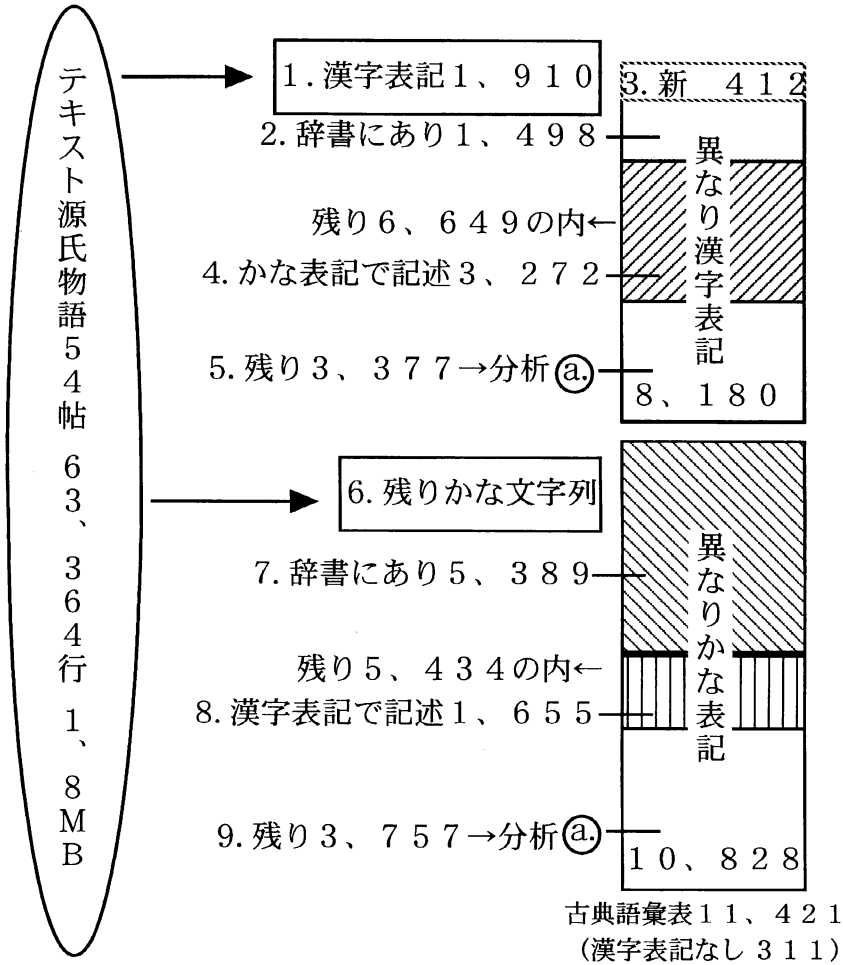

図３　語彙集との比較、分析の結果

5．かな表記を抽出するための分析

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)* 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（~をどしいふ）、

萎伏,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）**

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）

---

*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

**＋＋は濁点躍り字「とをどを」に戻す処理をしている

・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）*、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（s）1%

沖玉藻,おきつたまも（s玉藻）

王家裔腹,わかんどほりばら（~"s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（k惟光朝臣）、沖玉藻,おきつたまも（ks玉藻）

王家裔腹,わかんどほりばら（k~"sわかんとをり）

---

*異体字がJIS内にあるが，外字（■）にしたためマッチしなかった

・（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

b．残ったかな文字列の分析

残ったかな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、7割近くのかな表記語彙を抽出できた（リスト中の○）。また、語彙表にない新規の語彙として2割弱のかな表記語彙を抽出できた（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○

○3あいだれ：あいだれたり あいだれて

4あえ：あえなうおぼ →あえなし

5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 →あく

○6あけた：あけたてば あけたり →あく

○7あけて：あけてみた あけてみんよ あけてゐたり →あく sub

8あげ：　あげずは　あげも あげらるるを あげを →あぐ

9げさ：あげさせ あげさせて　　→あげさす

○10あげた：あげたり　　→ あぐ

○11あげて：あげてみ あげてみたて　→ あぐ

12あげの：あげのでうどめく

13あざ：　あざわらひて　→ あざわらふ sub

14あざむ：

○15あざむき：あざむきて あざむきゐてたて →あざむく

○16あざや：あざやぎ あざやぎて →あざやぐ

○17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 →あざる

18あそはせど： →゛ あそぶ

○19あそべば：→あそぶ

20あぢき： →sub

○21あぢきな：あぢきなの あぢきなふぞ →あぢきなし

○22あぢきなう：あぢきなうおぼ あぢきなうも →あぢきなし

○23あぢきなき：あぢきなきこ →あぢきなし

.....

