第2章　全文からの「位置情報付き用語」の抽出（谷口敏夫）

はじめに

昨今、インターネットや電子図書館の出現によってテキスト情報を簡便にかつ効果的に扱う要請が高まってきた（谷口敏夫、1998a）。社会学などでの調査や統計も、本格的に文章を処理する事例が増えてきた。アンケート調査の自由回答や聞き取り内容も、以前に比較して容易にパーソナルコンピュータを用いて、迅速に処理することができるようになってきた（川端亮、1998）。これらの多くは比較的若い研究者層によって、汎用的なソフトウェア、PEAL や SED などによって、随分効果的な処置をされるにいたっている。基本的にこれらは、文字列に対する正規化処理を伴ったパターンマッチングや、簡便なファイル管理能力の助けを得て進められている。

しかし、こういった汎用的なソフトウェアは強力な機能を持つと同時に、一般的な研究者にとっては習熟の上でなかなかに問題も多い。社会学者にとって SPSS や SAS などの基本的な統計処理ソフトウェアの利用は避けて通れない面もあるが、他の汎用的なソフトウェアにまでは手がのばせない、あるいはいたずらに習熟のための時間をとり、本来の目的からはずれてしまうことも、往々にして見かけられるところである。

ここで発表する KTCoder は、文章全体に対する文字列処理、単語処理などの専用目的を持ったアプリケーションである。小規模なパーソナルコンピュータ上で、比較的大容量の文章から、実用レベルの処理時間で特定の文字列を抽出し、その出現位置情報を添付するものである。結果として出される出現位置付き情報はデータベース化や、SPSS 等ほかのソフトウェアを使って本来の研究目的に資することが可能となる。本書での川端の論考（２部２章１節）は、データベース・アクセスや表計算・エクセル（マイクロソフトウェア社製）を利用したその一例である。

本章は、KTCoder の概略と簡単な事例とを記し、今後の検討を重ねる資料として提供するものである。付録１には KTCoder の使い方をまとめてあるので、そこを最初に参照していただいてもよい。

１．KTCoder の基本設計思想

まず KTCoder の目的は、テキスト中から必要とする単語（正確には「特定文字列」）を抽出することであり、抽出した単語を使った基本的な事例は本報告書に収めたが、応用部分は各研究者の裁量にまかされている。

つぎに現状の KTCoder は、各研究者にいくつかの事前準備を要求している。これらの多くは将来自動化する予定ではあるが、単純に文章から特定用語を取り出したいだけの研究者にとっては煩瑣なことでもある。そこで、「なぜ KTCoder は他の類似アプリケーションに比べて、そのようなことを必要とするのか」について、いくつか設計思想上での考えを述べておく。

字種切り方式の採用に起因する辞書作り

KTCoder は特定文字列を抽出するために「字種切り」という方式を用いている（佐竹秀雄、1987）。詳細は後述するが、このことによって研究者は KTCoder に熟練するにしたがって辞書作りを必要とする。字種切り方式は、基本原理が簡明なので、ソフトウェアを作ったり保守するのが容易である。また研究者側から見ると、能力の低いコンピュータであっても妥当な正確さで処理速度が得られるなどの利点がある。辞書作りは、確実に抽出したい文字列（例「教主さまが」）や、確実に棄却したい文字列を、標準テキストファイルに列挙するだけである。この労力は、当初何度か試行錯誤を必要とする場合もあるが、その結果完全に近い出力を得ることができるようになる。

抽出語の全文中での出現場所を精緻に求める意義

KTCoder は、各用語の正確な位置情報（ＩＤ）出力を可能としている。これには、文章ファイル区分、階層構造、段落や文番号、文内での一連番号などが含まれる。しかし次の項で記すが、そのために HTML 見出しタグを用いた事前の文章階層付けなどの作業を研究者に要求し、その労は軽微ではない。そこでなぜ、抽出語の出現位置を求めるのかについて、いくつか説明する。

KTCoder は元来学術図書や論文ないし小説などのような、階層構造を持った長文を対象に設計された事実がある（谷口敏夫、1994）。学術図書の目次が概念構造を明確に反映している事実からも、文章とは一般的に、概念表現の階層を伴っている。よってその中で使用された「用語」は、概念の階層の高低によって異なった意味を持っていることが判明している（長尾真・谷口敏夫、1994）。たとえば、出版社名「哲学社」とか、大ざっぱな書名レベルで使用される「人間と哲学」と、階層の低い「ギリシャ哲学」の章下での、「プラトン哲学」とでは、同じ「哲学」であっても意味内容が異なる。この違いを、文章批判などで正確に反映させるためには、ことばの使われた位置情報が重要である。単純に、言葉の頻度を出しただけでは、文章やその著者の意図を判定することはできない。

つぎに一般論として、非文ないし非文章でないかぎり、文章は一定の文脈を持っている。これは文章の階層構造問題以前のこととして、文脈がない限り文章は話者の意図を他に伝達できない。

話題に応じた文脈を、ひとまとまりの大海や川や池になぞらえると、強い意味を持った「言葉」はそこに浮かぶ島のようなものであり、必ずその島と他の島との関係を持ってい

る。これらの関係を把握することによって文章の十全な意味を理解することができる。文章から特定の言葉を抽出するということは、この「島」だけを抽出することともいえる。様々な島をひとまとまりの日本海から取り出してきても、日本海に「ある島」があるのかどうかとか、およびその島数の多寡は判明しても、島と島とがどのような配置にあったかは、その言葉の位置情報を持たなければ不明である。もしここに、各抽出語の位置情報が正確にあるならば、言葉と言葉の共起や、距離なども正確に判定することができる。

しかしながら、KTCoder が抽出語の位置情報を必要とする意図を説明するのは、これが限界であって、そのような距離や共起位置を正確に持つ抽出語が、各研究者にとってどのように有効であるのかは、テキストを扱う各研究者の固有の問題ともいえる。

文章の階層構造を定義つけること

KTCoder 研究者は文章の階層構造を HTML タグによって定義付けなければならない。文章をよほど精密に扱わない限り、このタグ付けは最少の労力で済ませることもできるが、最小限<BODY>文章</BODY>という形態を必要とする。この HTML タグ付け形式は、現代における文章階層構造表現の最も単純な手法があると考えて、採用したわけである。文章に階層構造があることによって得られる利点は上述した。それは、抽出語の位置情報を、章節項目段落などの上下関係とともに得られるからである。このことによって、研究者が任意ないし文脈に沿って決定した文章階層構造中での、用語の位置関係を正確に把握できるので、他の出現用語との相対的な距離、あるいは層の深さによる絶対的な比較などが可能になる。しかし前述したが、これらの研究応用は各研究者の固有の問題ともいえる。本章では５に事例を記した。

２. KTCoder の仕組み

ここでは上述した設計思想を KTCoder システムによって具体的に説明する。

字種切り方式

日本語文章は、漢字、英字、数字、平仮名、カタカナなどによって記されている。字種切りによる用語の抽出とは、比較的簡単な操作（すなわち処理が高速になる）によって、この字種が異なる文字の前後で用語を抽出する方式である。たとえば『日本のコンピュータと社会学』ならば、「の」によって「日本」、「と」によって「コンピュータ」および「社会学」が切り出される。

3種類の辞書と、本文に対する辞書作用の順序

KTCoder は字種切り方式を使っているので、保守取り扱いや実行速度に関しては後述の形態素解析方式に比較して優れた面を持つが、このままではノイズが大きく利用が困難となる。よって、そのノイズを低減するために、研究者が簡便に操作できる辞書を用意した。

辞書は３種類あり、これらは内部でシステムが適切な順序のもとに使用している。その順序にしたがって辞書の性格を述べる。これらの辞書内容は一般的な標準テキストファイルであるが、特定のファイル名を付け、KTCoder システムと同一フォルダー内に置いておく必要がある。なお、これらの辞書引きはハッシュド・バイナリー方式で、辞書照合は最長一致方式を用いた。

