論文

『源氏物語大成』の品詞情報つきフルテキストデータベースの作成について[†1]

上田英代[†2] 上田裕一[†3] 村上征勝[†4]

本論文は,「源氏物語」の統計的手法を用いた計量分析を行うために,『源氏物語大成』の品詞情報付きフルテキストデータベースを作成した過程での様々の成果と,いくつかの問題点について,今後の方向性も含めて明らかにしたものである.コンピュータを用いた文献の計量分析は端緒についたばかりであるが,データベース作成過程で試みた諸々の方法は今後も有効なものとなるであろう.

1 データベース作成の目的

「源氏物語」は千年も前に書かれた長編物語であり,日本の古典作品中の最高峰といわれる.紙の希少な平安時代にあって,書き直しなどは度々は行えなかったであろう.書き直しが度々行えないということは,すでに書き終えた部分は消去せずに,内容の追加や修正を行わなければならなかったということであり,作者の構想が当初の方針と変わった場合には,相当の準備のもとに,内容を少しずつ変化させながら追加していったと考えられる.こうした追加修正の仕方が作品の構成に影響を及ぼし,読者を困惑させる謎として残っているように思われる.また,流布の仕方も個々人が書写することによって少しずつ広まっていったために,様々な異本を生み出すことになった.「源氏物語」成立後およそ200年位たつと,こうした異本を整備し本文を校訂しようとする試みが始まり,同時に諸々の研究も始まった.以後,日本文学の代表的古典として,あらゆる面から研究が進められている.後半の宇治十帖の文体が前半と微妙に違い,和歌も前半よりは巧みなことから宇治十帖他作家説や,第2巻「帚木」の冒頭部分の叙述が不自然なことから全54帖の構成に疑問を投げかけ,この作品が複数の作家によって書かれたとする複数作家説,成立過程における後期挿入説,物語音読説らが出されているが,未だ明確な結論が出されているとはいえない.これに対し,従来の研究方法とは全く違った,コンピュータによる統計的手法を用いた計量分析を行うことによって,何らかの解を与えることが研究の目的である.

統計的手法を使った分析は昭和32年に,安本美典氏(現産能大学)が行っているが,著者らはこの分析を更に発展拡大させた.コンピュータによる情報の大量処理によって,今まで細かく手作業で行っていた仕事でも,一気に素早く結果が得られると同時に,手作業ではできなかった様々な分析が行えるようになる.

コンピュータによる計量分析を行う際には,まず最初に,機械可読のテキストデータベースを作らなければならない.次にそのプレーンなテキストデータベースを,始めから終わりまで統一された基準単位で分かち書きしなければならない.更に解析を深める為には品詞情報も付加したほうがよい.そこで著者等は「源氏物語」の一つの校訂本文を決めて,品詞情報つきフルテキストデータベースを作成することにした.

この一連の作業を,流れ図で表わすと図1のごとくである.ここでは本文の選定から,機械可読文献の作成,自動単語分割,自動品詞つけの過程と問題点について述べる.

---

†1 The Full-Text Database of Genji Monogatari Taisei with Codes for Parts of Speech

†2 Hideyo Ueda, 統計数理研究所外来研究員, The Institute of Statifical Mathematics

†3 Yasuichi Ueda, もとぶ野毛病院

†4 Masakatsu Murakami, 統計数理研究所

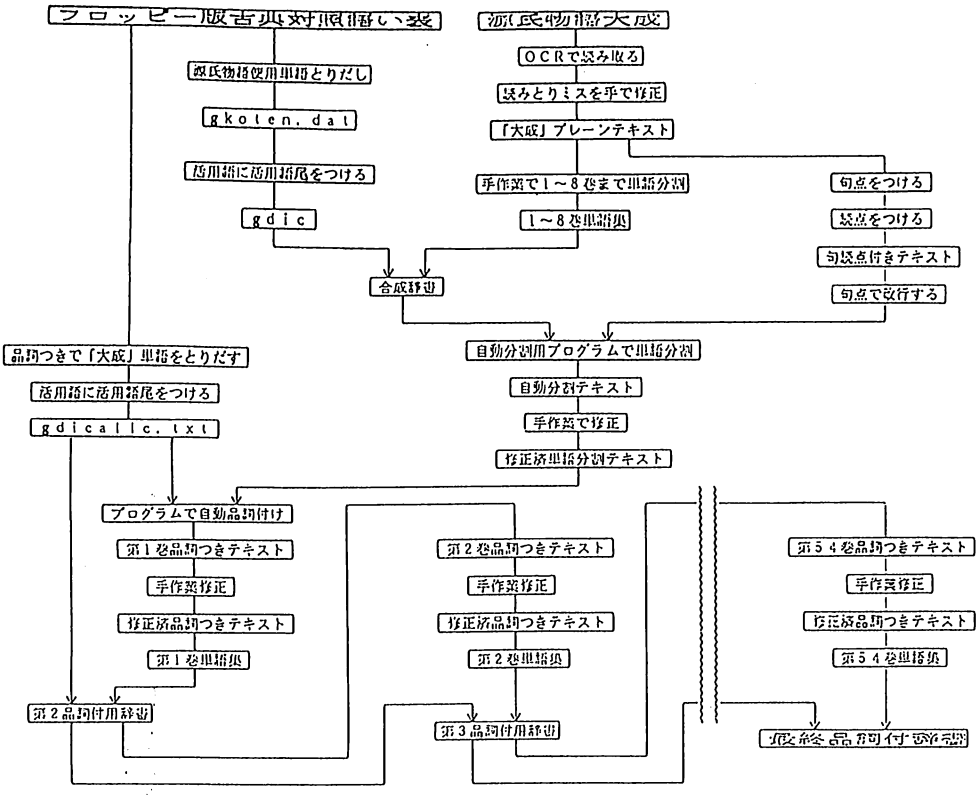


図1

2 テキストの選定と本文入力

本文の入力方法については，現在いくつかの方法が試みられている．影印本を画像認識させてそのままデータとする方法，活字になった校訂本文を手作業で入れていく方法，活字の校訂本文をOCR(Optical Character Reader)等で読み込む方法，などである．これらの入力方法は，使用目的，データ作成の人手，作業量，繁雑さなどを考慮して選択されなければならない．

