コーパス日本語学の射程

丸山　岳彦　　田野村　忠温
（国立国語研究所）　　（大阪大学）

キーワード
コーパス日本語学，現代日本語書き言葉均衡コーパス

要　旨

現在国立国語研究所において構築が進められている「現代日本語書き言葉均衡コーパス」が2011年に完成し，日本語初の大規模な均衡コーパスを誰もが利用できるようになる。これにより，諸外国，諸外国語に大幅な遅れを取っていた日本語のコーパス言語学的な研究は，新たな段階を迎えるものと期待される。

「コーパス日本語学の射程」と題した本特集の巻頭論文として，本稿では日本語研究におけるコーパスの利用の歴史を振り返り，将来の展望やコーパスの利用をめぐって注意すべきいくつかの問題について述べるとともに，特集に収めた各論文について簡単に紹介する。

1. はじめに

世界各国でさまざまなコーパスの構築・公開が進められている。1964年のブラウンコーパスに端を発するコーパスは，今や，書き言葉コーパス，話し言葉コーパス，学習者コーパス，多言語コーパス，Web コーパスと，多様な内容・形態において作成されている。そして，英語の研究を中心に，文法の考察や語の意味の分析，語彙表の作成，文法書・辞書の編纂，言語教育への活用，言語処理技術への応用と，コーパスを利用した広範囲な研究成果が挙げられている。今後の言語研究においてコーパスがますます重みを増すことは間違いなく，今なおコーパス利用の黎明期にある日本語研究もその動向に関して例外ではないはずである。

この『日本語科学』第22号では，コーパスに基づく日本語研究――これを「コーパス日本語学」と呼ぶことにする――の今後の発展を願い，その可能性のほどを問うために，「コーパス日本語学の射程」と題する特集を企画した。以下では，コーパス日本語学の背景や課題を明らかにするとともに，特集に収められた各論文の概要とその位置付けについて述べる。

2. 日本語研究におけるコーパス利用の歴史

過去半世紀余りの日本語研究を特徴付ける要素の１つは，研究方法論上の内省の重視であろう。特に文法の領域においては，1960年代半ば以降，生成文法の理論と手法にのっとり母語話者の内省や言語直感を駆使した日本語の分析が行なわれ，めざましい成果を収めた。有効性の示された内省重視の研究法は，生成文法の枠組みにはよらない記述的な日本語文法研究にも大きな影

響を及ぼした。

2.1. 第 1 期コーパス日本語学

そうした動向の一方で，現実に用いられた言語表現を大量に収集して統計的に分析・記述を行なう研究法が，国立国語研究所において早い時期から実践されてきた。本特集の宮島氏の論文に詳しく述べられているように，国立国語研究所の厳密なサンプリングに基づく統計的な語彙調査は，世界的に見ても同種の研究として極めて早い時期のものである。新聞や雑誌から大量の実データを収集して語彙調査を行ない，さらにそこで得られた用例から文法記述を行なうという一連の研究は，定量的な調査と定性的な言語記述を同時に実現しているという点で，現代のコーパス言語学の先駆的な存在であったと言える。

当初は紙媒体のコーパス（言語資料体）であったが，その後1960年代後半から1970年代にかけて，国立国語研究所では電子計算機と電子化した日本語資料を利用した文法や語彙の研究を集中的に行なっている。当時はまだ「コーパス」という語は使われていなかったものの，我が国におけるまさにコーパス言語学の嚆矢であった。

国立国語研究所の当時の取り組みはコーパスを利用した日本語研究，すなわちコーパス日本語学の第 1 期と位置付けることができるが，残念なことにそこで作成されたコーパスは公開されることがなく，国立国語研究所の取り組みも比較的短期間で終息した。このため，第 1 期コーパス日本語学は日本語研究の学界全体に対して大きな影響を与えることはなかった。

