古文の自動単位句切り

西端幸雄

最近、様々な作品の語彙索引が刊行されている。そのうちの一部は、大型のコンピューターを使って、その作成作業を行っているが、多くの場合、編集者の手作業によるものと思われる。また、大型コンピューターを用いた作業のシステムの詳細が公開されていないので、［注①］軽率な判断を下すことはできないが、大型コンピューターを用いた作業であっても、まだ、かなりの部分を人間によって補っている点があるのではないだろうか。ましてや、人間の手作業によって、語彙索引を作るとなると、多大の労力と時間を要することになる。そして、人間が関われば関わるほど、その作業過程で生じる誤りの数が多くなる。例えば、語句の単位認定の基準の揺れや見出し語に当てる漢字表記の不統一など、いろいろな問題点が挙げられる。いま、筆者の手元にある語彙索引のうち、同一の編集者の手になるものを見ても、それぞれに単位認定の揺れや仮名遣いの揺れが散見できる。こうした問題点が残存していては、索引を使う側にとって、使いにくいし、また、それを作った側にとっては、多大の労力と時間をかけたにも関わらず、その努力が十分に報われないことになる。

そこで、小稿では、語彙索引を作成する作業過程の中でも、人間が関われば、誤りや不統一が生じやすい語句の単位認定、見出し語の仮名遣いやそれに当てる漢字表記の統一といった作業をパーソナルコンピューター（以下、パソコンと略す）によって行い、語句索引を作成する作業をできる限り自動化しようという試みを報告する。

1．使用機器

今回の作業に用いたパソコン、及び周辺機器は、以下の通りである。

本体	ＰＣ９８０１ＶＸ２（ＮＥＣ）
ＣＲＴ	ＰＣ－ＫＤ８５４（ＮＥＣ）
ハードディスク	ＨＤＤ－２０（ＥＰＳＯＮ）
プリンター	ＶＰ－１３０Ｋ（ＥＰＳＯＮ）
ＲＡＭディスク	ＰＩＯ－９２３４－２ＭＤ（Ｉ・Ｏデータ機器）

上記の機器のうち、ハードディスクは、語彙索引作成という大量のデータを処理するのに絶対に必要なものである。今回用いたハードディスクは、２０ＭＢ（全角文字1000万文字分）のものであるが、将来性を考えれば、４０ＭＢのものの方が望ましいと思われる。また、ＲＡＭディスクは、後述する作業辞書用として使用するため、２ＭＢのものを用いたが、これも、作業辞書の内容が増加することを考えれば、４ＭＢのものの方が望ましいと思われる。

なお、本作業における各ディスクのドライブ割当とそれぞれの用途は、下記の通りである。

ドライブＡ（本体内蔵ドライブ１）
カレントドライブとして、システム関係と作業プログラムが入ったフロッピーを常駐

ドライブＢ（本体内蔵ドライブ２）
立ち上げ時は、作業辞書をＲＡＭディスクに転送。
作業時は、和歌、詞書、人名などのデータの入った
フロッピーを常駐
ドライブＣ（ハードディスク）
作業によって作成された索引用データを登録
ドライブＤ（ＲＡＭディスク）
作業中、作業辞書を常駐

２．作業辞書

以前、本誌23号（昭和61年1月刊）において、『マイクロ・コンピューターによる自動品詞認定の試み』という小論を発表したことがある。その時は、品詞認定を行うのに、主として、プログラム内で語形や構文によって判断を下し、それだけでは判断が下せない語句について、BASIC言語（FBASIC）のDATA文に辞書データとして若干の語句を掲げ、それらを配列変数に読み込み、SEARCH関数で検索させるようにした。

しかし、大量のデータを処理し、一つ一つのデータに、後述するような様々な情報を付加させようとしたとき、DATA文内のデータだけでは、不足してしまう。また、より多くのデータや情報をそのDATA文に掲げようとすると、非常に長いプログラムになり、実行速度が落ちるだけではなく、そのプログラムのサイズがメモリー内のプログラム領域を越えてしまう恐れもある。さらに、もう一つの問題点は、今回用いた機種、PC9801に使われているN88日本語BASICのSEARCH関数の受け付ける変数が整数型だけであり、文字型のものは受け付けないという点である。以上の理由で、辞書データをすべてフロッピー・ディスクに登録し、それを作業辞書として用いることにした。

