Windowsパソコンにおける中国語の検索
―EmEditorを例に

沈　　国威・氷野善寛

0　始めに

現在、言語の研究・教育は、もはやコーパスの利用なしでは語れないであろう。いわゆるコーパスとは、言語研究のために加工され、蓄積された言語データの集合であるが、実際の利用者にとって、データベースという方が場合によっては分かりやすい。

目標言語が実際に使用されていない環境において、その言語を学習する場合、正確かつ自然な例文に接し、正しい語感を養うことは、何よりも重要である。またその外国語を研究の対象に据えるとき、例えば論文を作成する際、ある語や文型、ないし文法事項に関して、その実際の使用状況、頻度などを把握しなければならない。このような場合、コーパスは大きな威力を発揮するのである。コーパスはまた、例文をチェックする母語話者（インフォーマント）の役割を担うことも出来る。

コーパスの利用に関しては、無料公開されたものは便利であろう。現代中国語研究の場合、台湾中央研究院の「平衡語料庫」や香港城市大学の「共時語料庫」などが公開されている。利用法などについて、沈国威2000を参照されたい。また関西大学サイバーキャンパス計画の一環として、「関西大学中国語語料庫」をウェブ上に公開した。現在では、『人民日報』をはじめ、多くのジャンルの言語データを取りそろえ、運用中である。このコーパスは、強力な検索インターフェースを備えており、教育と研究の両面に大きな威力を発揮している。特に学習者に均一的な言語資料を提供することにより、再現性を重んずる授業でも使える点は評価に値する。しかし、公開コーパスは、インターネットにアクセスする環境がなければならず、そして何よりも問題になるのは、自分で言語資料を用意することや自分の研究内容に応じて、言語資料をカスタマイズすることができない点である。このデメリットを解消すべく、沈・氷野は公開コーパスを補う存在として、研究者、あるいは学習者が、Windowsの環境において自分の研究テーマに合致するコーパスを構築し、簡単に検索できる方法を模索してきた。コーパスを構築し、利用するには基本的に、

1．データの確保
2．コーパスの構築
3．データの検索

という3つのハードルをクリアしなければならない。沈2000は、1と2についてWindows OSでの手順と方法を述べているが、3については、Mac OS 9.xによる解決案しか提示できなかった。しかし、同機種のユーザは非常に少ないのが現状である。一方、Windows XPが登場してからも長らく日本語環境では、エディタや検索ソフトによって中国語を検索することができなかった。最近にな

って、EmEditor（EMエディタ）を用いることで、目指す目的をほぼ達成できることが分かった。

言語データの収集、コーパスの構築に関しては、沈2000を参照することとし、ここでは、いままで難しいとされた日本語環境（Windows XP）において、EmEditorによる中国語データの検索、及び中国語研究に応用する可能性について報告する。

1　テキスト・エディタ：EmEditorについて

EmEditorは、㈱エムソフト社が開発したテキスト編集用のエディタで、多言語処理に秀でている。EmEditorは、シェアウェアで下記のサイトから直接購入できる（4,400円）：http://www.emeditor.com/jp/。また学生には在学中、アカデミックライセンスを登録することにより無料で使用できる制度もある。EmEditorは、強力な検索機能を備えているのみならず、中国語を含む多言語処理も得意とするテキストエディタである。ここではまず、EmEditorについて説明し、それからEmEditorによる中国語の検索方法、テクニック及び検索作業に欠かせない“正規表現”を詳しく見ていくことにする。

パソコンの上でデータに対して検索を行うには、基本的に次のような流れが考えられる。

1．検索対象の選択：特定のファイル（あるいはあるディレクトリにあるファイルや、ある拡張子をもつすべてのファイル）を指定し、検索する。
2．検索語句の入力：正規表現（後述）を用いる。
3．検索の結果の表示と保存。

以下、この流れを念頭に置きながら検索ツールと検索法について説明していきたい。

2　EmEditorの起動と設定

上記のサイトからダウンロードしたファイルのアイコンをダブルクリックすれば、インストールが始まる。指示に従って進めれば特に問題ないと思われる。インストール終了後、EmEditorのアイコンをダブルクリックしてみよう。すると次の画面（図1）が現れる。まずこの画面について説明しておこう。

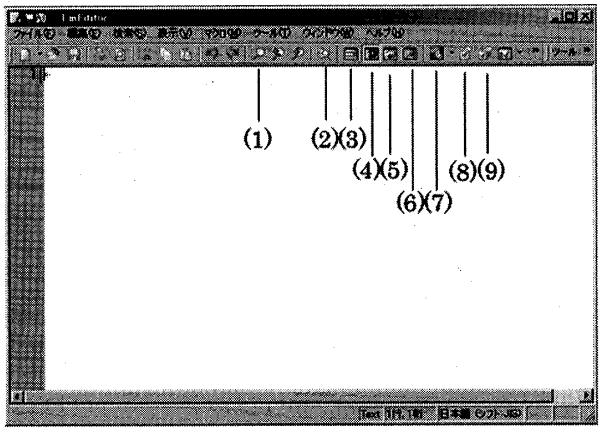

図1　EmEditorの起動画面

画面上部に一列のアイコンがある。左側にあるものは、ワープロソフトなどでお馴染みのもので特に説明は不用であろう。ここでは検索のアイコンより右側のものについてその機能を確認しておく。

