日本語の全体像を知るために

—国立国語研究所による言語資源整備—

大学共同利用機関法人 人間文化研究機構
国立国語研究所副所長／言語資源研究系長 教授
前川喜久雄

1．はじめに

産業日本語は日本語に対する領域限定的な規制の試みである。言語の規制に対する言語研究者の立場は、これを強く要請する者から全く不可とするものまで様々であるが、そもそも言語の規制というアイデアが浮かぶ原因が言語の多様性にあるという点については大方の一致が見られるものと思う。

言語の多様性というと世界に何千種類の言語があるという類の、言語間の多様性が注目されがちであるが、一言語の内部にも多様性がある。規制の問題と結びつくのは、この言語内多様性の問題である。

言語内多様性をもたらす要因には様々なものがある。そのうち言語の歴史的変化と地理的変化につては、古くから研究が行われてきており、データも不十分ながら整備されてきている。しかし、多様性をこのふたつの要因で説明しきることは到底不可能である。生きた言語を仔細に観察すると、歴史的変化とも地理的変化とも関係しない、あるいは双方と関係するように見える変異、いわば純粋に確率的な変異とでも呼ぶべき言語変異現象があり、変異の大部分はむしろそのような性質のものであると思われるからである。

そのような変異の研究は標準語を対象として行われることが多いが、現代の標準語は話し言葉としても書き言葉としてもデータサイズが膨大であるために、一個人の力でその全体像を把握することは極めて困難である。

国立国語研究所ではこの問題の解決に寄与するために、1990 年代末から一連のコーパスを構築し公開してきた。また 2016 年の 4 月から始まる次期中期計画期間(2021 年度までの 6 年間)においても、コーパスに代表される言語資源の整備が研究所の重要な活動目標のひとつとなっている。以下では言語資源の整備に関わる国語研のこれまでの成果と今後の活動計画を紹介する。

2．KOTONOHA 計画

KOTONOHA 計画とは国語研による言語資源整備計画の総称である。実際にこの名称を使い始めたのは 2006 年からだが、ここでは 1999 年に計画がスタートしたものとして説明する。最初に KOTONOHA 計画開始直前の状況を説明しておくと、現代語の書き言葉データとして、まとまった量を誰でも利用できたのは新聞各社の記事テキストデータベースのみであった。他に著作権の消滅した文芸作品を中心とする「青空文庫」のデータも利用でき

たがこれは現代というよりむしろ近代語のデータである。

２．１　『日本語話し言葉コーパス』(Corpus of Spontaneous Japanese: CSJ)

1999 年に構築をはじめ、2004 年に公開した話し言葉コーパスである。開発費は科技庁の科学技術振興調整費であり、東工大、情報通信研究機構との共同開発である。学会での研究発表や少人数の聴衆を前にしたスピーチなどのモノローグを中心に 662 時間 (750 万語) の音声を精密に転記し、短単位と長単位による二重形態素解析、節境界情報、印象評定情報などのアノテーションを施した。コアと呼ばれる 50 万語分のサブセットには X-JToBI による精密な分節音・韻律アノテーションも施した。CSJ はもともと自発音声の音声認識研究用に設計したコーパスであり、音声認識研究に大きく貢献したが、コアは音声学の領域でもよく利用されている。Google Scholar で検索すると引用件数は 800 件を超えている。

２．２　『現代日本語書き言葉均衡コーパス』(Balanced Corpus of Contemporary Written Japanese: BCCWJ)

2006 年に構築をはじめ、2010 年に公開した日本語初の均衡コーパスである。開発費は文科省科研費（特定領域研究）であった。規模は British National Corpus に倣って 1 億語であり、書籍、雑誌、新聞、白書、ブログ、ネット掲示板、法律、広報誌、韻文など 11 種のレジスターにわたる書き言葉のサンプルを可能な限りランダムサンプリングによって収集している。従来利用が困難であった各種書籍からのサンプルを大量に含んでいる点と、全サンプルに著作権処理が施されていて安心して利用できる点に特長がある。CSJ と同様、テキストには二重の形態素解析が施されている。

BCCWJ は DVD で全データが公開されているが、ウェブ上の検索アプリ（『中納言』）で形態論情報を検索することもできる。日本語学の広い領域で活発に利用されており、現時点での引用件数は 600 件を超えている。

２．３　『国語研日本語ウェブコーパス』(NINJAL Web Japanese Corpus: NWJC)

BCCWJ では量的に不足する研究がある。そのために母集団をウェブ上の日本語に限って 200 億語規模のコーパスを構築したのが『国語研日本語ウェブコーパス』である。開発費は運営費交付金（特別経費）である。2010 年に設計を開始して 2015 年に構築を終えた。現在は 2016 年の公開にむけて準備を進めている。短単位相当の形態論情報に加えて、文節係り受け構造の情報も提供する。

２．４　『日本語歴史コーパス』(Corpus of Historical Japanese: CHJ)

国立国語研究所は 1948 年に現代日本語を研究する国立試験研究機関として設置されたが、2009 年に大学共同利用機関法人に移管されるに際してこの制約が外された。そこで学界の要望に応えて過去の日本語を対象とするコーパスの構築を開始した。これが CHJ である。

現在は古典文学作品を対象として作業を進めており、本文テキストは㈱小学館のご厚意で同社の『新編日本古典日本文学全集』を利用させてもらっている。現時点で「平安時代編」約 73 万語と「室町時代編 I 狂言」約 24 万語が公開されている。テキストには短単位で形態素解析が施されており、BCCWJ と同じ『中納言』での検索が可能である。