2.2. 第 2 期コーパス日本語学

1990年代前半にパーソナルコンピュータの高性能化と普及に伴って新聞記事や文学作品の電子出版が始まり，それを日本語研究に利用する試みが一部の個人研究者によって始められる。これ以後，現在に至るまでの状況をコーパス日本語学の第 2 期と位置付けることができる。「電子化コーパス」という用語が使われ始め，その後「電子コーパス」そして「コーパス」と略称されるようになる。コーパスの利用は年々一般化し，1990年代後半にはインターネットやサーチエンジンの普及に伴い，Web 上のデータを日本語の考察に役立てる試みもなされるようになる。

第 2 期コーパス日本語学は，市販の電子出版物あるいは Web 上のデータといった，誰でも入手することのできる電子媒体の日本語資料を利用しているところに特徴がある。しかし，未解決の大きな課題が 2 つあった。1 つは，均衡コーパスの不在である。すなわち，研究者はたまたま入手できる電子資料を使っているに過ぎず，日本語全体を偏りなく代表させるという配慮に基づいて作られた言語研究資料としてのコーパスは存在しない。もう 1 つは，確立された研究法の不在である。電子資料を用例の収集に用いることは今や常識化したが，大規模な電子資料ならではの特性を生かした日本語研究ということになると少数の研究者が独自に模索を試みているに過ぎない。

2.3. 第3期コーパス日本語学

第2期コーパス日本語学の2つの問題はいずれもしばらく解決の兆しが見えなかったが，1つ目の問題については最近になって大きな状況の変化が生じた。国立国語研究所において，2006年度からの5年計画で，現代日本語書き言葉の大規模な均衡コーパスの構築が開始されたのである。詳細は前川氏の論文および本稿末に掲げるWebサイトに譲るが，2011年には大規模な日本語の均衡コーパスが完成し公開される見通しである。それをもってコーパス日本語学は第3期に移行すると言ってよいであろう。

大規模な均衡コーパスが利用可能になり，また，世界的な言語研究法の趨勢から考えても，これからの日本語研究においてコーパスの果たす役割はますます拡大していくであろう。しかし，第3期の到来を数年後に控えた今，コーパス日本語学の可能性について見通しを得る努力を始める必要がある。コーパスを使ってどのような研究ができるのか，それにはどのような方法が必要となるのかといったことについて，確かな理解がほとんど得られていないからである。

この特集は，そのような状況認識に基づき，コーパス日本語学の射程を明らかにし，その展望や課題を論じるために企画された。具体的には，コーパス日本語学をめぐる状況の回顧と展望，コーパスを利用した日本語研究の実践，コーパスの構築や利用に関わる研究，という3つのカテゴリーに属する論文を収めた。なお，射程とは今我々に予見可能な研究の可能性の範囲という程度の意味であり，射程をさらに伸ばす研究が今後次々に現れることが予想され，かつ期待されることは言うまでもない。

3. 収録論文の概要と位置付け

本特集には，寄稿論文3編，研究論文3編，調査報告1編，研究ノート2編の計9編の論文を収める。ここでそれぞれの論文の内容を簡単に紹介し，それが本特集あるいはコーパス日本語学において占める位置付けを明らかにしたい。

以下，収められた9編の論文について，その内容の観点から上述の3つのカテゴリーに分けて述べる。

3.1. コーパス日本語学をめぐる状況の回顧と展望

まず，前川喜久雄氏「コーパス日本語学の可能性――大規模均衡コーパスがもたらすもの――」では，国立国語研究所が構築を進めている「現代日本語書き言葉均衡コーパス」の設計に触れた上で，大規模均衡コーパスがこれからの日本語研究に及ぼす影響を論じている。「現代日本語書き言葉均衡コーパス」は，綿密な設計に基づく日本語初の均衡コーパスであり，今後のコーパス日本語学において中心的な役割を果たすことが期待される。前川氏は，大規模なコーパスの利用によって類義語やコロケーションの研究が一層進展するとともに，文法性判断の個人差を説明するモデルを構築するといった従来では考えられなかった研究が開花する可能性を指摘している。言語現象の分析・記述の精密化と新規の研究領域の開拓という2つの面において，現代日本語書き言葉均衡コーパスは今後の日本語研究に大きな恩恵をもたらすはずである。

