日本語文章における自動索引の試み

——保田與重郎における専門用語の考察をふまえて——

谷口敏夫

はじめに

先に「日本語文章からの知識の抽出」[1]を著したとき，「は」や「が」を含む文章の中で，名詞を中心とした知識の核となる用語を単に「重要語」とすませていたことに，あとで気づいた。言葉を綴れ錦に織っていった結果が文章ならば，その中にひそむ花や蝶が重要語なのだという淡いイメージを持ってはいたが，実際の文章からどれが花でどれが蝶なのかを選り分けるのは，それを意識しだすととたんに難しくなる。

本稿では，先に考えた助詞の堤題や主格関係から知識を抽出する試みを一層堅実にするために，一つ一つの用語をより詳細に見ることにした。詳細に見るために「テキストの索引化」という方法論を援用し，問題となる箇所をあらかじめ確認することに，本稿の目的を定めた。このことは同時に特定の専門的なテキストに対して，コンピューターによって現実的に処理できる部分と，解釈を人間が与えねばならない部分とを，区分した自動索引システムの基礎研究を意味している。すなわちここで述べる「自動索引」とは，コンピューターを用いることにより索引処理の過程を可能な限り再現できることを目指している。それはどこでコンピューターを用い，どこで人手による判断を挿入したかの全体の処理過程を明確にし，コンピューターと人の判断を適宜交えたインターラクティブな方式の試みである。そのための対象として保田與重郎『日本の文學史』[2]からその「序説」「神話」「神詠」の３章を選んだ。

以上のことを考察するために，次の５つの過程を踏んだ。

「**第１章．索引のとらえ方**」では，「**索引**」という言葉のとらえ方を，科学技術文献を主たる対象とする情報検索の世界と，日本語の古典テキストを主たる対象とする国語国文学の世界を例にあげ，対照的に論じた。またそれらを踏まえた上で，本稿における「索引」を索引語の採否に作成者の解釈を通常以上に含んだ用語索引であることの提示をおこなった。これは従来の印刷形式による索引が，テキストの機械可読形式によりその目的や様相を変えてきたことによるものである。

「**第２章．テキストの加工（マーク付け）**」では，今回の実験と将来の展開を図り，テキスト自体に対する加工をほどこした。テキストを加工する目的はテキスト構造の把握と，出現用語の位置づけをテキスト自身の文番号という絶対的な指標により，正確に定めるためのものである。もとより，テキスト自体の誤植などを訂正する意味は全くない。手法としては HTML（HyperText Markup Language）を用いた。

「**第３章．用語の抽出**」では，文末まで切れ目なく続く日本語文章からどのようにして用語を自動的に抽出するかの説明と実験をおこなった。手法としては，漢字やひらがなカタカナという文章中の字種に着目した「字種切り」を使用した。字種切りはシステムとして非常に簡便で高速ではあるが精度は高くないので，本実験では３種類の辞書を採用し，最長一致方式による照合を行っている。なおこのシステムは既発表[(3-1)]のものなので，ここでは辞書などの改良点や運用面に焦点をあてた。

「**第４章．索引語の作成**」では，第３章で抽出された用語のうち，複合語を自動的に分離合成するためのシステムと，用語に対する一般語彙からの読みの自動添付について実験をおこなった。ただし本稿での「複合語」とは単なる長い文字列もさす，便宜的な呼称である。

「**第５章．索引とその考察**」ここでは，４章までで作成された索引語を実際のシステムに組み込んだ事例を示した。また完成した索引からいくつかの問題点を列挙し，今後の展望を述べた。

第1章．索引のとらえ方

1．1．「索引」の諸相

「索引」という言葉は，一般用語や専門用語として見た場合その内容の理解に多少の不一致がある。情報検索の世界では，ある文献を代表する鍵となる語（キーワード）を抽出することが索引付けであるとなっている[4]。他方国文学の世界で索引といえば「用語索引」を指す場合が多い。また，コンコーダンスは用例索引として別扱いにもできるが，索引という言葉を用いている場合もある。一般的に情報検索の世界では索引という言葉が1つの文献に付けられた数個の代表的なキーワードを指す場合が多いようである。索引作業の具体例としては科学技術文献などの世界で，大量の文献に対して分野独自のシソーラスを用いて専門家が人手で付けている例などがある。この，シソーラスを用い独特のルールによって短時間で，ある程度普遍的な索引語付けを行う作業過程は，研究の対象として価値があり，詳しい分析と新たな索引プロセスの提案[5]がなされている。

さて国語国文学の世界での「用語索引」は一般的に「総合索引」と呼ばれており，これら人手による索引作成の歴史は厚い。たとえば平安時代の作品に例をとれば，そのほとんどに索引があるようである。『濱松中納言物語總索引』[6]などは昭和30年代に出版されているが，当時は同物語研究者の数も比較的少なかったようである。しかるに同書のような用語索引が公にされている。また，最近では源氏物語に関するデータベース作成[7]がいくつか行われており，それらの作業には多くの場合，索引という観点が強く現れている。本稿での「索引」とは，基本的に国語国文学での「用語索引」から派生している。

テキストから言葉や文字を探すことは，対象がコンピューター上にあるかぎり困難なことではない。すでに発表した「島探索」[3-2]方式によれば，1文字単位であっても，いくつかの用語の同時出現に対しても，妥当な速度で結果を得ることができる。このような状況にあって，索引をどのように捉えるのかについて，なぜ索引をつくるのか，それをどのように使うのかということに関して，