辞書データをディスクに登録して用いる利点は、データ量をディスクの管理できる限界まで増やすことができるという点と、前述のＲＡＭディスクを用いれば、非常に高速にデータを検索することができるという点と、さらにその辞書データを他の作業に容易に転用することができる点などが挙げられる。

では、作業辞書のレコード・フォーマットを示すと、下記の通りである。

語句	見出語	漢字	文法情報

ファイル形式は、ランダムファイル（256ﾊﾞｲﾄ）とし、フロッピー・ディスクの使用効率を高めるため、「語句」の文字数１～１４文字（2ﾊﾞｲﾄ～28ﾊﾞｲﾄ）のデータ長別に１４のファイル（DIC1～DIC14）にそれぞれ登録するようにした。登録データの合計数は、現在のところ、約35000語である。

このうち、「語句」とは、索引を作成する原文と語形の一致を判定し、単位句切りを行う基準となるものである。また、単位句切りを行った後、句切られた各単位語句との一致を判定し、各単位語句に、以下に説明する見出し語、漢字、文法情報を付ける基準ともなるものである。

現在、進行している作業が、八代集の各語彙索引を作るということであるので、この「語句」データは、小学館刊『古語大辞典』（中田祝夫編監修）所収の語句を中心とし、その他に八代集の本文中に実際に用いられている語形、また用いられているだろうと推定した語形、さらには、仮名遣いの間違いを犯す可能性のある語句をも間違った語形のまま、全角文字で登録した。なお、用言については、

各活用形で同形を示すものは、一方の活用形にまとめた。例えば、４段活用動詞の場合、終止形と連体形、已然形と命令形が同じ語形を示すので、それぞれ終止形と已然形として登録し、実際の活用形は、後述するように、プログラム内で判定した。

ところで、各辞書データの登録を行うために、辞書作成用のプログラムをBASIC言語で作り、特に、用言については、各活用形をいちいち登録するのではなく、終止形で入力すれば、プログラム内で各活用形が自動的に生成され、登録されるようにし、辞書作成にかかる時間を短縮するように努めた。［注②］

「見出語」は、索引としてできあがった全データを五十音順に並べ替え、同形同義の語句毎に整理するための基準となる、「語句」に対応する読み（歴史的仮名遣い）を情報として与え、索引上で見出しとして掲げられるものである。そのため、当然、用言については、すべて終止形で登録した。また、原文で仮名遣いの間違いがある場合も、この「見出語」データで正すことができるようにした。［注③］

この「見出語」データを全角文字で登録すると、ディスクの使用効率が悪くなるので、すべて半角文字（カタカナ）として登録し、実際の作業で用いる場合には、全角の平仮名に直して用いるようにした。［注④］

「漢字」は、「見出語」を基準として五十音順に並べ替えられたデータを、さらに意義によって下位分類する基準となるものである。この「漢字」情報は、既刊の索引においては、不統一な点が目立っている。そこで、今回の作業では、この作業辞書に登録する「漢字」情報を、小学館刊『古語大辞典』（中田祝夫編監修）によることとし、統一化を図った。

なお、助詞や助動詞については、漢字表記することがないので、このデータ部が空きになる。そこで、助詞や助動詞のデータの場合、この「漢字」データ部に接続情報を入れてある。そして、そのデータを作業プログラム内で読み取ることによって、上接の語句の品詞や活用形を判断できるようにした。

「文法情報」は、各レコードの先頭にある「語句」に対応する品詞・活用形を半角文字２ケタ（2ﾊﾞｲﾄ）のコード番号化したものである。［注⑤］

以上のようなレコード・フォーマットで作業辞書はできているのであるが、前述のように「語句」データが１～１４文字からなるものを合計約35000語も保存するためには、１ＭＢのフロッピー・ディスク１枚には収納できないので、現在のところ、２枚のフロッピー・ディスクに分散して収納している。

３．自動単位句切り

索引作成作業を人間が行う場合に一番揺れが生じやすいのが単位句切りの作業においてであろう。ここで言う揺れとは、単位認定に対する考え方の違いではなく、同一索引内、または同一編集者の手になる索引間で、例えば、同じ語構成の語句に対して、異なった単位認定を行っていることを言う。こうした単位認定の揺れを避けるためには、その認定基準を一覧表なりカード化して、手元に置いて、認定作業を行えばよいとも言えようが、その労力は相当なものが予想できるし、また、それでも不統一が生じる恐れはあろう。その点、前述の作業辞書を用いて、認定作業を行えば、認定基準の画一性という問題は残るが、統一性を持った基準で単位句切りができる。

