文字列解析ツールe-CSA ver.1.00 : 国語学・国文学研究者用マニュアル

南里, 一郎
純真女子短期大学国文科助手

竹田, 正幸
九州大学大学院システム情報科学研究院助教授

福田, 智子
九州大学大学院人文科学研究院助手

https://doi.org/10.15017/8971

出版情報：文獻探究. 42, pp.118-132, 2004-03-31. 文献探究の会
バージョン：

文字列解析ツールe-CSA ver.1.00
国語学・国文学研究者用マニュアル

南 里 一 郎　　竹 田 正 幸　　福 田 智 子

要旨

文字列解析ツールe-CSA“efficient character string analyzer，イークサ”は，テキストデータを単なる文字の連鎖として扱う立場で開発した，汎用のソフトウェアツールである。本稿では，とくに国語学・国文学研究者による利用を想定して，e-CSAの特長と使用法を解説する。

０．研究とツールの関係について

（ツールの使用法だけを知りたい方は，次節からお読みください。）

国語学・国文学の研究の基本は，まず「読む」ことである。文献をもとに研究する人はその作品や資料を手に取り「読ん」で分析する。フィールドワークの人は，インフォーマントのことばを録音し，それを「聴い」て分析する。そのような地道な作業が，すべての研究の基礎となり，研究者のカンを育てていくものであろう。

しかし，一人の研究者の人生は，あまりにも短い。20歳で学部に入り，大学院を経て，国公立大学（法人化で姿を消すそうである）に着任，退官後私立大学に移り，定年退職を70歳と考えると，50年しかない。そのうちの４分の１か３分の１は睡眠時間であるし，会議や学内の雑用，家事や趣味，病気や怪我などに取られる時間を差し引くと，研究に費やせる時間は，正味15～20年といったところではないだろうか。

では，どうやって研究に費やす時間を捻出すればいいのであろうか。第一著者の実家は自作農であるが，親がつねづね言うことには，「仕事は人手と道具やけん」であった。人手に関して言えば，科学研究費などを得て全国的に研究者を募り，大規模に共同研究するような例があろう**（注１）**。また道具（ツール）に関して言えば，おおよそ次のような捉え方ができないだろうか（文献研究の場合）。

(1) 原典の時代

写本や版本を入手できる限られた人たちだけが研究していた時代。定家や宣長。

(2) 影印本の時代

写本や版本の複製が入手できるようになり，研究者の裾野が広がった時代。

(3) 活字本の時代

ツール第１期。読みやすい活字本が出回るようになった時代。岩波書店の全102冊『日本古典文学大系』（旧大系）はその金字塔であろう。変体仮名やくずし字があまり読めなくても，それなりに研究できてしまう。時期的には，上記の(2)と相前後したか。

(4) 索引の時代

ツール第２期。さまざまな作品で自立語索引や総索引が作られ，研究の糸口が飛躍的に見

いだしやすくなった時代。1966年生まれの第一著者はこの世代。『古今集総索引』（西下経一・滝沢貞夫編，明治書院）で用例を拾って演習の発表をした。当時，便利なものがあるものだ，と感じていた。

(5) 電子テキストとPC-9801の時代

ツール第３期。パソコンといえば，NECのPC-98シリーズであった。整備され始めた電子テキストを全文検索して用例を収集していた時代。当時，修士課程に在籍していた第一著者は，先輩**（注２）**の手による「けんさく君」なるプログラムを使いたいがために，授業料を滞納して，NEC PC-9801NS/R**（注３）**を購入した。1994年10月のことであった。

(6) 発見科学とe-CSAの時代

ツール第４期。［文部省科学研究費補助金特定領域研究(A)「巨大学術社会情報からの知識発見に関する基礎研究」(平成10～12年度，領域代表者：有川節夫)］が開始され，計算機科学における「発見科学（Discovery Science)」という新しい研究領域が登場した時代。計算機プログラムによって研究者の発想を支援し，発見の糸口となりうるヒントを提示することを目指した。この思想に基づいて設計されたツール“e-CSA”は，「検索」という枠から脱却し，「抽出」という形で，国語学・国文学研究者の発想を支援する。

人類はその英知によって，時間短縮（人生の擬似的延長とも）のためのさまざまな道具（ツール）を編み出してきたが，国語学・国文学の分野に限れば，上記の「ツール第４期」に至って新たな局面を迎えたと言えるであろう。e-CSAは，機械に考えさせて研究者が楽をしようという発想で生まれたツールではない。あくまでも研究の主体は人間で，それを支援する計算機プログラムなのである。機械には語学も文学も解らない。「道具」という意味では，万年筆と同じなのである。

１．e-CSAの概要

1.1　どんなツールか

文字列解析ツールe-CSA“efficient character string analyzer，通称「イークサ」”は，テキストデータを単なる文字の連鎖として扱う立場で開発した，汎用のソフトウェアツールである**（注４）**。

