シンポジウム

「進化するWebコーパス：現状と課題」

投野由紀夫

まえがき

本シンポジウムは、近年急速に研究領域として拡大しつつあるWebコーパスの動向を紹介し、英語コーパス研究におけるその位置づけ、現状と課題を検討するという趣旨で、2010年4月24日、英語コーパス学会第35回大会（会場：兵庫県立大学）において開催された。

Webコーパスはこの10年ほどで圧倒的な速度で増え続けている。2006年9月にはGoogleが1兆語のWebデータをもとにしたn-gram頻度リスト、Google 1T 5-gram Version 1をLinguistic Data Consortiumから配布したのは記憶に新しい。Webコーパスを利用した代表的な検索システムがSketch Engine (http://www.sketchengine.co.uk) である。2010年10月1日現在、Sketch Engineのトップ画面からは52種類の多言語コーパスが検索可能であるが、そのうちの約20種類がWeb上のテキストをソースとして構築されたコーパスであり、特に1億語以上の大規模コーパスはほとんどがWebコーパスである。またSketch Engine以外にも、一般利用者がWebコーパスを簡単に作成できるツール群（例：BootCat[1]）や、Web全体をコーパスとして検索するような検索エンジンとコンコーダンサーの融合ツール（WebCorp[2]など）も4,5年前よりも格段に品質向上してきており、Webコーパスを取り巻く利用環境は加速度的に充実してきている。

本シンポジウムでは、講師として、田中省作（立命館大学）、大羽良（中央大学）、中村隆弘（ネットアドバンス）、星野守（小学館）の4氏をお招きし、投野がモデレーターおよび問題提起の発表を行った。まず投野が「Webコーパス概観」と題する導入的発表を行った後、田中講師が「Webコーパスの言語情報処理基盤」という題で、Webコーパス構築に関する自然言語処理の基礎技術の紹介および研究基盤としてのWebコーパス、個別研究におけるWebコーパスの位置づけを論じた。続いて、大羽講師が「Webコーパス研究におけるブログの可能性とその文体的特徴」というテーマで、Web上に存在するブログのもつコーパスとしての潜在的可能性とその話し言葉・書き言葉の両面を併せ持つ文体的な特徴について実証データに基づき論じた。中村・星野両講師は「小学館Webコーパス『Sekai Corpus』構築とその活用」と題して、小学

『英語コーパス研究』第18号（2011), pp.81-82

館コーパス・ネットワークの一環として構築した20億語規模のSekai Corpusの概要と、その専門分野英語比較の作業から得た活用可能性を論じていただいた。最後に再び投野による総括的な発表の後、パネリストおよびフロアとのディスカッションに入った。

以下に掲載した3編の論文は、田中氏、大羽氏、および投野の発表内容をもとに加筆修正したものである。中村・星野両氏の発表に関しては、残念ながら今回論文の形で掲載することはできなかったが、最後のディスカッションの内容などはできるだけ投野の論文に盛り込む形でまとめさせていただいた。

本シンポジウムの3編の総括論文が、英語コーパス研究におけるWebコーパスの利活用に少しでも参考になれば望外の幸せである。

注

[1] http://bootcat.sslmit.unibo.it/

[2] http://www.webcorp.org.uk/