→ sub：複合語またはその部分 → v：活用形の差 → ゛：異表記の差

7．大量コーパスの分析

ここでは、多種大量のテキストを分析するため、3．処理手順中の1.～3. 6.～7. 10.～11.の処理のみを行う。源氏物語と同様に、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。

い．同じ作品の語彙表との比較

語彙の異なりを調べ、異本間の語彙の差異、凡例による語彙の差異などを分析する。古典語彙表の中に語彙集がある作品で、二種類以上のテキストが利用できる源氏物語、古今和歌集について、両者を対象としてテキストによる差の分析を行う。

ろ．他の作品の語彙表との比較

い．の実験を行った作品について、ジャンルの違う作品の語彙表と入れ替えて分析を行い、ろ.の結果と比較して、異なるジャンルの語彙表を使った時の特徴を分析する。

は．14作品の総合語彙表との比較

古典語彙表の全体を使って、和歌集として二十一代集（古今和歌集以外は語彙表に納められていない)、物語、日記、随筆として、語彙表に納められた中から10作品を語彙抽出を試みる。*

い.ろ.は.の分析結果をそれぞれ表１.２.３.にまとめる。分析結果の表から以下のことが読み取れる。

い．同じ作品の辞書を使っても、テキストによって漢字表記／かな表記の割合が大きく異なる。翻刻時の凡例の差に大きく依存すると考えられる。

ろ．語彙数の少ない和歌の辞書を使用した場合、源氏物語と古今和歌集の間でマッチしたかな表記数は源氏物語が少し多いが、マッチした漢字表記数は殆んど差がない。逆に源氏物語の辞書を使って古今和歌集を処理すると、漢字表記は多く、かな表記も和歌の辞書の場合より多くマッチしている。単順に数だけを比較すると辞書の語彙数が多い方が優位に見えるが、語彙数と比例して増える訳ではない。やはり、同じジャンルの辞書の方が優位で、ジャンルによって使用される語彙が異なる傾向があると言えよう。

は．和歌集は語彙表に納められていない作品であるが、漢字表記、かな表記と

---

*表中、ボルド体作品名は中村康夫助教授、明朝体作品名は安永尚志教授により構築されたデータベース。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。

表１．同じ作品の辞書を使う

源氏物語語彙表（278 かな表記／11421 漢字表記）

古今和歌集語彙表（29／1994）

後撰和歌集語彙表（24／123）を使用

テキスト	テキスト行数／文字数	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	63366/945980	1498	412	5389	33258
校訂源氏物語	10762/529562	2031	530	1885	13109
古今和歌集	2474/ 55388	486	247	786	2535
校訂古今和歌集	2674/ 56379	375	789	1038	3639
後撰和歌集	3522/ 80008	463	313	754	3626

表２．他の作品の辞書を使用

テキスト	使った辞書(かな表記／漢字表記)	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	古今 (29/ 1994)	545	1012	976	41867
絵入り源氏物語	後撰 (24/ 1923)	554	1033	954	42986
古今和歌集	後撰 (24/ 1923)	455	271	516	3151
後撰和歌集	古今 (29/ 1994)	431	298	558	4166
古今和歌集	源氏 (278/11421)	615	110	786	2137
後撰和歌集	源氏 (278/11421)	627	115	818	3638