通常のキーワード検索ツールでは，使用者がキーワードをシステムに与えると検索が行われる。言い換えれば，使用者が具体的なキーワードを思いつかなければ，検索はできない。e-CSAは，そのキーワード探しの段階で，有用なキーワード発見を支援する。すなわち，これは，テキストデータにおける，あらゆる長さの，あらゆる部分文字列について，その生起頻度を計数し，その結果を効果的に表示する機構を備えることで，研究者の発想を支援するツールなのである。

1.2　e-CSAの特徴（興味のある方だけお読みください）

(1) 自然言語処理の手法を用いない

日本語の表記法には明確な単語の区切りがないので，単語ごとの統計頻度表を作成するた

めには，テキストを単語に分割する作業が必要となる。しかし，この作業には膨大な労力が必要である。自然言語処理技術を用いて自動的に単語分割を行う方法もあるが，精度が十分でないため，人手による修正作業は避けられない。よって，e-CSAでは，この手法を採らない。

(2) *n*グラム解析の手法を用いない

単語分割をいっさい行わず，テキストデータを単なる文字の連鎖として扱う文字列分析の立場がある。最近にわかに注目されている*n*グラム解析**（注５）**も，そのひとつである。ここでいう*n*グラムとは，長さ*n*の任意の部分文字列をさす。この*n*グラム解析では，扱う部分文字列の長さ*n*を，２，３，４，というように固定するのが普通である。だが，テキストの部分文字列の生起頻度を調査する際には，*n*の値をひとつに定めるわけにはいかないので，２，３，４……といった複数の*n*の値について処理を行うことになる。その結果，それぞれに得られた複数の部分文字列リストを集積すると膨大な量となり，利用者の作業効率を著しく低下させる。よって，e-CSAでは，この手法を採らない**（注６）**。

(3) Blumer(1987)らの文字列上の同値関係を利用する

e-CSAでは，上記のような問題を避けるため，*n*グラム解析を用いず，「同値」の部分文字列をグループにまとめてグループ単位で提示することにした。以下，一例を示そう。

『古今和歌集』において，「は・る・か・す・み」という文字列は21例存する。では，「は・る・か・す」「る・か・す・み」「る・か・す」という文字列はどうかというと，やはり各21例である。すると，これら三種類の文字列は，例外なく「は・る・か・す・み」の一部ということになる。このような関係を「同値」といい，e-CSAでは，同値の文字列をグループにまとめて示す。ちなみに，「は・る・か」は27例，「か・す・み」は31例が存在するので，「は・る・か・す・み」の部分文字列でない用例が，前者には6例，後者には10例あるということになるため，別グループとして扱う。

情報科学では，このグループを「同値類」と呼ぶ。Blumer（1987）らによって導入された文字列上の同値関係に基づくものである。長さmのテキストの部分文字列の個数は，mの二乗に比例するが，同値類の個数は，mに比例することが知られている。同値類にまとめることは，データ検証の能率化に効果的である。

なお，e-CSAでは，同値類の表示に際し，最長文字列と，極小部分文字列だけを並べて示した。●を付した文字列が，その同値類内の最長文字列である。

1.3　e-CSAの特長（必ずお読みください）

e-CSAにできることを，端的に述べると，

(1) あるテキストデータから，部分文字列を抽出し，どのような文字列が何回出現するか（出現頻度）を提示する。

(2) 複数テキストでの共通文字列の出現頻度を比較する。

ということである。

では，この機能を，どのように活用することができるか。たとえば，物語と歌集との間で，生起する共通文字列を抽出し，提示された結果を検討することによって，従来指摘されていなかった引歌が発見できる可能性があることはもとより，そこから，両者の表現の関連など

について，新たな研究の糸口を見いだすこともできよう。テキストは，研究者の目的や発想次第でさまざまなものが想定できる。抽出された文字列自体は，意味のない単なる文字の連鎖であるが，ある出現傾向を見いだせれば，研究者にとって有用な「意味」を持ってくるのである。それゆえ，作品相互の影響関係，成立年代，作者の同定，語彙研究，文法研究，文体研究，さらに，音声や音韻を文字化できれば音韻研究など，テキストデータさえあれば，様々な研究目的に応じて使用できる。

こうして，決して長くはない研究人生の間に視野に入れることができるデータ量を，飛躍的に増加させることが可能となった。もちろん，「読む」という研究の基本を軽視するわけではない。最後は必ず原典に戻る必要があろう。けれどもやはり，作業の能率化を図るための「道具」の使用も考えてみてよいのではないだろうか。総索引の見出し語を目で追いながら，漠然とした問題意識を具体化していくように，このツールで抽出した，一見無意味な文字列の中から，研究のきっかけを掘り起こすのである。文字列を閲覧しやすく整理した上で提示するこのツールは，研究者のカンを刺激するに足る機能を備えている。裏返せば，これまで以上に研究者の発想力が問われるツールなのである。