(1)複合語辞書:　01 複合 NI.WRD

この辞書内容は最初に作用する。一般には字種切りに対応しない異字種による複合語を強制的に抽出するために設けた。たとえば「パターン認識」。これをパターンと認識とに分離させては語の持つ意味内容が変化する。しかし字種切り方式だけではこれを分離してしまう。あるいは「ふたおやさま」。これは平仮名だけの言葉であり、KTCoder は原則として平仮名文字列を棄却する。よって平仮名だけの、抽出しなければならない文字列は、パターン認識と同様に複合辞書に格納する。

最初はファイル名だけの空ファイルで試すのがよい。

(2)字種切り

この部分は、辞書を必要とせずプログラム内部で自動的に行われる。ここでは複合語辞書に登録されていない平仮名文字列が全て捨てられる。また、連続した漢字文字列やカタカナ文字列は、全て抽出される。

(3)停止語辞書:　02 停止 NI.WRD

この辞書内容は本文からの用語の抽出を禁じ、その語を捨てるためのものである。(２)の字種切りで抽出された用語は、すべてこの辞書内容と照合し、辞書と同一の用語があればそれは捨てられる。状況によるが、最初から不要と思われる「一般用語」をセットする場合もあるし、あるいは後述する、結果として出力される「抽出単語ファイル」や「棄却ファイル」を分析することにより、この停止辞書を定める場合もある。

最初はファイル名だけの空ファイルで試すのがよい。

(4)復活語辞書:　03 復活 NI.WRD

この辞書は漢字１文字を復活させるためのものである。KTCoder は平仮名文字列と単独の漢字を棄却する。たとえば文「明るい朝の光」に対して、複合語辞書に「明るい」、「朝の光」があればこれら二つの用語は抽出されるが、もしなければこの文から抽出される用語は皆無となる。しかし、復活語辞書に「光」といれておけば、「光」だけは抽出される。

この辞書も最初はファイル名だけの空ファイルで試すのがよい。

形態素解析方式との違い

現代の一般的な自然言語処理システムは、形態素解析という方法によって日本語文章を正確に品詞分析している。ここでは、その「形態素」に関する言語学的な考えに言及しておく。これは言葉の全体的な成り立ちを知った上で、本章での KTCoder の基本構造の限界と、有用性を把握するためである。

言語学一般では次のような説明がなされている、

> ［6］形態法・統語法
> 言葉の慣習に従って分節される単位、すなわち言語記号は、音素（音声）と意味の両面が表裏一体の関係で結びついたものであり、その最小の単位は、「形態素」(morpheme)あるいは「記号素」(moneme)と呼ばれる。つまり、形態素（記号素）と呼ぶ言葉の単位は音素の連続と最小の意味素から成り立ち、その成立は当該言語の音素体系と意味体系を背景としている。形態素は１つあるいはそれ以上が結合して単語を形成する。単語の集合が語彙である。（西田龍雄、1986、pp6-7）

この引用によって、形態素と単語、単語と語彙との関係が明らかになる。これに従えば、格助詞「て」「に」「を」「は」も、名詞「猫」もそれぞれ形態素である。一般に日本語で単語と呼ばれるものは、助詞「は」のようにそれ自体が形態素である場合もあり、幾つかの形態素が、「焼豚」のように結合して単語となる場合もある。むろん明瞭な意味を持った単語「人間」は形態素である。しかしここでは簡単に、形態素とは「なんらかの意味を持った言葉の最小の単位」としておく。

現代日本語の自然言語処理では、この形態素をある程度の精度で自動的に分別することができるようになった。これに従って、意味解析などの後の仕事を自動的に行うことが研究されている。およそ言葉を扱う世界では、情報図書館学の自動索引や、社会学における自動コーディングなど、こういった自然言語処理の成果が用いられてきた。

しかし自然言語処理の用途によっては、厳密な形態素や単語の概念ではなく、研究者が求める形での「用例」抽出に目的を持つ場合がある。「御」などを伴った「御蔵」のような接頭（ないし接尾）辞つきの用例とか、あるいは「ふたおやさま」のような文節、あるいは「きょうびの子供ら」のような句など、その求められる用例は多様である。こういった場合には、高度な形態素解析システムは却って使いにくくなるものである。

KTCoder は字種の違いによって用語を抽出する。日本語では漢字や英字やカタカナないし平仮名の違いという字種の異なりが、ある種の意味を持っている。たとえば日本語は一般的に概念語や具体物を漢字で表現することが多く、「子ども」と表記してある場合「子」と「子ども」とは相互に形態素といえるが、漢字「子」だけでも正しく概念を表している。これは漢字の持つ強力な象徴性によるものである。もちろん例外は山のようにある。子供ひとつとってみても、「子供」「子ども」「こども」「コドモ」というように表記は様々である。KTCoder の字種切り方式は「私のコドモです」という文に対して、辞書なしで「私」、「コドモ」の２用語を高速に抽出することができる。また辞書にもないような未定義語であっても、字種の異なりによって分離し抽出することができる。

字種切り方式は最良でも最善でもないが、開発の容易さ、未知語への対応能力、処理速度の高速さ（非常に長い文章、図書数冊であっても辞書引きが簡略なのと、基本アルゴリズムが単純なので、高速である）、メンテナンスの容易さなど、それなりに用途がある。今回の事例では、漢字だけの事項・名詞抽出も多くあり、また「すべてのことを受け入れる」などのような文近似単位での抽出も多数あり、こういった「形態素」に留まらない不特定文字列の大量抽出には KTCoder の高速性が生かされる。

3. 本文テキストの準備

KTCoder を利用するには、対象テキストに事前の手当を行っておかねばならない。この内のある部分は将来自動的に処理する方策を考えているが、現状では以下の２点に関して準備しておかないと利用できない。

(1)すべての文字を全角にしておく

これは現状の KTCoder が２バイト文字（全角、すなわち一般的にワープロで使用する日本語文字）を対象にして高速処理を実現しているので、当面はこの制限が残る。例外は２つある。

その１は特定の HTML タグと呼ばれる埋め込み記号である。これには<BODY>、</BODY>、<H1>、</H1>、<H2>、</H2>、<H3>、</H3>、<H4>、</H4>、<H5>、</H5>があり、総て半角でなければ対応しない。

その２は、偶数個の半角文字列であり、これは多くの場合システムが無視するので例外的に使用することもできるが、結果の保証は低くなる。特に半角のスペースは必ず全角スペースに変換しておく必要がある。

変換作業は、ワープロやエディターでの文字置き換えないし文字種の全角変換機能を利用することで行える。たとえば 20 世紀末のワープロとして、一太郎 v9 では、全角にする文字列をマウスの左ボタンなどで範囲指定し、次に［書式-全角］→［半角変換-全角に変換］を選択すればよい。また Word98 では、範囲指定した後、［書式］→［文字種の変換］の選択で可能である。

(2)HTML タグの利用により文章の階層を整理する

これは文章構造を把握できる程度に HTML 記号を埋め込む作業である。文章構造を全く無視する場合でも、必ず文頭に<BODY>と記し、文末には</BODY>を置かねばならない。<BODY>タグは、KTCoder の処理対象文章であることを意味しているので、なければ何もしないでシステムは終了する。

<BODY>
　今朝は明るくて暖かい日和です。これから遠足にいくつもりです。
　あなたもご一緒しませんか。
</BODY>

<BODY>～文章～</BODY>だけあれば、とりあえずは処理を行うが、この場合段落（改行で終了する文集合）の数が 99 段落に制限され、それ以上の段落番号は不定となる。よって、文の構造を把握し、適当な HTML タグによって文章全体を細分化することが必要である。