表２．14作品総合彙表（828／23877 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)
拾遺和歌集	3328/ 70645	906	96	1354
後拾遺和歌集	3500/ 79156	1063	133	1255
金葉和歌集	3199/ 62257	1035	134	1163
詞花和歌集	1222/ 25760	676	83	836
千載和歌集	3525/ 71039	1334	178	1057
新古今和歌集	5261/ 97174	1462	188	1238
新勅撰和歌集	3602/ 66483	1355	187	1149
続後撰和歌集	3721/ 64109	1360	185	995
続古今和歌集	5368/ 91565	1486	177	1211
新後撰和歌集	4350/ 73331	1302	178	990
玉葉和歌集	7746/140180	1780	260	1367
続千載和歌集	5783/ 97070	1649	243	1078
続後拾遺和歌集	3762/ 63077	1417	177	969
風雅和歌集	6009/ 10218	1613	238	1133
新千載和歌集	6601/119164	1879	272	1150
新拾遺和歌集	5368/ 92283	1569	212	1139
新後拾遺和歌集	4206/ 67841	1419	192	958
新続古今和歌集	6175/106034	1677	256	1173
方丈記	159/ 3826	280	80	243
伊勢物語	463/ 13168	366	229	243
蜻蛉日記	251/ 13464	607	298	422
枕草子	2366/ 85166	8315	694	1187
紫式部日記	755/ 2735	675	433	714
大鏡	2418/158394	2011	1651	1336
更級日記	251/ 13464	439	298	422
竹取物語	232/ 7228	263	139	242
土佐日記	390/ 10824	32	36	520
徒然草	1315/ 34595	1277	1315	724

も非常に照合率が高い。和歌集で使われる語彙は近いことがわかる。また和歌集は同じ凡例に基づいて翻刻、電子化されているのも一つの大きな理由であろう。それ以外の作品は、語彙表に納められているものを選んだが、照合率は漢字表記、かな表記とも和歌集に比べて低い。電子化は同じ凡例だが、翻刻の凡例が作品ごとに異なっていることが理由に考えられる。

今回はこれ以上の詳細な分析には至らなかったが、ここまでの分析結果から
・辞書の語彙数は多い方が優位
・ジャンルが異なるテキストを処理する方が優位
・同じ作品、ジャンルでも異なる凡例により電子化されたテキストが優位
であると言える。常識的な所見しか得られなかったが、それが証明はできた。これは、コーパスとして使用するテキストを選択する時に役立つであろう。

9．考察

1．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、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は既存語彙であるという大雑把な判断を採用した。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から見て許容範囲と考える。また以下の点でも厳密性に欠けている。
・最長一致法で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

2．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。

3．当初文法を使わないでどこまで可能かを見極めようと考えた。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いので改善したい。

4．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（読みは不明のため）など表記上のシソーラスの整備が必要である。

5．今回はJIS第2水準までで電子化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。

6．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、電子化する際の凡例に依存する部分が大きい。

7．電子化する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の経験が、電子化時の凡例を決める際の参考になれば幸いである。

8．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。

10. 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電子化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、また電子化の際にも電子化する人が記述方法の凡例を決める。この凡例を計算機上でフィルタリングに利用できるよう記述の仕方を定め、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。

最後に、ここで紹介した処理プログラムは著者ホームページからダウンロード可能、また処理結果は近々の公開を目指して目下整理中である。ともに

URL　http：//www.nijl.ac.jp/~keikoを参照されたい。

謝辞：

快くテキストを提供してくださった当館安永教授、中村助教授のこれまでの長年の努力と偉大な成果なくしては本試みは実現せず、この報告は生まれなかった。ここに尊敬と感謝の意を表させて頂く。また、今回利用させて頂いた「フロッピー版古典対照語い表および使用法」の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する。

[参考文献]

[1] 市古貞次（代表）：国文学語彙検索システム及び索引誌の作成に関する研究、文部省科学研究費試験研究（2）＃581009研究報告書（1982）

[2] 国文学研究資料館：古典テキストデータ用データベースシステムの開発、国文学研究資料館報告第11号、(1983)

[3] 安永尚志：日本古典文学作品本文データベースの開発とデータ記述文法について、国文学研究資料館紀要、第18号、pp.1-18（1992）

[4] 安永尚志：日本古典文学作品フルテキストデータベースのためのデータ記述文法に関する実証的研究、文部省科学研究費一般研究（A）、＃03402051研究報告書（1995）

[5] 中村康夫（佐竹昭廣・立川美彦代表）：重層型情報時代に対応する国文学高機能情報形成手法の開発とその実用化に関する研究、文部省科学研究費基盤研究（A）(2)、＃07401014研究報告書（1998）