(1)「検索」：開いているファイルを検索するためのアイコン。
(2)「ファイルから検索」：このアイコンは同じフォルダ内にあるファイル群、或いは同じ名前や拡張子を持ったりする複数のファイルを一度に検索するのに用いられる。
(3)「折り返さない」：このアイコンを押せば、長い行でも折り返さない設定となる。

⑷ 「指定文字数で折り返し」：このアイコンを押せば、指定された文字数（文字数の指定は後述）で行を折り返す設定となる。
⑸ 「ウィンドウの右端で折り返し」：このアイコンを押せば、ウィンドウの大きさに応じて、行が折り返される。
⑹ 「ページの右端で折り返し」：このアイコンを押せば、ページに合わせて行を折り返して表示する。
⑺ 「フォント」：フォント選択のアイコン。
⑻ 「現在の設定のプロパティ」：開いているファイルに関する設定を行うアイコン。
⑼ 「すべての設定のプロパティ」：基本設定のアイコン。詳細は、次節を参照されたい。

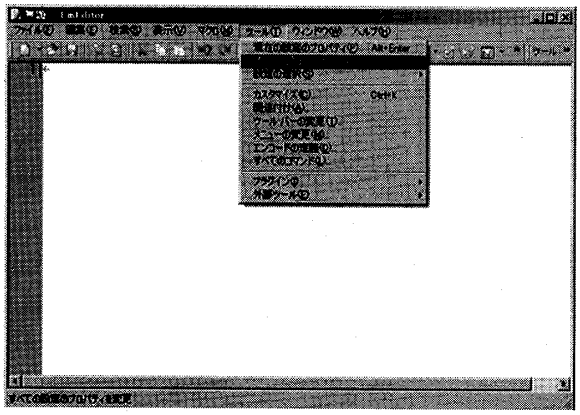

図２　設定画面を表示させる

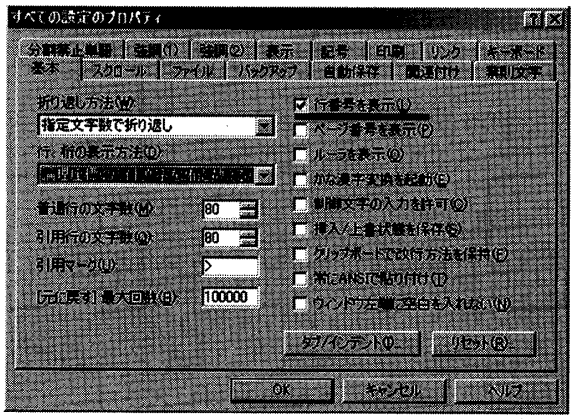

図３　すべての設定のプロパティ

次に、EmEditorを中国語検索専用のツールとして使うことを前提に、設定を見ていきたいと思う。まず上記のアイコン⑼、あるいは、図２のように［ツール（T)］→［すべての設定のプロパティ（O)］を選択し、設定画面を表示させておく。

タグ「基本」を選択する（図３)。ここでは「行番号を表示（L)」のチェックボックスにチェックを入れる必要がある。これにより画面の左に行数が表示される。この行数は、ヒットした検索語の個数である。つまり、検索画面の最後に行けば、ヒット数を確認することが出来るのだ。例えば、図４では、「老师」という語がYuwen（語文テキスト）というフォルダから146例ヒットしたことが示されている。

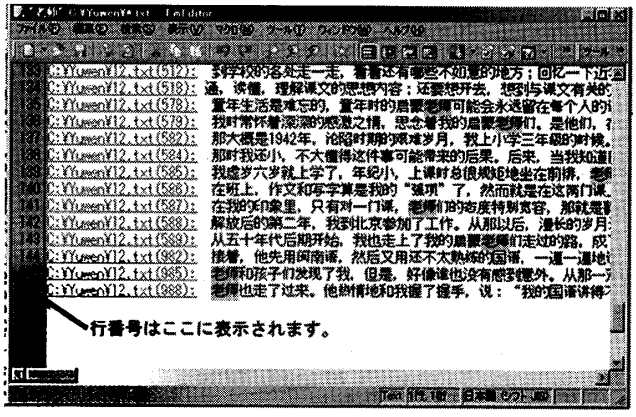

図４　検索画面

次に［ファイル］のタグを選択しておく（図５)。ここでは［開く時のエンコード（E)］で「簡体字中国語（GB2312)」を選択しておけば、検索するときの文字化けが解消される。