1.4　動作環境（必ずお読みください）

ＯＳ　　：日本語版Windows98/98SE/ME/2000/XPで，動作確認ずみ。

ＣＰＵ　　：動作周波数は，500MHz以上を推奨する。これより低速でも動作自体には支障ないが，高速であればあるほど抽出時間が短くてすむ。

メインメモリ：512MB以上を推奨する。他のソフトウェアを同時に起動しないのであれば，256MBでも問題ない。

ディスプレイ：1024×768以上の解像度であること。800×600でも動作自体には支障ないが，ウインドウがはみ出すため，操作がしにくい。

文字コード　：Shifted-JIS（SJIS）を想定している。Unicode，および外字には対応していない。

1.5　使用許諾条件など（必ずお読みください）

(1) 配布と使用について

e-CSAの配布は無償とする。ただし，開発者側でユーザを把握しておきたいので，使用される方は，電子メールなどでご連絡いただきたい。メールには，お名前，ご所属，現在のご専門，使用目的（具体的なものでなくとも，「……で試したい」などでも可）を明記されたい（メールアドレス：takeda@i.kyushu-u.ac.jp)。折り返しインストールに必要なパスワードをお知らせする。また，e-CSAを用いて得られた研究成果を学術雑誌などに発表される際には，これを使用した旨を明記されたい。

(2) 著作権

e-CSAの著作権は，開発者の竹田正幸が保有する。

(3) 免責事項

e-CSAを使用して出力した統計データの誤り，および，それによる学術的な意味での損害に

ついて，開発者は一切の責任を負わない。また，本ソフトウェアを使用することで生じたその他のあらゆる損害についても，開発者は一切の責任を負わない。

(4) バグの報告，機能拡張に関する要望

バグにお気づきの際は，ご一報賜れば幸いである。開発者側でも同じバグが再現できるように，状況を詳しくお知らせ願いたい。また，拡張機能に関するご要望も歓迎する。

２．e-CSAの入手とインストール

2.1 入手方法

入手の手順は以下のとおりである。

(1) インターネットに接続し，次のURLのWebページを表示する。

URL　http://www.i.kyushu-u.ac.jp/~takeda/software

(2) リンクをクリックし，ソフトウェアをダウンロードする。ダウンロードするファイルの一時保存場所には，デスクトップを推奨する。

(3) ダウンロードしたファイルは，自動解凍書庫になっている。ダブルクリックすると，解凍され，インストールが開始される。

2.2 インストール

コンピュータにインストールする手順は，以下のとおりである。

(1) 電子メールで，開発者の竹田に対してパスワードの発行を請求する。「氏名」「所属」「現在の専門」「このソフトウェアの使用目的（具体的なものでなくとも「……で試したい」などでも可）」を明記のこと（メールアドレス：takeda@i.kyushu-u.ac.jp）。

(2) ダウンロードして入手したファイルをダブルクリックして，インストールを開始する。途中でパスワードの入力が求められるので，発行されたパスワードを入力する。指示に従ってインストールを完了する。

2.3 起動

左下のスタートボタン→プログラム→eCSAとたどり，クリックすると起動できる。使用の便のために，適宜，デスクトップやクイック起動バーなどに，起動のためのショートカットを作成されたい。ショートカットの作成方法については，Windowsのマニュアルやヘルプを参照のこと。

2.4 アンインストール

コントロールパネルの「アプリケーションの追加と削除」でアンインストールできる。

３．使用方法（マニュアル）

3.1 テキストファイルの準備

次に挙げた『うつほ物語』俊蔭巻の冒頭の一節を使って，e-CSAで処理するテキストファイルの作り方を説明する。

むかし、式部大輔左大弁かけて、清原の王ありけり。皇女腹に男子一人持たり。その子、心のさときこと限りなし。父母、「いとあやしき子なり。生ひ出でむやうを見む」とて、書も読ませず、いひ教ふることもなくて生ほし立つるに、年にもあはず、たけ高く、心かしこし。

七歳になる年、父が高麗人にあふに、この七歳なる子、父をもどきて、高麗人と詩を作り交はしければ、おほやけ聞こしめして、あやしうめづらしきことなり。いかで試みむと思すほどに、十二歳にてかうぶりしつ。

