特集にあたって

投野由紀夫（東京外国語大学）

ことばの能力は人間のもつ最も高次の認知能力の1つであり，脳科学が日進月歩する中，言語中枢の解明は最後の難関ともいわれている．言語能力はソシュールのラングとパロールの区別のごとく，人間のこころ(脳）の中にあり，その力を外に現れる言語使用をもとに推定するしかない．

ことばの仕組みや力の解明に関して，過去にさまざまな試みがなされてきた．20世紀前半はヨーロッパでは記述言語学が，また米国では構造主義言語学の考え方が主流であった．それはフィールドワークによる言語の標本観察を通して，言語をできるだけ最小の有意味な単位に細分化していき，それらの項目の分類整理・階層化が中心的課題であった．それらの構造解析の成果を受けて，チョムスキーが20世紀半ばに，単なる構造の分類整理では不十分で，より生成的なメカニズムによる有限の規則から無限の文法的な文を作り出す文法規則のモデルが必要であると主張．さらに，そのような知識を一定期間でこどもが獲得可能にする言語獲得装置としての普遍文法（universal grammar）というものを提唱したのである．

この考え方は人間のことばの創造的な側面やこころの解明という「認知科学」と呼ばれる新分野を開拓するのに功績大であったが，一方で言語の仕組みを解明する方法論に変化が生じた．すなわち，文法は人間のこころの中にあり，実際の言語使用データを参照せずとも，人間が文法性の判断をもとにモデル構築を行えば事足りると考えたのである．よって生成文法の方法論的には内省（introspection）が重視され，実際の言語使用データは軽視されるという皮肉な結果となった．

チョムスキーの生成文法の試みは，その後さまざまな関連分野に影響を与えつつ今日に至っているが，その中で1つの大きな変化が，「利用できる言語資料の増大」であろう．1960年代，最も大きな英語コーパスは100万語のブラウン・コーパス(Brown Corpus)であった．この規模でチョムスキーがコーパスを非力だと思ったのも無理はない．しかし今日，コンピューターとインターネットの普及により，数十億から1兆語規模のコーパス・データを扱う研究も行われるようになってきた．

これだけの大量データが利用可能になると，言語の仕組みや力の解明に，大量テキストを利用した新しい方法論の期待が高まってくる．今回の特集で，我々が見るのはそのような大量テキストを基にした自然言語処理の最前線である．浅原氏は自然言語への情報付与を系列ラベリングのタスクと位置づけ，最新の形態素解析技術の動向を紹介する．松本氏には，統語解析の分野で近年注目を集めている統計的統語解析アプローチを解説していただく．さらに具体的な応用分野として，永田氏には統計的機械翻訳，相澤氏には共起に基づく類似性尺度とその利用法，工藤氏には頻出パターンマイニング，徳永氏には情報検索・抽出・要約・分類，QAといった知的情報アクセスの分野の動向をご紹介いただく．どれも，今日のインターネット社会の情報検索を裏側で支えている技術であるといえる．

私自身は，コーパス言語学の応用分野（特に言語教育）が専門であるので，今回の特集記事を書かれた自然言語処理の先生方の論考はどれも消費者・エンドユーザーの立場として読ませていただき興味深いものであった．ことばの能力とその獲得のメカニズムの解明という言語学の大目標に照らしてみると，大量言語使用データが利用可能な今日においては，チョムスキーが言っていた内省重視の方法論から，新たな次元での「言語使用からの推定」を行える時代に我々はいる，という認識を新たにするのである．

私の期待は，このような確率統計的な手法を駆使した大量言語データの分析により，ことばの仕組みの解明が進み，特に私が専門としている言語教育の分野で革新的な学習理論や学習シラバス，支援システムが構築されることである．そのような特集を次回本誌で出来ることを期して，本特集のプロローグとしたい．

本特集の構成や人選に関して奈良先端科学技術大学院大学の松本裕治先生にお世話になった．ここに謝意を表したい．