人文社会系

コーパス日本語学の挑戦

文学部行動科学科・教授　伝　　康晴

研究の背景

私たちは普段どんな語をどれくらい目にしているでしょうか。政府白書のような硬い文章とブログのような軟らかい文章の違いは何でしょうか。こういった疑問に答えることは、私たちの言語活動の実態を明らかにし、ことばを理解する心の仕組みを明らかにすることにつながります。

このような研究を進めるためには、多様なジャンルや文体をカバーした言語資料（コーパス）が必要です。大規模コーパスに基づく日本語研究はコーパス日本語学と呼ばれ、始まったばかりの分野です。コーパス日本語学の推進のため、平成 18 年度から国立国語研究所を中心とする特定領域研究がスタートし、1 億語を超える規模の現代日本語書き言葉コーパスの構築に着手しました。私はこのプロジェクトで「電子化辞書」班を組織し、コーパス構築に中心的な役割を果たしました。

研究の成果

「さくらの木が 1 本ありました」という文を考えてみましょう。この文が一体何語からなるか、一目見ただけではわかりません。英語のように空白で語を区切る習慣のない日本語では、研究利用のために、テキストを語に分割し、品詞などの情報を付与する必要があります（右図）。

これをコンピューターで自動的に実行する技術は形態素解析と呼ばれ、30 年以上前から研究されています。形態素解析の実行には、テキストに出現しうる語を列挙し、品詞などの情報を定義した辞書が必要です。これが電子化辞書です。従来は、表記のゆれ（「桜」「さくら」「サクラ」）をまとめたり、同表記異語（せいぶつ／なまもの【生物】）を区別したり、文脈によって発音を変化（「イチホン」→「イッポン」）させたりといったことができませんでした。私たちは、新しい考え方で辞書を設計してこれらの課題に対処し、語彙数 20 万語超・解析精度 98.5% 超（200 語に 3 語以下の誤り）の電子化辞書を開発・無償公開しました。この辞書は、特定領域研究のコーパス構築に利用され、語彙調査やジャンル特徴の分析といったコーパス日本語学研究に応用されるとともに、米国 Apple 社を始め、数社とライセンス契約（無償）を結び、商用でも利用されています。

今後の展望

書き言葉コーパスと比べ、話し言葉コーパスの構築は遅れています。特に、私たちの日常の言語活動の中心である会話については、書き言葉のように資料として残りにくく、また、言い淀みやあいづちなど無意識的な言語使用が多く、大規模な実データの収集・分析が不可欠です。この課題についても、特定領域研究と並行して基盤研究 (B) で検討を開始しており（平成 23 年度より後継課題を開始）、今後、コーパス日本語学の射程をもっと広げていきたいと考えています。

さくらの木が1本ありました

表記	発音	辞書見出し	品詞	活用型	活用形
さくら	サクラ	さくら【桜】	名詞-普通名詞-一般		
の	ノ	の	助詞-格助詞		
木	キ	き【木】	名詞-普通名詞-一般		
が	ガ	が	助詞-格助詞		
1	イッ	いち【一】	名詞-数詞		
本	ポン	ほん【本】	接尾辞-名詞的-助数詞		
あり	アリ	ある【有る】	動詞-非自立可能	五段-ラ行	連用形
まし	マシ	ます	助動詞	助動詞-マス	連用形
た	タ	た	助動詞	助動詞-タ	終止形

【支援を受けた科研費】
平成18～22年度　特定領域研究「多様な目的に適した形態素解析システム用電子化辞書の開発」
平成20～22年度　基盤研究(B)「対話における発話単位と機能の認定に関する研究」

【備考欄】

掲載論文：伝ほか (2007) コーパス日本語学のための言語資源: 形態素解析用電子化辞書の開発とその応用, 日本語科学, *22*, 101-123. Den et al. (2010) Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme, *Proc. LREC'10*, 2103-2110.
受賞：共同発表で情報処理学会から平成22年度山下記念研究賞を受賞
成果公開 HP：http://download.unidic.org