帝、ありがたき才なり。年の若きほどに試みむ、と思して、唐土に三度渡れる博士、中臣門人といふを召して、難き題を

（新編日本古典文学全集『うつほ物語』第１冊19頁）

上記の部分を，次のような形式でテキストファイル化した。

<1-019-01>むかし、しきぶのたいふさだいべんかけて、きよはらの
<1-019-02>おほきみありけり。みこばらにをのこごひとりもたり。
<1-019-03>そのこ、こころのさときことかぎりなし。ちちはは、
<1-019-04>「いとあやしきこなり。おひいでむやうをみむ」とて、ふみも
<1-019-05>よませず、いひをしふることもなくておほしたつるに、としにも
<1-019-06>あはず、たけたかく、こころかしこし。
<1-019-07>ななとせになるとし、ちちがこまうどにあふに、このななとせなるこ、ちち
<1-019-08>をもどきて、こまうどとふみをつくりかはしければ、おほやけきこ
<1-019-09>しめして、あやしうめづらしきことなり。いかでこころみむとおぼ
<1-019-10>すほどに、じふにとせにてかうぶりしつ。
<1-019-11>みかど、ありがたきざえなり。としのわかきほどにこころみむ、とおぼして、
<1-019-12>もろこしにみたびわたれるはかせ、なかとみのかどひとといふをめして、かたきだいを

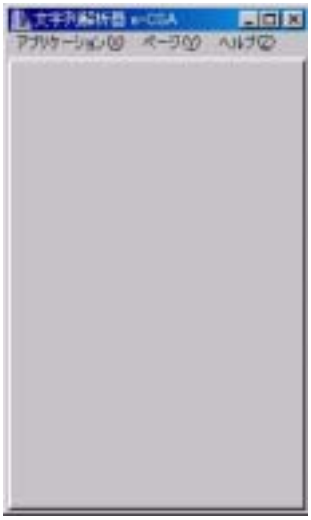
e-CSAは，ファイル中の改行コードを無視して処理する仕様になっているので，テキストファイル作成時には，見やすいように適宜改行を入れておく。また，後述するが，改行コードと同様にユーザが指定した文字や文字列（たとえば，句読点など）を無視して処理することもできる。

また，テキスト中のどこにでも，ラベルを貼り付けることができる。この『うつほ物語』の例では，新編日本古典文学全集における「冊数-頁数-行数」を< >で囲み，行頭に置いた。このようにラベルを埋め込んでおけば，用例をKWIC（KeyWord In Context）表示する際には，当該文字列の生起箇所を知ることができる。

e-CSAは，テキストを単なる文字の連鎖として扱う。したがって，「花」「華」「はな」といったような表記の揺れがある場合，それらは全く別の文字列として処理されることになる。表記を統一する作業は，必要に応じて，テキストファイル作成時にユーザの責任で行っていただきたい**（注７）**。ここでも，先の漢字仮名交じり文を，すべて仮名に直して，表記を統一した。ただし，片仮名から平仮名への変換や，清濁の区別の有無など，機械的に変換可能なものには，ある程度対応している（3.4 使用の具体例(14)参照）。

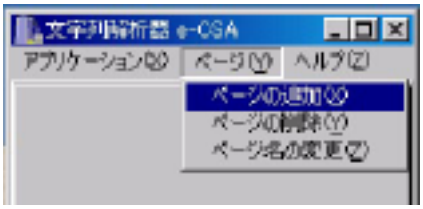
3.2 用意したテキストファイルの登録

さきほど用意した『うつほ物語』冒頭を登録する。ここでは，これらのファイルはすべて，「D:\Text」（Dドライブにあるtextフォルダ）の中にあるものとする。なお，ファイルを指定するこのような道筋を「パス」と呼び，Windowsにおける表記のルールである。これに関しては，本稿末尾の「パスについて」を参照されたい。



(1) e-CSAを起動すると，画面左側に制御ウインドウが現れる。

(2) メニューから「ページ」を選び，その中の「ページの追加」をクリックする。

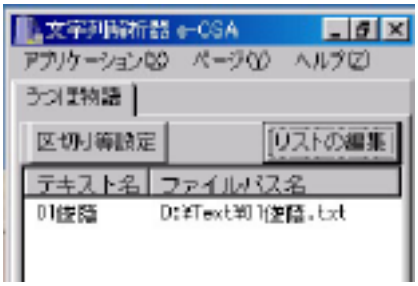


すると，「ページの追加」ダイアログボックスが現れる。

(3) ページ名を，たとえば「うつほ物語」というように入力する。入力し終わったら「OK」ボタンをクリックする。

(4) 制御ウインドウ上にページが現れる。ページタブには，先ほど入力したとおり，「うつほ物語」と書かれている。これから，このページに処理の対象であるテキストファイルを登録していくことになる。

(5) テキスト名とそのファイルのパスを登録する。ここでは，「テキスト名」と書かれたエディットボックスに「01俊蔭」と入力し，その下の「ファイルパス名」と書かれたエディットボックスに，対応するファイルのパス名「D:\Text\01俊蔭.txt」と入力する。パス（ファイルの所在）が判らない場合には，「参照」ボタンをクリックして，登録したいファイルを探す。テキスト名とファイルパス名を入力し，「追加」ボタンをクリックすると，入力したテキストファイルがリストに追加される。