国文学の観点[8]からいくつかの示唆を得た。

コンピューターなどの利用によりすべてが引ける状況にあっての索引とは，冊子体形式であれコンピューター画面上での表示であれ，引かれるに値する意味のある言葉を表示することが求められる。これは言葉を換えれば「専門用語」を抽出するということである。元来索引は中立公正なものである必要があった。利用者は用例を自らの考えに従って選び本文に当たるものである。もし索引に特殊なフィルターがかかっておれば，その利用者は選択の幅をせばめられてしまう。このことの穏やかな例では文節切りを元にした索引では，どのように文節を認定したかによって多少のゆれが生じる場合などにあてはまる。

しかし本稿での索引とはそのような中立公正なものではなくて，むしろ解釈の強く入った索引となる。一般的に自然科学系での専門用語は普遍性をもち，用語自体が共通の概念を持つ[9]といわれている。一方人文科学系ではテキスト自体が固有の世界を確立しており，およそ共通の概念を認めがたいという事例が多々ある。ここでの索引とは後者のような世界に対応するものである。

この共有概念の希薄な世界，すなわち自然科学系のような意味では専門用語が確立していないところで，いわゆる「自動索引システム」をつくることが今後大きな問題になってくる。たとえば情報検索の世界では「神と人」という用例に対してこれを「神＋と＋人」に分割して，神，人をキーワードと定め，検索する際には「神 and 人」ないし「神 or 人」というような指示をするのが一般的である。しかし「神と人」については，すくなくとも本稿での保田の索引を考えるにあたっては，見出し語として「神と人」がなければならない。これを神，人に対してそれぞれ分離したままにするのは粗雑であるといえる。このことばは，保田の索引においての専門用語である。解釈するならばこれは『戴冠詩人の御一人者』[10]に描かれた「神から人への下降の悲劇としての日本武尊」からきている。そこで自動索引というものを考えたとき，理想的には「神と人」を自動的に索引語として選定する過程を踏む必要があるのだが，本稿においては未だそこまでの考察はできていない。これについては後述の手技によるものとした。

以上のようなことを考慮に入れた上で，本稿での索引というものを専門用語索引とし，一般用語は『分類語彙表』[11]など別の辞書の利用により排除することも基本的に可能とした。これは，現代のテキスト検索が１文字単位で可能であるという事実から選択した結果である。

１．２．何を索引語として採るのか

テキストはコンピューター上にあり，１文字ずつの検索も可能となっているときに，何を索引語として残すのかについて，いくつかをまとめておいた。

（１）保田の思想表現上で鍵となる言葉

例：敷松葉

これは本稿の保田與重郎に対する見識，すなわちある種の独断において設定せざるを得ない。しかし，それは全く意味がないわけでもなく，たとえば例に挙げた「敷松葉」は保田の終の住処となった京都太秦鳴瀧三尾山の身余堂にある敷松葉であり，身余堂という人工の建築物と嵯峨野にかかる自然の中の敷松葉との対比として使われた言葉である[2]。文学作品を対象とするときは，テキストからこぼれ落ちるかもしれない言葉を，自らの感性と経験において拾い上げるという行為がなければ，逆に自動化すなわち普遍性にいたる客観的な作業は成り立たないものである。この場合索引行為において「敷松葉」という言葉を選定した事実は辞書というアルゴリズムに記述され，そこに客観性の証も残されるものと考えている。

（２）文学用語

例：詩歌，物語

詩と歌とは分離もするが，詩歌としても残す。また，物語も残す。これらは分類語彙表での1.321（体／人間活動－精神及び行為／創作・著述／芸術・文芸）に位置づけられる一般用語であるが，対象の性格により頻出しても必要であるとみなした。「物語と歌」をテーマの１つに持つ保田から，頻出語は採らない傾向にある自然言語処理での，一般的な統計的手法を適応させることはできない。

（３）人名，作品名，神名などの固有名詞

例：大伴家持卿，万葉集古義

人名や作品名を索引語の説明としてあえて特別にとりあげたのは，表記の多様性を多角的に検索することを目的とするための複合語処理にかかわるからである。たとえば，例の２つは４章（１）で詳述する記号によってそれぞれの名詞を細分している。

（大伴｜家持）（卿）　→　{大伴家持卿，大伴家持，大伴，家持，卿}
大伴｜卿　→　{大伴卿，大伴，卿}
家持｜卿　→　{家持卿，家持，卿}

「卿」の採択には迷うところもあるが，保田の用例では「卿」とか「翁」とか「先生」は，特別な人物に付く場合が多く不用意に棄てることはできない言葉である。

―万葉集｜古義　→　{万葉集古義，古義}

ここでは「万葉集」を採らない。古義が鹿持雅澄の「万葉集古義」をさす場合はあるが，「万葉集」と「万葉集古義」とは別の作品である。

（４）ひらがな表記の言葉

例：みそぎ（禊ぎ），くれなゐ（紅）

本稿の実験は技術的に字種による用語の分割を行っている。そこでは基本的にひらがなの文字連糸は排除される。しかし保田の場合には重要な名詞なども例のようにひらがなで表記される場合が多い。また，みゆじく（ミュジーク：音楽），みゆとす（ミュトス：神話）などのようにドイツ語をひらがなで表したものもあり多様である。これなどは折口信夫が外来語をひらがな表記している例に近いであろう。このために，最初の段階で分割した言葉のうち棄却されたものを悉皆調査することにより，おもにひらがなの名詞を抽出した。