著者等は，品詞情報つきデータベース作成の作業をできるだけ自動化することも研究課題としたため，活字の校訂本文をOCRで読み込むことにした．「源氏物語」の校訂本文は現在数多く存在するが，本文テキストとして，池田亀鑑編著の『源氏物語大成』(中央公論社)を選んだ．『大成』を選んだ理由は，写本の系統を明らかにした本文を使用し，他本との校異を載せ，語彙索引が完備しているからである．単語の自動分割を行う際，語彙索引が完備していることは基準単位の決定に役立つため，利用価値が高かった．

読みとりに使用したOCRは，富士電機(XP-50S)である．このOCRは読み込むときにルビ行も認識してしまうのだが，『大成』にはルビ行がなく，又，漢字表記が少ないという点でも，読みとりエラーが少ないと予想された．本文の入力はまず，『大成』の本文の見開き2ページ分をB4用紙1枚にコピーし，948ページ分をOCRで読みとり，手作業で修正を行った．読みとり誤読例は，図2のごとくである．『大成』の本文をOCRで読みとったものが図3である．

このOCRの漢字読みとり機能はJIS第一水準で2965字，第二水準で526字なので，第二水準の漢字に読みとれないものが多い．修正作業の主な

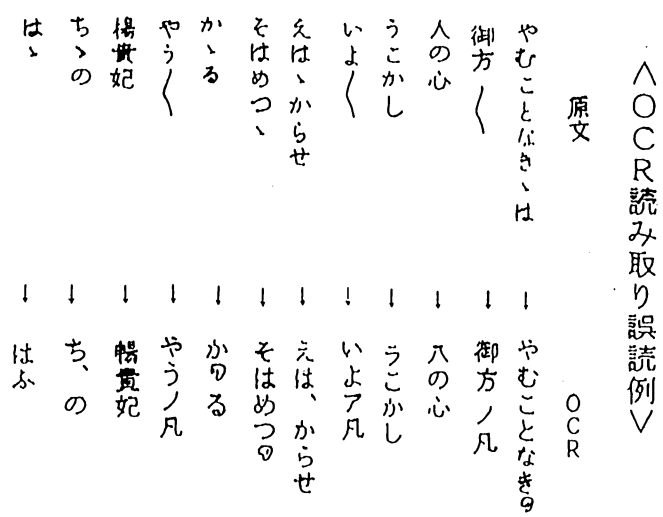


図 2

ものは，誤読漢字を修正入力すること，踊り字を入れること，繰り返し記号に文字数分の記号を入れることなどである．文字数分の記号を入れたのは，各巻における総文字数，文の長さ，文の数のデータが必要だったからである．繰り返し記号は，同じ単語でも使用されている場合と使用されていない場合がある．たとえば，「中～に」と「中中に」などである．この場合機械的に，前者は「中¥に」とし，後者は「中中に」のままとした．しかしコンピュータは，これらを別単語として別々に認識してしまう点が問題となったため，後に自動単語分割用辞書には「中中に」と「中¥に」の両方を登録し，異なり語数を数える時には，¥マークをすべて文字に変換して数えた．

3 自動単語分割

3.1 手作業による単語分割から自動単語分割へ

3.1.1 単語分割用辞書作りと自動単語分割プログラムの作成

まず『大成』の1～8巻までを，手作業で単語分割した．しかし同一人物がこの作業をしても，分割基準の揺れが生じたので，この作業を自動化することになった．『大成』のプレーンテキストは全部で約2MBあり，それを一巻毎に単語分割するとしても，分割するべき基準の単語が集まっている単語集，即ち分割用辞書と，分割されるべき巻の二つのテキストファイルが同時にオープンされていなければならないし，コンピュータ上にはその作業領域も必要である．作業領域の大きさと作業速度の速さ，作業プログラムがコマンドレベルで行える等々の利点を考慮し，自動単語分割はOSがUNIXのNEWS-1850を使用して行うことにした．

最初に，手作業で単語分割した1～8巻までの単語集を作る．この単語集をアイウエオ順で文字数の少ない単語順に並べ換えて，自動単語分割用辞書とする．自動単語分割プログラムは，Cシェルプログラムを使って作成した(図4)．このプログラムは，まず分割用辞書の中の最長文字列の単語150個ずつで本文を一行毎に検索し，その単語があれば“<=”と“->”で囲み分割する．次にその次の長さの単語を分割するが，すでに“<=”と“->”で区切られた単語の中は分割しない．引き続き一文字ずつ短い単語を，順に分割してゆく．“<=”と“->”で区切られていない部分が自動単語分割されていない単語であり，辞書にない単語である．単語分割を文字数の多い単語から行なっていくため，後半ほど検索する部分が少なくなってい

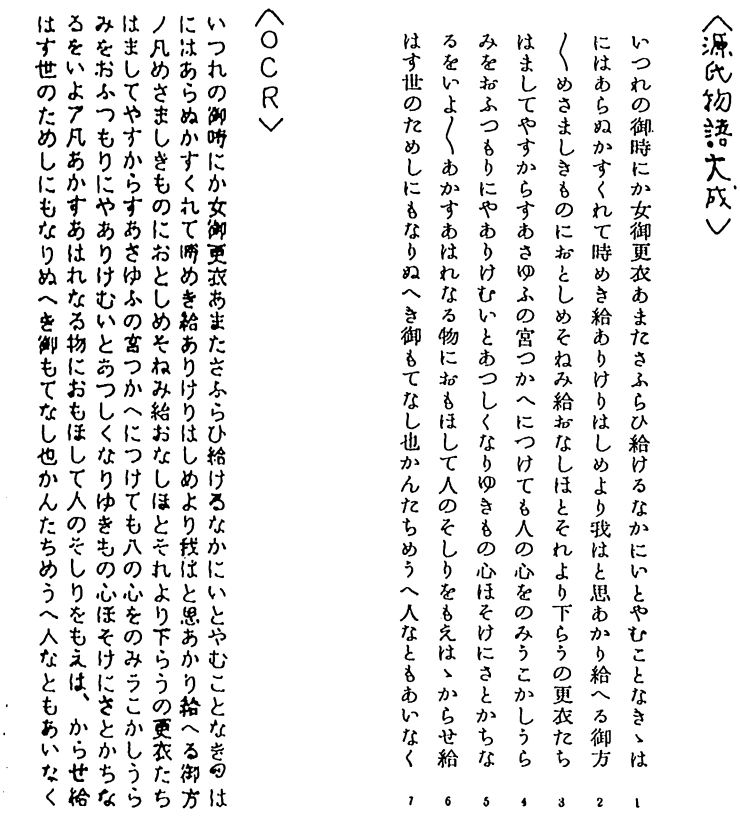


図3

く.