宮島達夫氏「語彙調査からコーパスへ」は，設立当初から続いてきた国立国語研究所の語彙調査を振り返りつつ，特に1962年の「雑誌九十種調査」が持っていた，そして今もなお持っている先見性について論じている。宮島氏の挙げる「見出し語立て」「標本抽出の方法」「代表性の確保」など，50年近くも前の調査で採用された方法論は，今後のコーパスを用いた語彙調査や言語研究においても有効である。宮島氏はまた，巨視的に言語の全体像を眺める研究と，微視的に言語事実を記述する研究とが，総合的に行なわれるべきであることを主張する。例えば，前者はコーパスによって語種や語彙の変遷の実態を明らかにすることであり，後者はコーパスに記録された文脈を利用して文法現象を個別に分析することである。定量的な調査と定性的な言語分析の両立という仕事は，今後のコーパス日本語学における大きな課題と言えるであろう。

後藤斉氏「コーパス言語学と日本語研究」は，英語を対象としたコーパス言語学が発展してきた経緯をたどった上で，日本語を対象としたコーパス言語学がこれまで定着してこなかった，あるいは散発的にしか行なわれてこなかった要因を分析する。その上で，現代日本語書き言葉均衡コーパスの構築が状況改善のきっかけになることへの期待を述べるとともに，日本語研究におけるコーパスの活用のためには，コーパスに関する知識とそこから必要な情報を引き出すための技術を得る主体的な努力が日本語研究者に求められること，特に付加情報の付与されたコーパスの利用の可能性の検討が今後重要な課題になることなどを指摘している。

以上３編の論文はいずれも，コーパスを用いた日本語研究の過去と現状を踏まえ，コーパス日本語学の将来を論じている。我々は，目先の個別的な調査や分析にのみ目を奪われることなく，コーパスを用いた日本語研究が現在どのような位置にあり，どこへ向かおうとしているのか，コーパスの特定の利用法あるいはコーパスの利用全般が日本語ないし言語の研究の流れの中でどのような意味を持つのかといったことを意識しつつ，研究を進めることが重要であろう。

3.2. コーパスを利用した日本語研究の実践

次に，曹大峰氏「多言語コーパスと日本語研究――『中日対訳コーパス』の利用研究例から――」は，今後質・量ともに増えるであろう多言語パラレルコーパスを用いた対照言語学的研究に関わる論考である。曹氏は，中日対訳コーパスの構築に参与した経験に基づき，多言語コーパスの「利用モデル」を提唱し，その分析事例を紹介している。自然言語処理の分野では，多言語パラレルコーパスから対訳部分を取得して機械翻訳に用いるなど，その有用性が早くから認識されているが，対照言語学的な研究に多言語コーパスを用いた研究例はまだ少ない。対照研究での利用のために，どのような多言語コーパスが望まれるか，どのような付加情報を付与すればよいかといった問題の議論が今後求められるであろう。

語学教育への応用という実際的な要請から，学習者が書いたり話したりした内容を収集した学習者コーパス(learner corpus)が世界各地で構築されている。陳曦氏「学習者と母語話者における日本語複合動詞の使用状況の比較――コーパスによるアプローチ――」は，学習者コーパスと母語話者コーパスの両方を用いて日本語の複合動詞の使用実態を分析している。２種類のコーパスから得られたデータを対比している点において，これも一種の対照研究と言える。使用された

コーパスの規模の問題もあってここで述べられた結論はより多くのデータに基づいて検討・補強される必要があるが，こうした研究の方向は言語教育の現場に有益な知見をもたらすものとして今後の展開が期待される。