（５）一般語の削除について

結果に対して分類語彙表に現れた用語を削除する試みも行った。しかし，この作業は補助的なものであり，索引としては上記（１）～（４）でほぼ満足な索引語が得られたと結論づけている。

第2章．テキストの加工（マーク付け）

本実験ではHTML[12]を使ってテキストに事前の加工をほどこした。本格的なテキストデータベース構築に関してはSGML[13][14]や国文学固有[15]のものがすでに開発され実施されている。本実験でHTMLを採用した理由は次のような点にある。

（1）HTMLは難解なSGMLのサブセットであり，比較的容易に修得できる。

（2）HTML文書はインターネット[16]を通して世界的な学術伝搬を容易にする。

（3）HTML文書はMosaic[16]によって，直ちにハイパーテキスト表示を可能にする。

本実験では，用語の位置決定をある程度普遍的にするためにこれを用いた。一般的に索引は，もとからある冊子を原本としてそれに対する頁の指示を基本としていたが，ここではテキストを絶対的にあつかい，その「章｜節｜段落｜文」を基に指示をおこなっている。もちろん，全集に対する総索引をつくるような作業であれば，全集自体の頁版面情報を表として作りそれに結合することは容易である。しかし，本稿での目的は索引作成上の用語の抽出過程の考察にあるので，現実的な版面作業は保留した。

さてこのような目的において，HTMLは章節段落ならびに引用のマーク付けにだけ用いた。そのために必要となった記号は表1のようなものであり，実際の文章記述例は図1に示した。また，図1をMosaic上で表示したものを図2にあげた。ただし図2は本稿での実験と直接の関わりはなく，ハイパーテキスト化した将来の姿として参考に提示した。

このうち特に頻繁に使用した記号についていくつかの気づいた点を記しておく。まず表1から，HTML、HEAD、TITLE、BODYの4種類は本実験では大きな意味を持ってはいない。これらの宣言はMosaicによって有効となる。〈H1〉～〈H6〉は章節項以下をマーク付けするためのものであり，具体的にはMosaic上での表示方法(文字フォントの種類や大きさ)を指示することができる。本実験では，〈H1〉と〈H2〉しか使用していないが，このマークによってテキスト内での章と節という位置情報を正確に定義している。次に〈P〉

の段落終了情報であるが，これはHTMLの規則では未だ明瞭には定義付けされてはいない。本実験では，段落の終了すなわち改行マークとして使用した。

表1　HTML文章埋め込み記号

このうち／で始まるものは終了タグである。〈Cite〉は本稿での仮設定である。

〈HTML〉	HTML文書宣言
〈HEAD〉 〈/HEAD〉	HEAD宣言
〈TITLE〉 〈/TITLE〉	タイトル宣言
〈BODY〉 〈/BODY〉	本体宣言
〈H1〉 〈/H1〉	表題階層 1
〈H2〉 〈/H2〉	表題階層 2
〈H3〉 〈/H3〉	表題階層 3
〈H4〉 〈/H4〉	表題階層 4
〈H5〉 〈/H5〉	表題階層 5
〈H6〉 〈/H6〉	表題階層 6
〈P〉	段落終了情報
〈Cite〉 〈/Cite〉	引用文
〈UL〉 〈/UL〉	リスト宣言
〈LI〉	リスト内容
〈BR〉	強制改行

〈HTML〉
〈HEAD〉
　〈TITLE〉日本の文学史〈/TITLE〉
〈/HEAD〉
〈BODY〉
〈H1〉序説〈/H1〉
　　〈H2〉一〈/H2〉
　私が日本の美術史を書き了へたのは，～（中略）～全く無なのである。〈P〉

