正規表現・文字コード

田野村 忠温

（2007年9月7日刊行、2011年9月30日更新）

本文書は田野村忠温・服部匡・杉本武・石井正彦『コーパス日本語学ガイドブック』（特定領域研究「日本語コーパス」日本語学班、2007年）に収録した解説に多少の加筆・修正を施したものです。文書全体で完結していますので、複製・配布に際しては、内容に改変を加えない、文書全体を用いる、の２点をお守りください。内容の補足などのために追加の資料を作成して同時に配布していただくことは差し支えありません。

本文書に大幅に加筆したものを、荻野綱男他編『講座ITと日本語研究3　アプリケーションソフトの応用』（明治書院、2011年）の「付録　正規表現・文字コード」として刊行しました。より正確で詳細な解説については、そちらを参照してください。

1　文字列検索と正規表現

エディタや検索ソフトウェアなどで検索文字列として普通の文字列ではなく正規表現を指定することにより、複雑な条件での検索が可能になる。ここでは正規表現の要素のうち、日本語テキストからの文字列検索上特に利用価値の大きいものについて説明する。

正規表現と一口に言ってもさまざまな異なる規格がある。正規表現が使えるかどうか、使えるとしてもどのような範囲の正規表現の要素が使えるかはソフトウェアによって異なる。Ruby、Perl、EmEditor、秀丸エディタ、拙作KWIC（http://www.tanomura.com/research/kwic/）などにおいては下記のすべての要素を使うことができる。

2　正規表現の主な要素

2.1　[...] ＝ ...に含まれる任意の１文字

検索文字列として「[探捜]す」という正規表現を指定して検索すれば、「探す」「捜す」の両方の用例――正確に言えば、それら２通りの文字列の現れ――を同時に検索することができる。同様に、「[見観看診]る」という正規表現は４通りの漢字表記の「みる」を表す。

「曲[がげ]る」「[曲ま][がげ]る」「[探捜][さしすせそ]」「行[かきくけこっ]」「長[いかくけ]」「[早速][いかくけ]」「で[はも]ない」のそれぞれがどのような用例を探すためのものか考えてみよう。

なお、[ ] および以下に出てくる正規表現の記号（^, |, ¥, (), {}, など）はいわゆる半角文字で入力する必要がある。

2.2　[^...] ＝ ...に含まれない任意の１文字

「のだから」という文字列の検索結果には、「ものだから」の用例が少なからず含まれる。そのようなものを除外したければ、「[^も]のだから」という正規表現を指定して検索すればよい。これにより、直前が「も」以外の文字である「のだから」だけが検索されることになる。

「のだから」と「んだから」を同時に検索し、「ものだから」「もんだから」を除外したければ、「[^も][のん]だから」とすればよい。

2.3 A|B ＝ A または B のいずれか

「せっかく」と「折角」を同時に検索したければ、「せっかく|折角」という正規表現を指定する。同様に、「しかし|けれども」「かなり|相当|ずいぶん|随分」などのようにして使う。

| は、[Shift]+[¥] ([Shift]キーを押し下げた状態で[¥]キーを押す) によって入力する (ただし、パソコンによってはこれと異なる可能性がある)。また、キーボード上の表示・画面表示・プリンタ印字が１本の縦棒になる場合と縦棒の中央部分がとぎれた形になる場合とがあるが、実体としては同じなので気にする必要はない。

2.4 (...) ＝ ...を１つのまとまりとして扱う

「〜のだから」の用例を集めるとき、「〜んだから」のように「の」が「ん」に変わる可能性も考慮する必要がある。丁寧体では「〜のですから」「〜のでございますから」などの形になる。これらを一括して検索したいときは、検索文字列として「[のん](だ|です|でございます)から」という正規表現を指定すればよい。「ものだから」などの用例を除外するには、冒頭に「[^も]」を加えて、「[^も][のん](だ|です|でございます)から」とする。