宮島氏の論文で論じられている語彙調査と並んで，文字調査もまたコーパス利用の成果が期待できる研究領域である。小椋秀樹・相澤正夫両氏「現代雑誌70誌における漢字の使用実態と常用漢字表――国語施策へのコーパス活用に向けた基礎調査――」は，国立国語研究所「現代雑誌200万字言語調査」に基づいて1994年時点における漢字の使用実態（常用漢字・表外漢字や，表外音訓・表外漢字の音訓の出現状況）を調査し，その結果を常用漢字表の見直しなど国語施策に活用する可能性を論じている。小椋氏らの論考は，今後大規模コーパスが国語施策に対してどのような審議資料を提供できるかを見据えるためのものであり，さらに現代日本語における漢字の使用実態を的確かつ迅速に捉えるためのコーパスのあり方についても言及されている。

以上 3 編の論考は，日本語研究の相異なる領域におけるコーパスを用いた研究の事例である。ほかにも，文法，語彙，意味，文体，音声・音韻などの研究，方言や日本語史の研究，言語と言語外の要因の関わりに関する社会言語学的研究など，日本語研究へのコーパスの応用可能性には限りがない。我々日本語研究者の今後の課題は，さまざまな着想に基づく事例研究の実践を通して，コーパス日本語学の射程を見定めていく，あるいはむしろ，その射程をさらに広げていくことにあると言えよう。

3.3. コーパスの構築や利用に関わる研究

コーパス日本語学の発展のためには，よりよいコーパスの構築と，コーパスを効果的に利用するための手法・ツールの開発が必要である。

伝康晴氏他「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――」は，新たに開発した電子化辞書「UniDic」の設計と実装，応用例を報告している。現在広く用いられている形態素解析用辞書では，解析結果として得られる単位が斉一的でない（「幾何学」「心理｜学」），語形や表記の変異（「やはり」「やっぱり」，「猫」「ネコ」）に対処できないなど，特に語を対象とする調査研究にとって不都合な点があった。UniDic では，解析単位に「長単位」「短単位」「中単位」という 3 段階の粒度を設け，階層的な単位設計を行なっている。また，「語彙素」「語形」「書字形」「発音形」という 4 階層により「見出し」を定義し，異表記や異形態に同一の見出しが与えられるよう工夫している。多様なテキストの安定的な解析にはなお辞書の拡充が必要になるが，語に着目した日本語研究に利益をもたらすものと期待できる。

小木曽智信・近藤明日子両氏「日本語研究のための XML タグ付けプログラム――その開発と活用例――」は，XML 形式のコーパスに対して利用者が独自のタグを埋め込むソフトウェアを開発した経緯とその活用例を述べている。現在，構築されるコーパスは XML 文書として公開されることが一般的になりつつあるが，日本語研究者のあいだには XML 文書を扱うためのノウハウが普及していない。小木曽氏らは，XML 文書に適切なタグを埋め込んで分析を行なうことで，調査結果の再利用や再調査が容易になることを提案する。その具体例として，コーパス中に出現

した特定の語にタグ付けを行ない，その語と共起する語の集計表を作成するまでの一連の手順が示されている。日本語研究者にコーパス利用技術を獲得する努力が必要だとの指摘が後藤氏の論文にあったが，そうした要請に応えるひとつの提案と言える。

深田淳氏「日本語用例・コロケーション情報抽出システム『茶漉』」は，コーパスからコロケーション（連語）情報を抽出するソフトウェア『茶漉』を紹介するものである。Web ブラウザ上の操作により，ある語とその前後文脈の語数などの条件を指定すれば，その語と強い結びつきを持つ語の情報（コロケーション情報）が表示される仕組みになっている。現状では検索対象となるコーパスが「青空文庫」のデータの一部と「名大会話コーパス」に限られており，また語をどのように認定して語数を数えるのが適切かなどの検討課題も残るが，今後そうした問題にも配慮したツールが開発されれば，日本語のコロケーション研究に大きく貢献するであろう。『茶漉』は日本語研究におけるそうした方向での試みとして興味深い。

