特集「日本語コーパス」にあたって

松本　裕治
（奈良先端科学技術大学院大学）

言語学の研究においては，チョムスキーによる人間の言語機能の言語能力（language competence）と言語運用（language performance）への区分と，文法研究が前者に関する研究であるとの立場を受けて，長く内省に基づく文法研究がその主流であった．言語処理研究においても，文法理論の研究や具体的な言語に対する文法規則の記述は，いわゆる文法的に正しい文の解析を対象とするものが主で，言語運用上に現れる文法的に不適格な文の解析は研究対象の中心ではなかった．しかし，現実には，従来の文法理論では説明できない言語現象や文法に逸脱したと考えられる文が少なからず用いられており，人間の言語使用の実体を客観的に眺める必要が生じてくる．言語処理研究において，旧来の箱庭的な文脈での言語解析ではなく，現実に存在する文の解析が実応用として期待されるようになり，いわゆる頑健な言語処理技術が重要なテーマになってきている．

現実にあるがままの文を対象にする言語学および言語処理研究において，言語の実際の使用を広くカバーする大規模なテキストデータの存在は極めて重要である．90年代半ばから飛躍的に発展したインターネットが世の中に浸透し，電子化された言語データの入手自体は困らない状況になった．しかし，著作権問題に接触せず追試が可能な形で共有できる言語データということになると，青空文庫の小説データや，有償で契約を交わしたうえで利用許諾が得られる新聞記事データなど，利用可能なテキストデータは限られていた．何らかの視点と目的をもって収集された言語テキストデータのことをコーパス（corpus）という．最も有名なのは英国のBritish National Corpus（BNC）で，話し言葉をも含む広い範囲からの英語の文例を集めた1億語規模のコーパスであり，幅広い分野や状況における英文が含まれている．このようなバランスを考慮したコーパスは均衡コーパス（balanced corpus）と呼ばれる．

本特集は，BNCに匹敵する，あるいはそれ以上の規模をもつ日本語の均衡コーパスの構築を目標として，2006年に発足した文部科学省特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築」について，その目的，関連領域の動向，現状，および今後の計画を解説するものである．国立国語研の前川喜久雄を代表者とする本プロジェクトは，二つの研究項目に分かれた八つの計画班，および，四つの公募班からなっている．計画班の構成と代表者は次のとおりである．

- 研究項目A01コーパスの構築
  - —データ班：山崎　誠（国立国語研究所）
  - —ツール班：松本裕治（奈良先端科学技術大学院）
  - —電子化辞書班：伝　康晴（千葉大学）
- 研究項目B01コーパスの評価
  - —日本語学班：田野村忠温（大阪大学）
  - —日本語教育班：砂川有里子（筑波大学）
  - —言語政策班：田中牧郎（国立国語研究所）
  - —辞書編集班：荻野綱男（日本大学）
  - —言語処理班：奥村　学（東京工業大学）

本プロジェクトの特徴は，「均衡」という性質をいかに達成するかという取組みと著作権処理の問題への対応を考慮したコーパス構築法，品詞タグ付けのための単語単位の認定基準とそれに準拠した辞書の構築，コーパス利用のためのツールの構築など，コーパスの構築と利用を視野に入れたコアグループ（A01）を構成していること，および，構築されるコーパスの多方面からの応用を想定して組織された評価グループ（B01）からなることである．これにより，単にコーパスを構築するだけでなく，その基になった電子化辞書，その辞書に従って未解析のコーパスを解析し，それをさまざまな視点から検索するツールが提供される予定である．コーパスへのフィードバックは，評価グループを構成する計画班や公募班からだけでなく，著作権処理の終了したデータをモニタ版として無償で公開することにより，一般の利用者からも積極的に取り入れるようにしている．本年度のモニタ版の情報は以下から入手可能である．

http://www.kokken.go.jp/kotonoha/ex_8.html

本特集は，前川氏によるプロジェクトの全体概要と意義の解説に続き，辞書編集班を除く各計画班の代表者によるそれぞれの分野でのコーパスに関する研究の現状と各班の取組みを解説している．読者には，まず前川論文で概要を把握することをお勧めする．山崎論文と伝論文はコーパスと辞書を連動させた構築を理解するため並行して読んでいただきたいと思う．その他の解説は比較的独立して読めるようになっている．