文の構造は一般的に、部・章・節・項・目の５段階が考えられる。KTCoder は５段階までの構造を認めている。ただし、実際には<H1>タグをタイトルに使用することを奨めるので、４段階までと考えるのが妥当である。以下の事例では、文構造把握のためにインデントを付けたが、インデントの有無は結果に影響を与えない。ただしインデントなどの空白部分は必ず全角スペースを使わねばならない。事例での～印は文を指すが、文の有無は結果に影響を与えない。極端な場合には、すべて見出し（HTML タグで囲まれた文）だけでも結果を出す。

ファイル名：　　0001.htm

```
<BODY>
<H1>文構造の把握事例</H1>
～
<H2>第１部</H2>
～
        <H3>第１章</H3>
        ～
                <H4>第１節</H4>
                ～
                        <H5>第１項</H5>
                        ～
                        <H5>第２項</H5>
                        ～
                        <H5>第３項</H5>
                        ～
                <H4>第２節</H4>
                ～
                <H4>第３節</H4>
                ～
        <H3>第２章</H3>
        ～
        <H3>第３章</H3>
        ～
<H2>第２部</H2>
～
<H2>第３部</H2>
～
</BODY>
```

この例は一般的な文構造を示したものであり、<H5>タグまでのすべてを使う必要はない。

次にこの（２）に関して補足的な説明をしておく。

位置情報(ID)と HTML 化

HTML タグを用いて文章の階層化をはかる理由は、文ないし用語の位置情報を構造的に得るためである。文の位置が階層的に把握できると、後々抽出した用語の文章内での出自が明白になる。これは部章節項目名で用いられた用語と、地の文で用いられた用語には差異があることから、場合によっては有効なことがある。この事例としては本稿での４（２）応用例「日本の文學史」を参照願いたい。

HTML タグを用いることにより、プログラム処理をした結果として得られる抽出単語ファイルや、文番号付きファイルには、３つのブロックからなる番号が付けられ、これは単語や文の位置情報を表している。

［事例：単語］

0400,0103000000,0000001,内外
0400,0103000000,0000002,ニュース
0400,0103010000,0000001,救う
0400,0103010000,0000002,宇宙
0400,0103010000,0000003,荘厳
0400,0103010000,0000004,元旦会法要
0400,0103010000,0100101,昭和六十年元旦
0400,0103010000,0100102,霊気
0400,0103010000,0100103,希望
0400,0103010000,0100104,新年

［事例：文］

0400,0103000000,00000,<H2>内外ニュース</H2>
0400,0103010000,00000,<H3>宇宙を救う荘厳にむけ　元旦会法要に誓いをこめる</H3>
0400,0103010000,01001,　昭和六十年元旦、澄みわたる霊気のなか希望にみちた新年は明けた。

ここで、「0400,0103010000,0100104,新年」を例にとってみてみる。

第１ブロックの「0400」とはファイル名から得た４バイト文字である。整数値でファイル名をつけていくと 0000 から 9999 の１万種類を区別できる。特殊な 16 進表記を用いれば、約６万５千を区別できる。ここには全角の日本語も使用できるが、総数が４バイトであるから、日本語では２文字までが限界である。たとえば「村 01.HTM」、「大学.HTM」、「AZ99.htm」などのように使える。この第１ブロックは、ファイル名の基底部から直接文字列が採取される。

第２ブロックの「0103010000」とは、先頭から２桁毎に意味をもち、おのおの H1～H5、すなわち章立て位置を示している。この数字列の場合なら、部としては 01 番目、部の中の章としては 03 番目、章内での節としては 01 番目、項・目については未定義となる。

第３ブロックの「0100104」は、抽出単語ファイルでは７桁あるが、文番号ファイルでは、最後の２桁の無い５桁となる。

例では、01｜001｜04 というように区分して考える。最初の 01 は、ある部章節項目内での段落番号であり、ここでは 01 番目の段落を意味する。次の 001 は同一段落内での文の順序数であり、この場合には 001 番目の文となる。最後の 04 とは、同一文内での抽出単語の番号を表しているが、これは文内での順番ではなく、同一文内で処理された順番である。すなわち、同一文内での同一用語の弁別を計るために設けたものであり、一般研究者にとっての意味は低い。

なおこの第３ブロックにおける先頭５桁が総てゼロの文および抽出単語は、その位置が見出し内であることを意味する。

文章の階層構造についての理解

H1～H5 の記号によって文章の階層関係を設定するのは、小さな範囲の中では容易にできるが、さまざまな文章を多数扱う際には、共通の階層関係を把握するのは不可能である。たとえば、私が一冊の図書をまとめるとき、私が自身でその図書に統一的階層を持った目次を付けるのは可能であるが、その図書と他人の書いた図書の目次による階層関係を同一レベルで比較するのは意味がない。すなわち、あるまとまりを持つ文章集合を処理する人が、そのまとまりという小さな世界のなかで首尾一貫して階層構造を確定するのは可能であるが、それを他の文章集合全部に当てはめるのは困難である。このことから、局地的な階層構造から得られる用語の位置関係は絶対的であり、他の階層構造（たとえば他の図書全文）と比較する場合には、それが相対的になるという理解に立つ必要がある。

４．KTCoder の画面構成

KTCoder の画面構成は、大きく４つに分けてあり、図２－１にみるように「（１）辞書窓」「（２）本文窓」「（３）抽出単語窓」「（４）文番号付き文窓」となる。

（1） 辞書窓

1. [全自動]すべての処理を一括して行う。最初はこのボタンだけを使用するのがよい。
2. [9:全終了]プログラムの終了。
3. [辞書読み込み]三種類の辞書をまとめて読み込む。
4. 各[辞書保管]各辞書内容を独立して保管する。

（2） 本文窓

1. [テキストを読む]任意のテキストを処理対象とする。
2. [テキスト上書保管]編集されたテキストを保管する。以前のテキスト内容は変更される。
3. [検索]テキスト内の文字列出現箇所を順次に検索する。

（3）抽出単語窓

1. [1:単語切り出し]
　単語切り出しを独立して行う。対象テキストは本文窓に表示されているファイルである。
　ただし、本文内容や各辞書に編集が行われているときは、それを各［辞書保管］ないし［テキスト上書き保管］した後でなければ、新たな内容を抽出することはできない。