3.1.2 最初の自動単語分割の試み

残念ながら『大成』の本文には句点が付いていない. 3.1.1 で述べたプログラムを実行して自動単語分割をする際, 行毎の処理をするためと, 正確さを増すためには句点の情報が必要なので写本の系統は違うが『源氏物語(日本古典文学大系)』(岩波書店)を参考にして句点を付けた. このときに『大系』で終止形でも『大成』で終止形でないものは, 句点を付けず,『大成』で終止形のもののみ句点をつけた. この時点で読点を同時に付けなかったのは, 句点だけの情報でどの程度正確に区切れるかを知るためである.

次に 3.1.1 でできた 1～8 巻までの単語集を分割用辞書として, 句点のついた第 1 巻「桐壺」の巻を自動単語分割し, 手作業による分割と比較した. 若干の相違はあるが, かなりの同一性を確認できたのでプログラム上は問題ないことがわかった.1～8 巻までの単語集を分割用辞書として自動単語分割した結果が図 5 である. そこで, 次に手作業分割していない第 9 巻「葵」を, おなじ分割用辞書を使って自動単語分割したものが図 6 である.

3.2 辞書用語彙の追加

『フロッピー版古典対照語い表』(笠間書院)が入手できたので, その中の「源氏物語」使用語彙を分割用辞書に加えることにした.『古典対照語い表』は,『源氏物語大成総索引』より使用単語を収録している. 見出し語は自立語のみ収録しており, 活用する語は終止形だけ載り, 濁音, 半濁音を含んですべてひらがな表記となっている.『大成』の本文には, 濁音, 半濁音がないので, 見出し語すべてを清音に直した.『古典対照語い表』中の「源氏物

```
#源氏物語の分割 sed command 使用
#gsplit_s <源氏物語> <辞書>      と使用
cp $1 zz1;cp $2 zz2
echo "END" >>zz2
cat zz2|tr -d '¥012'|tr '。' ' '  '>t$2
set w = 'wc t$2'
@ w[1] = $w[2] / 150; @ w[1]++ ; @ w[3] = $w[2] % 150; @ w[3]--
while( $w[1] )
        @ w[2] -= 150
        if ( $w[2] > 0 ) then
                cat t$2|tr '  ' '¥012'|tail +$w[2]|head -150 > s$2
                set i = 150
        else
                cat t$2|tr '  ' '¥012'|head -$w[3] > s$2
                set i = $w[3]
        endif
        set l = 'cat s$2|tr '¥012' ' ''
        cp commandf cmf
        while( $i )
                echo -n '/^<=.*->$/¥!' >> cmf
                echo "s/$l[$i]/<=$l[$i]->/g"  >>cmf
                @ i--
        end
        cat zz1 |sed -f cmf > yy
        cat yy |sed 's/</ </'g |sed 's/>/> /g' |tr ' ' '¥012'
               |sed '/^$/d' > xx
        rm zz1 yy cmf s$2
        mv xx zz1
        @ w[1]--
        echo $w[1]
end
rm zz2 t$2
```

図4

<=いつれ-><=の-><=御時-><=に-><=か-><=。<=女御-><=更衣<=あまた-><=さふらひ-><=給-><=
ける-><=なか-><=に-><=<=いと-><=やむことなき-><=きは-><=には-><=あら-><=ぬ-><=
か-><=すくれ-><=て-><=時めき-><=給-><=あり-><=けり-><=。<=はしめ-><=より-><=我-><=
は-><=と-><=思あかり-><=給へ-><=る-><=御方<=<=めさましを-><=もの-><=に-><=おとし
め-><=それみ-><=給-><=。<=おなし-><=ほと-><=それ-><=より-><=下らう-><=の-><=更衣-><
=たちは-><=さして-><=やすから-><=す-><=。<=あさ-><=ゆふ-><=の-><=宮つかへ-><=に-><
=つけて-><=も-><=人-><=の<=心-><=を-><=のみ-><=うとかし-><=うらみ-><=を-><=おふ-><
=つもり-><=に-><=や-><=あり-><=けむ-><=いと-><=あつしく-><=なり-><=ゆき-><=もの
心ほそけに-><=さと-><=かち-><=なる-><=を-><=い-><=よ-><=<=く=あ-><=かす-><=あはれ
なる-><=物-><=に-><=おもほし-><=て-><=人-><=の-><=そしり-><=を-><=も-><=え-><=は
ゝから-><=せ-><=給は-><=す-><=世-><=の-><=ためし-><=に-><=も-><=なり-><=ぬ-><=へ
き-><=御もてなし-><=也-><=。<=かんたうめ-><=う へ人-><=なと-><=も-><=あいなくめ-><=
を-><=そはめ-><=つ-><=いとさはゆき-><=人-><=の-><=御おほえ-><=なり-><=。<=もろと
し-><=に-><=も-><=かゝる-><=こと-><=の-><=おとり-><=にり-><=に-><=こそ-><=世-><=
も-><=みたれ-><=あしかり-><=けれ-><=と-><=やう-><=<=く=あめのした-><=に-><=も-><=
あうさなう-><=人-><=の-><=もてなやみくさ-><=に-><=なり-><=て-><=楊貴妃-><=の-><=
ためし-><=も-><=ひき-><=いて-><=つ-><=へく-><=なり-><=ゆく-><=に-><=いと-><=はし
```

図5

語」使用語彙は，11421 語である．この見出し語の　　みで分割用辞書をつくり自動単語分割した結果が
```