図1　HTML文章例

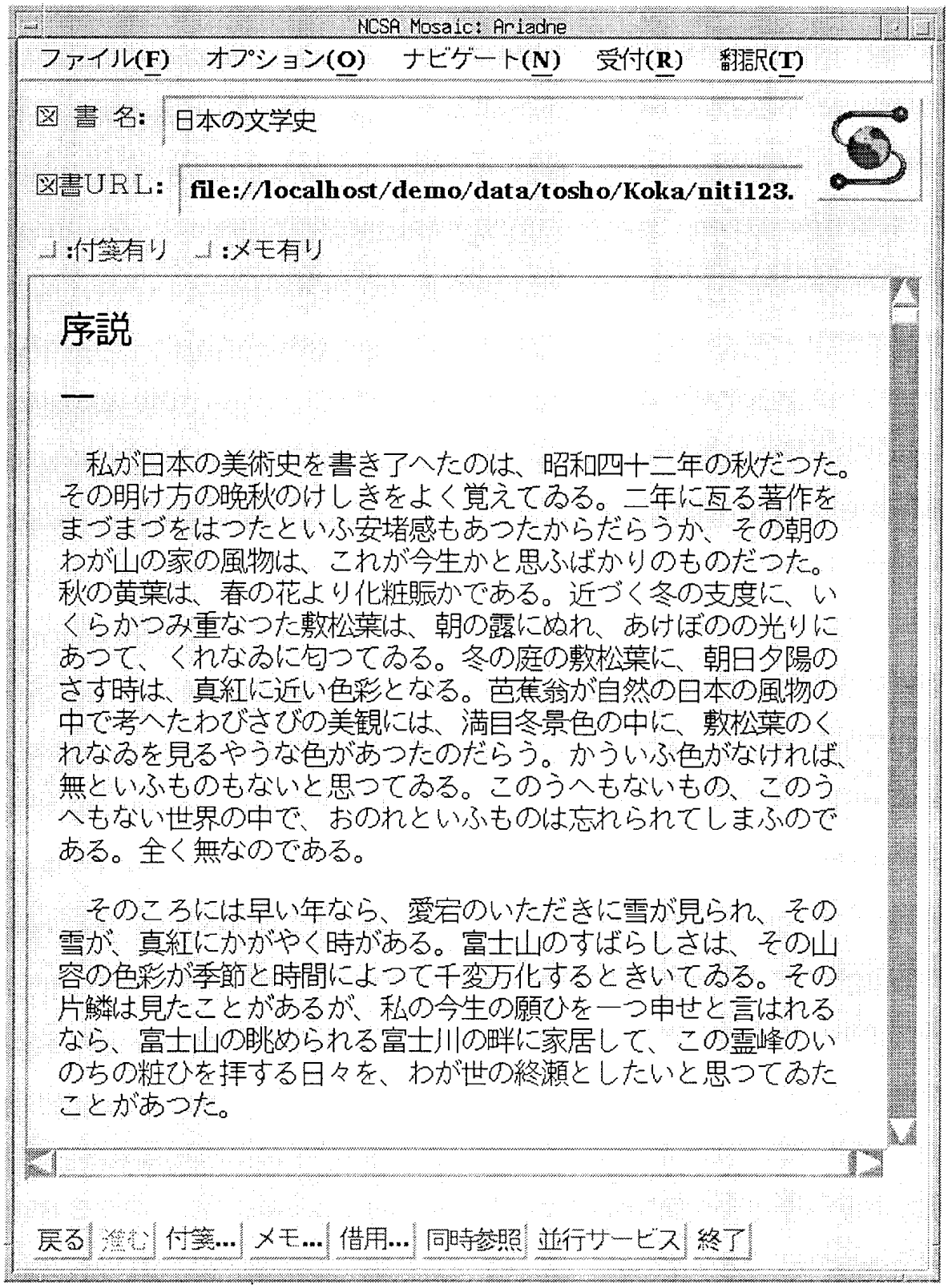

図 2　Mosaic による文章表示

第３章．用語の抽出

２章でのHTML文書をもとにして，まず粗い用語の自動抽出をおこなった。図３によりこの処理は用語抽出プログラムを使った。用語抽出プログラムには３つの辞書がある。まず複合語辞書は分割されては困るひとまとまりの語を持つ。たとえば「神と人」や，ひらがなでしるされた「みそぎ」などが含まれて