もちろん、専ら繁体字のデータを検索する人は、「繁体字中国語（Big５)」を選択するとよいだろう。

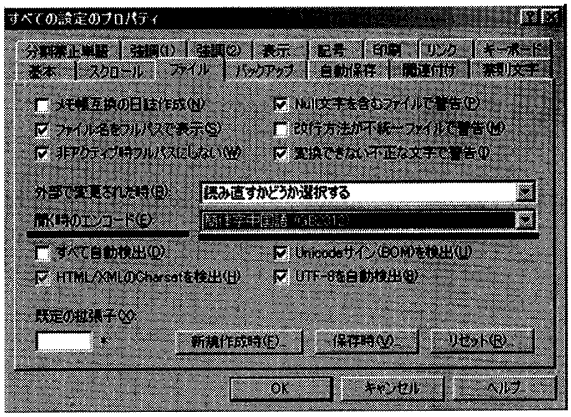

図5　エンコードの設定

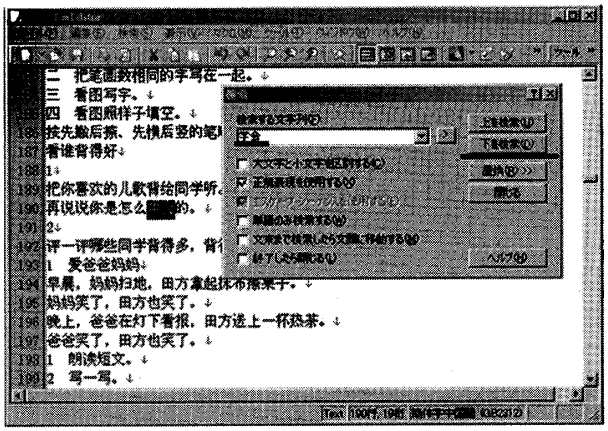

図6　検索画面

3　検索の実際Ⅰ：ファイルを開いて検索

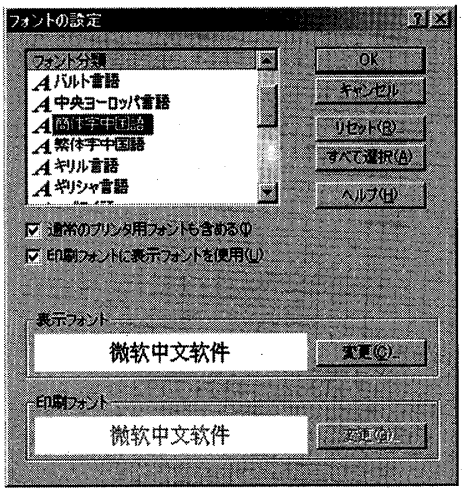

図7　フォント指定

1つの特定のファイルを開き、そのファイルに含まれている語句を検索する場合、この方法を取る。EmEditorの画面で、検索のアイコン(1)をクリックするか、メニューバーから［検索（S）］を開き、コマンド［検索（F）］を選ぶ。すると上の図6の画面が現れる。

［検索する文字列］の欄に検索したい文字列を入力し(※中国語を入力する場合にはIMEを中国語に切り替えておく)、［下を検索］を押していくと、順番に該当する文字列にジャンプしていく。なお検索でマッチした語は強調色で表示される（強調の色は変更が可能)。これが最も基本的な検索方法である。正規表現を使う特殊な検索法もあるが、この検索法については、後に例をあげて詳しく述べる。正規表現を用いる場合には、［正規表現を使用する（X）］というところにチェックを入れておく必要がある。

なお、デフォルトで中国語のフォントを指定していない場合は、フォント指定のアイコン（上記の(7)）をクリックするか、メニューバーから［表示］→［フォントの分類］、あるいは［フォントの設定］を選び、検索する際のフォントの指定をすることができる。簡体字中国語のデータを検索する場合には、「簡体字中国語」を指定すればよいだろう（図7）。

4　検索の実際Ⅱ：フォルダごと検索

EmEditorには複数のファイルを一括して検索する機能がある。この機能により特定のフォルダ(サブフォルダを含む）に納められたテキストファイルや共通した特徴、例えば、共通したファイル名（一部でもO.K.）や拡張子を持つファイルを開かずに一度に一括して検索することができる。対象ファイルを開かずに検索できる機能は、語彙研究などでは非常に便利である。現在、データフ

ァイルのサイズは、ますます巨大化してきた。例えば、新聞1年分のデータは、ゆうに100MBを超えてしまう。それを開くことは、普通のワープロソフトやエディタでは非現実的である。

5　検索の手順

メニューバーから［検索］→［ファイルから検索］を選択すれば（図8）、フォルダを一括検索するための検索ウィンドウが開く（図9）。

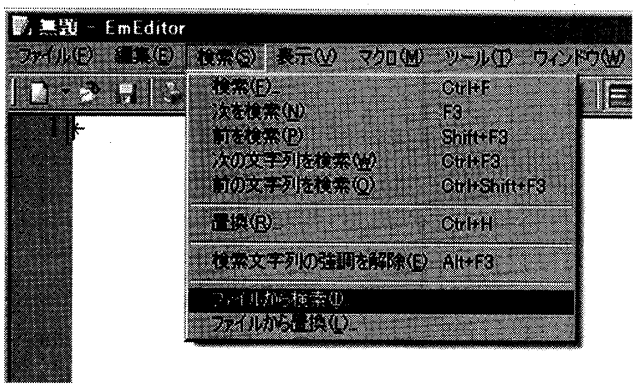

図8　検索方法の選択

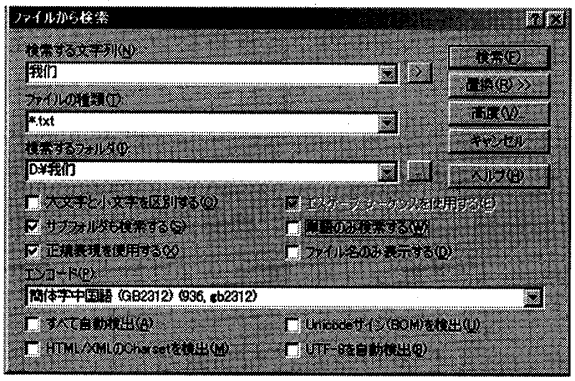

図9　検索ウィンドウ

また、ツールバーにある検索のアイコン（上記の(2)）をクリックしても、図9のウィンドウが開く。以下、検索画面について簡単に説明することにしよう。

検索語の入力

［検索する文字列（N）］の中に検索する文字列を入力する。なお入力ボックスの右側の三角印をクリックすると、これまで検索した語の履歴一覧がプルダウンメニューに表示される。