4. 日本語研究の将来のために

2.2.で述べた，確立された研究法の不在という，第 2 期コーパス日本語学のかかえる 2 つ目の問題は，本特集に収めた各論文のような考察や研究の実践を通して徐々に解決されていくはずのものであろう。ただし，そのことを無条件に楽観視できる状況にはない。昨今日本語の電子テキストの入手が容易になったにもかかわらず，その特性を生かした日本語研究の試みはあまり増えていない。そのことを思えば，大規模な均衡コーパスが構築・公開されれば直ちに日本語研究におけるコーパスの利用が盛んになるという予想は立てがたいところがある。第 3 期コーパス日本語学がいつ満足な意味で第 3 期と呼べるものになるか，それはひとえに我々日本語研究者の創意と努力にかかっていると言うべきであろう。

また，コーパス日本語学の今後を考えるとき，あわせて 2 つの点を明確に意識しておくことが重要である。その 1 つは，コーパスの利用が日本語研究に対して全面的に有利に働くわけではないということである。コーパスを使えば個別の表現あるいは言語現象の実態を精密に把握することは可能になるが，そのような作業に専心していると，日本語全体，文法全体といった大局的な視点を失いがちである。また，コーパスを使って研究しやすいテーマもあればそうでないテーマもあるという事実もあり，ともすれば“木を見て森を見ぬ”どころか“1 本の木を見て隣の木も見ない”研究に陥りかねない。実際，コーパスを使ったおびただしい量の研究成果が発表されている英語の研究状況に照らしてみても，そうした懸念は非現実のものではない。

もう 1 つの点は，当然のことではあるが，コーパス日本語学といっても，日本語の研究においてコーパスの利用が自己目的化されるべきではないということである。コーパス利用の目的はあくまでも日本語あるいは言語に関する我々の理解を深め知見を増すことにあるのであって，コーパスを中心に据えて日本語の研究を考えるとすればそれは主客の転倒である。日本語研究に関する理解の裏付けを伴わないコーパスの皮相な利用は論外として，コーパス日本語学という枠を固定的で排他的なものと捉えて他の研究法に関心を示さないとか，他の研究法に対して批判的になるといったことも賢明な態度ではない。コーパス日本語学の真の対象は日本語であり言語である

ということをあえて確認して，本稿の結びとしたい。

付記：現在国立国語研究所で構築中の「現代日本語書き言葉均衡コーパス」の詳細については，次の Web サイトを参照されたい。

国立国語研究所コーパス整備計画 KOTONOHA： http://www2.kokken.go.jp/kotonoha/

文部科学省科学研究費補助金特定領域研究「日本語コーパス」： http://www.tokuteicorpus.jp/

丸山　岳彦（まるやま　たけひこ）

国立国語研究所研究開発部門

190-8561　東京都立川市緑町10-2

田野村　忠温（たのむら　ただはる）

大阪大学大学院文学研究科

Japanese corpus linguistics:

Its aims and prospects

Takehiko Maruyama
National Institute for Japanese Language

Tadaharu Tanomura
Osaka University

Keywords

Japanese corpus linguistics, Balanced Corpus of Contemporary Written Japanese

Abstract

Japanese corpus linguistics will soon come of age. The Balanced Corpus of Contemporary Written Japanese (BCCWJ), which is being built in a five-year project of the National Institute for Japanese Language at present, will be completed and published in 2011. As the long-awaited first ever balanced corpus of Japanese, BCCWJ is expected to open up a new era of corpus linguistic studies of the Japanese language, which admittedly have been far behind the times on international standards.

In this editorial essay to the special issue devoted to Japanese corpus linguistics, we briefly sketch the history of the use of electronic corpora in the study of Japanese, discuss some of the issues which will need to be borne in mind when we use corpora, as well as summarize each of the papers included in the special issue.