「([探捜]|さが)[さしすせそ]」(または「(探|捜|さが)[さしすせそ]」)「で(は|も|すら)ない」「(なければ|なくては|ないと)(ならない|いけない)」(または「な(ければ|くては|いと)(なら|いけ) ない」)「[ただ](ほう|方)が[いよ良]」が何を探そうとするものか考えてみよう。

2.5 A? ＝ Aの0回または1回の現れ

「でない」と「ではない」を同時に検索するには、「では?ない」とすればよい。「は?」は、? の直前の「は」があってもなくてもよいことを表す。「で(は|も|すら)?ない」とすれば、「ではない」「でもない」「ですらない」に加えて、「でない」も検索される。「けれども」「けれど」「けども」「けど」の４つは、「けれ?ども?」として表すことができる。

「をみたようだ」「みたようだ」「みたいだ」の３つを同時に検索するには、「をみたようだ|みたようだ|みたいだ」として表すこともできるが、共通部分をまとめて「(を?みたよう|みたい)だ」とすることもできる。(もっとも、冒頭に「を」があってもなくてもよいという条件は実質上意味を持たないので、「(みたよう|みたい)だ」あるいはそれをまとめて「みた(よう|い)だ」として検索しても同じ結果になる。)

2.6 . ＝任意の１文字（改行を除く）

「.っ.り」という正規表現を指定して検索すれば、「しっかり」「そっくり」「はっきり」「びっくり」「やっぱり」といった形の語の用例を探すことができる。

「日本語.?研究」という正規表現は、「日本語研究」および「日本語の研究」「日本語を研究」「日本語学研究」「日本語史研究」などに合致する。

2.7　A* ＝ Aの0回以上の繰返し

2.8　A+ ＝ Aの1回以上の繰返し

これらを ．や [...]、[^...]と組み合わせることによって、表現の共起・呼応のパターンを探すことができる。

例えば、「(よほど|よっぽど|余程).+うと思った」という正規表現を指定すれば、「よほど私は入ってみようと思った」のような用例を検索できる。「.+」は任意の文字の連続を表す。ただし、その正規表現では、「よほど〜〜。〜〜。〜〜うと思った。」のように「よほど」と「うと思った」が文境界をまたがり、係り受けの関係にないものまで検索にかかってしまう。そこで、「(よほど|よっぽど|余程)[^。]+うと思った」という正規表現を指定すれば、「よほど」と「うと思った」のあいだに句点が介在するものは除外できることになる。「[^。]+」は句点以外の文字の連続を示すからである。

2.9　A{m} ＝ Aのm回の繰返し

2.10　A{m,} ＝ Aのm回以上の繰返し

2.11　A{m,n} ＝ Aのm回以上n回以下の繰返し

「よほど」と「うと思った」が係り受けの関係にないものを除外するために、「(よほど|よっぽど|余程).{1,10}うと思った」という正規表現を使って検索することも考えられる。「.{1,10}」は任意の文字が1字以上10字以下連続したものを表し、「よほど」と「うと思った」が遠くかけ離れたものは除外されることになる。

もちろん、係り受けの関係にある「よほど」と「うと思った」のあいだに介在し得る文字の数には原理上の上限はないから、検索漏れを減らすにはnに大きめの値を指定する必要がある。しかし、nの値を大きくすれば「よほど」と「うと思った」が係り受けの関係にないものまで検索にかかりやすくなる。上で用いた「[^。]」と組み合わせて、「(よほど|よっぽど|余程)[^。]{1,30}うと思った」と指定することも考えられる。

しかし、この方法では「よほどのことがなければ放っておこうと思った」のようなものは排除できない。このように、正規表現を使っても、求める用例だけを漏れなく検索することのできない場合は多い。そのような場合、求める用例を漏れなく検索したければ、やや緩い正規表現を利用して検索し、検索結果に含まれるゴミ（求める用例でないもの）を手作業などによって取り除くという方法によるしかない。