検索するファイルの指定

［ファイルの種類（T）］は、検索の対象となるファイルを指定するための入力ボックス。その中に検索したいファイルの名前を入力すれば、当該のファイルが検索の対象となる。しかしそうすれば、複数のファイルから検索という利点が生かされない。ふつう、この入力ボックスに、「*.txt」のように入力しておく。「*.txt」の「*」は、任意の文字列を表す記号なので、該当フォルダの中にあるすべての「.txt」という拡張子を持つ（つまりテキストファイル）ファイルが検索の対象となる。また、例えば「hongloumeng*.txt」と入力すれば、hongloumeng01.txt、hongloumeng02.txtのように『紅楼夢』という小説のすべての章が検索の対象となる。

検索するフォルダ

ここでは、検索ファイルが納められているフォルダを指定する。入力ボックスの右の［…］をクリックすると、［フォルダの参照］のウィンドウが開くので、直接検索したいフォルダを指定することが出来る。なお、下の［サブフォルダも検索する（S）］にチェックを入れておけば、下

位に位置するフォルダも検索の対象となる。入力ボックスの右側の三角印をクリックすると、これまで検索した語の履歴一覧がダウンメニューに表示される。入力ボックスの下にある［大文字と小文字を区別する（C）］［単語のみ検索する（W）］などは、中国語の検索に使わない機能だが、［正規表現を使用する（X）］は、チェックしておいたほうがよいだろう。

検索言語の選択

［エンコード（P）］は、検索の言語を指定するものだが、すでに中国語をデフォルトに設定しているので、「設定されたエンコード」という表示が出ても構わない。

検索語の入力、ファイルの指定などが一通り終われば、［検索（F）］を押す。例えば次のような検索結果画面（図10）が表示される。図10について、簡単に説明しておこう。画面は、2つの部分から構成されている。左側は、マッチした語のあるファイルの在りかをフルパスの形で示している。右側は、マッチした語を含む文が表示されている。つまり、①は、CドライブのDocumentというフォルダの中にあるサブフォルダ「語文教科書」の中の「08第八冊.txt」というファイルの633行目に当該の語が使用されていることを示している。同じ要領で、②の意味は「09第九冊.txt」というファイルの510行目に、検索語が使用されていることになる。左側の部分をダブルクリックすれば、該当のファイルが開き、前後の文脈の確認が出来る。マッチした語は③のように色が強調され表示される。ここでは「中国」という語を検索したので、「中国」が強調されている。

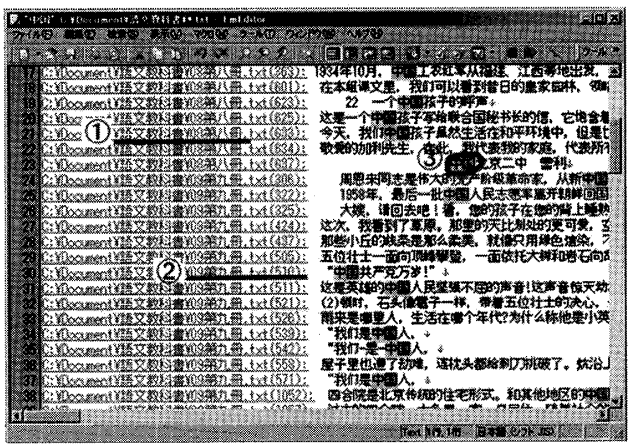

図10　検索結果画面

6　検索結果の保存

検索の結果は、以下の手順でワープロ文書のように保存することが出来る。つまりメニューバーの［ファイル（F）］→［名前を付けて保存（A）］を選択すれば、図11のような画面が表示される。

［ファイル名（N）］の入力ボックスに名前を入力する。この場合は、漢字ではなく、アルファベッドを用いた方が無難だろう。［ファイルの種類（T）］［エンコード（E）］なども確認して、［保存（S）］をクリックすれば、ファイルが保存される。