（4） 文番号付き文窓

1. [2:ID 付き文作成]
　本文を、文番号付きの文単位で表示する。

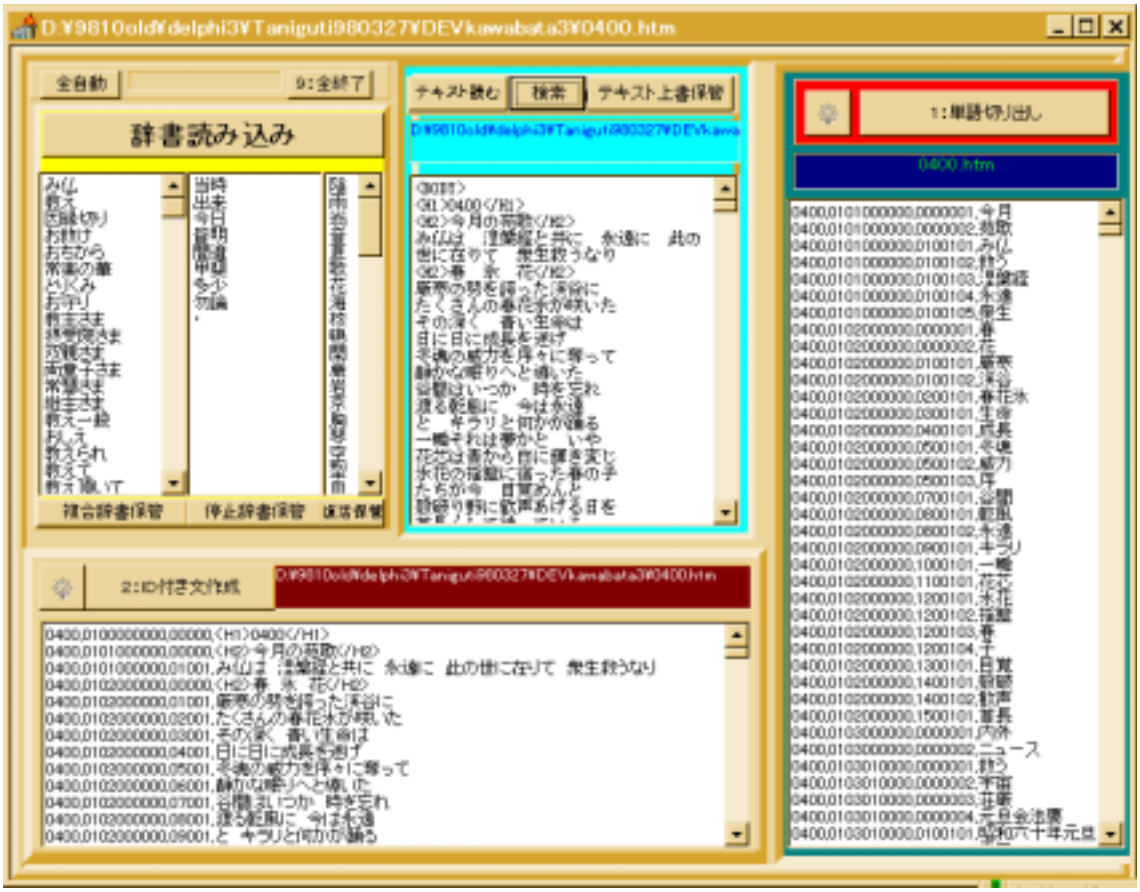

図２－１　KTCoder の画面構成

(1)辞書窓

この窓には、３種類の辞書内容が表示され編集も可能である。編集結果は各辞書の保管ボタンの押下によって決定される。各辞書に対する編集は、３辞書間（相互複製可能)、および本文窓、抽出単語窓、文番号付き文窓からの文字列を挿入できるので、ドラッグ＆ドロップ機能によって、効果的に編集することができる。

この窓が持つ機能として以下の４つがある。

１．［全自動］すべての処理を一括して行う。最初はこのボタンだけを使用するのがよい。辞書の読み込み、本文表示、抽出単語表示、文番号付き文表示までの一連の処理を総て行う。

２．［９：全終了］プログラムの終了。

３．［辞書読み込み］三種類の辞書をまとめて読み込む。辞書を修正し保管した後でこのボタンを押すと、KTCoder は新しい辞書内容を用いるようになる。

４．各［辞書保管］各辞書内容を独立して保管する。修正した各辞書を個々に保管し、［辞書読み込み］によって修正内容が有効になる。

このうち［全自動］ボタンは非常に有用で、一般的な処理はこのボタンを押し、ファイル名選択の問いかけに応じて、対象 HTML（本文）ファイルを選択するだけで処理は終わる。

(2)本文窓

１．［テキストを読む］任意フォルダーにある HTML テキストを処理対象とする。この場合テキストのあるフォルダーには３種類の辞書とシステムがなければならない。

２．［テキスト上書保管］編集されたテキストを保管する。以前のテキスト内容は変更される。テキストは、編集されたあと上書き保管することによって有効となる。

３．［検索］テキスト内の文字列出現箇所を順次に検索する。長い本文などで、目当てとする箇所を探す際に用いる。

(3)抽出単語窓

１．［１：単語切り出し］単語切り出しを独立して行う。対象テキストは本文窓に表示されているファイル内容である。ただし、本文内容や各辞書に編集が行われているときは、それを各［辞書保管］ないし［テキスト上書き保管］した後でなければ、新たな内容を抽出することはできない。

画面左側にある３つのブロックは、以下の通りである。

ファイル名（4バイト、0400)、
階層表示（00|00|00|00|00 は、左側から HTML タグの内、H1～H5 に対応する）
段落文および単語識別番号（00|000|00 は、段落番号、文番号、単語識別番号）

(4)文番号付き文窓

1.［2：ID 付き文作成］本文を文番号付きの文単位で表示する。ただし抽出単語窓での文番号とは異なり、ここでは第3ブロックが5桁になり、段落と文番号だけになっている（00|000)。この内容は「ファイル名.DAT」に保管されているので、他のソフトウェアに読み込み、別の処理をすることができる。読み込み方法は、「カンマ区切り読み込み」が多くのソフトウェアで対応している。以下の事例のように、各カンマで区切られたデータ内容の意味は、【ファイル名（4バイト)、階層表示（10 バイト)、段落・文（5バイト)、文（「。」ないし段落を終端とする比較的長い有限の文字列)】となる。

事例：　ファイル名　0400.Dat
0400,0103000000,00000,<H2>内外ニュース</H2>
0400,0103010000,00000,<H3>宇宙を救う荘厳にむけ　元旦会法要に誓いをこめる</H3>
0400,0103010000,01001, 昭和六十年元旦、澄みわたる霊気のなか希望にみちた新年は明けた。

(5)その他

現在の KTCoder(v0.5)におけるドラッグ＆ドロップは制限のある中で使用できる。制限とは、各窓で選択した文字列を別の窓に「移動」ではなく「複製」に限るということである。また4つの窓総てにおいて文字列の選択は、マウス左ボタンによるマウスドラッグが可能ではあるが、（3）抽出単語窓、（4）文番号付き文窓、の二つは書き込み禁止にしてあるので、この窓へのドロップはできない。

通常のマウス右ボタン操作および、Shift キーないし Ctrl キーによる複製・移動の操作には対応していない。

5. 事例と応用例

ここには KTCoder を用いた2つの事例をあげておく。

(1)事例「内外時報」

「内外時報」（川端亮、２部２章１節）四百号を対象にして、KTCoder の流れを記しておく。

１．「内外時報」400 号を WWW ブラウザで表示した例。HTML タグの作用で見出しの違いが見られる。このファイル名は 0400.htm である。

0400

今月の苑歌

み仏は　涅槃経と共に　永遠に　此の世に在りて　衆生救うなり

春　氷　花

厳寒の勢を誇った渓谷に
　　～中略～

内外ニュース

宇宙を救う荘厳にむけ　元旦会法要に誓いをこめる

　昭和六十年元旦、澄みわたる霊気のなか希望にみちた新年は明けた。　未明の真澄寺奥の院に清浄な大護摩の聖火はあがり、午前十時、真如教主さまのもと、両法嗣・真聰さま、真玲さまが第一精舎、第二精舎の導師座につき、元旦会法要を執行。おわって、教主さまのますますのご健祥を念じ、一層のご教導を願い新年のご摂拶を申しあげた。
～以下略～

２．ファイル 0400.htm を一般的なワードパッドやメモ帳でテキストとして編集対象にしている例。ここには、<BODY>などの半角 HTML タグが見られる。

```
<BODY>
<H1>0400</H1>
<H2>今月の苑歌</H2>
み仏は　涅槃経と共に　永遠に　此の世に在りて　衆生救うなり
<H2>春　氷　花</H2>
厳寒の勢を誇った渓谷に
～中略～
<H2>内外ニュース</H2>
<H3>宇宙を救う荘厳にむけ　元旦会法要に誓いをこめる</H3>
　昭和六十年元旦、澄みわたる霊気のなか希望にみちた新年は明けた。
　未明の真澄寺奥の院に清浄な大護摩の聖火はあがり、午前十時、真如教主さまのもと、両法嗣・真聰さま、真玲さまが第一精舎、第二精舎の導師座につき、元旦会法要を執行。おわって、教主さまのますますのご健祥を念じ、一層のご教導を願い新年のご摂拶を申しあげた。
～以下略～
</BODY>
```