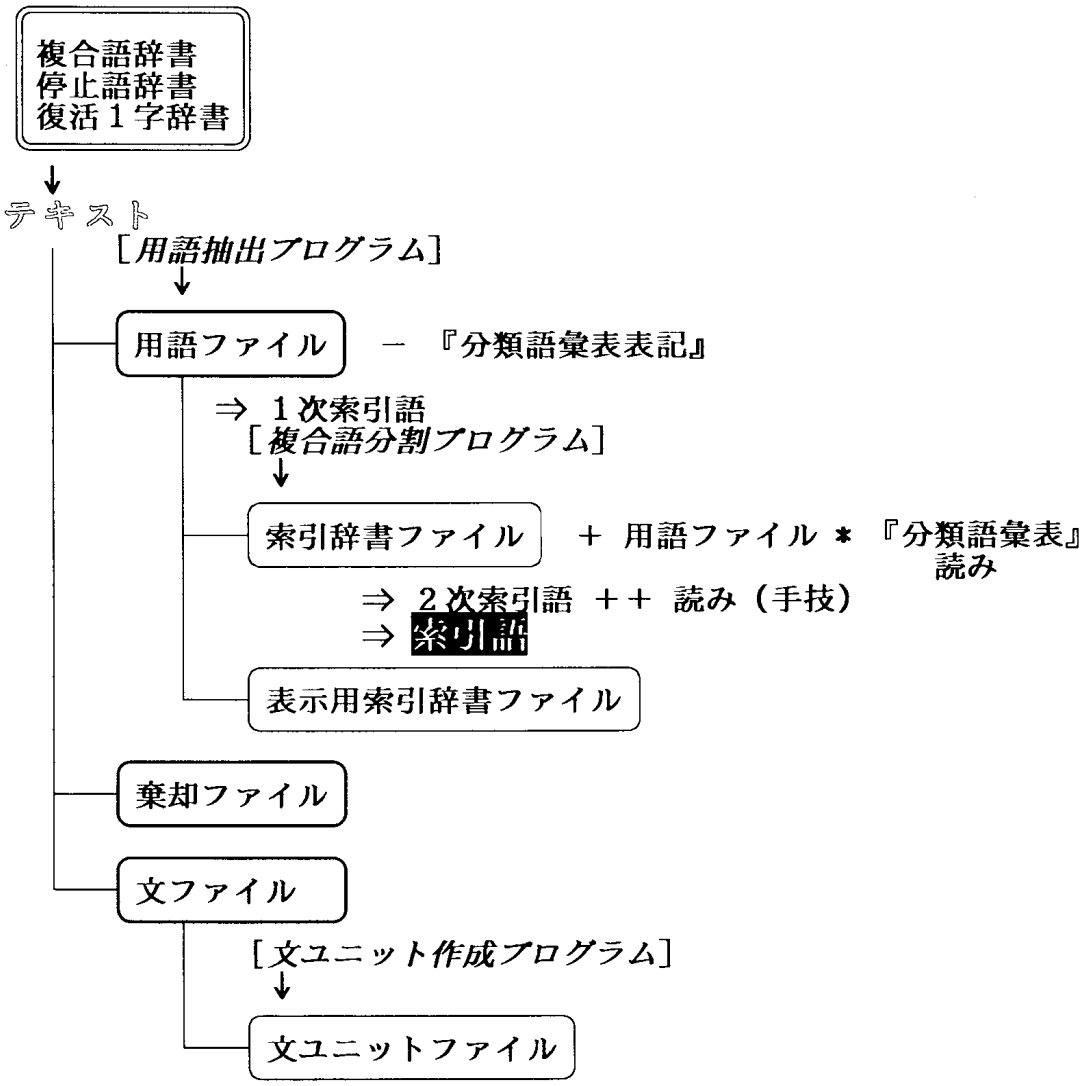

図３　索引語作成処理過程

図は**テキスト**に始まって**索引語**で終了する。記号｛＋，－，＊｝は関係演算での和集合，差集合，積集合演算子をあらわす。＋＋は手技による加工である。[***斜体太字***]は実行プログラムを意味する。同図の実行システムでの詳細は本稿末での付図３－１にあげた。

おり，本実験では157件を数えた。次に停止語辞書はたとえば「その時初めて」から「時初」という語などが切り出されるのを停止するために用い，ここでは46件を用意した。最後の復活１字辞書はアルゴリズム上，１文字出現の漢字を排除しているので，これをさけるために「神」や「社」など85種の漢字を用意した。ここで用いた３種類の｛複合語，停止語，復活１文字｝辞書はテキスト全域に影響を与えるものであり，一旦設定すれば全テキストに作用する。これは４章で述べる複合語の分割が局所的に影響を与えることとは異なる。

用語抽出プログラムによって作成されるファイルは｛用語，棄却，文｝の３種類である。用語ファイルは表２のように，先頭にテキスト中での位置記号が自動的に付けられている。「日0101×××010106,秋」を例にとると，先頭の「日」は仮に「日本の文學史」を意味する。他の多数のテキストと比較するときはさらに精密な記号化を図る予定である。次の0101は数字２文字単位で章節などの階層を現している。×は使用していない階層を意味し，全体として６階層まで表現できる。これはHTML文書におけるマーク付けに連動している。後尾の６文字の数字は，これも２文字単位で設定してあり｛段落，段落内文，文内用語｝連番となっている。段落連番は章節項などが変化するたびに１からはじまる。例での「秋」は「日本の文學史」第１章第１節第１段第１文の６番目に現れた用語「秋」となる。

表２　用語ファイル

日0101×××010106,秋
日0101×××010201,明け方
日0101×××010202,けしき
日0101×××010203,晩秋
日0101×××010301,二年
日0101×××010302,著作
日0101×××010303,安堵感
日0101×××010304,山の家
日0101×××010305,風物
日0101×××010306,かりのもの
日0101×××010307,今生
日0101×××010401,秋
日0101×××010402,黄葉
日0101×××010403,春
日0101×××010404,花
日0101×××010405,化粧賑
日0101×××010501,冬
日0101×××010502,支度
日0101×××010503,敷松葉
日0101×××010504,朝の露
日0101×××010505,あけぼの
日0101×××010506,くれなゐ
日0101×××010601,冬

次の棄却ファイルには用語ファイルから抜け落ちたすべての文字が保管される。これは特にひらがなの悉皆調査のために使う補助ファイルである。３番目の文ファイルは文番号を持った１文単位の記録である。これは第５章で説明するが，実際の検索システムにおいて文のかたまり（文ユニット）を高速に呼び出すために用意した。

次にこれらファイルの運用については，用語ファイルを中心として，棄却ファイルを調査し，３つの辞書を調整するという循環を数回繰り返した。専門辞書がないと考えられる分野ではこうした作業は必須のこととなる。その処理過程を３つに分けてみると次のようになる。

（1）字種による抽出

漢字ひらがなカタカナの変化するところで文字列を切断し主に漢字列を抽出する。この抽出結果は最終的に3658例あった（異なりは1772件）。これを「用語ファイル」とする。このとき採取されなかった文字列は棄却ファイルとして保管する。

（2）とりこぼした用語を視検により採取する

ここでは棄てられた，おもにひらがな文字列から用語の採取を手技で行った。これは最終的に157例あった。

（3）採用用語の調査

用語ファイルに採用された用語のうち「以来時」,「沢山立」などを停止語に組み入れた。これは最終的に46例あった。

第４章．索引語の作成

第３章で作成された用語ファイルは荒削りのものである。ここではこれにいくつかの加工をほどこし，さらに分類語彙表を使い読みを自動添付した。その一連の過程は次のようになる。

（1）１次索引語に対する加工

まず用語ファイルから分類語彙表表記の差分をとり，これを１次索引語とした。これは異なりで523件あった。ここで分類語彙表は一般語彙と仮定してい

るので，その差分523件とは保田固有の｛特殊な用法，専門用語，用言の語尾変化｝などが考えられる。

次に表３のように，この523語をすべて特殊な複合語であると仮定して，４種類の記号「－，(,)，｜」によって，棄却，統合，分割等の指示をマーク付けした。ここで－は直後の語を不採用とし，(　) は語を区切る範囲を設定し，｜は語を区切る。たとえば，「－古｜神道」は｛古神道，神道｝の２つの索引語として分割される。このマーク付けは自由度が高く，１つの複合語（文字連糸）を次のように自由に分離組み合わせることができる。