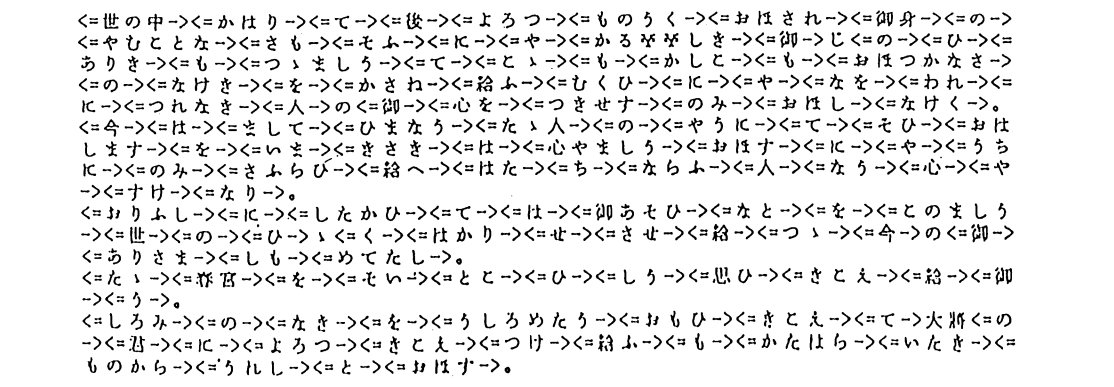


図6

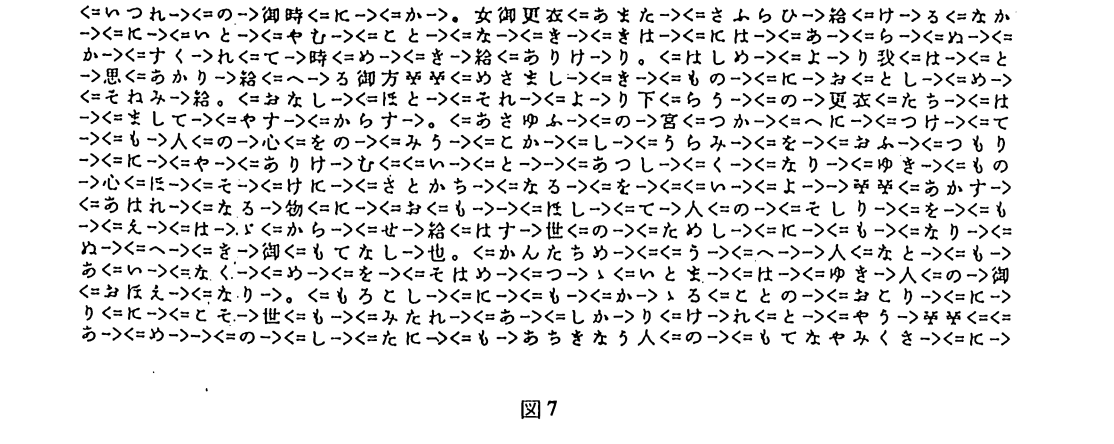


図7

図7であるが，見出し語から作ったこの辞書には漢字が含まれていないため，漢字を含む単語が分割されていない．

次に，『語い表』から採った『大成』単語集の活用する単語すべてに活用形をつけ，更に助動詞もすべての活用形を含めて追加した．この活用形を追加した『語い表』による『大成』単語集と，手作業分割による1～8巻までの単語集とを合成した辞書を作成した．その単語集を小文字数からアイウエオ順に並び変え，最初の自動単語分割用辞書とした．この時一文字の単語は，分割が不正確になりやすいので削除し，二文字の単語から収録した．

3.3 自動単語分割の工夫

3.3.1 読点情報の付加と自動単語分割

単語の自動分割をより正確なものとするため『大系』本を参考に句点をつけたテキストに更に読点をつけた．即ち読点のところでは必ず単語が切れるからである．この読点つきのテキストを,3.2でできた合成辞書で自動単語分割した結果が図8である．巻1の「桐壺」の巻を自動単語分割するのに1時間27分かかった．

この「桐壺」の巻を手作業で正確に修正し，この巻の異なり単語集を作り，分割用辞書にない単語を元辞書に追加する．追加した辞書で次の巻を自動単語分割して修正する．一巻ごとに新出単語は辞書用単語として，元辞書へ追加されていく訳である．この方法で順に正確な分割を行なってゆく．巻9「葵」をここまでの合成辞書で自動単語分