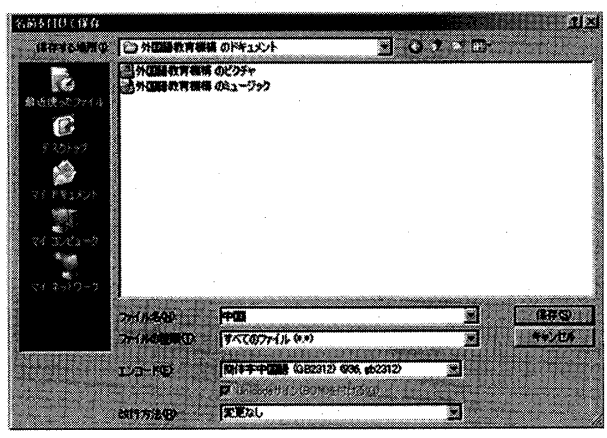

図11　名前を付けて保存

7　正規表現について

以上、検索の方法について簡単に説明した。次に、正規表現による検索法について、見てみよう。

検索ツールを紹介する中で、何回も「正規表現」ということばに触れた。正規表現とは、grep、sedやPerlなど、UNIX上の多くのソフトで採り入れられている文字列の条件表示方法である。DOS環境のワイルドカード（wild card）をご存じだろうか。もともとはトランプのジョーカーにあたる万能札のことだが、MS－DOSやUNIXなどでは、任意の文字列として利用できる「?」「*」などの特定の文字を言う。この場合、「?」は任意の１文字を、「*」は任意の文字列を表す。正規表現はワイルドカードよりも強力なものである。文字数や可能性のある文字列の範囲指定などが行える。たとえば「任意の１文字」や「文字の繰り返し」などを記号や文字で指示し、法則性のある文字列の検索に使われる。正規表現の中で用いる文字には、特殊な意味を持つものと持たないものがある。特殊な意味を持たないもの（たとえばすべての漢字や仮名、アルファベットの大部分）は、当該の文字そのものを表す正規表現になる。特殊な意味を持つもの（たとえば“$”や“[”）は、プログラムで制御記号として用いられるので、その文字自体を表さない。特殊な意味を持つ文字をメタキャラクタ（メタ文字）と呼ぶ。メタキャラクタ自体を表示したい場合には、その直前にバックスラッシュ（“\”）を置く。(Windows日本語の環境では“¥”)。たとえば、“[”を表す正規表現は“\[”となる。このような表現を“\でエスケープする”と言う。つまりバックスラッシュは、次の文字の特殊の意味を取り除くという意味のメタキャラクタなのだ。したがって、メタキャラクタ以外の文字の直前にバックスラッシュを付けてもその文字の意味は変わらない。“\A”は、単なる“A”と同様に文字“A”を表す。

8　正規表現による検索

単語がスペースによって区切られておらず、また“**一…就…**”のような不連続成分による表現型の多い中国語では、正規表現を用いる検索が大きな力を発揮できる。この節では、まず具体例を示しながら、基本的なメタキャラクタについて説明しよう。次節ではより高度な中国語の検索法を考えてみる。なお、ここで説明している方法は、多くの場合、同じく２バイト文字の日本語にとっても有効である。

^　（カラット）

^は、文字列の先頭、行の先頭にマッチする。

^ABCは、行頭にあるABCにマッチ

^**这是**は、行頭にある**这是**にマッチ

ブラケット［］の中にある“^”は、特殊な意味を持つが、詳しいことはブラケットの項で説明する。なお、いわゆる行頭は、強制改行の後の行頭のことである。

$　（ドル記号）

$は、文字列の終わり、行の終わりにマッチする。
ABC$は、行末にあるABCにマッチ
回来。$は、行の末尾にある“**回来。**”にマッチ
。$は、行の末尾にある句点“。”にマッチ（それ以外の“。”にマッチしない）
なお、いわゆる行の末尾は、強制改行の直前の末尾である。

.（ピリオド）
.は、改行コード（\n）以外の任意の1文字にマッチする。...なら任意の3文字にマッチすることになる。
a.cは、abc、acc、adc…などにマッチ
例えば**这.是**は、**这不是，这也是，这都是，这还是，这个是，这本是**…にマッチ
中国語の中に“离合词”と呼ばれる一群の動詞がある。“离合词”の使用実態を調べるのに、ピリオドによる検索が有効である。たとえば、**结.婚**、**生..气**はそれぞれ、**结了婚**、**结过婚**；**生他的气**、**生儿子气**などの表現をピックアップすることができる。また、**一.二.**は、**一干二净**、**一清二白**、**一来二去**、**一穷二白**…などにマッチする。ピリオド1つは、改行記号以外の任意の1文字を表すことができる。また後述するように、ピリオドは*、＋、?などとの併用によってもっと効率的な検索も可能である。

*（アスタリスク）
*は、直前の1文字（または正規表現）の0回以上の繰り返しにマッチする（0回も含む）。
ab*cは、ac、abc、abbc、abbbc、…のいずれかにマッチ
.*は、空文字列を含む任意の文字列にマッチ
例えば**结.*婚**は、**结婚**、**结了婚**、**结完了婚**、**结过一次婚**、**结过一次有名无实的婚**…
などにマッチ
ただし「.*」は、“结”と“婚”の間に挟まれている任意文字列が適合の条件になっているので、“…**结果，婚礼**没能按时举行。”のような例にもマッチしてしまう。

+（プラス）
+は、直前の1文字（または正規表現）の1回以上の繰り返しにマッチする（0回は含まない）。
ab+cは、abc、abbc、abbbc、…のいずれかにマッチ（acにはマッチしない）
.+は、任意の文字列にマッチ
中国語の例を挙げれば、**回.＋来**は、**回家来**、**回北京来**、**回到了阔别已久的故乡来**、…などにはマッチするが、“**回来**”にはマッチしない。

?（疑問符）
?は、直前の1文字（もしくは正規表現）の0回か1回の出現を表す。?は、繰り返しのメタ文字といわれるが、実際は2回以上の繰り返しはしない。
ab?cは、ac、abcのいずれかにマッチ