(6) 「OK」ボタンをクリックすると，ファイル登録が完了する。

3.3 テキストの「見方」の設定

テキストごとに「見方」を設定する。同一のファイルでも，ここの設定内容を変えることにより，異なる分析の視点が得られる。

(7) まず，「区切り等設定」ボタンをクリックする。すると，「区切り文字列等の設定」ダイアログボックスが表示される。

(8) 区切り文字（列）の指定

たとえば，句点「。」など，テキストを区切るために用いたい文字（列）を指定する。テキスト文字列は，区切り文字列の前後で分断され，それをまたいだ文字列は計数されない（KWIC表示においては，これらの区切り文字列も表示される）。句点「。」をエディットボックスに入力し，「追加」ボタンを押すと，リスト中にその「。」が現れる。

(9) 「無視する文字列群」の指定

たとえば，読点「、」など，テキストファイル中には存在するが，ないものとして処理したい文字列を指定する。指定の方法は，「区切り文字列群」の指定と同様である（それらの文字列は，処理時にテキスト文字列から削除するため，KWIC表示時にも表示されない）。

(10) 「ラベル文字列」の指定

テキスト中にラベルを埋め込んだ場合には，「ラベル文字列」の欄の「指定する」をチェックし，ラベルの先頭と末尾の文字列をそれぞれ入力する。ここでは，先頭に“＜”を，末尾に“＞”を，それぞれ指定する。

(11) 「対象文字列」の指定

これは，例に示した『うつほ物語』のファイルでは不用である。チェックをはずしたままでよい。この機能は，次に示すような形式の歌集の和歌本文のテキストファイルを扱うことを想定したものである。

```
<D>
<R>
<A>二巻　4古六帖　</A>
<N>　　1</N>
<P>[年のうちに][春はきにけり][一とせを][こぞとやいはん][ことしとやいはん]</P>
<K>[としのうちに][はるはきにけり][ひととせを][こそとやいはむ][ことしとやいはむ]</K>
</R>
<R>
<A>二巻　4古六帖　</A>
<N>　　2</N>
<P>[袖ひちて][むすびし水の][こほれるを][春たつけふの][風やとくらん]</P>
<K>[そてひちて][むすひしみつの][こほれるを][はるたつけふの][かせやとくらむ]</K>
</R>
　……中略……
<R>
<A>二巻　4古六帖　</A>
<N>　827</N>
<P>[つれづれの][はるひにまよふ][かげろふの][かげ見しよりぞ][人は恋しき]</P>
<K>[つれつれの][はるひにまよふ][かけろふの][かけみしよりそ][ひとはこひしき]</K>
</R>
<R>
<A>二巻　4古六帖　</A>
<N>　828</N>
<P>[てにとれど][たえてとられぬ][かげろふの][うつろひやすき][君が心よ]</P>
<K>[てにとれと][たえてとられぬ][かけろふの][うつろひやすき][きみかこころよ]</K>
</R>
</D>
```

ここでは，『古今和歌六帖』の和歌本文のテキストを示した。上段（<P>と</P>に挟まれた行）は，『新編国歌大観』本文に拠り，下段（<K>と</K>に挟まれた行）は，それを清音平仮名表記に改めたものである。

上のファイルにおいて,和歌を清音仮名で表記した部分だけを処理の対象としたい場合,「指定する」にチェックを入れた上で,「先頭」に「<K>」を，末尾に「</K>」を，それぞれ指定する。こうすると,「<K>」と「</K>」に囲まれた部分以外は処理の対象外となる。

なお，このデータでは，５-７-５-７-７の各の句を，おのおの[　]で括ってある。そこで，“[”“] ”を「区切り文字列群」に指定した場合には，文字列の計数は，それぞれ句ごとに行われる。逆に,「無視する文字列群」として指定すると，１首全体をまとめて１つの文字列として扱うので，複数の句にまたがった文字列も計数されることになる。

以上，(8)～(11)の指定が終了したら「OK」ボタンをクリックする。なお，この作業は，ページごとに行う。したがって，異なったタグの付け方をしたテキストファイルは，別ページに登録しておくことになる。(10)(11)の設定内容は保存されるので，次回から設定の必要はない。もちろん，変更は可能である。

3.4 使用の具体例

(12) 部分文字列統計プログラムの起動

メニューバーの「アプリケーション」→「文字列統計」の順に選択すると，作業ウインドウが開く。本ソフトウェアは，テキスト群ＡとＢに対し，それぞれにおける文字列の生起頻度を比較することを基本機能とする（単独のテキストファイル群における生起頻度を見る場合には，作業ウインドウ右上のラジオボタンで「テキストＡ群における生起頻度を調査」を選ぶ)。