2.12　¥1、¥2、¥3、〜＝ 後方参照

「(..)の¥1」という正規表現を使えば、「友達の友達」や「先生の先生」のように２文字の同一語が「の」をはさんで前後に現れる用例を検索することができる。「(.)不¥1」という正規表現は「運不運」「要不要」「適不適」といった表現を探すのに使える。¥1は、当該の正規表現で最初に現れる左括弧とそれに対応する右括弧とのあいだの要素――今の２つの例ではそれぞれ .. と . ――に合致した文字列を表す。ただし、¥1、¥2、¥3、～は後方（文字列の左のほう）の要素を参照するものであり、「¥1の(..)」「¥1不(.)」のようには書けない。

同様に、「(..)¥1」という正規表現を使えば、「いろいろ」「ますます」「飽き飽き」「一語一語」「毎日毎日」のような２文字の同一表現の反復の用例を検索することができる。しかし、. は任意の文字に合致するので、検索するテキストに例えば空白の４文字以上の連続が多数入っているとそうしたものまで検索にかかってしまうという問題がある。そのようなゴミが多い場合には、「([ぁ-ん亜-熙]{2})¥1」のように文字種を限定して検索すればよい。ここで、「[ぁ-ん亜-熙]」は平仮名または漢字の１文字を表す（３節を参照）。

１文字や２文字のように決まった長さの表現でなく、任意の長さの文字列に対して後方参照を行うときは検索の効率について注意が必要である。例えば、「人の人たる所以」「信長の信長たるところ」のように「XのXたる～」という形の用例（Xの長さは任意）を探すには、理屈のうえでは「(.+)の¥1たる」という正規表現を使えばよい。しかし、これではパソコンは条件に合わない膨大な数の可能性を試すことになり、実際に試してみれば分かる通り、検索の効率が著しく悪い。そこで、「([ぁ-ん亜-熙]+)の¥1たる」あるいは「([^、。]+)の¥1たる」などのようにする。こうすれば、Xが句読点を含む長い文字列の可能性までマッチングを試みるという無駄を省けることになる。あるいは、Xの長さは高々8文字だろうと見積もり、「(.{1,8})の¥1たる」のようにして検索するという方法も可能である。もしその結果Xが8文字ないしそれに近い長さの用例が出て来た場合は、「(.{1,12})の¥1たる」のように可能性を広げて検索し直せばよい。長めのＸの用例が出て来ないのであれば、検索し直す必要は小さい。

¥2（¥3、～）は、当該の正規表現で２番目（3番目、～）に現れる左括弧とそれに対応する右括弧とのあいだの要素に合致した文字列を表す。例えば、「(.{1,5})[がも]¥1なら、?(.{1,5})も¥2」という正規表現を使えば、「狸が狸なら、赤シャツも赤シャツだ」「向こうが向こうならこっちもこっちという気になる」といった表現の用例を探し出すことができる。

3　文字コードと[...]、[^...]における範囲指定

3.1　文字コードとは

コンピュータは個々の文字・記号をそれに割り当てられた固有の整数値として処理している。その整数値を文字コードと呼ぶ。

文字と整数値の対応付けにはいくつかの異なる方式――文字コード体系――がある。拙作KWICでは、Windowsで広く使われている文字コード体系Shift_JISの使用を前提としている。

Shift_JISの文字コード表の一部を本稿の最後に資料として掲げる。

文字コード表を見ると、例えば長音を表す「ー」は、左端に8150と書かれた行の0Bの列にある。これは、「ー」の文字コードが16進表記の815Bであることを意味する。16進表記のA～Fは10進表記の10～15であり、815Bは10進表記では33115になる（$8\times16^3+1\times16^2+5\times16+11=33115$）。同様に、Shift_JISの文字コード体系で最初に現れる漢字である「亜」のコードは16進表記の889Fであり、これは10進表記の34975に相当する。

3.2　文字コードの範囲を用いる指定

さて、2.1、2.2で説明した[...]、[^...]という表記法においては、複数の文字を列挙する指定に加えて、「[x-y]」という形での文字コードの範囲による指定が可能である。