つまり、ある文字があるかどうか分からないという場合に使うのだ。たとえば「マネージャー?」は、「マネージャ」と「マネージャー」の両方にマッチする。

中国語の例を挙げれば、“**看一?看**”は、“**看看**”と“**看一看**”の両方にマッチする。

| （選択）

|は文字列の選択を表す。

“あるいは|或いは”は、「あるいは」と「或いは」のどちらにもマッチ

通例、|は、パーレン（）と組み合わせて使うと効果的である。詳しくは、（）の項で説明する。

[ ]（ブラケット）

[ ]は、文字クラスと言い、[ ]内の任意の1文字にマッチする。範囲指定を使うこともできる。集合の指定方法は2つある。1つは集合の要素を並べて記述する方法である。

[abcdef]は、“a～f”のいずれか1文字にマッチ

[あいうえお]は、「あ～お」のいずれか1文字にマッチ

走[らりるれろっ]は、「走る」のすべての活用形にマッチ

もう1つは、ハイフン（マイナスとも言う）“-”による範囲指定である。“-”は文字クラス内では特殊の意味を持ち、[a-z]のように範囲指定することができる。

[ぁ-ん]　　ひらがな1文字にマッチ

[0-9]　　数字1文字にマッチ

[A-Za-z]　　英字1文字にマッチ

ただし、[一-十]のような指定はできない。漢数字は何千何万ある漢字の中で、他の漢字と関係なく定義されているからだ。

カラット“^”は、[ ]内では先頭に用いた時のみ特殊な意味を持ち、文字クラスの否定を表す。つまり、

[^0-9]は、数字以外の1文字にマッチ

[^A-Z]は、英字大文字以外の1文字にマッチ

また、[ ]の中では先頭の“^”、文字の間の“-”以外のメタ文字は、メタ文字としてではなく、その文字字体の意味になる（前にバックスラッシュを置く必要はない）。

（）（パーレン）

（）には2つの意味がある。1つは正規表現をグループ化するものである。

例えば**李**（**先生**|**同志**|**师傅**）は、**李先生，李同志，李师傅**にマッチ

（**高兴**）+は、**高兴高兴，高兴高兴高兴**…にマッチ

もう1つの使い方は、後方参照（back reference）とよばれるものである。\1～\9で引用する部分を指定する。数字は、n番目の（　）に対応することを示す。

a. (.)\1　は、AA、BB、**看看**、**说说**、**多多**…にマッチ

b. (.+)\1　は、**喝喝**、**说明说明**…にマッチ

つまり、aの正規表現の意味は、任意の1文字をもう一度引用するということで、2字の畳語に

マッチし、bの正規表現は、1文字以上の文字列をもう一度引用することで、AA、ABAB、ABCABCのような文字列（日本語ではころりころり、ごろんごろんなど）にマッチするということである。

{}（繰り返し）

{} は、ある一定回数以上の繰り返しを指定するためのメタキャラクタである。{n} は、直前の1文字（または正規表現）のn回の繰り返しにマッチする。{1,3} と記述する場合、直前の1文字（または正規表現）の1回から3回までの繰り返しにマッチする。

[0-9] {5} は、5桁の数字にマッチ

a{1,3} は、a、aa、aaaにマッチ

{min, max} は、直前の1文字（または正規表現）のmin回～max回の繰り返しにマッチする。minの省略は0回、maxの省略は∞回（無限大）の指定と解釈される。

*、+、?、{min, max} は、繰り返しパターンとして最大回数の繰り返しマッチを試みることになっているが、直後に?を追加することで最小回数の繰り返しでうち切ることができる。

*?　直前の正規表現の0回以上の繰り返し（最小回数、つまり0回を優先的に）にマッチ

+?　直前の正規表現の1回以上の繰り返し（最小回数、つまり1回を優先的に）にマッチ

??　直前の正規表現の0回あるいは1回の繰り返し（最小回数、つまり0回を優先的に）にマッチ

{min, max}?　直前の正規表現のmin回? max回の繰り返し（最小回数）にマッチ

9　中国語の検索テクニック

◇過不足なくデータを集めよう

できるだけ多くの用例を集めるのが大事だが、時には用例が集まりすぎて困るケースもある。検索の結果が1000例を超えたら一々目を通すことは難しいだろう。まさに孔子のいうように「過ぎたるは猶ほ及ばざるがごとし」である。たとえば、量詞“道”“条”について検索してみたら、ヒット数は忽ち10,000を超えてしまった。これは、情報資源が巨大になったのが一因だが、中国語は、英語のようにスペースによって単語が区切られておらず、ひとつの文字が、語と複合語の両方に用いられているという特有の事情による。常用語ほどこの現象が顕著で、使用頻度の高い1文字の名詞、動詞、形容詞、副詞は、ヒット数が多すぎて、検索が意味を為さないことが多い。しかし、先に述べた正規表現を使用することによって効率的にデータを収集することが可能となる。以下、2つのケースについて見ていきたいと思う。

◇1字語の検索：特徴を掴んで絞ろう

1字語検索のコツは、条件を付けて検索の範囲を狭めることである。たとえば量詞“道”について、“**[这那一两三四五六七八九十]道**”の形で検索すれば、非量詞の用法は完全ではないがかなりの程度で排除することができる。ちなみに、この正規表現の意味は、ブラケットの中の任意の1字に“道”が続いている文字列にマッチするということだ。