(13) テキストファイルの選択

登録したテキストファイルの中から，処理の対象とするファイルを選び，テキスト群Ａ，Ｂとする。すなわち，制御ウインドウ内でページを選び，その中のテキストファイルを選択し，その状態で作業ウインドウの「追加」ボタンを押す。テキストファイルは，複数選ぶことができる。テキストが連続して並んでいる場合は，Shiftキーを押したままテキスト名をクリックすると，複数のテキストファイルを一括して選択できる。また，表示上連続していないテキストを一括して選択したい場合は,Ctrlキーを押したままテキスト名をクリックする。

(14) 清音かな表記への変換指定

作業ウインドウ上部の「清音かな表記へ変換」をチェックすると，カタカナ・ひらがなについては，すべて清音ひらがなへ自動的に変換する。その際，濁点・半濁点の情報は失われる。

(15) 部分文字列統計の計算

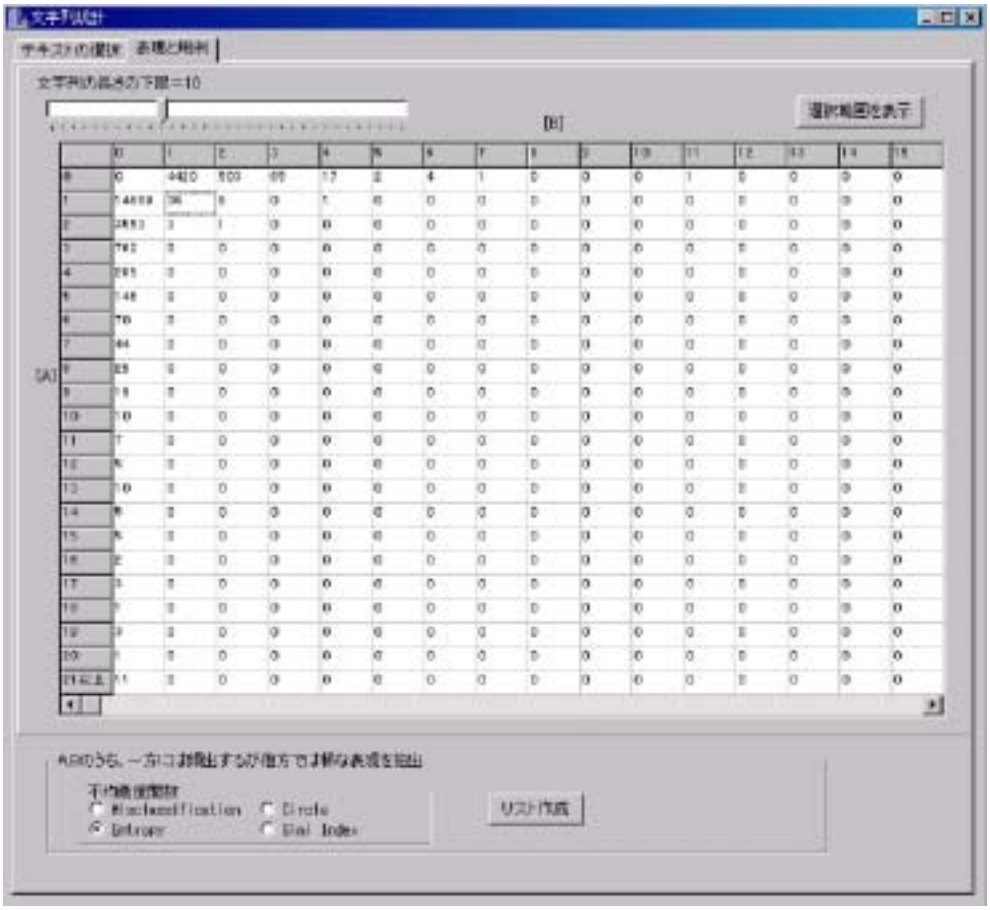
「実行」ボタンを押すと，(13)で選択したテキストのすべての部分文字列に対する生起頻度の計算が開始される。計算が終了すると作業ウインドウ上で「テキストの選択」ページから「表現と用例」ページに自動的に切り替わる。なお，この計算に要する時間は，テキストファイルの大きさとＣＰＵの処理速度，搭載メモリの容量に強く依存する。

(16) 頻度に関する度数分布表示

テキスト群Ａ，Ｂにおける部分文字列の生起頻度が，度数分布表として示される。各セル内の数は，該当する頻度をもつ部分文字列の個数を表す。たとえば，テキスト群Ａには全くないが，テキスト群Ｂには10回出てくる文字列が，何種類あるかを確認する場合，縦軸（テキスト群Ａ）「０」と横軸（テキスト群Ｂ）「10」の交わるセル内の数字が，求める文字列の数である。本ソフトウエアでは，同値な部分文字列を一括して扱うので，その同値類の個数が示されることになる。

(17) 文字列長の下限の指定

e-CSAは，*n*グラム解析ツールとは異なり，扱う文字列の長さ*n*を固定しない。任意の長さの部分文字列が対象となるため，短い部分文字列を排除する場合には，作業ウインドウ左上のスライダーバーを動かし，文字列の長さの下限を設定する。なお，ひとつの同値類には長さの異なる文字列が含まれるが，設定する文字列の下限は，同値類中の最長文字列に適用される。