３．あらかじめ設定しておいた複合語辞書ファイル(01 複合 NI.WRD）の内容。ただしこの内容は、川端亮が数度にわたり分析した結果辞書化したものであり、最初はこのようにまとまった量の辞書を使うわけではない。初めての際には、ファイル名だけの空白辞書が望ましい。

複合辞書の内容

教主さま
摂受院さま
双親さま
両童子さま
継主さま
雍主さま
常慧さま
大宇
創源
還着
済摂
歩み
あゆみ
歩ま
歩む
歩め
歩んで
あまねく
足もと
甘え
あるがまま
温か
命を
生かされる
いただく
いただいた
いただけた
至らな
慈しみ
一如の道
怒り
いかり
祈り
祈ら
祈って
容れる
受け入れ
うけいれ
怨む
嬉し
うれし
畢え
教え
おしえ
お心
お力
おつとめ
お救け
おたすけ
お役
お仕え
おつかえ
おまかせ
お任せ
お護り
お導き
思いあがり
お腹
おかげ
行い
限りない
重ね
賭け
還られ
還る
省る
キリスト教
きよめ
清め
浄め
浄ま
清らか
気持ち
厳し
苦しみ
こころ
心ぐせ
この身
定め
悟ら
悟り
悟る
悟ろ
捧ぐ
捧げ
支え
ささえ
親しく
識る
すべて
すなお
救い
救う
救って
救はで
すくわれ
救わ
救え
捨て切る
捨て切ら
捨て切り
捨て切れ
捨て
切っ
澄み
澄む
すばらし
すがた
添う
添い
添わせ
育てる
具え
譬え
正す
断つ
立て替え
足りな
救かる
救け
足りな
誓った
誓う
誓い
仕え
貫いて
貫か
貫き
貫い
貫く
貫け
尽くす
努め
積む
集い
とらわれ
執ら
執れ
執わ
とりくみ
とりくませて
取り組み
取り組ませて
尊さ
尊んで
尊い
尊き
尊く
悩み
になう
担われ
担い
担う
担って
念じ
遺された
はからい
ばく進
育んで
育まれ
運ぶ
運び
運ん
腹立ち
人任せ
人の心
ぼだい
奉ずる
迷い
まげて
曲げて
まがらない
曲がらない
まどわされ
まかせ
ま心

まこと	導き	結ばれ	歓び
み仏	磨き	結ぶ	歓んで
みほとけ	磨く	結び	歓ばす
み心	磨いて	結んで	喜ば
まこと	みかえり	むすぶ	喜び
まつり	みかえる	むすび	慶び
祀り	見返る	目ざめ	喜んで
護ら	見返り	滅し	悦び
護り	みつめ	滅する	弱い
護る	見つめ	滅って	依り処
み親	見きわめ	委ね	わだかまり
み心	導いて	浴し	わがまま
み旨	導く	邪な	
自ら	導き	よろこび	

４．停止語辞書ファイルの例（ファイル名、02 停止 NI.WRD）

内外	年頭	スリランカ	次第
ニュース	解説	当日	一番
今月	今度	ヨーロッパ	ドイツ
苑歌	ハワイ	両日	ロサンゼルス
〇〇	一人一人	インド	以上
今年	一度	以前	一瞬
本年	アメリカ	アジア	明日
昨年	何度	翌日	ベルギー
二人	同時	午後	タイ
一年	一方	カリフォルニア	翌年
司会	必要	イタリア	ニューヨーク
フランス	突然	パリ	二度
以来	文字	台湾	サンフランシスコ
一日	様子	最近	
今回	ページ	中国	

５．復活一字漢字辞書ファイルの例（ファイル名、03 復活 NI.WRD）

死　私　信　心　神　子　姿　業　声　咒　徳　我

６．ファイル名 0400.mei。全自動処理によって得られた「抽出単語」。

```
0400,0101000000,0100101,み仏
0400,0101000000,0100102,救う
0400,0101000000,0100103,涅槃経
0400,0101000000,0100104,永遠
0400,0101000000,0100105,衆生
0400,0102000000,0100101,厳寒
0400,0102000000,0100102,渓谷
0400,0102000000,0200101,春花氷
0400,0102000000,0300101,生命
0400,0102000000,0400101,成長
0400,0102000000,0500101,冬魂
0400,0102000000,0500102,威力
0400,0102000000,0700101,谷間
0400,0102000000,0800101,乾風
0400,0102000000,0800102,永遠
0400,0102000000,0900101,キラリ
0400,0102000000,1100101,花芯
0400,0102000000,1200101,氷花
0400,0102000000,1200102,揺籃
0400,0102000000,1200103,子
0400,0102000000,1300101,目覚
0400,0102000000,1400101,殻破
0400,0102000000,1400102,歓声
0400,0102000000,1500101,首長
0400,0103010000,0000001,救う
0400,0103010000,0000002,誓い
0400,0103010000,0000003,宇宙
0400,0103010000,0000004,荘厳
```

0400,0103010000,0000005,元旦会法要
0400,0103010000,0100101,澄み
0400,0103010000,0100102,昭和六十年元旦
0400,0103010000,0100103,霊気
0400,0103010000,0100104,希望
0400,0103010000,0100105,新年
0400,0103010000,0200101,教主さま
0400,0103010000,0200102,未明
0400,0103010000,0200103,真澄寺奥
0400,0103010000,0200104,清浄
0400,0103010000,0200105,大護摩
0400,0103010000,0200106,聖火
～後略～

７．ファイル名 0400.Dat。文番号付き文。

0400,0100000000,00000,<H1>0400</H1>
0400,0101000000,00000,<H2>今月の苑歌</H2>
0400,0101000000,01001,み仏は　涅槃経と共に　永遠に　此の世に在りて　衆生救うなり
0400,0102000000,00000,<H2>春　氷　花</H2>
0400,0102000000,01001,厳寒の勢を誇った渓谷に
～中略～
0400,0103000000,00000,<H2>内外ニュース</H2>
0400,0103010000,00000,<H3>宇宙を救う荘厳にむけ　元旦会法要に誓いをこめる</H3>
0400,0103010000,01001,　昭和六十年元旦、澄みわたる霊気のなか希望にみちた新年は明けた。
0400,0103010000,02001,　未明の真澄寺奥の院に清浄な大護摩の聖火はあがり、午前十時、真如教主さまのもと、両法嗣・真聰さま、真玲さまが第一精舎、第二精舎の導師座につき、元旦会法要を執行。
0400,0103010000,02002,おわって、教主さまのますますのご健祥を念じ、一層のご教導を願い新年のご摂拶を申しあげた。
0400,0103010000,03001,　輝かしい一年の出発に、教主さまより、まがらぬ信をつらぬき、和合の精進をもち総合道場をめざす　とご親教をいただき、一大荘厳へ邁進を誓った。
0400,0103010000,04001,　ついで、四日午前十時、怨親平等の初廻向が教主さまおん導師のもと執行され、顕幽一如の救いに浴し、先祖と喜びをともにした。
0400,0103010000,04002,なお、この日が新年の修行はじめとなっており、接心に仏智を磨き真剣な精進に踏み出した。
～後略～

８．応用例：エクセルによる処理

ファイル 0400.Dat の一部をエクセルに読み込んだもの。読み込み形式は、カンマ区切りである。カンマ区切りファイルであることをエクセルが判別できるように、Windows のエクスプローラ上で、0400.Dat の末尾に「.csv」を追加し、「0400.Dat.csv」と名称変更する。同ファイルをダブルクリックすると、エクセルを設定してある一般的なコンピュータシステムでは、データが自動的にカンマ区切りとして、エクセル内に取り込まれる。

表２－１　エクセルに読み込まれた文（ファイル 0400.Dat.csv)