例えば、「[ま-も]」は「[まみむめも]」と等価で、マ行の平仮名１文字を表す。「飲[まみむめも]」とする代わりに「飲[ま-も]」と表現することができる。

しかし、カ行の平仮名１文字を「[か-こ]」として表すことはできない。これは、文字コード表に見るように、文字コードが「...かがきぎくぐけげこご...」という順序で定義されていることによる。つまり、「[か-こ]」は、「[かきくけこ]」ではなく、「[かがきぎくぐけげこ]」と等価になる。したがって、カ行の平仮名１文字は「[かきくけこ]」のように列挙指定するしかない。１つの行の仮名１文字を範囲指定で正確に指定できるのは、濁音文字などのないナ・マ・ラの３行だけである。

数字についても同様の注意が必要である。アラビア数字１文字は問題なく「[０-９]」で表すことができる。これは文字コードが「０１２３４５６７８９」の順に定義されていることによる。それに対して、漢数字１文字を「[一-九]」のような形で表すことはできない。漢字は漢数字であるかどうかに関わりなく、読みや部首に従って配列されているからである。したがって、漢数字１文字は「[一二三四五六七八九]」ないし「[一二三四五六七八九十百千万億兆]」などのように列挙によって指定する必要がある。

3.3　列挙指定と範囲指定の組合せ

列挙指定と範囲指定を組み合わせることも可能である。例えば、「[ばびぶべぼま-も]」はバ行またはマ行の平仮名１文字を表す。「[だぢづでどな-のばびぶべぼま-も]」のように範囲指定を複数組含めることもできる。「飲[ま-もん]」は「飲[まみむめもん]」と等価で、「飲む」のすべての活用形を表す。

範囲指定の始点と終点の順序は文字コード表のそれに従わなければならないが、それ以外の順序は自由である。例えば、「[ま-も]」（＝マ行の１文字）を「[も-ま]」とすることはできないが、「[あいうえお]」（＝ア行の１文字）を「[あえいおう]」としたり、「[な-のま-も]」（＝ナ行・マ行の１文字）や「[ばびぶべぼま-も]」（＝バ行・マ行の１文字）をそれぞれ「[ま-もな-の]」「[ぼま-もばびぶべ]」としたりしても差し支えない。

3.4　仮名１文字、漢字１文字を表す正規表現

任意の平仮名１文字は「[あ-ん]」で表せる。ただし、文字コード表から分かるように「あ」の直前に小さい「ぁ」があるので、これも含めるには「[ぁ-ん]」とする必要がある。

任意の片仮名１文字は、長音を表す「ー」を加えて、「[ア-ンー]」として表す。「ー」は文字コードにおいて平仮名や片仮名とは離れた位置に定義されているので、範囲による指定はできない。また、「ア」の直前に「ァ」、「ン」の直後に「ヴ」があるので、これらも含めるなら「[ァ-ヴー]」となる。

任意の漢字１文字は、「[亜-熙]」で表せる。文字コード表に見るように、「亜」が最初の漢字、「熙」（文字コード EAA4）が最後の漢字となっているからである。この「熙」については入力に際して注意を要する。酷似した字形の異体字「熙」（文字コードE086）が別に定義されており、仮名漢字変換で「康熙字典」を出すとその異体字が出ることが多いからである。したがって、文字コードによって入力するのが確実である。なお、パソコンやプリンタの機種によっては、「熙」よりもさらに後ろのコードに追加の文字が割り当てられている場合がある。

よく使う正規表現は次のような具合に適当な読みで辞書登録しておくと便利である。

正規表現	登録読み
[ぁ-ん]	かなこーど
[ァ-ヴー]	かたかなこーど
[亜-熙]	かんじこーど※

※EmEditorで正規表現を使うときは、漢字１文字は「[亜-熙]」ではなく、「[一-龠]」で表す。これは、EmEditorはたとえShift_JIS文字コードのテキストを扱うときでも内部的には別の文字コード体系であるUnicodeで処理していることによる。平仮名・片仮名については上記の通りの正規表現で問題ない。