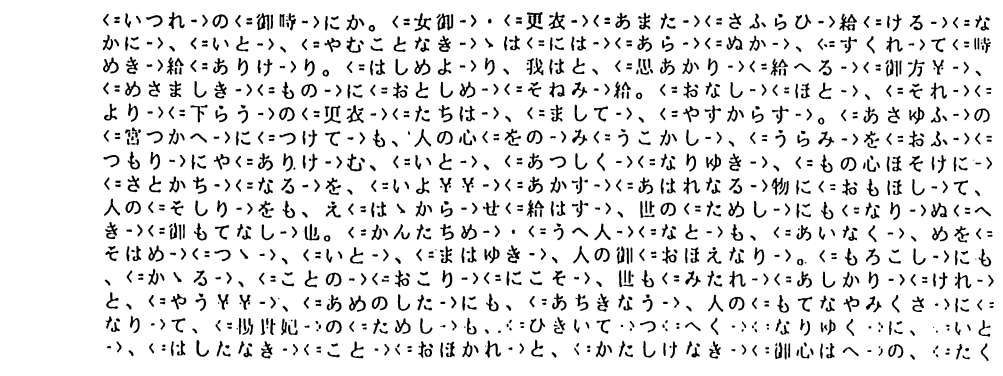


図8

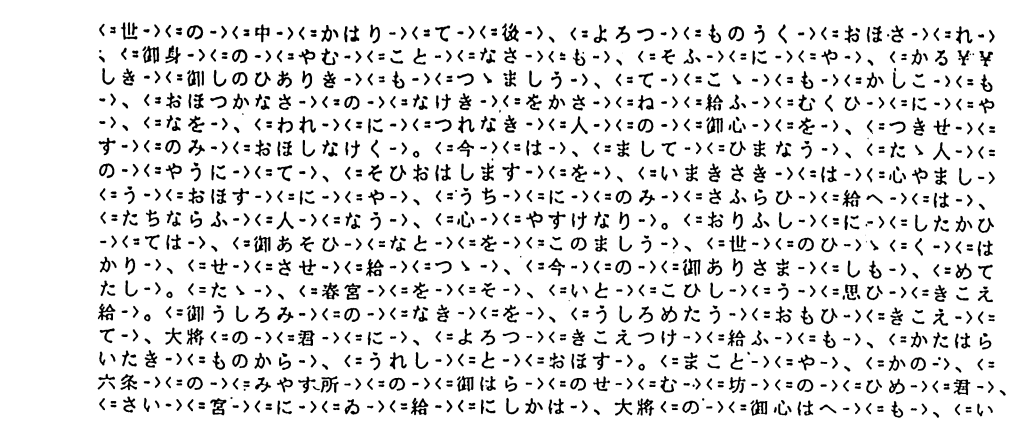


図9

割した結果が，図9である．『語い表』の見出し語はひらがなのみなので,9巻以降は辞書中単語に漢字混じりの単語が増えていくことになる．こうした工夫によって自動単語分割は正確さを増していったが，一文字や二文字の助詞，助動詞などの単語分割は，まだ不完全さが残った．これらの単語は自動単語分割の後，手作業で修正した．最終の「夢の浮橋」の巻は80%の正確さで自動単語分割が行なえた．さらに単語分割の精度を上げるために，単語の前後関係から判断して分割箇所を決定するプログラムなどの開発がのぞまれる．

4 自動品詞付け

4.1 テキストの修正と品詞つけ用辞書作り

プログラムによる自動単語分割で不正確な箇所は，手作業で正しく分割した(図10)．巻の長さにもよるが，たとえば桐壺の巻で約3時間程度の作業時間である．次に，正確に分割された単語に品詞情報をつける．そのために品詞つけ用辞書を作成したが，ここでも『古典対照語い表』を利用した．まず「源氏物語」使用単語を品詞つきで取り出し，品詞つけ用辞書とする(図11)．次に，活用する自立語は終止形で載っているので，その語幹にすべての活用語尾をつけた単語を追加し元辞書とする．ただし，たとえば動詞四段活用の終止形と連体形

/いつれ/の/御時/に/か/。/女御/・/更衣/あまた/さふらひ/給/ける/なか/に/、/いと/、/やむことなき/きは/に/は/あら/ぬ/か/、/すくれ/て/時めき/給/あり/けり/。/はしめ/より/、/我/は/と/、/思あかり/給へ/る/御方々/、/めさましき/もの/に/おとしめ/そねみ/給/。/おなし/ほと/、/それ/より/下らう/の/更衣たち/は/、/まして/、/やすからす/。/あさゆふ/の/宮つかへ/に/つけ/て/も/、/人/の/心/を/のみ/うこかし/、/うらみ/を/おふ/つもり/に/や/あり/けむ/、/いと/、/あつしく/なりゆき/、/もの心ほそけに/さとかちなる/を/、/いよ々々/あか/す/あはれなる/物/に/おもほし/て/、/人/の/そしり/を/も/、/え/はゝから/せ/給は/す/、/世のためし/に/も/なり/ぬ/へき/御もてなし/也/。/かんたちめ/・/うへ人/なと/も/、/あいなく/、/め/を/そはめ/つゝ/、/いと/、/まはゆき/、/人/の/御おほえ/なり/。/もろこし/に/も/、/かゝる/、/こと/の/おこり/に/こそ/、/世/も/みたれ/あしかり/けれ/と/、/やう々々/、/あめのした/に/も/、/あちきなう/、/人/の/もてなやみくさ/に/なり/て/、/楊貴妃/の/ためし/も/、/ひきいて/つ/へく/なりゆく/に/、/いと/、/はしたなき/こと/おほかれ/と/、/かたしけなき/御心はへ/の/、/たくひなき/を/たのみ/に/て/、/ましらひ/給/。/ちゝ/の/大納言/は/なくなり/て/、/はゝ/北の方/なん/、/いにしへ/の/、/人/の/よし/ある/にて/、/おや/うちくし/、/さしあたりて/世/の/おほえ/はなやかなる/御方々/に/も/いたう/おとらす/、/なにこと/の/きしき/を/も/、/もてなし/たまひ/けれ/と/、/とりたて/て/、/はかゝ々しき/うしろみ/

図 10

あ,〔代名〕
ああ,〔感動〕
あいきやう,〔名詞〕
あいきやうつく,〔動詞〕
あいしふ,〔名詞〕
あいたちなし,〔形容〕
あいたる,〔動詞〕
あいなし,〔形容〕
あいなたのみ,〔名詞〕
あいなたのめ,〔名詞〕

図 11

あか,〔動詞〕〔名詞〕〔連体〕
あかき,〔形容〕〔名詞〕
あかし,〔形容〕〔動詞〕〔名詞〕
あかり,〔動詞〕〔名詞〕
あかれ,〔動詞〕〔名詞〕
あき,〔動詞〕〔名詞〕
あけくれ,〔動詞〕〔名副〕
あさき,〔形容〕〔名詞〕
あさけ,〔形容〕〔名詞〕
あさけれ,〔形容〕〔動詞〕

図 12

は同じであるなど，活用語尾が同じものは一種類だけ採る．同音異義語で同一品詞のものは一語だけ採り，異なる品詞のものは一つの語に可能性のある品詞をすべてつけ，複数の品詞をつけた多品詞語とした(図 12)．図 11 から図 12 までの作業はいくつかのプログラムをかけることによって作成したので，手作業の時間はさほどかかっていない．

