人間文化研究情報資源
共有化研究会報告集1　抜刷

2010年3月26日　発行

国立国語研究所の言語資源

国立国語研究所言語資源研究系

前　川　喜久雄

国立国語研究所の言語資源

国立国語研究所言語資源研究系　前　川　喜久雄

1．言語資源

近年、言語情報処理ないし言語研究の世界では、研究に用いる大規模データとそれを処理するためのツールをまとめて言語資源（language resource）と呼ぶ。本稿では国立国語研究所、わけても筆者が属する言語資源研究系（2009年９月末までは研究開発部門言語資源グループ）における日本語言語資源開発の動向を紹介する。

国立国語研究所で開発されている言語資源には、①シソーラス、②コーパス、③形態素解析用電子化辞書の３種類がある。以下本稿ではこのうち②に焦点をあてるが、そのまえに①と③についても簡単に説明する。

シソーラス（thesaurus）とは類義語辞書のことである。『分類語彙表』は国立国語研究所の出版物としてもっとも広く世に知られたものであろう。1964年の初版は約4万語を収録していたが、2004年の二版では約10万語にまで拡張されている。収録語数だけをみれば、『講談社類語大辞典』のように一層大部なものも出版されているが、初版にさかのぼる『分類語彙表』分類番号は、日本語学のみならず日本語情報処理の世界においても、意味処理における基本分類システムとして広く利用されており、この領域における共通財産となっている。

形態素解析用電子化辞書は、電子化テキストを自動解析するために構築された日本語辞書である。日本語のテキストは分かち書きされていないので、コンピュータによる日本語処理の第一歩は、これを語に分割して、語の品詞を確定することからはじまる。この作業を形態素解析と呼ぶが、そのために構築されたコンピュータ上の辞書が形態素解析用電子化辞書である。従来、この種の解析用辞書は形態素解析ソフトウェア本体同様、自然言語処理研究者が主導して構築されてきた。Juman, Chasenなどの有名な形態素解析フリーウェアに付属の辞書がその例である。しかしながら、この種の辞書を日本語学の観点から評価すると、語の規定について十分な配慮があったとはいいにくい面があった。

例えばChasenとその標準辞書であるIPADicを組み合わせた場合、「国立国会図書館」は「国立＋国会図書館」と２語に解析されるが、「国立公文書館」は「国立＋公文書＋館」と３語に分割される。このような問題を解消するためには、解析用辞書を言語学的に一貫した方法で開発する必要がある。

国立国語研究所では千葉大学と共同で新しい解析用辞書UniDicを開発している。UniDicは国立国語研究所が語彙調査のために開発してきた各種調査単位のうち「短単位」と呼ばれる単位に基づいて斉一的に構成された辞書であり、これを利用した形態素解析では上述のような問題は多くの場合解消される。ちなみに上の例は、「国立＋国会＋図書＋館」「国立＋公文書＋館」と分割され、「館」は接尾辞、他の短単位は名詞に分類される。

UniDicも2007年から一般に無償公開しているが、最新版（Ver. 1. 3. 12）の登録ユーザー数は2300名を越えている。

2．コーパス

2.1　コーパスと代表性

以下では本稿の主題であるコーパスに触れる。コーパス（corpus）とは「特定の言語を研究するために、実際に用いられた用例を組織的かつ大量に収拾したもの」である。「実際に用いられた」という限定が奇異にひびくかもしれないが、研究者が自分の母語を研究対象とする場合、データを集めることをせず、内省にたよって研究をすすめることが少なくないので、このような制約をくわえるのである。

コーパスの対象は、現代語であっても過去の言語であってもよく、書き言葉であっても話し言葉であってもよい。ただし、用例をただ大量に集めればコーパスになるわけではなく、研究対象の全体像を可能なかぎり正確に反映したものになっている必要がある。これをコーパスの代表性（representativeness）と呼ぶ。

コーパスの代表性をいかにして確保するかについては、様々な議論があるが、過去の言語のように、そもそもデータの総量が限られているのであれば、その全体をコーパス化してしまえば、代表性は完全に保証されることになる（そのようなコーパスは全文コーパスと呼ばれる）。しかし近代語や現代語になるとデータ量の総体は膨大なので、その一部をもって全体を代表させることになる。

問題はそのサンプルの選び方であるのだが、この問題は統計学の立場からすれば標本調査におけるサンプリング問題に他ならず、調査の対象となる集団（母集団）の特性を最も正確に反映させるためには、サンプルを無作為に抽出すればよいことがわかっている。

実際、国立国語研究所では1950年代から統計的標本調査理論に立脚した語彙調査を実施してきたし、下に紹介する『現代日本語書き言葉均衡コーパス』でも可能なかぎりサンプルを無作為抽出している。

2.2　『太陽コーパス』と『日本語話し言葉コーパス』

国立国語研究所では、1948年の創設以来、種々の言語資源を開発してきたが、その大部分は一般公開を前提としたものではなかった。先に言及した『分類語彙表』はむしろ稀なケースに属する。公開を前提とした言語資源が立案されるようになったのは、1990年度に入ってからである。

1990年代半ばに、言文一致運動の完成期にあたる明治末から昭和冒頭にかけての総合雑誌の記事を対象とした『太陽コーパス』（700万語規模）の構築がまず始まり、1999年には現代語の話し言葉を対象とする『日本語話し言葉コーパス』（750万語）の構築が始まった。そして2004年にはまず『日本語話し言葉コーパス』の一般公開が始まり、翌2005年には『太陽コーパス』も公開にこぎつけた。

これらのコーパスはそれぞれ目標とした研究領域における標準的な言語資源の地位を獲得している。『太陽コーパス』は一般の書籍付属のCD-ROMとして頒布されたために、普及の実態が正確に把握できていないので、『日本語話し言葉コーパス』の利用