ファイル	00|00|00|00|00	00|000	文
0400	0100000000	00000	<H1>0400</H1>
0400	0101000000	00000	<H2>今月の苑歌</H2>
0400	0101000000	01001	み仏は　涅槃経と共に　永遠に　此の世に在りて　衆生救うなり
0400	0102000000	00000	<H2>春　氷　花</H2>
0400	0102000000	01001	厳寒の勢を誇った渓谷に
			～中略～
0400	0103000000	00000	<H2>内外ニュース</H2>
0400	0103010000	00000	<H3>宇宙を救う荘厳にむけ　元旦会法要に誓いをこめる</H3>
			～後略～

同様の処理をファイル 0400.mei に適用した結果を次に示す。ファイル名は 0400.mei.csv である。

表２－２　エクセルに読み込まれた抽出単語（ファイル 0400.mei.csv）

ファイル	00|00|00|00|00	00|000|00	抽出単語
0400	0101000000	0000001	今月
0400	0101000000	0000002	苑歌
0400	0101000000	0100101	み仏
0400	0101000000	0100102	救う
0400	0101000000	0100103	涅槃経
0400	0101000000	0100104	永遠
0400	0101000000	0100105	衆生
0400	0102000000	0000001	春
0400	0102000000	0000002	花
0400	0102000000	0100101	厳寒
0400	0102000000	0100102	渓谷
0400	0102000000	0200101	春花氷
0400	0102000000	0300101	生命
0400	0102000000	0400101	成長
0400	0102000000	0500101	冬魂
0400	0102000000	0500102	威力
0400	0102000000	0500103	序
0400	0102000000	0700101	谷間
0400	0102000000	0800101	乾風
0400	0102000000	0800102	永遠
			～後略～

表はユーザー定義機能によって、先頭にゼロを付加し桁数を調整している。

（2）応用例「日本の文学史」

図書『日本の文學史』は評論家保田與重郎によって、昭和 47 年に新潮社から出版されたものである。ここで紹介する内容は KTCoder による文学テキストに対する索引作成や用語の散布をあつかったものである（谷口敏夫、1998b）。図２－２は、同図書全文中における「芭蕉」関連用語を位置情報付きで抽出したものを、エクセルによって散布図にしたものである。

表２－３　芭蕉用語集合

芭蕉(0)用語集合

総数:124　異なり:10

芭蕉	10101	芭蕉	160203	芭蕉	200402	芭蕉	210401
芭蕉	10110	芭蕉	160206	芭蕉	200403	芭蕉	210401
芭蕉	10210	芭蕉	170102	芭蕉	200403	蕉翁	210404
芭蕉	20208	芭蕉	180103	芭蕉	200403	蕉翁自筆	210404
芭蕉	20301	芭蕉	180304	芭蕉	200404	芭蕉	210405
芭蕉	50209	芭蕉	190201	芭蕉	200404	芭蕉	210406
芭蕉	50210	芭蕉	190205	芭蕉	200404	芭蕉	210406
芭蕉	50210	芭蕉	190308	芭蕉	200404	蕉門	210406
芭蕉	50210	芭蕉	200102	芭蕉	200404	蕉門	210406
芭蕉	50403	芭蕉	200102	芭蕉	200404	蕉門	210406
芭蕉	50406	芭蕉	200102	芭蕉	200406	蕉門関係	210406
芭蕉	70413	芭蕉	200102	芭蕉	200406	蕉門	210408
芭蕉	80102	芭蕉	200107	芭蕉	200407	芭蕉	230402
芭蕉	90302	蕉風	200202	芭蕉	200407	芭蕉	230402
芭蕉	90401	芭蕉	200203	芭蕉	200407	芭蕉	240205
芭蕉	100301	芭蕉	200203	芭蕉	200408	芭蕉	240205
芭蕉	110206	芭蕉	200203	芭蕉	200408	芭蕉	240205
芭蕉	120203	芭蕉	200203	芭蕉	200408	芭蕉	240205
芭蕉	120203	深川芭蕉庵	200206	芭蕉	200408	芭蕉	240205
芭蕉	120203	芭蕉	200206	芭蕉	200409	芭蕉	240205
芭蕉	120203	芭蕉	200206	芭蕉	200409	芭蕉	240205
芭蕉	120203	芭蕉	200206	芭蕉	200409	芭蕉	240205
芭蕉	120302	芭蕉	200206	芭蕉	200409		
芭蕉	130101	芭蕉	200206	蕉門	200409		
芭蕉	130101	芭蕉	200207	蕉門顕彰流布	200409		
芭蕉	130101	芭蕉	200213	蕉門無数	200409		
芭蕉	130103	芭蕉	200213	芭蕉	210102		
芭蕉	130204	蕉門	200213	蕉門	210102		
芭蕉	140204	蕉門	200213	江戸座蕉門	210103		
芭蕉	140204	蕉門	200213	芭蕉	210302		
芭蕉	140204	芭蕉	200401	芭蕉	210302		
芭蕉	140301	芭蕉	200401	芭蕉	210302		
芭蕉	160203	芭蕉	200402	芭蕉	210305		
芭蕉	160203	芭蕉	200402	芭蕉	210305		

異なり	頻度
芭蕉	107
蕉門	9
江戸座蕉門	1
蕉翁	1
蕉翁自筆	1
蕉風	1
蕉門関係	1
蕉門顕彰流布	1
蕉門無数	1
深川芭蕉庵	1

124

表２－３　芭蕉用語集合は、芭蕉およびその関連語が、図書全文中のどの部分に出現したかを現したものである。たとえば表中左端の最下部にある「芭蕉 160203」とは、先頭から２桁毎に章節段を表しており、16 章２節３段落に用例があることを示している。

図２－２　芭蕉散布図は、この表内容を散布図としたものであり、３つに分けてより詳細にしるしている。各表は縦軸に章節段をあらわし、「芭蕉（１）章」での 20000 とは 20 章であり、「芭蕉（２）20 章節」での 200400 とは 20 章４節、「芭蕉（３）20 章４節段落」での 200409 とは 20 章４節９段落を意味している。横軸は、各章節段内での用語集合の頻度である。この散布図は本論から外れるが、章節段内での最高頻度をたどり、最後の段落内容を調査対象とした実験の事例である。他に 20 数件の用語集合を散布図とし、各用語集合の散布図によるパターンの類似性も考究した。

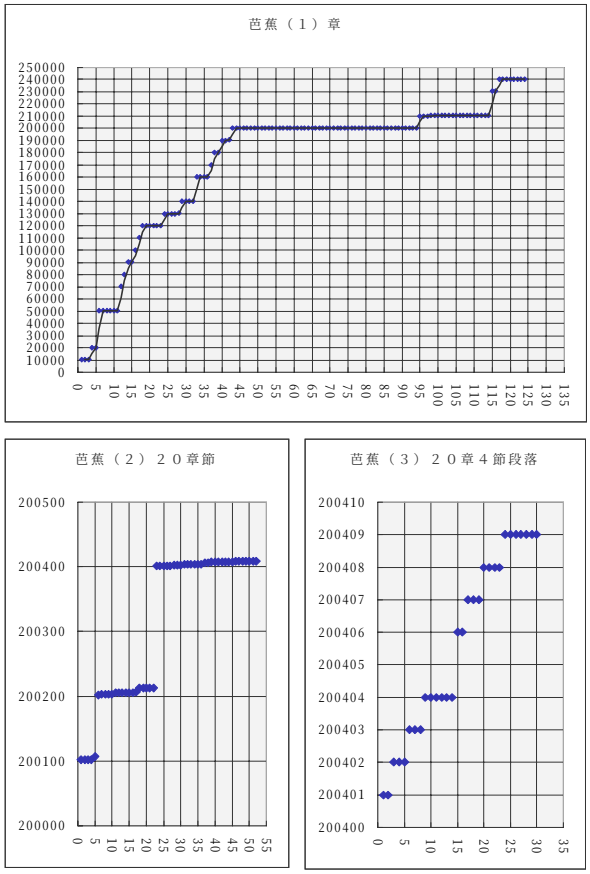

図２－２　芭蕉散布図

６ 改良と拡張構想

KTCoder(v0.5)は実用に供しているが、改良の余地がある。この点に関していくつかの論点をまとめておく。

辞書ファイル取り扱いの改良

辞書のファイル名を対象テキスト毎に変えられるようにすること。KTCoder は３種の辞書を持ち、これを標準テキスト扱いしているので研究者は辞書メンテナンスを簡便にできる。しかし、辞書を仕事（対象テキスト）毎に使い分ける研究者にとっては、一つの仕事に３種あるということは、仕事が増えるたびに辞書メンテナンスに複雑さや面倒さを味わうようになる。現在はフォルダーの利用による仕事毎辞書の分離を促しているが、これについては再考する必要がある。