仮名１文字や漢字１文字を表すための正規表現については、状況や目的によってさらに考慮の余地がある。例えば、日本語の電子テキストで長音が「ー」ではなくハイフン（マイナス）の「－」を使って入力されていることがよくある。そうした可能性にも対処する必要があれば「[ァ-ヴー－]」としなければならない。平仮名での表記にも長音の「ー」が使われているテキストを検索するとき、目的によっては平仮名１文字は「[ぁ-んー]」と表すことが必要となる。また、漢字１文字を表す正規表現には「々」という踊り字を加えて「[亜-熙々]」としたほうがよい場合も考えられる。ほかにも、「三ヵ所」「三ヶ月」のような「ヵ」「ヶ」や、「ゝ」「ゞ」「ヽ」「ヾ」などの踊り字をどう扱うかといった問題がある。

（資料）２バイト文字（いわゆる全角文字）の Shift_JIS 文字コード表（JIS 1990 年改訂）

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F	
8140		、	。	,	.	・	:	;	?	!	゛	゜	´	｀	¨	^	←冒頭の8140は全角スペース
8150	￣	_	ヽ	ヾ	ゝ	ゞ	〃	仝	々	〆	〇	ー	―	‐	／	＼	←長音の「ー」や踊り字はここにある
8160	～	∥	｜	…	‥	‘	’	“	”	（	）	〔	〕	［	］	｛	
8170	｝	〈	〉	《	》	「	」	『	』	【	】	＋	－	±	×		
8180	÷	＝	≠	＜	＞	≦	≧	∞	∴	♂	♀	°	′	″	℃	￥	
8190	＄	¢	£	％	＃	＆	＊	＠	§	☆	★	○	●	◎	◇	◆	
81A0	□	■	△	▲	▽	▼	※	〒	→	←	↑	↓	＝				
81B0									∈	∋	⊆	⊇	⊂	⊃	∪	∩	
81C0									∧	∨	￢	⇒	⇔	∀	∃		
81D0											∠	⊥	⌒	∂	∇	≡	
81E0	≒	≪	≫	√	∽	∝	∵	∫	∬								
81F0	Å	‰	♯	♭	♪	†	‡	¶					◯				
8240																０	
8250	１	２	３	４	５	６	７	８	９								
8260	Ａ	Ｂ	Ｃ	Ｄ	Ｅ	Ｆ	Ｇ	Ｈ	Ｉ	Ｊ	Ｋ	Ｌ	Ｍ	Ｎ	Ｏ	Ｐ	
8270	Ｑ	Ｒ	Ｓ	Ｔ	Ｕ	Ｖ	Ｗ	Ｘ	Ｙ	Ｚ							
8280		ａ	ｂ	ｃ	ｄ	ｅ	ｆ	ｇ	ｈ	ｉ	ｊ	ｋ	ｌ	ｍ	ｎ	ｏ	
8290	ｐ	ｑ	ｒ	ｓ	ｔ	ｕ	ｖ	ｗ	ｘ	ｙ	ｚ					ぁ	┐
82A0	あ	ぃ	い	ぅ	う	ぇ	え	ぉ	お	か	が	き	ぎ	く	ぐ	け	│
82B0	げ	こ	ご	さ	ざ	し	じ	す	ず	せ	ぜ	そ	ぞ	た	だ	ち	│
82C0	ぢ	っ	つ	づ	て	で	と	ど	な	に	ぬ	ね	の	は	ば	ぱ	├ 平仮名
82D0	ひ	び	ぴ	ふ	ぶ	ぷ	へ	べ	ぺ	ほ	ぼ	ぽ	ま	み	む	め	│
82E0	も	ゃ	や	ゅ	ゆ	ょ	よ	ら	り	る	れ	ろ	ゎ	わ	ゐ	ゑ	│
82F0	を	ん															┘
8340	ァ	ア	ィ	イ	ゥ	ウ	ェ	エ	ォ	オ	カ	ガ	キ	ギ	ク	グ	┐
8350	ケ	ゲ	コ	ゴ	サ	ザ	シ	ジ	ス	ズ	セ	ゼ	ソ	ゾ	タ	ダ	│
8360	チ	ヂ	ッ	ツ	ヅ	テ	デ	ト	ド	ナ	ニ	ヌ	ネ	ノ	ハ	バ	├ 片仮名
8370	パ	ヒ	ビ	ピ	フ	ブ	プ	ヘ	ベ	ペ	ホ	ボ	ポ	マ	ミ		│
8380	ム	メ	モ	ャ	ヤ	ュ	ユ	ョ	ヨ	ラ	リ	ル	レ	ロ	ヮ	ワ	│
8390	ヰ	ヱ	ヲ	ン	ヴ	ヵ	ヶ										┘