同じく単音節の形容詞、たとえば“大”の場合も、“**[很真太]大**”で検索するほうが目当ての用例が集まりやすい。

それでは、“这是”の間に生起する単音節の副詞を調べるにはどうすればよいのだろうか。“**这[也不都可倒]是**”のように行えばよい。あるいはブラケット中の“^”が否定の意味を持っていることを思い出すとよい。“**这[^个次人回]是**”のように“这”の後に生起しそうな量詞などを除外する方法もよいだろう。

一方、単音節の動詞の場合は、助辞の“**着、了、过**”といっしょに検索するほうが絞りやすいだろう。また、“**看一看、尝了尝**”もよく用いられる形である。次のような正規表現を使えば、これらの形式は、網羅的に検索することができる。

正規表現	検索語
(.)一\1	看一看，尝了尝...
(.)了\1	想了想，说了说...
(.)不\1	是不是，大不大...

◇2字語の検索：漏れはないか

2字語（双音節）の名詞（動名詞を含む）は、検索上特に難しいことはないだろう。

しかし動詞と形容詞の場合は、事情が少し複雑である。“**说明说明、介绍介绍、简简单单、高高兴兴**”のような重ね型については、次節で見ることにして、ここでは、“离合词”と呼ばれる一群の語の検索法について、考えてみたいと思う。

いわゆる“离合词”は、語と連語（短语）の性質を合わせ持つ動詞性の成分で、結合が緩やかで、中間に他の成分が入ることができる。

たとえば“**结婚、生气、请客、洗澡**”などはそのままの形で用いられる一方、“**结了婚、结过一次婚、生闷气、生孩子的气、请他客、请老王的客、洗完澡、洗一个热水澡**”のようにも用いられる。このような分離した形の用例まで視野に入れなければ、“离合词”の検索は不完全となる。しかし、“结”と“婚”を別々に検索すれば“**结果、结合、总结；离婚、婚礼、金婚**”などもピックアップされてしまう。非能率的だし、時には検索の結果が乱雑すぎて役に立たない。このような事態は、正規表現を使うことで防げる。たとえば次のように、

正規表現例	正規表現の意味	マッチしたもの
结.婚	任意の1文字が入る	结了婚，结过婚，结完婚…
结.?婚	0文字か任意の1文字が入る	结婚，结了婚，结完婚…
结..婚	任意の2文字が入る	结不起婚，结不了婚…
结…婚	任意の3文字が入る	结了两次婚…
结.*婚	0文字～任意の文字列	结婚，结过婚，结了一次很草率的婚…
结.+婚	1文字～任意の文字列	结了婚，结过三，四次婚…

ただし、正規表現でも“婚已经结完了，但是，…”のような倒置のケースに対しては無力で、“**婚.*结**”の形で検索する必要がある。

中国語は基本的に単音節語か、それとも徐々に多音節化に向かっているかという中国語の本質論

については、意見の分かれるところである。しかしたとえば、

结婚　生气　请客　洗澡　鞠躬　上当　敬礼　睡觉　撒谎
发财　吃亏　捧场　帮忙　造谣　发言　毕业　离婚　鼓掌

などの使用頻度の高い“离合词”について、その使用実態を一定規模のコーパスで調査することは決して無意味なことではないだろう。

◇不連続成分の検索

“结婚”のような不連続成分の存在は、中国語の１つの特徴と言えるかも知れない。このようなパターンについて、正規表現は強力な検索手段を提供してくれることは前節で見た通りである。この節では、不連続成分の検索についてもう少し詳しく見ていこう。

不連続成分には、２種類あると考えられる。１つは、“**因为…所以**”，“**虽然…但是**”，“**虽然…可是**”，“**虽然…不过**”のように、それぞれ単独でも使えるが、前後呼応して使用される場合もある。検索法としては、それぞれ個別に検索することができるが、共起情況を調べるのに“**因为.＋所以**”というように正規表現を用いればよいだろう。つまり、“**因为**”と“**所以**”の間に１字以上の文字列が存在しているケースである。以下は、幾つかの語群について沈のコーパス（5000万汉字）で検索した結果である。

検索文字列	ヒット数	分布
因为	10,000超	910ファイル
所以	7,084	852ファイル
因为…所以…	621	359ファイル
虽然	4,840	820ファイル
但是	7,227	768ファイル
可是	7,834	888ファイル
不过	5,883	874ファイル
虽然…但是…	422	255ファイル
虽然…可是…	455	238ファイル
虽然…不过…	78	72ファイル

もう１つは、“**是…的…**”，“**一…就…**”のように呼応（搭配）してはじめて所定の形式的な意味を表すものである。この種類のものは正規表現を使って検索することが必須になっている。以下、“**是…的…**”を例にして少し詳しく見てみよう。

“**是…的…**”構文は、“田中是在北京大学学的中文。”のように完了した動作について、動作に関わる時間、場所、道具、相手などの副次的成分を強調して説明する文型で、使用頻度が非常に高い。しかし検索による用例の収集は、意外に難しい。たとえば“是.+的”という正規表現で検索したら、忽ち10,000例を超えてしまい、しかも次例のような“**是…的…**”構文ではないものまでピックアップされてしまう。