今回の作業で用いた単位認定方法は、最近の日本語フロントエンドプロセッサが用いている後部一致法ではなく、基本的には前部一致法によることとした。さらに、品詞・活用形の決定については、後接語句の文法的性格、および助詞や助

動詞の接続情報によることとした。

では、その単位認定方法について、具体的に説明する。［後掲のフローチャートを参照］

単位認定の基準は、基本的には作業辞書に収められた「語句」データによって定めているので、下記のように、まず、読み込んだ本文データより取り出すデータ長を、作業辞書の「語句」データの最長文字数１４文字分から最短文字数１文字分まで１文字ずつ減らしながら、作業辞書との一致を判定する。

はるたつといふ許にや三吉野の山もかすみてけさは見ゆらむ（拾遺集　１）

14文字　はるたつといふ許にや三吉野の
13文字　はるたつといふ許にや三吉野
12文字　はるたつといふ許にや三吉
（中略）
４文字　はるたつ
３文字　はるた
２文字　はる　←－－作業辞書の「はる（春）」と一致

一致しなければ、データを取り出す起点を次の文字として、例えば、上の例の場合、もし、「はる」でも「は」でも一致しなければ、「は」の次の「る」を起点として、上記の作業を繰り返す。なお、現在の作業用プログラムや作業辞書において、語頭にラ行音が立つのは、助動詞である可能性が高いとか、語頭に「ば」が立つのは、助詞である可能性が高いという様な細かい配慮は行っていない。その点については、今後検討の余地がある。

一方、一致した場合は、次のデータを取り出すため、一致したデータの起点から（一致したデータ長＋１）のところへ起点を移して、例えば、上の例の場合、「はる」で一致したので、次のデータを取り出す起点を「た」に移して、上記の作業を繰り返す。なお、実際の作業においては、句切り点を視覚的にも確認するため、下記のように、本文データの各文字間に数字を入れて、一致した箇所の前後には句切り符号「｜」を表示し、さらに、その句切り符号の付けられた番号を整数型の配列変数にそれぞれ格納した。そして、以後、一致した場合、重複した番号が格納されないように、SEARCH関数で検索して、配列変数内に存在しない番号だけ格納するようにした。

1　2　3　4　5　6　7　8　9　10　11　12　13　14　15　16　17
は　る｜た　つ｜と｜い　ふ｜許　に　や｜三　吉　野　の　山｜も｜か

18　19　20　21　22　23　24　25　26　27
す　み｜て｜け　さ｜は｜見　ゆ｜ら　む｜

上記のように、正しく句切れれば、各単位を出現順に文字型配列変数に格納し、次項で述べる各種の情報を付加する便を図った。

なお、単純に文頭から順次作業辞書との一致を判定していくと、特に助詞が取り出したデータの頭にきたときに誤った判定結果になることがある。例えば、上記の例では、句切り番号１５以下の「もか」という語形が助詞の複合形として作業辞書に存在すると、句切り番号１５と１７で句切られてしまう。こうした問題を排除するために、助詞である可能性の高い「て・に・の・は・も・や・を」が、

取り出したデータの頭にきたときは、次の文字に強制的に起点を移すようにした。そして、次の文字を起点にして、判定した結果、作業辞書と一致しなかった場合、起点を元に戻し、同じように判定を行うようにした。

ただ、最近の日本語フロントエンドプロセッサに備わっている逐次自動変換方式においても、あまり高い正答率が得られていないように、今回の作業辞書を用いての単位認定にも、限界がある。現在のところ、正答率は、約90パーセントである。そのため、パソコンによっての一応の単位認定の後、人間の手作業による修正が必要になる場合もある。その修正作業には、余分な句切りを削除するのと、逆に句切りを追加するという二面がある。今回用いた作業用プログラムでは、まず、余分な句切りがあった場合、ディスプレー上に表示されているデータを見ながら、人間がその句切り番号を指示すると、プログラム上でその句切り番号を配列変数内から削除するようにした。また、追加すべき句切り箇所がある場合、同じく人間がその句切り番号を指示すると、プログラム内でその句切り番号を配列変数内に追加し、さらに句切り番号を昇順に並べ替えさせるようにした。

