

## 第4章 自分のプログラムでウェブにアクセスできると 小木曾智信 137

1. はじめに.....	137
2. ウェブ上のデータをダウンロードしてミニコーパスを作る.....	138
ウェブ上のデータを利用するには…138／データのダウンロード …140／ダウンロードに際しての注意点…148／ダウンロードし たファイルを加工して利用する…151／ミニコーパスを「ひまわ り」で利用する…157／形態素解析の必要性…168	
3. ウェブのデータと形態素解析・データベースを利用した研究例.....	169
調査の対象…169／UniDicを使った形態素解析…172／データ ベースを使った集計…173／分析の例…173	
4. ウェブ API .....	177
ウェブ API とは…177／ウェブ API を使ったプログラミング …180	
5.まとめ.....	182

## 第5章 ウェブと他のコーパスとの比較

前田広幸 185

1. ウェブと他のコーパスとの比較にあたって.....	185
2. コーパスのタイプ分類の諸観点.....	187
3. ウェブコーパスが一般的傾向として有する特性.....	189
4. ウェブコーパスと他のコーパスの利用について.....	190
大規模コーパス利用で収集されるようになった用例について …191／使用度数の時系列推移・書き手の属性とのかかわりの分 析…203／複数のコーパスからえられる検索結果をもとにした比 較…208	
5. ウェブと他のコーパスとの比較のまとめ.....	210

## 第1章

小野 正弘

### ウェブ検索概論

#### 1. はじめに

本章では、検索エンジンを利用して行う、ウェブにおける日本語に関する研究について概説する。

ウェブで用いられている日本語は、日本語研究のためのコーパスとして十分に魅力的なものであるということは、荻野（2008）、黒田他（2008）等に述べられているとおりであるが、その最大の魅力は、なんといっても、データの規模が極めて大きく（量）、しかも多様であること（質）であろう。

一般に、扱ったデータの量が大きければ大きいほど、そこから得られる情報に基づいた結論の信頼性は高くなる。例えば、ある日の新聞の1ページだけを見て出した結論と、すべてのページを見て出した結論のどちらが信頼できるかを考えてみれば、答えは明白であろう。さらに、それが、1カ月分、1年分、10年分…と増えていったら、信頼性はますます高くなっていくに違いない。とはいえ、10年分の新聞を独力で読み通すには、個人の力では絶望的な労力と時間を要する。それを一気に調査し終えるような手段があったら、と思う。検索エンジンとは、まさに、その夢をかなえるものなのである。

一方、データの量が仮に十分なものであっても、それがまったく均一的なものであったら、データとしての価値は低下してしまう。

まったく同じ例が10回あるよりも、10種類の例があって10回になっているほうが、いや、それよりも、a - 3回・b - 2回・c - 2回・d - 1回・e - 1回のように、aからeの5種類あって全部で10回になっているほうがあ

りがたい。質の多様性と、それぞれの勢力が同時に計れるからである。先ほどの新聞の例で言えば、政治面だけをずっと見続けて出した結論よりも、政治面・社会面・家庭面・スポーツ面等を総合的に見て出した結論のほうが信頼できる、ということである。もちろん、データの質が多様であるということは、否定的な観点からすれば雑多であるということでもあり、果たして、その結果が日本語のどこの何を表しているのかは慎重に見定める必要はあろう。しかし、だからといって、これだけのものを見逃すという手はないだろう。

このような魅力のあるウェブで用いられている日本語を自在に検索して、おおまかな見通しをつけ、さらに、本格的な研究にまで発展できたなら、日本語研究の方法に関して、有力で力強い選択肢を1つ獲得したことになるのである。

## 2. ウェブにおける日本語

ウェブにおける日本語の特性については、本巻、また、他の巻でも触れられるところがあろうが、ここでも少し整理しておきたい。

ウェブにおける日本語というのは、つまりは、インターネットにアクセスしたときに、Internet Explorer（IEと略される）等のブラウザの画面に現れてくる日本語のことである。この日本語は、さまざまなもので用いられている。さまざまなサイトで用いられているということは、さまざまな分野で用いられているということである。これが、前節で述べた「多様性」の1つである。

例えば、ある語を検索してみると、たちどころに、演劇関係のブログ、高速バスツアーの案内、ローカルテレビ局の番組の宣伝、雑貨商のブログ、書籍販売のサイト、インターネット上の辞書、動画のサイト等、実際にさまざまなページが、結果として表示される。場合によっては、ウィキペディアという、インターネット上の百科事典が検索される場合もある。

ウェブにおける日本語という括りは、青空文庫や国會議事録検索システ

ム、各新聞社の記事データベースのような、内容や範囲がある程度限定されたものに対して、限界がなく、均質性も保証されないものなのである。

そのような多種多様なページにおける用例から帰納した結論は、それだけ偏りのない一般性がある、とも言える反面、あまりに多種多様すぎて、全然結論が見えてこないという場合もありうることを考慮しなければならない。また、多種多様なものから帰納したかに見えて、実は、その中のある限られた部分だけからの結果が反映しているにすぎない、という場合もありうる。

また、それとは別の多様性もある。

ウェブの日本語は、小学生ぐらいの低年齢層から、かなりの高年齢層まで、広い範囲の年齢層によって書かれたものである。低年齢層が使用する語彙・語法と、高年齢層が使用する語彙・語法は微妙に異なることが多い。また、同じ語を使っていたとしても、それをどういう意味で使うかも異なる。だから、例えば、「やばい」を検索して調査したとして、ウェブ上では、〈追い込まれて、困った状態だ〉の例が何回あって、〈おいしい〉の例が何回あった、などということを平板的に報告しても、あまり面白いものにはならないだろう。

また、ウェブの日本語は、多種多様な職業の人間によって書かれたものもある。職業によって、ひどく堅い言い回しが好まれたり、ある職業独自の言い方や意味・用法があったりすることは、日常的に経験することもある。そのほか、性別、住んでいる土地なども異なる。意識して、また、意識することなく、方言的な意味・用法で書き込まれたものもある。

場合によっては、書き込まれたものに、筆者の性別や年齢、居住地などが示されているものもある。ブログやSNS（Social Networking Service）などの自己紹介欄がその1つである。中には信用できるものもあるかもしれないが、いわゆる「なりすまし」の可能性もある。多様性のほかに、このような不確実性も考慮しておかなければならない。例えば、80歳男性を自称している人の書き込みに、「このスイーツ、やばい！」などとあったからといって、「やばい」の〈おいしい〉の意味は、すでに、高年齢層にまで広がっている、などと結論づけることは早計であろう。