辞書内用語の記述

辞書適用のケース区分や、用語のグループ使用を可能にすること。適用のケース区分とは、ある程度の正規化処理を可能にすることであり、たとえばある辞書にセットした用語について、ある条件のもとにある作用をするというようなルール作りを可能とすることなどである。また、用語のグループ処理とは、一つの代表語を辞書にセットすることにより、その語に関連する語をまとめて処理対象とすることなどである。一種のシソーラス処理であるが、シソーラスを利用者定義部分とシステム固定部分に分離する必要もある。以上の工夫については、本報告書佐藤の論考（１部１章）が参考になる。

KTCoder の出力データに関する取り扱い

抽出単語ファイル（例　0400.mei）や文番号付き文ファイル（例　0400.DAT）をどのように効果的に研究に役立てれば良いのかなどについて、詳しい説明文書をヘルプファイルとして内蔵させる必要がある。つぎに、SPSS やデータベース・アクセスなどとの連携に必要な組み込みツールも用意する必要がある。

文番号付き文ファイルの利用

特定用語（例「教主さまが」）が含まれている文集合に対して、それを２次、３次、ｎ次的に KTCoder で再分析できるようにすること。さらに用語の共起まで分析出来る方が実際の研究には役に立つ。こういった柔軟な構造を初期仕様の内に組み込む必要がある。これには、文番号のｎ重化など、テキスト中での絶対位置と、各フェーズでの相対位置設定の課題もある。

階層構造の自動設定

おそらく研究者の多くは、HTML タグによるテキストの階層構造記述に無駄や難渋を覚える可能性がある。これをある程度自動的に行える機能が必要である。現代のワードプロセッサソフトウェアや、その他テキストを扱うシステムには、HTML 文書化を自動的に行う機能を持っているが、階層構造の判別には多々問題もある。しかし、KTCoder がある程度の範囲で階層構造を自動化することは、必須のことであると考えている。

用語の位置情報（ID）

現在の KTCoder は、語の出現位置を３つのブロック（ファイル識別、階層識別、段落文）で表しているが、現状のままでは後の処理が複雑である。これは変換ツールを作成し、必要な部分だけを抽出したり、いろいろな基本的応用ソフトウェア（たとえば、エクセルやアクセスや SPSS）で即座に処理できるできるようにしておく必要がある。たとえば、文章の階層構造には一切考慮しない研究者にとっては、文番号や用語番号は、単純な生起順番号だけの方がよいかもしれない。

インターフェースの改良

現在の KTCoder は、一枚の大きなキャンパスに、各機能が定位置を占めた固定的な画面構成であるが、習熟してくるとこれは最良の使い勝手とはいえない。研究者の指示によって、各ウィンドウをすべて独立したものとして操作できるような機能が必要である。こういった些事は、しかし、実際の運用では大きな違いをもたらすので、考慮する予定である。

まとめ

KTCoder は社会調査のツールとして、まだ熟成度が高いとはいえない。だが単一目的に限った場合、基本的にパーソナルコンピュータの限界まで、ほとんど無制限の大量テキストに対して、妥当な正確さと速度で用語を自動抽出することができる。今後は、「６．改良と拡張構想」に記したような方向で、多くの研究者に利用していただき、不備を繕っていきたい。

謝辞

本稿を記すにあたり、大阪大学人間科学部川端亮氏、富山大学人文学部佐藤裕氏とのディスカッションに負うところが大きい。記して感謝する。

参考文献

川端亮、1998、「社会調査の歴史」高坂健次・厚東洋輔編『講座社会学　1　理論と方法』東京大学出版会、239-270 頁

佐竹秀雄、1987、「表記辞書と仮名／漢字変換」水谷静夫編『文字・表記と語構成』(朝倉日本語新講座;1)朝倉書店、76-108 頁

谷口敏夫、1994、「日本語文章における自動索引の試み」『光華女子大学研究紀要』第 32 号、光華女子大学、43-63 頁

谷口敏夫、1998a、『電子図書館の諸相』白地社

谷口敏夫、1998b、「日本語文章における重要語の出現位置に関する分析」『光華女子大学研究紀要』第 36 号、光華女子大学、19-98 頁

長尾真・谷口敏夫、1994、「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」長尾真他著『研究情報ネットワーク論』勁草書房、161-174 頁

西田龍雄、1986、「言葉のしくみ」西田龍雄編『言語学を学ぶ人のために』世界思想社、3-30 頁

開発環境

パーソナルコンピュータは、FM-V DESKPOWER T20M(富士通社製)で、Windows95/98 を使用した。内部メモリー64Mb、外部記憶装置 2Gb の標準的な AT 互換機である。開発ソ

フトウェアは Delphi４（インプライズ社）を使った。他に、エクセル 97（マイクロソフトウェア社）を応用例に使用した。

付録１：KTCoder(v0.5 9903)の簡単な使い方

内容は重複するが、本文中で述べた骨子を作業の流れに沿ってまとめておく。問題が生じた場合には、kawabata@hus.osaka-u.ac.jp（川端亮）あるいは taniguti@koka.ac.jp（谷口敏夫）にメールで問い合わせていただきたい。

１. HTMLテキストを用意する

適当なワープロやエディターで、対象文章を以下の用例に従い作成し、ファイル名末尾を「.htm」として保管する。ファイル名は４バイト（全角２文字）の固定長とする。ファイル名は文番号を構成する要素なので、正確に４バイトとする。

ファイル名例示

0400.HTM、2001.htm、原 32.htm

文章例示

<BODY>
<H1>0400</H1>
<H2>今月の苑歌</H2>
み仏は　涅槃経と共に　永遠に　此の世に在りて　衆生救うなり
<H2>春　氷　花</H2>
厳寒の勢を誇った渓谷に
たくさんの春花氷が咲いた
～
<H2>内外ニュース</H2>
<H3>宇宙を救う荘厳にむけ　元旦会法要に誓いをこめる</H3>
昭和六十年元旦、澄みわたる霊気のなか希望にみちた新年は明けた。
～
<H3>厳寒に信を磨く　五十回目の寒修行</H3>
きたる一月二十日より、昭和六十年度の寒修行がはじまる。
双親さまは昭和十年の暮れに運慶作と伝わる不動明王を勧請。翌年の正月早々から寒三十日の行を修め、み仏のみ心のまま、道ひとすじに立たれた。
～
</BODY>

(1) 文章の先頭に<BODY>を記入し、末尾に</BODY>と記す。
(2) HTML のヘッダー指定タグ（<H1></H1>～<H5></H5>）を使用して、文章の階層を決定する。階層構造はファイルのすべての文章で統一してもよいが、１ファイル内で完結した階層構造としてもよい。不特定多数の文章全体に階層関係を統一的に持たせることは不可能に近いが、調整のきく文章ならば一般的な「部、章、節、項、目」を全文章に設定し、それぞれに HTML タグの記号付けを行うのがよい。