4．各種情報付加

単位認定の過程が終了すると、次に、同じ作業辞書を用いて、（2．作業辞書）の項で述べた「見出語」「漢字」「文法」の各情報を、および助詞・助動詞については、それぞれの「接続」情報をも、単位認定したデータ毎に付加する作業に移る。

この作業の第一段階は、単位句切りによって得られた各データの長さに対応する作業辞書を検索して、そのデータが作業辞書に存在すれば、作業辞書内に収めてある各情報をデータの順序と同じように文字型配列変数に収めるといったところまでを行う。

ただ、作業辞書から得られた各情報、特に「文法」情報については、（2．作業辞書）の項で述べたように、用言の語形が同じ活用形の場合、一方の活用形でしか作業辞書に登録していない。そのため、助詞・助動詞が後接しているデータについては、助詞・助動詞データの「接続」情報によって、プログラム上で、品詞や活用形の補正を行うようにした。例えば、4段動詞の已然形と命令形は、同形なので、作業辞書内には、已然形だけが登録してある。そのため、4段動詞の後に完了の助動詞「り」が後接している場合、補正を加えなければ、［4段動詞の已然形＋完了の助動詞「り」］となってしまう。こうした不都合が生じるのを避けるため、上記のような補正が必要になるのである。

さらに、逆に前接する語句によって、後接する品詞や助詞・助動詞の別を判定することも、プログラム上で行うようにした。これは、例えば、完了の助動詞「ぬ」の連用形と格助詞「に」とは、同形なので、前接のデータが体言の場合、格助詞と認定するといった程度の補正を加えるようにした。ただ、この「接続」情報による補正作業をプログラム上で完璧に行おうとすると、プログラムが長大なものになってしまう。そのため、かなり大雑把な補正にとどめざるを得ないのが現状である。

作業の第二段階が、人間の手作業による修正作業である。まず、単位句切りによって得られたデータが作業辞書内に存在しない場合、「見出語」「漢字」「文法」の各情報は、人間の手によって、補わざるを得ない。また、すべてのデータに各情報が付加できたとしても、現在のところ、作業辞書を用いての各種情報付加の段階での正答率が、単位句切りとほぼ同率の約90パーセントであるため、上に述べたような各情報、特に「文法」情報の修正は、どうしても行わざるを得な

い。また、作業辞書内には、同音異義語については、一方のデータしか登録していないので、少なくとも「漢字」情報の修正も必要になることがある。こうした修正作業を正確に行うため、各種の情報を付加した各データの正誤判断を、下記のように表としてディスプレー上に示し、視覚的に正誤判断が下せるようにした。なお、修正作業を、下記の表中で行えるように、各データが表示されている座標をもとにして、カーソル移動キーで修正箇所にカーソルを移し、その箇所において修正できるようにした。そうした配慮をしたのは、例えば、「文法」情報の修正を行う場合、前後の品詞、特に後接語の品詞が大きく関わる。そのため視覚的に前後関係を判断しやすくするためには、表中での修正作業が最適であろうと考えたからである。

表①

	語　句	見　出　語	漢　字	文法
1	はる	はる	春	００
2	たつ	たつ	立	１３
3	と	と		０６
4	いふ	いふ	言	１４
5	許にや	ばかりにや		０６
6	三吉野の山	みよしののやま	三吉野山	００

（以下、略す）

作業の第三段階は、掛詞の有無の判定である。現在のところ、掛詞を自動的に判別するプログラムやデータは作っていないため、この段階の作業は、完全に人間の手作業による。

まず、掛詞の有無をディスプレー上で問いかけてくる。それに対して、「ＹＥＳ（有）」と答えると、次にどの語句が掛詞になっているのかを問いかけてくるので、それに表の左の番号で答えると、その後、作業辞書を検索して、とりあえず、その番号の語句の各種データを表の下に表示する。そこで、その各データを、掛詞として対になる語句の形式に修正して、その修正が正確であれば、表中の各語句データを収める配列変数の後に加えるようにした。