異型伝承　→　異型｜伝承　→　｛異型伝承，異型，伝承｝
　→　－異型｜伝承　→　｛異型伝承，伝承｝
　→　－（異型｜伝承）　→　｛異型，伝承｝

なおこのマーク付けによる結果はこの複合語にあらわれるときだけに有効である。たとえば，以下の例では同じ「自動」という言葉があっても「自動分類」からは「自動」が生成されるが，自動索引からは「自動」が生成されないという違いが生じる。これは局所的に言葉の採否を処理するためのものであり，第３章での３種類の辞書とは適用範囲が異なる。

自動｜分類　→　｛自動分類，自動，分類｝
－自動｜索引　→　｛自動索引，索引｝

以上のマーク付けをされた１次索引語に独自に作成した複合語分割プログラムを使用した。この結果は表４および表５である。ここで表４はチェックや説明表示のために作成した。表５は同表４を関係データベースの処理に適した形にあらためたものである。

（２）２次索引語の完成（元の切り出し用語との併合）

この表５と元来ある用語ファイルとを併合しさらに重複した見出し語を削除し，これに分類語彙表から読みを自動添付したものが２次索引語である。これは異なりで2217件となった。これらの作業は関係データベースシステム（Rbase）によって行われた。

（３）「読み」の添付と索引語の完成

２次索引語では2217件のうち1142件が分類語彙表から「読み」を自動添付できた。この中には，「後世（こうせい，ごせ)」のように，読みが複数マッチした例も含まれる。残りの1075語については手技で読みをふった。またHTML文章では保田自身が歴史的仮名遣いでふった読みを「{カタカナ}」でしるしてあるので，これを別途自動抽出し添付し，完全な索引語とした。

（４）索引作成

索引語を索引辞書ファイルと併合しさらに用語ファイルと併合することによって，用語の出現位置を添付した。先述したように，本稿では印刷版面の問題は保留しているので，この作業でもって一連の操作は終了したことになる。

表３　１次索引語に対する加工（複合語分割マーク付け）

実際の１次索引語は1496件あったが，ここから見出し語の重複を削除したので，複合語分割マーク付けをしたものは523件であった。

（建速）（須佐之男｜命）
元禄｜六年
現世｜仮相
現世｜現実
（個人｜－的）（立場）
古今｜源氏
－古｜神道
古事記｜－下巻
古事記｜－三巻
古事記｜冒頭
固定｜観念
誇張｜変貌
五十鈴｜川
後水尾｜院
後鳥羽｜院
（後鳥羽｜院）（－以後）

表４　表示用索引辞書ファイル

複合語分割の結果生じた全件数は1471であった。表３の523件を処理したのだから新規に生成されたものは948件となる。また表中の記号 {!，＋，・} および５章で使われるマーク◎はそれぞれ次のような意味を持っている。

！：対象複合語そのままの語
・：複合語分割の結果生成された語
＋：対象複合語の分割結果であるがさらに分割されるもの
◎：複合語処理の対象とはなっていない語（第５章図４のmk1欄参照）

後鳥羽院以後隠遁詩人，！　（後鳥羽｜院）（－以後）（隠遁｜詩人）
後鳥羽院，＋
　後鳥羽，・
　院，・

隠遁詩人,＋
　隠遁,・
　詩人,・

表5　索引辞書ファイル

左方の「建速須佐之男命」は用語ファイルに出現した元の形である。右方の語は分割された形である。

建速須佐之男命,！,建速須佐之男命,！
建速須佐之男命,！,建速,・
建速須佐之男命,！,須佐之男命,＋
建速須佐之男命,！,須佐之男,・
建速須佐之男命,！,命,・

表6　2次索引語

2次索引語は2217件あった。このうち1142件は分類語彙表から「読み」が照合され自動添付された。その中には「後世（こうせい，ごせ)」のように，読みが複数マッチした例も含まれる。

"古事記伝",こじきでん
"古事記冒頭",こじきぼうとう
"古心",こしん
"古神道",こしんとう
"古人","こじん"
"古典","こてん"
"古伝",こでん
"古道",こどう
"固守","こしゅ"
"固定","こてい"
"固定観念",こていかんねん
"孤城",こじょう
"誇張","こちょう"
"誇張変貌",こちょうへんぼう
"五十鈴",いすず
"五十鈴川",いすずがわ
"後水尾",ごみずのお
"後水尾院",ごみずのおいん
"後世","こうせい"
"後世","ごせ"
"後代","こうだい"
"後鳥羽",ごとば
"後鳥羽院",ごとばいん

第5章．索引とその考察

第3章で言及した「文ユニット」と第4章で完成した「索引語」とをデータベースシステム（Paradox）に組み込み，具体的な索引を完成した。例として，

図4の「みだし」の「伊勢」をマウスで指示すると左方の「文」欄に「伊勢」の出現文が表示される。図には現れていないが，コンピューターの画面上では検索するための見出しや読みの言葉を受け付ける欄や，索引全体を移動できる垂直スクロールバーなどが設定してあるので，多量の索引語の出現箇所を自在にながめることができる。

日本の文學史／保田與重郎

文
国の初めに出来た伊勢の皇大神宮の建物が、一番うつくしい建物だといふことが、絶対的な観念となる時、一体われわれの造形美術の歴史は何だつたかと、かうしたことを考へると、われわれの心は一種のあやしさにまぎれておぼれる感がする。・・・

