本報告書の目的と概要

著者	田中 牧郎
雑誌名	近代語コーパス設計のための文献言語研究 成果報告書
ページ	7-10
発行年	2012-10-31
シリーズ	国立国語研究所共同研究報告 ; 12-03
URL	http://doi.org/10.15084/00002761

本報告書の目的と概要

田中　牧郎　（国立国語研究所言語資源研究系）[1]

１．本報告書の目的

国立国語研究所では、2006 年から『現代日本語書き言葉均衡コーパス』の開発に取り組み 2011 年にこれを完成させたが、2009 年 10 月の大学共同利用機関法人人間文化研究機構への移管を機に、日本語の史的研究に幅広く活用できる通時的なコーパスを構築することにも手を広げることになった。移管に際して始まった、基幹型共同研究プロジェクト「通時コーパスの設計」（プロジェクトリーダー：近藤泰弘客員教授）では、古代から近世までを対象とした「通時コーパス」の設計を行う研究を進めている。一方、ここに研究成果を報告する、独創・発展型共同研究プロジェクト「近代語コーパス設計のための文献言語研究」（プロジェクトリーダー：田中牧郎）においては、明治初年（1868 年）以後を「近代」と扱い、近世までのコーパスと『現代日本語書き言葉均衡コーパス』とをつなげる役割を持った「近代語コーパス」を設計するための論点を整理し、その設計図を描く道筋を付けることを目指して、近代の文献資料とその言語の研究を進めてきた。なお、『現代日本語書き言葉均衡コーパス』の開発以前に、国立国語研究所では最初の近代語のコーパスとして『太陽コーパス』を 2005 年に完成させている。このコーパスの実績を踏まえながらも、最新のコーパス研究の成果を取り込んだ、次世代の近代語コーパスを設計するための研究に主眼を置いてきた。３年のプロジェクト期間が終了するにあたり、その主要な研究成果をまとめて公表するのが、この報告書の目的である。

２．本報告書の概要

２．１　全体の構成

本プロジェクトは、今後構築されるべき近代語コーパスをどのように設計するのかについてと、構築される近代語コーパスをどのように活用するのかについての両側面から、近代の文献資料と言語の研究を行った。本報告書は、その二つの側面に即して「第 1 部 コーパスの設計」と「第 2 部 コーパスの活用」の２部構成とし、共同研究者による論文を集成する形で編集した。第１部では、「資料選定」「文字処理」「形態素解析」「モデルコーパス」の四つに分類し、２編ずつ計８編の論文を掲載した。また、第２部では、「語彙研究」「文法研究」「日中韓対照研究」の三つに分類し、３編ずつ計９編の論文を収録した。各論文の扱っている内容を簡単に紹介しながら、本プロジェクトの成果を概説しよう。

[1] mtanaka@ninjal.ac.jp

２．２　「第１部　コーパスの設計」の概要

第１部の「資料選定」に掲げた２編のうち、田中論文は、近代語コーパスにおける資料選定においては、当時の言語に対する代表性と各資料が有している独自性の両側面が尊重される必要があることを述べ、数千件からなる資料リストに基づいた資料選定の実施例を示し、実際に選定を行う際の論点について研究している。もう１編の岡島論文は、近代語コーパスの対象とすべき資料を探索する演習を、大学院生とともに行った記録を著したもので、院生から提案された資料の特徴とコーパス化する意義を資料ごとに具体的に示している。この２編で扱っているように、今後、近代語コーパスのための資料選定を実際に行う際には、多様な資料の性質を分類整理することと、多くの研究者が参加できる形で幅広い議論を行うことが、重要になるだろう。

「文字処理」には、近代語文献を電子化するために必要な文字セットと、異体字処理について論じる２編を掲載した。文字セットを扱う高田論文は、先行する『太陽コーパス』が、JISX0208（JIS 第１水準・第２水準）で電子化した際に外字となっていた文字が、その後普及した JISX0213 による JIS 第３水準・第４水準でどこまで電子化できるようになったのかを調査し、依然として外字になるものとともに、文字一覧を示している。モデルコーパスとして作成した『明六雑誌コーパス』の文字処理の基準と実際を記す須永論文は、外字をできるだけ減らすために行った、包摂規準の追加と別字による代用の全容を記録している。近代語の資料の多くは活字文献でありながら異体字が非常に多いため、この問題への対処法を明確化しておくことは、コーパス構築の基本問題としてきわめて重要である。この２編で打ち出された方向性は、『太陽』と『明六雑誌』だけにとどまるものではなく、近代語コーパス全体に適用されるべきものであり、さらには、近世以前を扱う「通時コーパス」の設計にも直接役立つものになろう。

「形態素解析」の論文としては、近代語テキストに形態素解析を施す実際の作業とその問題点を述べる小木曽論文と、形態素解析を実現させるのに必要となる形態論情報付与の規程を説明する須永・近藤論文の２編を載せた。それぞれ『太陽コーパス』開発当時は不可能であった近代語テキストへの形態素解析を実用化させるための技術開発を行い、その処理のために必要になる単語や品詞の認定基準を立てる研究である。そこには、自動形態素解析結果に人手修正をかけて形態素解析辞書と機械学習用データを整備していくことで、近代語の形態素解析が十分に実用化可能であることの見通しが明確に示されており、本プロジェクトで最も成果があがった部分である。近代語の言語状況は複雑であるため、今後多くのテキストを対象に人手による作業を重ねることが求められるものの、この技術が確立しつつあることによって、『太陽コーパス』の仕様を大きく進めた、次世代の近代語コーパスを設計できることが確実になったと言ってよい。