4.2 自動品詞つけ作業経過

4.2.1 自動品詞つけ

4.1 でできた辞書を使って，正確に分割されたテ

/いつれ〔　　〕/の〔助詞〕〔名詞〕/御時〔　　〕/に〔助詞〕〔助動〕〔動詞〕〔名詞〕/か〔助詞〕〔代名〕〔名詞〕/。/女御〔　　〕/更衣〔　　〕/あまた〔副詞〕/さふらひ〔助動〕〔動詞〕〔名詞〕/給〔助動〕/ける〔助動〕/なか〔動詞〕〔名詞〕/に〔助詞〕〔助動〕〔動詞〕〔名詞〕/いと〔副詞〕〔名詞〕/やむことなき〔形容〕/きは〔名詞〕/に〔助詞〕〔助動〕〔動詞〕〔名詞〕/は〔助詞〕〔助動〕〔名詞〕/あら〔動詞〕/ぬ〔助動〕〔動詞〕/か〔助詞〕〔代名〕〔名詞〕/すくれ〔動詞〕/て〔助詞〕〔助動〕〔名詞〕/時めき〔　　〕/給〔　　〕/あり〔動詞〕〔名詞〕/けり〔助動〕/。/はしめ〔動詞〕〔名詞〕/より〔助詞〕〔動詞〕/我〔　　〕/は〔助詞〕〔助動〕〔名詞〕/と〔助詞〕〔助動〕〔副詞〕〔名詞〕/思あかり〔　　〕/給へ〔　　〕/ろ〔助動〕/御方々〔　　〕/めさましき〔形容〕/もの〔名詞〕/に〔助詞〕〔助動〕〔動詞〕〔名詞〕/おとしめ〔動詞〕/そねみ〔動詞〕〔名詞〕/給〔　　〕/。/おなし〔形容〕/ほと〔名詞〕/それ〔代名〕〔動詞〕〔名詞〕/より〔助詞〕〔動詞〕/下らう〔　　〕/の〔助詞〕〔名詞〕/更衣たち〔　　〕/は〔助詞〕〔助動〕〔名詞〕/まして〔副詞〕/やすからす〔連語〕/。/あさゆふ〔名詞〕/の〔助詞〕〔名詞〕/宮つかへ〔　　〕/に〔助詞〕〔助動〕〔動詞〕〔名詞〕/つけ〔動詞〕〔名詞〕/て〔助詞〕〔助動〕〔名詞〕/も〔助詞〕〔名詞〕/人〔　　〕/の〔助詞〕〔名詞〕/心〔　　〕/を〔助詞〕〔名詞〕/のみ〔助詞〕〔動詞〕/うこかし

図 13

/いつれ〔代名〕/の〔助詞〕〔名詞〕/御時〔名詞〕/に〔助詞〕〔助動〕〔動詞〕〔名詞〕/か〔助詞〕〔代名〕〔名詞〕/。/女御〔名詞〕/更衣〔名詞〕/あまた〔副詞〕/さふらひ〔助動〕〔動詞〕〔名詞〕/給〔助動〕/ける〔助動〕/なか〔動詞〕〔名詞〕/に〔助詞〕〔助動〕〔動詞〕〔名詞〕/いと〔副詞〕〔名詞〕/やむことなき〔形容〕/きは〔名詞〕/に〔助詞〕〔助動〕〔動詞〕〔名詞〕/は〔助詞〕〔助動〕〔名詞〕/あら〔動詞〕/ぬ〔助動〕〔動詞〕/か〔助詞〕〔代名〕〔名詞〕/すくれ〔動詞〕/て〔助詞〕〔助動〕〔名詞〕/時めき〔動詞〕/給〔助動〕/あり〔動詞〕〔名詞〕/けり〔助動〕/。/はしめ〔動詞〕〔名詞〕/より〔助詞〕〔動詞〕/我〔代名〕/は〔助詞〕〔助動〕〔名詞〕/と〔助詞〕〔助動〕〔副詞〕〔名詞〕/思あかり〔動詞〕/給へ〔助動〕/ろ〔助動〕/御方々〔名詞〕/めさましき〔形容〕/もの〔名詞〕/に〔助詞〕〔助動〕〔動詞〕〔名詞〕/おとしめ〔動詞〕/そねみ〔動詞〕〔名詞〕/給〔助動〕/。/おなし〔形容〕/ほと〔名詞〕/それ〔代名〕〔動詞〕〔名詞〕/より〔助詞〕〔動詞〕/下らう〔名詞〕/の〔助詞〕〔名詞〕/更衣たち〔名詞〕/は〔助詞〕〔助動〕〔名詞〕/まして〔副詞〕/やすからす〔連語〕/。/あさゆふ〔名詞〕/の〔助詞〕〔名詞〕/宮つかへ〔名詞〕/に〔助詞〕〔助動〕〔動詞〕〔名詞〕/つけ〔動詞〕〔名詞〕/て〔助詞〕〔助動〕〔名詞〕/も〔助詞〕

図 14

キストの巻 1 から自動品詞つけを行なった結果が図 13 である．この自動品詞つけは，C 言語によるプログラムで処理は UNIX である．処理時間を短くするために，品詞つけ用辞書の単語をア行イ行ごとのグループにわけ，本文でア行の単語に品詞つけするときは，辞書中のア行の単語グループより探すという方法をとった．このため処理時間はかなり短くなった．最初の品詞つけ用辞書は，見出し語がひらがなのみで，元辞書中に該当単語がないときは，その単語には品詞がつかず[ ]内が空欄となる．次に，[ ]内が空欄の単語を集めてファイルにし，この[ ]に品詞を入れる．新たに品詞をつけたこの単語集は，最初の品詞つけ用辞書になかった単語集である．それまでの品詞つけ用辞書に，この新異なり単語集を加える．新異なり単語集を加えた辞書で，その巻をもう一度品詞つけする．すると[ ]内が空欄の単語がなくなる(図 14)．こうした作業が何度も行えるのも，「桐壺」の巻で約 20 分という短い処理時間のためであり，作業の試行錯誤を行う際には，処理速度が速いということは好都合である．次に多品詞の単語を文脈から判断して手作業で品詞を決定する(図 15)．この巻がすべて正確に品詞つけされてから，この巻の品詞つき異なり単語集を作り元の辞書に加える．重複同一単語は除き，同音異義語は多品詞語とする．新たに異なり単語が追加された辞書を使って次の巻の自動品詞つけをする．次の巻で，[ ]内が空欄の単語を集め，前巻と同様の作業をする．修正の後，再びその巻の異なり単語集を作り，新異なり単語を元辞書に加え次の巻の自動品詞つけを行なう．その作業を 54 帖分続ける．