(3)　HTML タグは半角英大文字数字によって、<H1>タイトル</H1>、<H2>大見出し</H2>、<H3>中見出し</H3>、<H4>小見出し</H4>、<H5>細区分</H5>、の５階層までとする。
(4) 各見だし記号の直後で改行し、地の文があれば見出しの次行から記す。文章は段落単位で改行する。目には見えない段落記号によって、KTCoder は段落を判定している。
(5) 上記で使用した記号以外は、すべて全角文字とする。スペースも全角スペースとする。
(6) 文章を保管したフォルダーに、KTCoder.EXE、01 複合 NI.WRD、02 停止 NI.WRD、03 復活 NI.WRD の４ファイルをセットする。

２. KTCoderを動かす

KTCoder.EXE をマウスで指示し、左ボタンをダブルクリックする。添付 CD-ROM には例として、KTCoder のショートカットアイコンがあるので、これをマウスクリックすればよい。

３. 全自動を指示する

KTCoder 画面左上部の[全自動]を左ボタンでクリックすると、３つの辞書を読み込む。

４. テキストを選択する

次に対象文章ファイル（例：0400.HTM、原３.HTM）を選択するように求めてくる。画面上でどれかを選択する。しばらく後に作業が完了する。

５. 同一フォルダーに置くファイルの説明

文章がおかれてあるフォルダーには、実行プログラム（KTCoder.EXE）と以下の３辞書が必要で、これらと HTML 文書ファイルが同一フォルダーにある限り、他の環境（フォルダー）から独立している。すなわち、３辞書とセットにすれば、多数のフォルダーに分散したり、場合によってはフロッピーに入れて別の Windows95/98 下へ移動することも可能である。これは KTCoder が Windows の複雑な環境から独立して使用できるからである。

01 複合 NI.WRD ： 複合語のための強制抽出辞書

「パターン認識」とか「ふたおやさま」のように字種切りでは抽出できない文字列を入れておく。

02 停止 NI.WRD ： 抽出しない用語の辞書

用語として抽出をさけたいものをこの辞書に入れておく。ただし、強制抽出辞書内容以外の用語を対象とする。

03 復活 NI.WRD ： 1文字漢字のための強制抽出辞書。

KTCoder は１文字独立漢字を捨てるので、捨てた内、強制復活させるためにある。

0400.HTM ： 対象文章ファイル（0400 はファイル名の一例）

全体が<BODY>と </BODY>とで囲まれた、HTML 文章の一種であり、HTML 記号は階層構造を定義するためだけに使われている。タグは<H1></H1>～<H5></H5>を、この範囲で使用する。

６. 出力データファイルの利用

KTCoder を実行したあと、さらに４種類のファイルができあがる。研究者が実際に利用対象とするのは、下記例での先頭２種類（0400.mei、0400.DAT)である。以下、各ファイル名先頭の 0400 は任意の４バイト文字である。

――

0400.mei ： 抽出単語（文字列）ファイル

これは、字種切りや辞書によって抽出された単語ファイルである。ただし本章及びシステム画面で「単語」という言葉を使っているが、これは正確には文字列といった方がよい。

［事例］

0400,0101000000,0000001,今月
0400,0101000000,0000002,苑歌
0400,0101000000,0100101,み仏
0400,0101000000,0100102,救う
0400,0101000000,0100103,涅槃経
0400,0101000000,0100104,永遠
0400,0101000000,0100105,衆生
0400,0102000000,0000001,春
0400,0102000000,0000002,花

――

0400.DAT ： 文番号付き文ファイル

これは、抽出単語を含む文のファイルである、下記事例により、先頭の４桁（**0400**）はファイル名から採られる全角２文字である。次の 10 桁（**0103020000**）は２文字単位で H1 から H5 までの階層を現している。仮に、部章節項目の５単位とするなら、この場合、第１部、第３章、第２節となり、末尾４桁の項と目は使っていないことになる。最後の５桁(**03001**)は、最初の２桁が段落番号（部章節項目内での)、末尾３桁が文番号（部章節項目内での）である。

［事例］ **0400,0103020000,03001, 本年は、数えて五十回目の寒修行である。**

0400.bun ： 文番号付き文作成予備ファイル

作業用のファイルであり、一般には利用しない。

――

0400.xxx ： 棄却ファイル

棄却ファイルの内容は、全文から抽出文字列を引き去った残りの文字列である。辞書類を整理し成長させる際に用いるのがよい。

［事例］

の は と 共 に に 此 の 世 に 在 りて なり 氷 の 勢 を 誇 った に たくさんの が 咲 いた その 深 く 青 い は 日 に 日 に を 遂 げ の を に 奪 って 静 かな 眠 りへと

付録２：KTCoder(v0.5 9903)と対象データの格納形式その他

Windows95/98 の環境のもとで、事前にいくつかの工夫をすることによって KTCoder を扱いやすくすることができる。

１．フォルダーの階層化

対象とする文章ファイル群は、フォルダー単位に管理し、そのフォルダー単位で４つの基本システムファイルを複製しておく。

フォルダー格納事例

C:¥Data
C:¥Data¥資料１
C:¥Data¥資料１¥KTCoder.EXE **基本プログラム**
C:¥Data¥資料１¥01 複合 NI.WRD **複合辞書**
C:¥Data¥資料１¥02 停止 NI.WRD **停止辞書**
C:¥Data¥資料１¥03 復活 NI.WRD **復活辞書**
C:¥Data¥資料１¥0001.HTM 文書ファイル群
～
C:¥Data¥資料１¥9999.HTM

C:¥Data¥原村資料B
C:¥Data¥原村資料B¥KTCoder.EXE
C:¥Data¥原村資料B¥01 複合 NI.WRD
C:¥Data¥原村資料B¥02 停止 NI.WRD
C:¥Data¥原村資料B¥03 復活 NI.WRD
C:¥Data¥原村資料B¥A01.HTM
～
C:¥Data¥原村資料B¥Z99.HTM

２．文章の事前整形と、非文のあつかい

事前整形

KTCoder v0.5 では、文章の自動整形は行っていないので、事前に Windows95/98 に備わっているメモ帳やワードパッドないし使い慣れたワープロなどで、次の原則に従って整形しておく。
文章はスペース、英数字を含めてすべて全角にする。
半角が必要なのは、
<BODY></BODY>,<H1></H1>,<H2></H2>,<H3></H3>,<H4></H4>,<H5></H5>
の文字列、および偶数文字半角（偶数個の連続した半角文字列：aa00，123456，00）だけである。この条件にあわない、たとえば半角奇数個のスペースなどがあると、結果がでない。

非文ないし非文章のあつかい

現代文で非文と呼ばれる文の体裁をなさない文、すなわち意味不明の未定義語の連続や、状況情報がなく日本語文法を大きくはずれた会話文や、古語雅語俗語が入り交じった文章は、あらかじめ辞書を用意しても無駄になることが多い。

こういった文章に対しては、最初辞書を空のまま使用することが望ましい。フォルダーに空で名前だけの複合、停止、復活の３辞書を置いておく。
実行の後、棄却ファイルと抽出単語ファイルとを十分に検査し、その上で複合辞書ファイルを追加していくと作業がはかどる。

３．辞書の処理の流れ

KTCoder は、内部で次のような順序で言葉を処理している。
(1) 最初に**複合辞書**と文章を相互参照し、複合辞書ファイルにある文字列は強制的にすべて抽出する。
(2) 任意の連続した漢字文字列（ないしカタカナ文字列）を自動抽出する。
(3) この漢字文字列のうち、**停止語辞書**に含まれる漢字文字列を排除する。
(4) 棄却された文字列のうち、**復活辞書**にある単漢字を復活し抽出する。