CHJ とは別に近代語（明治・大正期の日本語）のコーパスも構築している。総合雑誌のテキストを対象としており、『太陽コーパス』『近代女性雑誌コーパス』『明六雑誌コーパス』『国民之友コーパス』を公開しているが、将来的には CHJ と統合する予定である。

3．今後の開発計画

以上のコーパス以外にも国立国語研究所内で構築作業が進められているコーパスがいくつかある。そのうち三つを紹介する。

第一のコーパスは、日本語学習者による日本語のサンプルを集めた『多言語母語の日本語学習者横断コーパス』(International corpus of Japanese As a Second language: I-JAS)である。日本を含む 20 の国と地域で、異なる 12 言語を母語とする日本語学習者 1000 人の話し言葉と書き言葉を収集することを目標としている。データの一部は今春公開の予定であるが、最終的な公開は次期中期計画期間年度末（2021 年度）を予定している。

第二のコーパスは「方言コーパス」（仮称）である。これは 1977～85 年に文化庁が実施した「各地方言収集緊急調査」で収録された方言談話資料（全国 224 地点×30 時間の録音）を活用したコーパスであり、方言談話中の語形を標準語で検索可能とし、転記テキストと方言音声を聴取可能とする予定である。

第三のコーパスは「大規模日常会話コーパス」(仮称)である。CSJ では対象から外れた日常の話し言葉を対象とするコーパスであり、首都圏の日常会話を個人の生活に密着する手法と特定の場面を想定する手法とで合計 1000 時間分収録する予定である。「方言コーパス」「大規模日常会話コーパス」ともに、やはり 2021 年度の公開を目標としている。

ここまで紹介してきたコーパスは、それぞれ独自の経緯をもって開発されたものであり、検索系はそれぞれに独立している。しかし、日本語全体を研究する立場からすれば、各コーパスをバラバラに利用するのではなく、一括して検索できることが望ましい。これは理論的にも技術的にもいくつかの難しい問題を孕んだ要請であるのだが、コーパス利用の利便性に大きくかかわる問題であるから、包括的な利用を可能にする検索環境を実現して公開したいと考えている。

国立国語研究所が開発した言語資源でコーパス以外に広く活用されているものに形態素解析用辞書 UniDic がある。これは BCCWJ のために開発された短単位辞書で、2005 年以降更新されていない。UniDic の拡張も次期中期計画期間の目標のひとつに掲げている。

4．現代標準語の言語内多様性

すでに構築が完了している CSJ、BCCWJ、NWJC などを検索することで、現代標準語の

言語内多様性を客観的に把握することが可能になってきた。本稿では紙幅の関係で省略に従うが、講演時には以下のような実例を示す予定でいる。①「来れる」などのラ抜き言葉の使用に関する意識と実態の乖離、②文法書では常に「～ている」の形で使われるとされる動詞類（「そびえる」等）の例外的用法、③「形容詞＋です」「動詞＋です」の用例、④「書けれる」のように可能動詞に可能の助動詞が付加されたと解釈できる二重可能の用例、⑤「～するべきでない」の意味で「～しないべき」を用いる例などである。

５．おわりに

最後に指摘しておきたいことがある。それは、少なくとも日本語の場合、言語コーパスや音声コーパスの構築に最初に取り組んだのは人文系の言語研究者ではなく、音声認識や機械翻訳などを研究する情報処理領域の研究者たちであったという事実である。

本稿では国立国語研究所の活動ばかりに触れたが、音声コーパスであれば、音声認識研究用に構築された種々の朗読音声コーパスが CSJ 以前に構築され広く利用されていた。また書き言葉についても、新聞記事を素材とする『京都大学テキストコーパス』が 90 年代に開発されている。

KOTONOHA 計画で開発した各種コーパスはこのような先行研究の成果に立脚して設計されたものであるが、わけても自動形態素解析に代表されるアノテーションの自動化技術は日本語コーパスを質量ともに一変させる効果があった。国立国語研究所におけるコーパス開発は今後とも情報科学の成果に多くを負う形で進められるであろう。反対に言語研究者が情報科学にできる貢献としては、計画的に設計され、品質管理されたコーパスの開発が重要であり、現に CSJ や BCCWJ は情報科学領域でも活発に利用されている。

今後、人文系の言語研究における諸問題、特に言語内多様性の分析にコーパスを活用するにあたり、階層ベイズモデルなどによる複雑かつ柔軟な統計モデリングが非常に重要な役割を果たすものと予想している。そのような研究において、言語研究者と情報科学研究者による相方向的な共同研究が実現できれば、言語の本質解明に到る新しい可能性が拓けるにちがいないと期待している。産業日本語も、そのような分析の成果に立脚して設計を進めれば、よりよい成果を得られるのではなかろうか。

６．本稿で紹介した言語資源の URL

BCCWJ: http://pj.ninjal.ac.jp/corpus_center/bccwj/
CHJ: http://pj.ninjal.ac.jp/corpus_center/chj/
CSJ: http://pj.ninjal.ac.jp/corpus_center/csj/
I-JAS: http://ninjal-sakoda.sakura.ne.jp/lsaj/
NWJC: http://pj.ninjal.ac.jp/corpus_center/nwjc/
UniDic: https://osdn.jp/projects/unidic/
各種音声研究用コーパス: http://research.nii.ac.jp/src/list.html
各種近代語コーパス: http://pj.ninjal.ac.jp/corpus_center/cmj/
京都大学テキストコーパス: http://nlp.ist.i.kyoto-u.ac.jp