00章 正規表現って,どうゆう意味?
——それは言葉の規則性,パターンを表します

0.1 正規表現をご存じですか

正規表現は日常的によく使うようになると手放せない便利な道具ですが,私は最初のころ言葉の意味に悩みました。"**正規表現**"って,どういう意味じゃー?。何がどう"**正規**"なのよー?。

言葉の意味がよくわからないまま使っていると,心の隅に不安が残ります。でも,最近はやっとわかってきました。いきなり「正規」という意味不明な日本語を相手にすると,疑問はなかなか晴れませんが,元の英語は比較的わかりやすいです。

正規表現という日本語に訳されている元の英語は,regular expression(レギュラー・エキスプレッション……レギュラーな表現)です。レギュラーな表現。これなら,正規表現の性質を少々体験して知っている者にとって,なんとかわかりそうです。片仮名の「レギュラー」は,すでに日常の日本語としても定着しています。たとえばテレビのクイズ番組の「レギュラー回答者」は"**毎回決まって出演する回答者**"という意味です。それは,同じ人物が何度も何度も繰り返し出演ことを指しています。つまり,その人たちの出演の仕方には『**規則性**』があります。

レギュラーの反対のイレギュラー(irregular)も,「イレギュラー」という日本語としてよく使われます。野球でイレギュラー・バウンドといえば,ゴロの打球のバウンドが不規則なことです。ですから,レギュラーなバウンドは,弾み方に規則性があり,一定のパターンがあるので,野手が捕球位置を事前に正確に予測できるゴロです。そうするとイレギュラー・バウンドは,ボールの不規則な弾み方,一定のパターンに当てはまらない弾み方と言えますね。

レギュラーなバウンドをするゴロは,パターンを事前に読んで捕球できますが,イレギュラーにバウンドするゴロは名野手でも取り損なうことが多いです。この,パターンがある,ない,"**捕らえられる**"と"**捕らえられない**"の違いは,正規表現を理解するためにも重要です。

英語のレギュラーの意味を理解していくと,正規表現すなわちレギュラー・エキスプレッションの意味もわかってきます。つまりそれは,言葉の「規則性」を表現する方法です。最近は規則性という硬い言葉よりも,パターンという片仮名語がよく使われます(上の文ですでに使ってしまいました)。そこで正規表現は,言葉の一定のパターンを表現する方法だと言えます。

たとえばこんな正規表現※1は:


```
[ア-ン] {3}ビール
```

〈仮名3文字の直後にビールがある〉,というパターンや規則性を表現(express)しています。**〈ビールの直前に片仮名3文字がある〉**と言っても同じです。このパターンには,アサヒビールやキリンビールは当てはまりますが,サッポロビールは当てはまりません。サッポロは片仮名4文字ですからね。和歌山県の「くまのビール」は,ビールの直前が片仮名でなく平仮名3文字なので,これもこのパターンには当てはまりません。

※1 `[....]`は文字の種類を指定します。この中の`'-'`は文字の範囲を指定します。たとえば`[a-z]`と書けば英語のアルファベットの小文字を指定します。`[ア-ン]`は五十音表の中のほとんどの片仮名を指定しますが,ァ,ヶ,ヴなど,一部の特殊な片仮名はこの`ア-ン`の範囲に入りません。`{}`の中の数値は,その直前の正規表現の出現回数を指定します。そこで `[ア-ン] {3}` は"**片仮名3文字**"という意味になります。

0.2 正規表現を試してみませんか

正規表現はふつう,大きな文書の中に特定の文字列を見つけるために使います。たとえば,ある小説の中に日本人の女性の名前がいくつ登場するか知りたい,という文学研究の目的があったとします。では,"**日本人の女性の名前**"は,どんなパターン,つまりどんな正規表現で指定できるでしょうか?。

この問題は,本格的に考えはじめるとかなり難題です。ここでは例として,いちばん単純なパターンを挙げておきましょう。こんなのです:

```
[一-龠]{,3}.{,2}子
```

この正規表現は,`[一-龠]`で"漢字",`{,3}`で"それが3文字以内",`.`(ドット)で"何らかの文字(どんな文字でもよい)",`{,2}`で"それが2文字以内",子はそのまま"子"という文字が最後にあることを指定しています。つまり,苗字(姓)として漢字3文字以内,名前として○子または○○子(1〜2文字に続く'子')を想定しています。日本人の女性の名前を指定する正規表現としてはかなり貧しいといえますが,正規表現の単純な例としてご理解ください。

この正規表現を使って文学作品の分析をしていると,「宇多見ゆめ子」は,漢字3文字(宇多見),その次に何かの文字が2文字(ゆめ),最後が子ですから,パターンに当てはまり,無事に拾われます。しかし「大子守山加奈子」は(こんな女性の名前が実際にあったとすると!),この正規表現の網からはこぼれて,捨てられてしまいます……姓が「大子守」だとすると名前が「山加奈子」(**3文字+子**)になりダメ,名前が「加奈子」だとすると姓が「大子守山」と計4文字になり,これもダメです。

ここで2つの簡単な例で経験したように,正規表現は言葉に関する一定の規則性,一定のパターンを表す表現です。ですから,「正規表現」という日本語よりも,レギュラー・エキスプレッションという元の英語のほうが,すでに多くの片仮名日本語(外来語)が日常語の中に定着している今日の日本では,用語の意味を理解しやすいでしょう。

本当は,レギュラー・エキスプレッション(regular expression)の「レギュラー」は,形式言語学という特殊な言語学の分野で昔から「正則」と訳されている用語です。しかし私たちふつう人にとっては,日本語が正規表現でなく正則表現であっても,わかりづらい点では同じでしょう。「レギュラー(regular)」という形容詞のガクモン的な意味に関心のあるかたは,形式言語学※2と呼ばれる分野をすこしかじってみてください。

※2　ただし今日の正規表現の言語パターン指定機能は,もともとの形式言語学で言うレギュラー(正則)の意味を大幅にはみ出て多機能化しているので,「レギュラー・エキスプレッション」という用語を廃語にしようという動きもあります。たとえばPerlという,正規表現を多用するプログラミング言語の最近のドキュメンテーション(説明文書)は,regular expressionではなくregexpという新しい用語を使っています。私たちは,"レギュラー"を特定のガクモン用語でなく"言葉の一定のパターン"だと理解しておけば,用語をめぐる騒動にとりあえず巻き込まれずにすむでしょう。