人文・社会系

現代日本語の全体像を把握するための コーパス（データベース）を構築

国立国語研究所研究開発部門言語資源グループ長　前川　喜久雄

【研究の背景】

私たちは『現代日本語書き言葉均衡コーパス』を構築しています。コーパスとは言語研究用のデータベースのことですが、対象となる言語の全体像を歪みなく反映しているものを特に均衡コーパスと呼び、定量的な言語研究に欠かすことのできない大切なインフラです。しかし、日本語にはこれまで本格的な均衡コーパスが存在せず、大きな問題となっていました。

【研究の成果】

『現代日本語書き言葉均衡コーパス』は「出版」、「図書館」、「特定目的」の3種のサブコーパスから構成される書き言葉コーパスで、全体で約1億語のデータを格納します（図1）。

出版サブコーパスは2001年から2005年の間に出版された書籍、雑誌、新聞を母集団とした3500万語のコーパス、図書館サブコーパスは東京都下の公共図書館に所蔵されている全書籍のうち13館以上に所蔵されているものを母集団とした3000万語のコーパスです。これら二つのサブコーパスは、統計的なサンプリング（データの無作為抽出）を行っている点で、従来世界各地で開発されてきたコーパスにはない特長を具えたものとなっています。

特定目的サブコーパスは、コーパスの研究遂行上必要とするデータを格納した3500万語のコーパスで、検定教科書、白書、ベストセラー、法律、論文、ウェブ上のテキストなどのサンプルを格納しています。

今回（平成19年6月）、著作権処理を終了したデータのうち1000万語分をインターネットで公開しました（図2）。検索したい語句を入力すると、それを含んだ用例が最大で500件まで表示されます。是非試してみてください（http://www.kotonoha.gr.jp/demo/）。

【今後の展望】

『現代日本語書き言葉均衡コーパス』は2011年に一般公開する予定です。このようなコーパスが普及することによって、国語教育や辞書の編集なども、識者の経験や勘に頼るのではなく、定量的なデータに基づいて行われるようになることが期待されます。

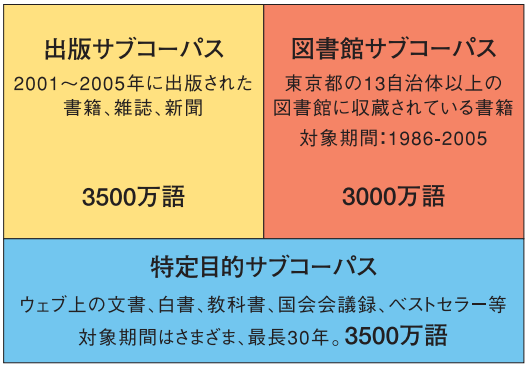

図1　『現代日本語書き言葉均衡コーパス』の構造

図2　検索デモンストレーションサイト

【交付した科研費】

平成18ー22年度　特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築」
:21世紀の日本語研究の基盤整備