●老先生的上身穿着件短蓝布袄，下身可只是件很旧很薄的夹裤。　《四世同堂》

そこで考えられる解決策は、検索の対象となるデータ量を減らすことと、もう少し条件を付けて

検索することだ。たとえば“是.{5,8}的(。|,)”のように検索条件を指定すれば、“是”と“的”の間の文字数を５～８に、“的”の直後に句読点が来るものに限定されることになる。ヒット数がある程度絞られるだろう。また“是.{4,8}的.+[,。？]”のように指定すれば、

- 他是昨天去的北京。
- 我说，“前天不是我们一起打的电报？”

など、目的語が“的”の後に置かれている用例も検出することができる。このように不連続成分の検索は、試行錯誤と工夫が必要である。

不連続成分には、いわゆる“连词”（接続詞）や接続性のある副詞が多く、個々の意味用法をきちんと記述し、文型として整理していくのが、中国語に関する研究の基本作業と言えよう。検索法の復習を兼ねて、下記のパターンの用例を集めてみよう。

既…又…
又…又…
一…就…
才…就…
越…越…
连…都（也）…
既然…那么（就）…
无论…（还是）…都（也）
不论…（还是）…都（也）
不管…（都）也…
只有…才…
只要…就…
即使…也…
尽管…可是…

◇重ね型の検索

中国語には日本語の活用形のような形態変化は存在しない。しかし、中国語の動詞、形容詞は、“看看、说明说明；大大的、高高兴兴”のような重ね型と呼ばれる用法があり、文法的な意味を表している。重ね型について“看一?看”（“看看”と“看一看”の両方にマッチ）、“简简单单”のように個別に検索することは難しいことではないが、しかし、ある作品、あるいはある範囲内のデータに使用されている重ね型を全体的に把握したい、ひいては重ね型の形式的、語用的意味を考察したいときは、どのようにすれば効率的にデータを集めることができるだろうか。ここでは重ね型にマッチする正規表現について説明する。

まず検索語入力ボックスに“(.)\1”と入力しておこう。この正規表現の意味は、括弧の中のもの（ここでは任意の１文字）をもう１回引用して、マッチするという意味で、AAのような畳語を見つけだすことが可能だ。検索結果から分かるように、“马马虎虎”，“太太”など動詞、形容詞の重ね型ではないものまで拾い上げられてしまった。これは、品詞標識がついていないコーパスの宿

命的限界で、手動で不適格な用例を削除していくより他ない。

それでは、重ね型を検索する正規表現を下表に整理しておく。

正規表現	マッチパターン	例語
(.)\1	AA	茫茫，尝尝，说说
.(.)\1	ABB	白胖胖，恶狠狠
(..)\1	ABAB	介绍介绍，热闹热闹
(.)\1(.)\2	AABB	热热闹闹，高高兴兴
(.+)\1	A(B)A(B)	说说，说明说明
(.)一\1	A一A	看一看，写一写
(.)一?\1	A(一)A	看看，看一看
(.)了\1	A了A	尝了尝，看了看
(.)了?\1	A(了)A	尝尝，尝了尝
(..)了\1	AB了AB	解释了解释
(.)不\1	A不A	是不是，来不来
(..)不\1	AB不AB	可能不可能，热情不热情
(.)\1.	AAB	开开心，点点头
(.)\1[了的地]	AA[了的地]	谢谢了，红红的，狠狠地

このように重ね型の使用は、作者、地域、文体等によって変わるものかどうか、重ね型の“一”、“了”の省略は、どのように条件付けられているかなどは、本当に興味深い問題と言えよう。

10　終わりに

以上、EmEditorによる中国語の検索、及び中国語研究に応用する具体的な方法について簡単に説明してみた。EmEditorには、便利なマクロやプラグインが多くある。例えば、文章の整形や繁体字・簡体字の変換などである。紙幅の関係で省略に付すが、興味のある方は、中国語教材研究会のホームページ（http://www.we.fl.kansai-u.ac.jp）を参照されたい。

電子メディアの出現により、今までに蓄積してきたアナログの情報がデジタル化され、データベースに集約されるようになった。その結果、情報資源がとてつもなく巨大化してしまった。またインターネット環境の普及によって、瞬時にして世界中の情報資源にアクセスできるようになった。このように文献、ないし情報の存在形態とそのアクセスの方法の変化により、情報収集の環境は、従来に比べて飛躍的に改善された。したがって、如何に情報資源へアクセスするか、そして情報資源からどのようにして必要な情報を取り出すかという知的活動に関する古典的な問題のうち、アクセスの問題より、巨大化した情報資源を対象に、どのようにして必要な情報だけを過不足なく抽出できるかは、今までになく重要度を増してきた。情報資源の巨大化の結果、その中に蓄積している情報内容の全てに目を通すことはもはや不可能になったからである。情報の選別的入手は、今日の情報化社会における個人のパワーを拡大するための重要なテクニックである。いわゆる情報格差は、このような情報を扱うテクニックの有無を指していると言える。巨大な情報資源からの情報抽

出こそは，情報処理の基本的なスキルとして学ぶべき重要な課題であることを研究者も学習者も認識すべきである。

参考文献

沈国威2000 『電脳による中国語研究のススメ』 白帝社