4.2.2 辞書の工夫

辞書の単語が増加すると同時に，品詞が増加してゆく単語も出てくるので，約 5 巻ごとに辞書の点検をすることにした．たとえば次のように簡略化する．「いと」には最初[副詞][名詞]と自動品詞つけされるが，「いと」の[名詞]は，源氏物語 54 帖中 6 例(複合名詞も含む)しかないので，[副詞]のみとする．また「とし」は[形容][名詞]とついてくるが，「とし」[形容]は「源氏物語」には用例がないので「とし」[名詞]のみとする．「に」は[助詞]

いつれ〔代名〕/の〔助詞〕/御時〔名詞〕/に〔助詞〕/か〔助詞〕/。/女御〔名詞〕/更衣〔名詞〕/あまた〔副詞〕/さふらひ〔動詞〕/給〔助動〕/ける〔助動〕/なか〔名詞〕/に〔助詞〕/いと〔副詞〕/やむことなき〔形容〕/きは〔名詞〕/に〔助詞〕/は〔助詞〕/あら〔動詞〕/ぬ〔助動〕/か〔助詞〕/すくれ〔動詞〕/て〔助詞〕/時めき〔動詞〕/給〔助動〕/あり〔動詞〕/けり〔助動〕/。/はしめ〔名詞〕/より〔助詞〕/我〔代名〕/は〔助詞〕/と〔助詞〕/思あかり〔動詞〕/給へ〔助動〕/る〔助動〕/御方〻〔名詞〕/めさましき〔形容〕/もの〔名詞〕/に〔助詞〕/おとしめ〔動詞〕/そねみ〔動詞〕/給〔助動〕/。/おなし〔形容〕/ほと〔名詞〕/それ〔代名〕/より〔助詞〕/下らう〔名詞〕/の〔助詞〕/更衣たち〔名詞〕/は〔助詞〕/まして〔副詞〕/やすからす〔連語〕/。/あさゆふ〔名詞〕/の〔助詞〕/宮つかへ〔名詞〕/に〔助詞〕/つけ〔動詞〕/て〔助詞〕/も〔助詞〕/人〔名詞〕/の〔助詞〕/心〔名詞〕/を〔助詞〕/のみ〔助詞〕/うこかし〔動詞〕/うらみ〔名詞〕/を〔助詞〕/おふ〔動詞〕/つもり〔名詞〕/に〔助動〕/や〔助詞〕/あり〔動詞〕/けむ〔助動〕/いと〔副詞〕/あつしく〔形容〕/なりゆき〔動詞〕/もの心ほそけに〔形動〕/さとかちなる〔形動〕/を〔助詞〕/いよ〻〻〔副詞〕/あか〔動詞〕/す〔助動〕/あはれなる〔形動〕/物〔名詞〕/に〔助詞〕/おもほし〔動詞〕/て〔助詞〕/人〔名詞〕/の〔助詞〕/そしり

図 15

いつれ〔代名〕/の〔助詞〕/御時〔名詞〕/に〔助詞〕〔助動〕/か〔助詞〕/。/女御更衣〔名詞〕/あまた〔副詞〕/さふらひ〔動詞〕〔助動〕/給〔動敬〕/ける〔助動〕/なか〔名詞〕/に〔助詞〕〔助動〕/いと〔副詞〕/やむことなき〔形容〕/きは〔名詞〕/に〔助詞〕〔助動〕/は〔助詞〕/あら〔動詞〕/ぬ〔助動〕〔動詞〕/か〔助詞〕/すくれ〔動詞〕/て〔助詞〕〔助動〕〔名詞〕/時めき〔動詞〕/給〔動敬〕/あり〔動詞〕/けり〔助動〕/。/

図 16

[助動] [名詞] [動詞] とついてくるが, [名詞] [動詞] の用例は [助詞] [助動] に比して非常に少ないので最初から削っておく. ただし多品詞語はプリントするときに大文字化し, 見逃しのないようにする (図 16).

5 検索作業の容易化

検索作業を容易にするために, 品詞情報つきデータに『源氏物語大成』と同じページと行番号をつけた. この作業もプログラムを作り, 単語分割する前の『大成』の行構成と同じ行構成のテキストの行末 5 文字と, 品詞情報つきテキストを対応させて自動的に改行して, ページと行番号をつけた (図 17).

6 今後のデータベースの利用

今回,『源氏物語大成』の品詞情報つきフルテキストデータベースを完成したことによって得られる成果は, 計り知れない. 宇治十帖他作家説や複数作家説, 成立過程に関する諸説や物語音読論等々の詳細な検討が文法的側面からも, 使用単語の面からも行える.「源氏物語」の文体を構成する諸々の要素について, 一つ一つ検証してゆくことができる.

たとえば各巻毎の品詞の出現率が得られるので, 比較検討できる (図 18). また, ある品詞のなかで, どういう単語が多いのか少ないのかがわかる. ここでは桐壺の巻の助詞に, どのような語があるかを示した (図 19). こうして作成したデータを用いて主要 7 品詞 (名詞, 動詞, 形容動詞, 助詞, 助動詞, 形容詞, 副詞) の出現率による主成分分析を行ったところ, 初期の文体と宇治十帖の文体が少し異なっていることがわかった (図 20). ただし, 3000 語以下の巻は, 分析からはずしてある.

