Vol.2017-CH-115 No.11
2017/8/4

国語学と情報技術

高田智和[†1]

概要：音韻・文法・語彙・表記といった伝統的な国語学の諸分野にあって，情報技術の導入によって大きく進んだ分野は語彙研究である。言語計量の手法は計算機と極めて相性が良く，大規模な語彙調査を可能とし，基本語彙の策定などに対して知見を与えた．また，電子化テキストの出現と流通は，用例取得を主目的とした文法研究からも歓迎された．これらを土台として，現在のコーパス開発と研究利用が展開されている。

Study of Japanese Language and Information Technology

TOMOKAZU TAKADA[†1]

Abstract: Among the traditional Japanese linguistic fields of study of phonemes, grammar, vocabulary and orthography, vocabulary research is an area that has made great progress due to the introduction of information technology. Since quantitative linguistic methods are extremely compatible with computers, they enable large-scale survey of vocabulary and have conferred insight on such matters as the formulation of basic vocabulary. Furthermore, the appearance and distribution of digitized texts is welcomed also from the viewpoint of grammar research with the main purpose of acquiring examples. Current corpus development and research use are in progress based on these circumstances.

1. はじめに

本発表では，国語学における情報技術の導入とその歩みについて，私見を述べる。

国語学は，日本語の音韻，文法，語彙，表記（文字）などについて，主に実証的観点から研究を行う分野である．歴史的変遷を扱うことが多いため，表記（文字）が重要視されるが，言語研究一般として見た場合，中心は音韻と文法になるであろう．音韻研究と文法研究に比べて，語彙研究は比較的若い分野である．ある単語やその関連表現の使われ方の歴史をたどる研究は，語史あるいは語誌と呼ぶべきもので，単語のあつまりやまとまりを扱う語彙研究とは一線を画するものである．

語彙研究は，基本語彙の選定という，言語教育への応用を主な目的として進展してきた. この点, 実学志向である．ほかに，語彙の量的構成の把握といった，言語の語彙的実態を捉える目的がある．これらの目的を満たすためには，基礎的・客観的なデータが必要である．分析データを得るために語彙調査を行う. 分析には計量的手法を用いるため，計算機と極めて相性が良い．国語学における情報技術の本格的な導入は, 語彙調査を始まりとするのが妥当であろう．

†1 国立国語研究所
National Institute for Japanese Language and Linguistics.

2. 語彙調査

筆者が務める国立国語研究所は，1948 年の創立以来，60 年にわたって，対象を替えて語彙調査をやり続けてきた．1949 年 6 月発行の朝日新聞 1 紙の語彙調査が最初で，1994 年発行の雑誌 70 種の語彙調査が現時点での最後である（報告書の刊行は 2005 年）．

- 朝日新聞 1 紙（1949 年 6 月発行，全数調査）
- 婦人雑誌 2 種（1950 年発行，標本調査）
- 郵便報知新聞 1 紙（1877 年発行，標本調査）
- 雑誌 13 種（1953 年発行，標本調査）
- 雑誌 90 種（1956 年発行，標本調査）
- 朝日・毎日・読売新聞 3 紙（1966 年発行，標本調査）
- 高等学校教科書（1974 年度使用，全数調査）
- 中学校教科書（1980 年度使用，全数調査）
- 中央公論（1906-1976 年発行，標本調査）
- テレビ放送（1989 年 4-6 月放送，標本調査）
- 雑誌 70 種（1994 年発行，標本調査）

語彙調査への計算機の導入は，1966 年発行の朝日・毎日・読売新聞 3 紙の語彙調査である. 計算機を使うことで，調査語数が大きく増大した．計算機導入直前の，1956 年発行の雑誌 90 種の語彙調査では，延べ語数約 50 万語であったのに対して，新聞 3 紙の語彙調査の延べ語数は約 300 万語と 6 倍になっている．新聞 3 紙の語彙調査では，独自の

漢字コードを導入したり，検索文字列の前文脈と後文脈を記した KWIC による分析も行われている．新聞 3 紙の語彙調査の報告書は『電子計算機による新聞の語彙調査 1〜4』（1970-1973）である．同時期に，新聞以外を対象とした計算機利用の報告として,『電子計算機による国語研究Ⅰ〜Ⅹ』（1968-1980）がある。

新聞 3 紙の語彙調査は，計算機を導入した初めての日本語語彙調査として特筆すべきものであるが，同語異語判別をせずに出現した表記形のみで集計を行っている。したがって「いき（行き)」と「いき（粋)」などは語彙表で区別されず，言語データとしては使いにくいものである．語彙調査としての到達点は,計算機導入直前の雑誌 90 種であるとする評価が多く，計算機を用いた語彙調査の始まりは順調ではなかった．

3. 電子化テキストの利用

パソコンが大学に導入されると，自らの研究で使ってみようとした研究者が，国語学分野でも現れた．電子化テキストを作り，プログラムを書いて分析したり，索引やフルテキストデータベースを作成するのが目的である．筆者の知っている研究者では，近藤泰弘，豊島正之，池田証寿，伊藤雅光，當山日出夫，金水敏などがパソコン使って何かしようとした最初の世代であり，現在も第一線で活躍している．ちなみに筆者は，国語学の研究者が，個人の研究でパソコンを使い始めた世代の教え子にあたる世代である．恩師池田証寿の影響を大きく受け，今のようになった．同世代の一般的な国語学の研究者と違っていると，自覚している．

さて，筆者が大学に入学した 1995 年は，Windows 95 が発売され，パソコンが家電により近づいた始まりの年である．大学・大学院に在学中に，CD-ROM やインターネットで利用した主な電子化テキストは次の通りである．

- 青空文庫
- 日本古典文学大系本文データベース
- CD-ROM 版新潮文庫の 100 冊
- CD-HIASK 朝日新聞記事データベース
- CD-毎日

これらは，国語学の研究者が多く利用するものである．語法や文法の用例を採集するための検索対象としての利用である．これらの全文テキストに，各種言語アノテーションを付与したコーパスが加わることになる．

4. おわりに

現在，国語学分野では，コーパスに対する関心が高い．国立国語研究所がコーパス開発を進めるのは，前身に語彙調査があるためである．国語学の研究者が，コーパスに関心を寄せるのは，用例検索の対象として，電子化テキストの後継にコーパスを捉えているのかもしれない．

コーパスが，従来の研究手法に，どのような変革をもたらすのかは，未知数である．コーパス構築に携わる者としては，作り手が一番の使い手でありたいと念じている．