この掛詞を処理するための辞書やプログラムを作成することは、基本的にはさほどの困難はないと思われる。ただ、今日、掛詞についての総合的、体系的研究が十分になされていないため、辞書データとして、掛詞を収集するのに大変な労力が必要になる。そのため、この点については、今後の課題として留保しておく。

作業の第四段階は、複合語の処理である。この作業については、当初、作業辞書内の複合語データに各語構成要素をも入れ、そのデータによって、判別する予定でいたが、作業辞書が膨大なものになる恐れがあったため、現在のところ、人間の手作業によって、各語構成要素を句切るようにしている。

複合語の処理をするため、下記のように、各用言や助動詞を終止形に変換した、

単位句切り済みの平仮名書き本文データをディスプレーに表示させ、人間の判断によって、複合語の各語構成要素間にある番号を指示するようにした。その際、自動単位句切りの場合と同じく、句切り位置の番号を指示すれば、その番号に該当する文字間に句切り符号「＝」を表示させ、その句切り位置を視覚的にも確認できるようにした。（下記の例の「ばかり=に=や」「み=よしの=の=やま」参照）

```
 1  2  3  4  5  6  7  8  9  10 11 12 13 14 15 16 17
 は  る｜た  つ｜と｜い  ふ｜ば  か  り=  に=  や｜み=  よ  し  の=  の=

 18 19 20 21 22 23 24 25 26 27 28 29 30 31
 や  ま｜も｜か  す  む｜て｜け  さ｜は｜見  ゆ｜ら  む｜
```

以上のように、複合語の各語構成要素間に区切り符号がすべて付けられたら、その後、その複合語の各語構成要素を一語ずつ表示し、それを索引中に採用するかどうかの判定を人間の手によって行うようにした。そして、採用するとした語構成要素だけ、本節の冒頭で述べた各種情報付加の手順と同じように、作業辞書を検索し、そのデータが存在すれば、各情報を表①と同形式の表として表示し、修正が必要な箇所には修正を加えるようにした。

5．索引用データ

前項で述べた各種情報付加が終われば、そこで得られたデータを索引用データとして、ハードディスクに登録する。

そのレコード・フォーマットは、以下の通りである。

番号	行数	語句	見出し	漢字	参照	文法	掛詞	所在	索引種	作品

ファイル形式は、ランダムファイル（256ﾊﾞｲﾄ）の１レコード１９５バイトに上記の各種データを登録するようにした。以下に示すデータ長からも分かるように、「語句」から「参照」までの４データの長さに余裕をもたせているのは、散文や人名を処理することを考慮するとともに、一方、このデータをそのまま大型コンピューターに引き渡し、それを元に、並べ替え・印刷といった作業を大型コンピューター側で行うにあたり、読み取りやすさを考慮したためである。

「番号」とは、本文の和歌に付けられている通し番号（基本的には、新編国歌大観番号）を半角４ケタで示したものである。（データ長は、4ﾊﾞｲﾄ。以下、同じ）

「行数」とは、長歌や詞書の複数行にわたるものの中での語句の所在を示すためのものである。前述の「番号」とこの「行数」は、本文データに、すでに付けられているものを読み取り、そのまま用いた。（3ﾊﾞｲﾄ）

「語句」とは、本文データを単位句切りしたデータそのものであり、索引中には直接必要ないものであるが、索引用データのチェックを行う場合の判断基準になり、また、索引用データをデータベース化する際に何等かの有用性があるとも考えられるので、あえてデータとして取り入れた。（40ﾊﾞｲﾄ）

「見出し」とは、索引データを五十音順に並べ替える基準となり、索引中では見出語として掲げられるものである。前項で述べた各種情報付加の作業中で、単位データに対して付けられた「見出語」データや複合語を各語構成要素毎に句切ったそれぞれの構成要素データがここに収められる。そのため、当然、これは、表記としては、全角の平仮名からなり、用言や助動詞については、すべて終止形

で統一されている。（50ﾊﾞｲﾄ）

「漢字」とは、「見出し」を基準に五十音順に並べ替えられたデータを、さらに意義別に下位分類する基準となるもので、前項で述べた各種情報付加の作業中で、単位データに対して付けられた「漢字」データがここに収められる。（40ﾊﾞｲﾄ）

