特集「コーパスに基づく音声・自然言語処理」にあたって

田中　裕一*1　松本　裕治*2　平川　秀樹*3

昨今，音声情報処理や自然言語処理の分野ではコーパスを用いた研究が花ざかりである．例文集としてのコーパスそれ自身は古くから研究対象あるいは研究上の資料として用いられていたが，量的にもバリエーションとしてもそれほど多くはなかった．しかし，この数年ほどのコーパスの蓄積と，これに基づく研究の広がりは目をみはるばかりである．これは特に海外で顕著であるが，何年か遅れて日本でもこの傾向は強まってきた．

その理由として，音声情報処理の分野では，例えば音声認識におけるHMM（Hidden Markov Model）など，確率・統計的手法が成果を上げてきているといったことがあり，自然言語処理の分野では，これまで主流であったルールベースの処理のボトルネックを解消するために，EBMT（Example-based Machine Translation）などの研究が盛んになってきており，それに近年の電子化テキストの発行・流通が輪をかけているということなどが考えられるだろう．

また，こうして集められた，あるいは集めようとしているコーパスデータに関しては，その内容，労力の重複を避けるために共有化の動きがでてきている．米国ではLDC（Linguistic Data Consortium）が中心になって進めており，日本でもLRSI（言語データ共有計画）が立ち上がろうとしている．このような時期にあたり，言語に関連した広い見地からコーパスに基づく処理の研究を見渡そうという目的で今回の特集が企画された．またこれは，人間の言語という同じものを対象とする分野でありながら，これまで交流が必ずしも盛んとはいえなかった音声情報処理と自然言語処理の両分野の研究者に対して，コーパスという共通の基盤を通じて互いの研究内容を紹介しあうことも併せて意図している．

さて，以上のような趣旨で，本特集においては，まず音声・言語コーパスの現状の概観を竹沢寿幸氏，末松　博氏にお願いした．この「音声・テキストコーパスとその構築技術，標準化動向」においては，両分野のコーパスに対して，諸元，入手方法などに関する詳細なリストを作成いただき，それに対して開発動向，標準化動向，コーパス構築技術などに関する解説をつけていただいた．この資料からも，冒頭に述べたコーパスデータの広がりのようすがうかがい知ることができることと思う．

次に音声情報処理に関して，中村　哲氏，北　研二氏，永田昌明氏に2編の解説をお願いした．まず「音声言語の確率モデル」では，統計的ゆらぎを伴う情報源である音声のモデル化に対する有力な方法としての隠れマルコフモデル（HMM）を中心として，音声言語処理における確率モデル，それに対する最ゆう推定法やEMアルゴリズムを基本としたモデルのパラメータ推定法について述べられている．ここで紹介された方法は，統計的ゆらぎに対して頑健なシステムを提供するが，そのためには大量の音声データを用いたモデルパラメータの推定が必要であり，大規模な音声コーパスの利用例として興味深いものである．

引き続いて，同じ著者に「確率・統計モデルの音声言語処理への応用」というタイトルで，上記のモデルが実際の音声認識や自然言語処理でどのように用いられているのかを中心に解説していただいた．前半は音声言語処理におけるHMMを用いた連続音声認識の基本的な考え方およびアルゴリズムの紹介，後半は自然言語処理への応用例である．ここでは，確率・統計モデルに基づく品詞づけや形態素解析，確率文脈自由文法を用いた構文解析法の紹介の後で，確率文脈自由文法の弱点である文脈情報や語彙情報の記述能力の不足を解消する新しい構文解析モデルについても触れていただいた．

一方，自然言語処理の側からの解説として，まず宇津呂武仁氏と編集委員の一人，松本とで，「コーパスを用いた言語知識の獲得」を執筆した．ここでは，自然言語処理の課題のなかでも，動詞の格フレームや語の

＊1　(株)富士通研究所 マルチメディアシステム研究所 ソフトウェア研究部
＊2　奈良先端科学技術大学院大学 情報科学研究科
＊3　(株)東芝 研究開発センター 情報・通信システム第三研究所

階層構造といった，より意味的な情報をコーパスや辞書から獲得する手法を解説した．

最後に，工藤育男氏と井ノ上直己氏に，「コーパスに基づく共起知識の獲得とその応用」をお願いした．この論文では，前半で自然言語処理におけるコーパスの利用に関する一般論と，そのためのテキストコーパスのあるべき姿とが述べられ，後半では，特に共起関係に関する知識をコーパスから獲得する手法，およびその応用が述べられている．

以上，5編の解説論文によって，音声・自然言語コーパスに関する話題を広く取り上げたつもりである．まだカバーしきれないテーマは多々あるが，それは読者諸賢がそれぞれ参考文献に当たって理解を深められたい．本特集を実現するために，著者の皆さんにはご多忙中のところ快く解説をお寄せくださり，深く感謝する次第である．この分野に興味を持つ研究者が増え，理解が深まって，音声情報処理・自然言語処理研究の発展に大きく資することを期待する．