「モデルコーパス」においては、明治７（1874）～明治８（1875）年に発行された学術啓蒙雑誌『明六雑誌』の全文を対象とした『明六雑誌コーパス』を作成することを通して、今後構築していく近代語コーパスのモデルを提示する２編を執筆した。コーパスの仕様を扱う、近藤・田中論文は、上述した文字セット、包摂規準、形態素解析をはじめとした、『太

陽コーパス』から大きく発展させた仕様の全体を網羅的に記述したものである。コーパスの概要を記す近藤論文は、形態素解析が実現したことで明らかになった、語種構成・品詞構成をはじめとした、『明六雑誌』の詳細な言語状況が示され、形態論情報付きコーパスが持つ高い価値を印象づけるものになっている。この『明六雑誌コーパス』は、本報告書の公開と同時に、本プロジェクトのホームページを通してダウンロード公開を開始した。コーパス検索ツール『ひまわり』に搭載できるデータも公開することによって、誰でもが容易に利用できるようにした。

２．３　「第２部　コーパスの活用」の概要

よいコーパスを設計するには、コーパスをどのように活用してどのような研究を展開するのかを考えながら研究することが不可欠である。そのような考え方に立ち、本プロジェクトでは、コーパスを活用した新しい研究領域の開拓にも力を入れた。

「語彙研究」に収録した田中論文は、『太陽コーパス』に形態素解析を施したデータを用いて年次別の語彙頻度調査を行い、明治後期から大正期にかけて漢語が減少し和語が増加していく実態を明らかにした上で、個々の語が語彙全体の中に占める位置がどう変わっていくのかという観点から語彙を類型化している。また、小野論文は、『明六雑誌コーパス』から作成した漢語リストをもとに、『日本国語大辞典第２版』の初出時代と比較することで、語史の視点から漢語を階層化する研究である。さらに、近藤論文は『明六雑誌コーパス』から一人称代名詞を抽出し、統計的指標を用いて代名詞一つ一つの性格を詳細に明らかにしている。これら３編はいずれも、従来のテキストコーパスだけでは行えなかった、形態論情報付きコーパスの利点を生かした語彙研究の事例となっている。

「文法研究」にも３編をおさめた。まず田中論文は、やはり『太陽コーパス』の形態素解析データを使って、言文一致が進行するのにともなって、文語助動詞から口語助動詞への推移がどのように進んでいくのかを記述したものである。推移の過程は、個々の助動詞によって異なり、同じ助動詞でも活用形によって異なる事実が種々発見されており、文体変革期における書き言葉の文法変化の記述という、新領域の開拓が期待できる。島田論文は、終止形による準体法が近代語において多様に発展していた事実を、コーパスからの豊富な事例によって明らかにしている。ここでも、品詞や活用形を指定して抽出・検索できるようになった形態素解析データが真価を発揮しており、コーパスが重要テーマでの集中的な議論を可能にする利点が示されている。そして、小島論文は、東北出身の宮沢賢治と濱田廣介の文学作品のコーパスを独自に作成し、標準語のコーパスである『太陽コーパス』と比較して、地方出身者の言語に方言的な特徴がどの程度見られるのかを分析している。標準となるコーパスとは別に特定目的のコーパスを作成して両者を比較する応用的な研究は、現代語のコーパス研究でも行われているが、近代語コーパスにおいてもそのような研究の広がりが見込めることを教えてくれる。

最後に「日中韓対照研究」としてまとめた３編について述べたい。近代に西洋からの新概念を受容するにあたり、多くの漢語が作られたり意味を変えたりしたが、その漢語の変

容は、日本語だけに起こったのではなく、同じ漢字文化圏を形成していた中国語や韓国語にも起こり、相互に語彙の貸し借りを行っていた。この語彙交流の研究は従来から盛んであったが、三つの言語の近代語コーパスを連携して作ることができれば、この方面の研究を一層充実させることができる。そのような考えから、中国語や韓国語の近代語と日本語の近代語とを対照した研究をここにおさめた。朱論文は、日本語の漢文系の複合辞（「～に基づいて」など）とそれに相当する中国語の語彙との関係を、日中双方の近代語コーパスの調査によって明らかにしようとし、陳論文は、日中の語彙交流の記述を、多くの文献資料をもとに行う際の問題点を整理している。いずれも、両言語の多様な文献資料の性質を見きわめた上で比較することの重要性を指摘している。張論文は、近代に日本語から韓国語に訳された文語体の資料と口語体の資料の２組をそれぞれコーパス化し、語種構成や品詞構成を対照し、近代日本語と近代韓国語の比較研究を行っている。これらの研究は、近代語コーパスの構築において、東アジア言語との関わりにも留意することの重要性を示している。

３．今後に向けて

本報告書におさめた 17 編の論文は、本プロジェクトの共同研究発表会、国立国語研究所主催のコーパス日本語学ワークショップ、各種学会の口頭発表などで発表した内容に基づいているものが多いが、学術誌や著書などでは未発表のものばかりである。いずれも、現段階では、各研究者による新領域開拓の途上にあるものであり、この報告書への執筆を経て、さらに研究の段階を進めて、学術誌や著書としてより完成されたものへとまとめられるべきものである。このような性格の論文が集まったことは、コーパス設計のためという目的を共有して研究することで、共同研究者の目が自ずと新領域へと向いていった結果だと考えられる。

近代語コーパスの設計図そのものは、まだ描かれていないが、本報告書の各論文が指し示す方向のすぐ先には、その作業に着手できる場が見えているはずである。近代語コーパスの構築に本格的に着手するには、開発予算や開発体制の検討作業が不可欠であるが、そうした実務的な検討作業に際しても、この報告書が役立てられることを願うものである。