「参照」とは、前項で述べた、複合語を語構成要素毎に句切った結果、得られた各語構成要素を索引中の見出しに掲げる場合、その構成要素を含む複合語を示し、検索の便を図るためのものである。よって、索引中で、和歌番号や行数を掲げるデータは、この「参照」データの箇所を空白にし、逆に、索引中で、和歌番号を掲げずに、参照項目だけ掲げるデータは、「語句」データの箇所を空白にすることによって、両者を識別できるようにした。（50ﾊﾞｲﾄ）

「文法」とは、「見出し」「漢字」で並べ替えられたデータを、さらに品詞や活用形別に下位分類する基準となるもので、「語句」や、「見出し」に収められた複合語の語構成要素の品詞や活用形を２ケタのコード番号で示したものである。（2ﾊﾞｲﾄ）［注⑤］

「掛詞」とは、掛詞となっている語句の内、主たる意味を表すデータの方は、ここを空白とし、副次的な意味を表す語句の場合、それを識別し、索引中で、和歌番号や行数に傍線を付ける手がかりとするためのもので、半角の記号「＊」をここに収めた。（1ﾊﾞｲﾄ）

「所在」とは、索引中には、直接必要のないデータではあるが、当該の「語句」データが１行中の何文字目に存在するかを示すためのものである。今回の作業では、得られた索引用データは、ＫＷＩＣ形式をとらないが、この「所在」データと本文データを用いることによって、ＫＷＩＣ形式に容易に移行できるものと思われる。（2ﾊﾞｲﾄ）

「索引種」とは、公刊予定の索引中で、和歌、詞書・左注・序文、作者名それぞれを別々に掲げる予定であるので、その３種の索引の別を識別するためのものである。和歌については「Ｗ」を、詞書・左注・序文については「Ｋ」を、作者名については「Ｓ」を、それぞれ半角で収めた。（1ﾊﾞｲﾄ）

「作品」とは、『古今和歌集』から『新古今和歌集』までの八代集の成立順に１から８までの数字をあて、それぞれの作品を識別できるようにしたものである。ただ、将来、二十一代集の索引を作成することを考慮し、半角２ケタの数字（01～08）で表した。（2ﾊﾞｲﾄ）

前述のように、これらのデータをランダムファイル（256ﾊﾞｲﾄ）に１レコード195バイトとして、収めるのであるが、いま、２０ＭＢのハードディスクを用いていると、（２０ＭＢ÷２５６Ｂ）で約８００００データが登録できることになる。しかし、将来、和歌数の多い『新古今和歌集』を処理したり、蓄積されるデータが増加してくると、２０ＭＢの容量でも不足してくるという問題が残る。

６．問題点

以上、述べてきたところが、今回試みた自動単位句切りの作業であるが、まだ、いくつかの問題点が残存しているのも事実である。その内のもっとも大きな点は、作業辞書の大きさであろう。単位句切りの精度を高めようとすればするほど、この作業辞書の量を増加させていかなければならない。しかし、現在、１４ファイルに約３５０００語のデータが登録してあるが、これが占める容量は、計１３７２４１６バイトにもなっていることからすると、これ以上のデータの増加は、ディスクの容量を越える恐れもあり、また、処理速度を低下させる一因にもなる。

その一方で、作業辞書のデータを増加させることによって、かえって単位句切りの精度を低下させるという二律背反的な現象も見られる。というのは、本作業においては、前述のように、本文データを１４文字から１文字へと順に取り出して、作業辞書との一致を判別して、句切り点を付けようとしているため、例えば、本文データとして、「てかくばかりに（手書くばかりに）」とあって、作業辞書中に「かくばかり」が登録されていると、別に短い単位で「かく」が登録されていても、単位句切りとしては、長い単位の「かくばかり」が優先されて、「て＋かくばかり＋に」となってしまう。その意味では、今後は、作業辞書をあまり増加させることなく、プログラム上での、古文解析の方法に改良を加えることによって、単位句切りの精度を高めるようにすることが必要となろう。

問題点の二つ目は、表題において、「自動単位句切り」としたものの、まだ、人間の関わる作業が多すぎるという点である。特に、複合語の語構成要素毎に句切る作業においては、複合語自体の判別と語構成要素の句切り点の判別、さらには、それをデータとして採用するかどうかといったところまで、人間の判断によっているということは、その判断にかなりの揺れが生じる恐れもある。そのため、現在、作業辞書から複合語を抽出することによって、複合語辞書を作成し、少なくとも、複合語の判別とその語構成要素の句切り点を付けられるようにする計画である。