このようなデータベースを作ったことによって, 54 帖すべてに関しての単語や品詞の情報が, コンピュータによって敏速に取り出せ, 巻毎の数量的な比較検討が容易になり, 視覚的にわかりやすく

0005-01
いつれ〔代名〕/の〔助詞〕/御時〔名詞〕/に〔助詞〕/か〔助詞〕/。/女御〔名詞〕/更衣〔名詞〕/あまた〔副詞〕/さふらひ〔動詞〕/給〔動敬〕/ける〔助動〕/なか〔名詞〕/に〔助詞〕/いと〔副詞〕/やむことなき〔形容〕/きは
0005-02
〔名詞〕/に〔助詞〕/は〔助詞〕/あら〔動詞〕/ぬ〔助動〕/か〔助詞〕/すくれ〔動詞〕/て〔助詞〕/時めき〔動詞〕/給〔動敬〕/あり〔動詞〕/けり〔助動〕/。/はしめ〔名詞〕/より〔助詞〕/我〔代名〕/は〔助詞〕/と〔助詞〕/思あかり〔動詞〕/給へ〔動敬〕/る〔助動〕/御方
0005-03
〔名詞〕/めさましき〔形容〕/もの〔名詞〕/に〔助詞〕/おとしめ〔動詞〕/そねみ〔動詞〕/給〔動敬〕/。/おなし〔形容〕/ほと〔名詞〕/それ〔代名〕/より〔助詞〕/下らう〔名詞〕/の〔助詞〕/更衣たち〔名詞〕
0005-04
/は〔助詞〕/まして〔副詞〕/やすからす〔連語〕/。/あさゆふ〔名詞〕/の〔助詞〕/宮つかへ〔名詞〕/に〔助詞〕/つけ〔動詞〕/て〔助詞〕/も〔助詞〕/人〔名詞〕/の〔助詞〕/心〔名詞〕/を〔助詞〕/のみ〔助詞〕/うこかし〔動詞〕/うら

図 17

品詞	(%)	0 5 10 15 20 25 30
名詞	22.162	----------
動詞	17.166	----------
形容詞	1.691	---
助動詞	30.696	----------
接頭語	0.000	--
接尾語	0.019	--
形容動詞	5.618	--------
感動詞	0.019	--
助動詞	12.578	----------
連体詞	1.205	---
副詞	4.238	------
接続詞	0.156	--
連語	0.019	--
名詞-副詞	0.000	--
動詞-動詞	4.238	------
補助動詞	0.194	--
代名詞	0.000	--
枕詞	0.000	--
複合動詞	0.000	--
複合形容	0.000	--

図 18 「桐壺」

提供されるようになったことである．更に従来の説の計量的な検証を行うなかで，新たな分析方法も次々に工夫，開発してゆくことができる．

現在『源氏物語大成』と同様の手順で「紫式部日記」(日本古典文学大系)，本居宣長自筆本「手枕」(本居宣長全集第 15 巻，筑摩書房)の品詞付けが終わり，「山路の露」(日本古典全書第 7 巻所収)，「雲隠六帖」(『源氏物語の研究』巻末付録)の単語分割が終わっている．これらのデータベースを使って，何種類かの計量分析もすでに行われており，興味深い結果も提出されている．今後，品詞情報つきデータベースが増えることによって，各文献の比較等はいうまでもなく，日本語のより精緻な分析が，可能となってゆくであろう．

(1993 年 11 月 2 日受付)
(1994 年 9 月 2 日採録)

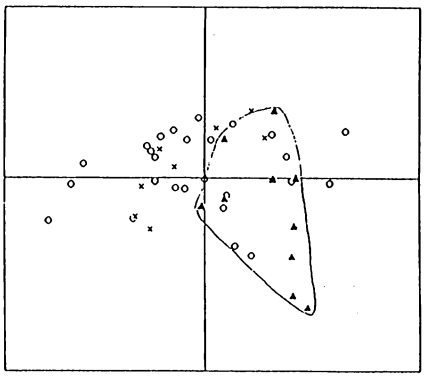
・・・・ 助　　詞 ・・・・

番号	かな漢字	(%)	0 10 20
1		1.013	--
2	し	0.063	--
3	なそへ	0.317	--
4		0.760	--
5		0.253	--
6		0.697	--
7		1.203	--
8	に、	0.760	--
9		0.760	--
10		10.766	------
11		9.246	------
12	もから	0.633	--
13		0.443	--
14		2.470	---
15		0.697	--
16		1.330	---
17		15.643	---------
18	て	2.027	---
19		16.529	----------
20	み	0.950	--
21		10.576	------
22	り	0.127	--
23		0.127	--
24		0.887	--
25		10.260	------
26		0.950	--
27	り	1.267	---
28		8.930	------
29		0.127	--

図19 「桐壺」

主要7品詞による分析

横軸は第1主成分
縦軸は第2主成分



各記号の意味は以下のようになっています.

○:第一部
×:第二部
▲:宇治十帖

図20

著者紹介

上田英代(正会員)
昭和47年東京教育大学国語国文学科卒業，同年，都立大崎高校勤務，昭和48年同校退職，平成4年より文部省統計数理研究所にて外来研究員として「源氏物語」の計量文献学的研究を共同で行なっている，著書「パトグラフィー紫式部-解読『源氏物語』」，「源氏物語語彙用例総合索引-自立語編-」

上田裕一
昭和41年東京大学医学部医学科卒業，昭和51年獨協医科大脳神経外科助教授，昭和53年，医博，(頭部CTscanによる頭蓋内疾患の自動定量診断)，昭和58年文部省統計数理研究所特別研究員，昭和62年琉球大学工学部電子情報工学科講師，昭和63年もとぶ野毛病院院長，昭和63年沖縄県医師会医療情報システム委員，平成2年沖縄県医師会医療情報システム担当理事，著書「パトグラフィー紫式部-解読『源氏物語』」，「源氏物語語彙用例総合索引-自立語編-」

村上征勝(正会員)
昭和43年北海道大学工学部精密工学科卒業，昭和48年スタンフォード大学大学院修士課程修了，昭和49年北海道大学大学院博士課程修了，工博，同年，統計数理研究所入所，現在，同研究所領域統計研究系人文社会科学研究部門及び総合研究大学院大学数物科学研究科教授，計量文献学，計量考古学，社会調査などを主として人文社会科学の分野への統計理論の応用に関する研究に従事，著書「工業統計学」，「真贋の科学」，「源氏物語語彙用例総合索引-自立語編-」など