(18) セルの選択

度数分布表において，具体例を表示したいセルをダブルクリックすると，「文字列閲覧」ウインドウが開く。上部中央に，該当する文字列のリストが表示される。複数のセルを同時に指定する場合には，複数セルを矩形状に選択した上で，「選択範囲を表示」ボタンを押す。リストには，特に指定しない場合，同値類中の最長文字列のみが表示されるが，「同値類中の最長文字列のみ表示」のチェックをはずすと，同値類中の極小文字列も併せて表示する。

(19) 用例の表示

「文字列閲覧」画面において，テキスト群A，Bにおける用例を表示するには，次のような方法がある。

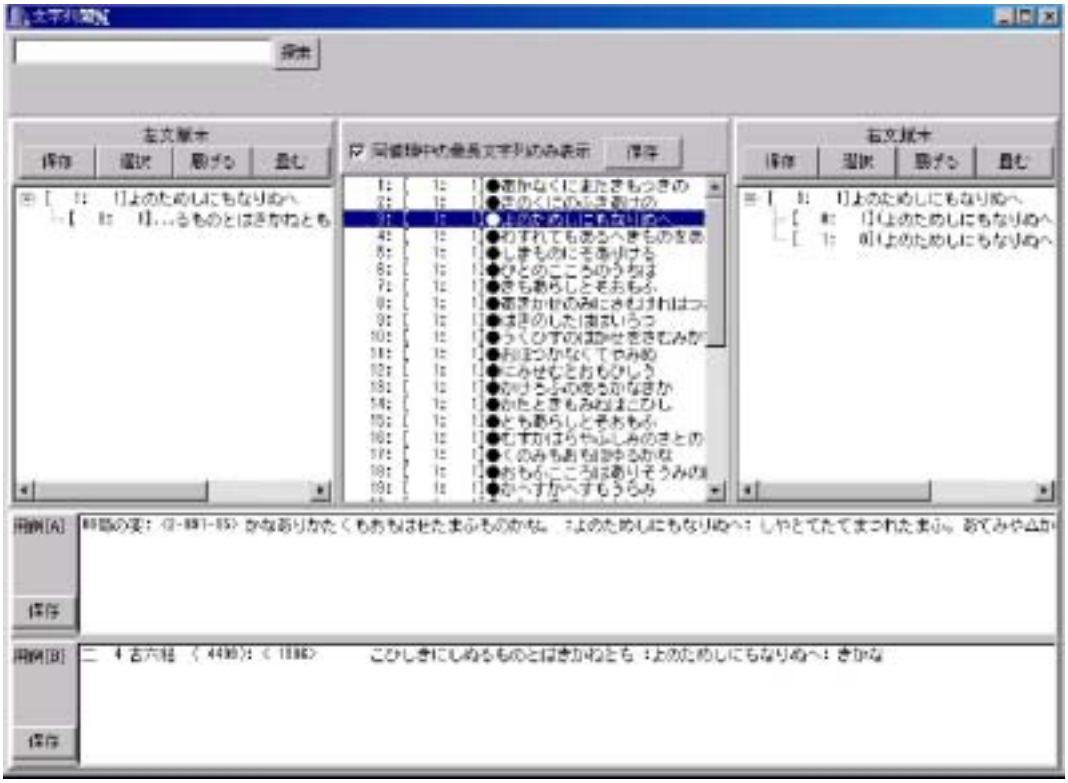
ｉ．部分文字列リストからの選択

リストから特定の文字列を選択する。前後の文脈が，それぞれ左右に木のかたちで示され

る。同時に，ウインドウ下部に，テキスト群A，Bにおける用例がKWIC表示される。

ⅱ．左文脈木・右文脈木からの選択

左右の文脈木から，特定の文字列をクリックしても，ウインドウ下部に，テキスト群A，Bにおける用例がKWIC表示される。



ⅲ．付録機能（文字列直接指定による選択）

キーワードが発想できれば「検索」も可能である。ウインドウ左上のエディットボックスに文字列を入力し,「探索」ボタンをクリックすると，その文字列の左右の文脈木が表示され，ウインドウ下部には用例がKWIC表示される。

４．結語

用例収集という国語学・国文学研究の基礎作業は，研究者が特定の語句に着目するところから始まる。どのような語句に着目するかが研究の成否を分ける鍵となるため，研究者は，慎重に研究の糸口を模索し，しかる後に用例収集に入る。その段階で，e-CSAの「付録機能・文字列直接指定」は，ある程度の利用価値があるといってよいであろう。けれどもそれは，このソフトウエアの中心的機能ではない。