みだし	読み	mk1	位置記号
伊邪那美	いざなみ	・	日0203×××010105
伊邪那美	いざなみ	・	日0203×××010202
伊邪那美神	いざなみのかみ	!	日0203×××010105
伊邪那美神	いざなみのかみ	!	日0203×××010202
伊勢	いせ	◎	日0202×××020103
伊勢	いせ	◎	日0304×××030702
依然	いぜん	◎	日0201×××040302
偉人	いじん	◎	日0101×××100505
夷振	ひなぶり	◎	日0205×××060205
意志	いし	◎	日0301×××070903
意志	いし	◎	日0302×××010505
意志	いし	◎	日0303×××030103
意識	いしき	◎	日0101×××041704

図4　文表示機能をそなえた索引システム

完成した索引の延べ語は5750，異なり語は2097となった。また見出し語の頻度の上位部は表7に示した。この表で「4．日本 102」，「5．文学 85」と「40．日本文学 15」との関係はすでに4章（2）で説明したように相互に重なっている。この場合「日本」や「文学」のうちには図4のmk1欄の◎マークのように，分割対象ではない語や，・マークである分割語も含まれているわけである。それぞれのマークは内部保存しているので統計はある程度細かくとれるが煩雑になるので記さない。

表7を見てみると，一瞥してカタカナ外来語をほとんど使用していないことがわかる。本実験範囲ではアジア，シンガポールと「コギト」だけであった。「コギト」は初期日本浪曼派の関係同人誌名である。頻出する「美」「自然」「物語」「歌」「風景」「神」「国」とは，およそ保田が「日本の橋」以来一貫し

表7　索引語の使用頻度

1．私	294	19．古	23	37．幽契	16
2．神	121	20．わが国	22	38．色	15
3．国	112	21．古事記	22	39．石	15
4．日本	102	22．空	21	40．日本文学	15
5．文学	85	23．美	21	41．まこと	14
6．歌	80	24．志	20	42．自然	14
7．天皇	48	25．天	20	43．世界	14
8．歴史	47	26．伝へ	20	44．天地	14
9．女	44	27．思ひ	19	45．文芸	14
10．心	44	28．子	18	46．考へ方	13
11．詩	36	29．女神	18	47．神詠	13
12．書	34	30．美術	18	48．代	13
13．ことば	33	31．万葉	18	49．美術史	13
14．史	28	32．命	18	50．風景	13
15．時代	27	33．声	17	*1．物語	13
16．わが	26	34．山	16	*2．民族	13
17．翁	24	35．無	16		
18．神話	24	36．名	16		

て語り継いできた言葉である。その他として，あえて「私」や「わが」を索引にとりいれたのは別の論考(1)との関係によるものである。予測にすぎないが彼にとって「私」や「わが」は特殊な専門用語の位置を占めるのではないかと考えている。その言葉は文学，美，伝統の唯一継承者としての自負に満ちた鍵語となるものであるのかもしれない。しかし本稿では，これらの言葉への深い言及は措く。

第6章．結　　論

本稿では保田與重郎の専門的索引を構築するために，3つのプログラムを使用した。1つは，HTML 文書から用語を切り出すためのものであり，これを用語抽出プログラムとした。2つめは同プログラムによって抽出された原始的な1次索引語を，こまかく分離統合するためのプログラムであり，これを複合語分割プログラムとした。このプログラムの使用により，複合語を局所的に分割する方式を可能とし，索引作業をより精密にすることができた。3つ目は文

ユニット作成プログラムであり，これは文番号と文とをユニット化し，実際の索引システムを動かすために用いた。これらは独自にプログラムを作成し，その他の部分，たとえば索引語群に分類語彙表から「読み」を与えるところなどは，関係データベースシステム（Rbase）の持つ関係操作を用いて行った。以上の一連処理の後，一般的なデータベース（Paradox）上で「文表示機能をそなえた索引システム」を構築した。これに使われた索引語は延べ語で5750，異なり語で2097語となった。

本稿の目的のひとつは「特定の専門的なテキストに対して，コンピューターによって現実的に処理できる部分と，解釈を人間が与えねばならない部分とを，区分した自動索引システムの基礎研究」にあった。結論として，中心となる用語と特殊な用語のために判断と解釈とを含めた辞書を人手で作れば，索引語の自動抽出が可能となり，また一般的な用語に対しては分類語彙表などの既存の辞書を利用することにより，読みの添付などの自動化が可能であることが得られた。

なお，本稿では印刷版面処理，ならびに結果に現れた専門用語と一般用語との諸関係の深い分析は措いた。

謝　　辞

国文学における索引の全般的な様子をお話くださり，貴重な図書資料をここころよくお貸し願えた大阪大学文学部伊井春樹先生に感謝の意を記します。

また本稿は「文部省科学研究費『目次の構造と索引等を利用した日本語文献のハイパーテキスト化による高次検索』一般研究（C）課題06680385」に一部（索引の基礎的考察部分）関わるものである。記して感謝する。

参考文献と註記

（1） 谷口敏夫「日本語文章からの知識の抽出――保田與重郎における「は」と「が」との機能分析を中心に――」『光華女子大学研究紀要』31, 1993. p. 49–79.

（2） 保田與重郎『日本の文學史』新潮社，昭和47年.

（3）　長尾眞他著『研究情報ネットワーク論』勁草書房，1994.

（3-1）長尾眞，谷口敏夫「第9章　目次情報に基づく図書検索とOCRによる目次入力の実用可能性」p. 161–174.