その他にも、問題点は、山積しているが、今後、テストを繰り返しながら、一つひとつの問題点を少しでも解決していきたい。

７．おわりに

以上、自動とは名ばかりの作業を報告してきたが、今回の報告の中心は、あくまで、「単位句切り」にある。その作業過程で、作業辞書を用いて、約90パーセントの正答率が得られたという点と、各作業過程の内、パソコンに依存する「単位句切り」「各種情報付加」における処理時間が、単位語句データの量にもよるが、それぞれ平均１０秒以内である点、さらに、全作業に要する時間も、修正・補正を必要とするデータの量にもよるが、１本文データにあたり、平均３分以内である点からすると、一応の使用に耐えうるのではないかと思う。ただ、前述のように、まだ様々の問題点もあるので、今後は、できる限り作業辞書を増加させることなく、プログラム上での、古文解析の方法を改良するつもりでいる。

なお、本文中では触れなかったが、本作業に用いた本文データは、『新註八代集』（和泉書院刊）用に電算写植機によって入力されたものをパソコン用に変換して用いた。また、作業によって得られた索引用データの並べ替え・印刷は、株式会社コトブキ企画において、大型コンピューターにより処理される。

最後に、本システムを公開する用意がある。予定では、作業用プログラム（N88日本語BASIC）・作業用辞書計３枚のディスケット（５インチ２HD）で提供する。ただし、機種としては、PC9801のみとし、それにＲＡＭディスクを装備されていることが望ましい。ご希望の方は、本学国文学科西端宛に、代金３０００円を同封の上、お申し込みください。

注①　『「ぎやどぺかどる」の読解に於ける電子計算機の利用の試み』（1983年3月・科学研究費研究報告書・代表者　風間喜代三）
『連歌資料のコンピュータ処理の研究』（昭和60年3月・国文学研究資料館共同研究報告３・明治書院刊）

②　このプログラムは、『人文科学データベース情報（仮称）』（昭和63年10

月刊行予定）に掲載し、紹介する予定である。

③　例えば、「おと（音）」を、本文において、「をと」と表記されていても、「語句」データとして「をと」、「見出語」データとして「おと」と登録しておけば、正しい仮名遣いが得られる。また、「あしひきの」か「あしびきの」かの濁点の有無の問題も、「語句」データとして「あしびきの」、「見出語」データとして「あしひきの」と登録しておけば、「あしひきの」に統一することができる。

④　拙稿「文字処理のためのサブルーチン」（『樟蔭国文学』第24号　昭和62年3月）に掲載のプログラム２参照

⑤　本作業で用いた、各品詞や活用形のコード番号は、以下の通りである。

（品詞）	（コード番号）
名詞	００
副詞	０１
連体詞	０２
接続詞	０３
感動詞	０４
接頭語	０５
助詞	０６
接尾語	０７
枕詞	０８
連語	０９

	未然	連用	終止	連体	已然	命令
４段動詞	１１	１２	１３	１４	１５	１６
上１段	２１	２２	２３	２４	２５	２６
上２段	３１	３２	３３	３４	３５	３６
下１段	４１	４２	４３	４４	４５	４６
下２段	５１	５２	５３	５４	５５	５６
カ変	カ１	カ２	カ３	カ４	カ５	カ６
サ変	サ１	サ２	サ３	サ４	サ５	サ６
ナ変	ナ１	ナ２	ナ３	ナ４	ナ５	ナ６
ラ変	ラ１	ラ２	ラ３	ラ４	ラ５	ラ６
補助動詞	ホ１	ホ２	ホ３	ホ４	ホ５	ホ６
ク活用形容詞	ク１	ク２	ク３	ク４	ク５	ク６
シク活用	シ１	シ２	シ３	シ４	シ５	シ６
ナリ活用形容動詞	Ｎ１	Ｎ２	Ｎ３	Ｎ４	Ｎ５	Ｎ６
タリ活用	Ｔ１	Ｔ２	Ｔ３	Ｔ４	Ｔ５	Ｔ６
助動詞	９１	９２	９３	９４	９５	９６

フローチャート