本ソフトウエアは，研究者が着目すべき語句を模索している段階で，その語句の発見を支援するツール（道具）として開発した。すなわち，作品そのものを通読しても気づかなかった問題点を，計算機プログラムが提示したデータによって見いだすことを目的とする。情報

科学の一分野として近年誕生し，注目を集めている「発見科学（Discovery Science)」はまさに，このような形での機械による発見支援を目指すものである。

おひとりでも多くの方々が本ソフトウエアをご利用になり，研究に役立ててくださるよう，お願い申し上げる次第である。

注

(1) 近年では，たとえば，科学研究費補助金「特定領域研究(A)」「環太平洋の「消滅に瀕した言語」に関する緊急調査研究」（平成11～14年度，領域代表者：宮岡伯人）が記憶に新しい。

(2) 鹿児島大学の内山弘氏。もうお一方，国文学研究資料館の入口敦志氏も，第一著者，第三著者のパソコンの先生である。この世代前後の大学院生は，みなお世話になった。

(3) あれでも，キューハチ「ノート」と呼ばれていた。i486SX(16MHz)，メモリ4MB×2を順次追加。また，別途購入して組み込んだHDDは190MBだった。本体は重いがＯＳはじめソフトウェアは軽く，たいへん使いやすかった。10年経って，第一著者が使っているＰＣは，AthlonXP+2500(1.83GHz)，メモリ512MB，HDD120GBになった。単純計算で処理速度は約120倍になったが，仕事は120倍は捗らないのである。なぜか。

(4) 「汎用」を謳うからには，文字列で表現できるものは全て処理できる。国語学・国文学関連のテキストに限らず，「A」「G」「C」「T」の配列で表現されるDNAも，旋律を表すオタマジャクシも，処理可能である。なお，「e-CSA」（イークサ）の命名は，九州大学大学院システム情報科学研究院の篠原歩氏による。

(5) たとえば，近藤泰弘氏　「《文化資源》としてのデジタルテキスト――国語学と国文学の共通の課題として」（『国語と国文学』第77巻11号，《文化資源》としての国文学　2000年11月），「Linuxによる言語処理――高速文字列検索を例として」（『日本語学』特集<コンピュータによる日本語研究の新展開>2001年12月），また，近藤みゆき氏「*n*グラム統計処理を用いた文字列分析による日本古典文学の研究―『古今和歌集』の「ことば」の型と性差」（千葉大学人文研究29，2000年3月），「*n*-gram統計による語形の抽出と複合語―平安時代語の分析から」（『日本語学』20-9，特集　複合語・連語の文法　2001年8月）など。

(6) テキストを単なる文字列として処理するという点では，*n*グラムの手法と共通する。しかし，*n*グラムでできることは，このツールでもできる。むしろ，効果的に提示できる手法を用いている。

(7) 研究に用いるテキストは，その目的に応じた校訂基準に貫かれていなければならない。e-CSAによる抽出結果に明らかな誤謬が見出される場合は，まずテキストデータの精密度を疑う必要がある。つまり，他者が作成したテキストファイルを無批判に利用することはできないし，常に本文批判の手間を惜しんではならない。

主要参考文献

・「和歌データベースにおける特徴パターンの発見」
竹田正幸，山崎真由美，福田智子，南里一郎
（「情報処理学会論文誌」Vol.40，pp. 783-795，平成11年3月）

・「古典和歌における表現分析の新手法―類似歌発見のために―」
南里一郎，福田智子，竹田正幸

（「古典学の現在Ⅰ」<文部省科学研究費補助金特定領域研究「古典学の再構築」>，pp．53-81，平成12年3月
・「古典和歌における類似表現の自動抽出の試み」
南里一郎，福田智子，竹田正幸
（『純真紀要』41号，pp．79-87，平成12年12月）

附記

本稿は，国語学会2002年度秋季大会（徳島大学）において行ったデモンストレーション「文字列分析ツールを用いた散文作品の解析」の内容を骨子とし，第二著者がWebで公開しているマニュアルをもとに加筆したものである。

■パスについて

「パス」とは,「通り道」の意であり，Windowsにおいてファイルやフォルダを指定する表記法と考えてよい。

たとえば，Cドライブにあるwindowsフォルダの中に「aaaa.txt」というファイルが存在するとすると，その所在は次のように指定される。

C:\Windows\aaaa.txt

本稿では，処理対象のテキストファイルを，Dドライブの中の「Text」フォルダの中に格納したが，その中にあるテキストファイル「01俊蔭.txt」の所在は，次のように指定される。

D:\Text\01俊蔭.txt

この表記法とWindowsにおけるフォルダ階層の見せ方の関係に慣れれば，e-CSAで処理するテキストファイルの登録は容易であろう。

（なんり　いちろう・純真女子短期大学国文科助教授）
（たけだ　まさゆき・九州大学大学院システム情報科学研究院助教授）
（ふくだ　ともこ・九州大学大学院人文科学研究院助手）