（中略） ── ギリシャ文字・キリル文字など

8890																亜	┐
88A0	唖	娃	阿	哀	愛	挨	姶	逢	葵	茜	穐	悪	握	渥	旭	葦	│
88B0	芦	鯵	梓	圧	斡	扱	宛	姐	虻	飴	絢	綾	鮎	或	粟	袷	│
88C0	安	庵	按	暗	案	闇	鞍	杏	以	伊	位	依	偉	囲	夷	委	│
88D0	威	尉	惟	意	慰	易	椅	為	畏	異	移	維	緯	胃	萎	衣	│
88E0	謂	違	遺	医	井	亥	域	育	郁	磯	一	壱	溢	逸	稲	茨	│
88F0	芋	鰯	允	印	咽	員	因	姻	引	飲	淫	胤	蔭				│

（中略）

漢字（第一水準）

97D0	厘	林	淋	燐	琳	臨	輪	隣	鱗	麟	瑠	塁	涙	累	類	令
97E0	伶	例	冷	励	嶺	怜	玲	礼	苓	鈴	隷	零	霊	麗	齢	暦
97F0	歴	列	劣	烈	裂	廉	恋	憐	漣	煉	簾	練	聯			
9840	蓮	連	錬	呂	魯	櫓	炉	賂	路	露	労	婁	廊	弄	朗	楼
9850	榔	浪	漏	牢	狼	篭	老	聾	蝋	郎	六	麓	禄	肋	録	論
9860	倭	和	話	歪	賄	脇	惑	枠	鷲	亙	亘	鰐	詫	藁	蕨	椀
9870	湾	碗	腕													

漢字（第二水準）

9890																弌
98A0	丐	丕	个	丱	丶	丼	丿	乂	乖	乘	亂	亅	豫	亊	舒	弍
98B0	于	亞	亟	亠	亢	亰	亳	亶	从	仍	仄	仆	仂	仗	仞	仭
98C0	仟	价	伉	佚	估	佛	佝	佗	佇	佶	侈	侏	侘	佻	佩	佰
98D0	侑	佯	來	侖	儘	俔	俟	俎	俘	俛	俑	俚	俐	俤	俥	倚
98E0	倨	倔	倪	倥	倅	伜	俶	倡	倩	倬	俾	俯	們	倆	偃	假
98F0	會	偕	偐	偈	做	偖	偬	偸	傀	傚	傅	傴	傲			

（中略）

EA40	鵝	鵞	鵤	鵑	鵐	鵙	鵲	鶉	鶇	鶫	鵯	鵺	鶚	鶤	鶩	鶲
EA50	鷄	鷁	鶻	鶸	鶺	鷆	鷏	鷂	鷙	鷓	鷸	鷦	鷭	鷯	鷽	鸚
EA60	鸛	鸞	鹵	鹹	鹽	麁	麈	麋	麌	麒	麕	麑	麝	麥	麩	麸
EA70	麪	麭	靡	黌	黎	黏	黐	黔	黜	點	黝	黠	黥	黨	黯	
EA80	黴	黶	黷	黹	黻	黼	黽	鼇	鼈	皷	鼕	鼡	鼬	鼾	齊	齒
EA90	齔	齣	齟	齠	齡	齦	齧	齬	齪	齷	齲	齶	龕	龜	龠	堯
EAA0	槇	遙	瑤	凜	熙											