「字種切り方式」での処理時間例に言及するならば，たとえば新書1冊，HTML文書で約270 Kbのものに約600件の辞書を備えて，4分30秒で処理をする。1 Kb（日本語500文字）を1秒で処理したことになる。

（3-2）谷口敏夫「第10章　新しい全文検索アルゴリズム——島探索（アイランドサーチ）方式——」p. 175–182.

（4）　丸山昭二郎他著『主題組織法概論』紀伊国屋書店，昭61年.

（5）　倉田敬子，神門典子「索引作成者の認知とテキスト構造との関連から見た索引作成過程」『書誌索引展望』17（4），1993. p. 1–22.

（6）　池田利夫編『濱松中納言物語總索引』武蔵野書院，昭和48年（再版，初版は昭和39年）

同書の凡例では単語の取り扱いについて4頁にわたる説明が，12の観点から詳しく記されている。中でも複合語の扱いについては本稿においても多大の示唆を得た。

（7）　国文学研究資料館情報部編「第5回『国文学とコンピュータ』シンポジウム講演集」，国文学研究資料館，1994. 109p.

同集中，伊井春樹「源氏物語データベースの構築」では『古典文学総合事典』の説明があり，索引とシソーラスの関係が記されている。

（8）　村上學「中世の語り物の総索引作成に関する考察——幸若舞曲・平家物語・浄瑠璃物語をめぐって」『人文科学データベース研究』4, 1989. p. 56–66.

（9）　長尾眞「専門用語辞典作成法試論」『専門用語研究』5, 1993. p. 10–19.

（10）　保田與重郎「戴冠詩人の御一人者」『保田與重郎選集第一巻』講談社，昭和46年.

「神と人」ならびに「神人分離」についてさらに言及する。本稿対象テキストで7例あった「神と人」は倭建命の運命を記述するに際して使われ，そこでは「神人分離」という概念（専門用語）を表している。その意味は，倭建命が記紀で悲劇的に描かれたことの根底には，当時の人たちがそこに神から人への下降をみているからであるとなる。神ともあがめられる一族の皇子が，人として軍事政争のただ中で傷つき敗北していく過程に悲劇性をみたものといえよう。人としての衣装だけが残り，実体は白鳥として飛び去っていく姿は，一方で昇華を見せるが，次々と白鳥の陵を経巡る姿には，神→人→白鳥→神という変身の苦渋を印象つける感が深い。それを保田のテキストから，単に「神（名詞）＋と（並列助詞）＋人（名詞）」と分離して扱うだけでは実態から外れるものである。なお，

その先触れとしては崇神天皇時代の「同殿共床」があり，ここでの主役である倭姫命は倭建命の叔母にあたる。もちろん「同殿共床」は保田にあっては決定的な専門用語となる。これらは初期作品「戴冠詩人の御一人者」（昭和１１年）に典型的に描かれている。参考までに同書から２つ引用を示す。下線は谷口による。

「日本武尊の悲劇の根本にあるものは，武人の悲劇である。神との同居を失ひ，神を畏れんとした日の悲劇である。言あげと言霊の関係をつくる，神を失つてゆく一時期の悲劇として，この説話は古事記中でも重大な意味を言霊したのである。」，「いづこで神と人との上代に於ける分離がかくも美事に言霊されてゐるか。」

(11) 国立国語研究所『分類語彙表』秀英出版，1964.
本稿では同上「フロッピー版，国立国語研究所，1993年」を使用している。ここに国立国語研究所に対し，記して感謝する。

(12) "A Beginner's guide to HTML", National Center for Supercomputing Applications / pubs @ ncsa. uiuc. edu 1994年５月インターネット上からの抽出による。

(13) 長瀬真理「日本語―英語対照『源氏物語』のテキスト・データベースの作成に関する基礎的研究」『情報知識学会誌』1 (1)，1990. p. 40–53.

(14) 石塚英弘「SGML 形式による学会誌全文データベースの構築と印刷」『情報知識学会誌』2 (1)，1991. p. 23–48.

(15) 安永尚志「日本古典文学の本文データベース」『情報処理』35 (7)，1994. p. 642–650.

(16) 谷口敏夫「インターネットを利用した情報サービス」『第１回京都大学高度情報化フォーラム』京都大学学術情報ネットワーク機構、1994. p. 40–47.

開発環境

（１） NEC PC 9821 Ap
（２） Turbo Pascal for Windows / Borland International
（３） R : Base 4.0 / BCon systems and Microrim
（４） Paradox for Windows / Borland International

複合語辞書
停止語辞書
復活１字辞書
↓
テキスト（niti123. htm）
[***用語抽出プログラム*[TANGYASU. exe]]
↓
用語ファイル(niti123. mei) －『分類語彙表表記』
⇒ １次索引語(niti123. idx)
[***複合語分割プログラム*[SPLITIDX. exe]]
↓
索引辞書(niti123. dic) ＋ 用語ファイル ＊ 『分類語彙表読み』
⇒ ２次索引語(yomi. txt) ＋＋ 読み（手技）
⇒ **索引語**
表示用索引辞書(niti123. lix)

棄却ファイル(niti123. xxx)
文ファイル(niti123. bun)
[***文ユニット作成プログラム*[BUNDAREC. exe]]
↓
文ユニット(niti123. dat)

付図３―１　索引語作成処理過程の詳細

図は**テキスト**に始まって**索引語**で終了する。記号｛＋，－，＊｝は関係演算での和集合、差集合、積集合演算をあらわす。＋＋は手技による加工である。[***斜体太字***]は実行プログラムを意味する。(niti123. idx) などの英小文字はファイル名をさす。本文図３を参照。