状況を紹介しよう。このコーパスは、これまで国内外の大学、研究機関、企業研究所を中心に450本以上を頒布しており、音声自動認識研究を中心とした音声情報処理研究の領域において、少なくとも800篇以上、おそらくは1000本に近い数の研究論文で利用もしくは参照されている、このコーパスの内容に依拠した博士論文も少なくとも５編提出されている。さらに『日本語話し言葉コーパス』で採用した韻律ラベリング方式X-JToBIは、日本語の韻律ラベリングの実際上の標準として、理化学研究所をはじめとする国内外の研究機関で利用されている。

３．KOTONOHAと『現代日本語書き言葉均衡コーパス』

3.1　KOTONOHA計画

『太陽コーパス』と『日本語話し言葉コーパス』の成功は研究所内に言語資源開発に特化した組織を設立しようという気運を醸成した。折から独立行政法人として第２期中期計画期間（2006～2010年）を迎えるにあたり、研究開発部門に言語資源グループが設立され筆者を含む９名の常勤研究員が配置された。これは国立国語研究所としては格別に大きな研究グループであった。

この研究グループの活動目標を明確化するために、爾後20年程度の期間に実施すべきコーパス開発計画を立案したのがKOTONOHA計画である（図１参照）。KOTONOHAは明治から現代にいたる近現代日本語の全体像を把握するためのコーパス群の総称であり、多数の要素コーパスから構成されている。図中には『太陽コーパス』と『日本語話し言葉コーパス』も位置づけられていることに注意してほしい。

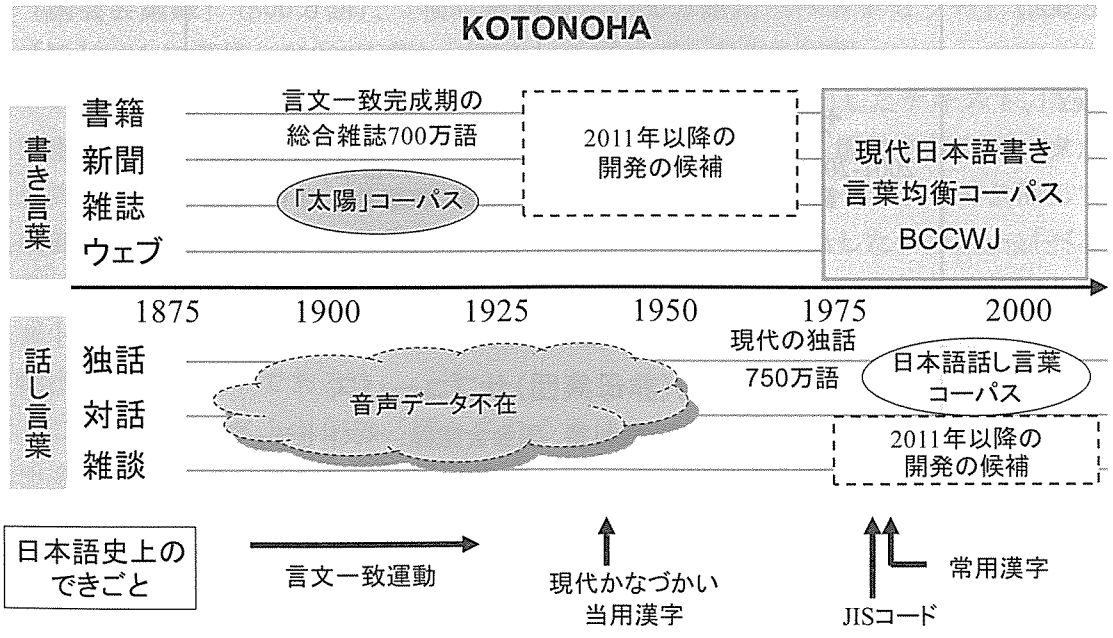

図１．KOTONOHA計画の概略図

図１についてもう少し詳しく解説しよう。図中央には時間軸が示されている。時間軸の起点（左端）は明治初年であり、終端は現在である。

図１の時間軸より上が書き言葉であり、その中は「書籍」「新聞」「雑誌」「ウェブ」に区分されている。時間軸の下は話し言葉であり、「独話」「対話」「雑談」に分類されている。これらは均衡コーパスが対象とすべき主要な変種ないしジャンルを示したものである。

図１には今後うめるべき空隙が多数存在しているが、筆者らがKOTONOHA計画を立案した2005年当事に焦眉の急と思われたのは、現代語の書き言葉を対象とした均衡コーパスの開発であった。現在、日本語の研究として発表される研究の多くは現代語の研究であり、その大半が実際上書き言葉データを対象とした研究である。そのため、この領域に係る信頼性の高いコーパスを欠くことは、日本語研究の品質を支えるインフラ構造に問題があることを意味するからである。このような問題意識をもって設計されたのが『現代日本語書き言葉均衡コーパス』である。

3.2 『現代日本語書き言葉均衡コーパス』

図２にこのコーパスの構成を示す。『現代日本語書き言葉均衡コーパス』は３種類のサブコーパスから構成されている。図左上の「出版（生産実態）」サブコーパスは2001年から2005年の期間に出版された書籍、雑誌、新聞の文字の総体を母集団とし、そこから約3500万語相当のテキストを無作為に抽出したコーパスである。書籍の母集団は国立国会図書館蔵書目録を電子化したJ-BISCに登録されている書籍の全体、雑誌の母集団は(財)日本雑誌協会に加盟している出版社による定期刊行物の全体、新聞の母集団は全国紙、ブロック紙、地方紙にまたがる16紙とした。

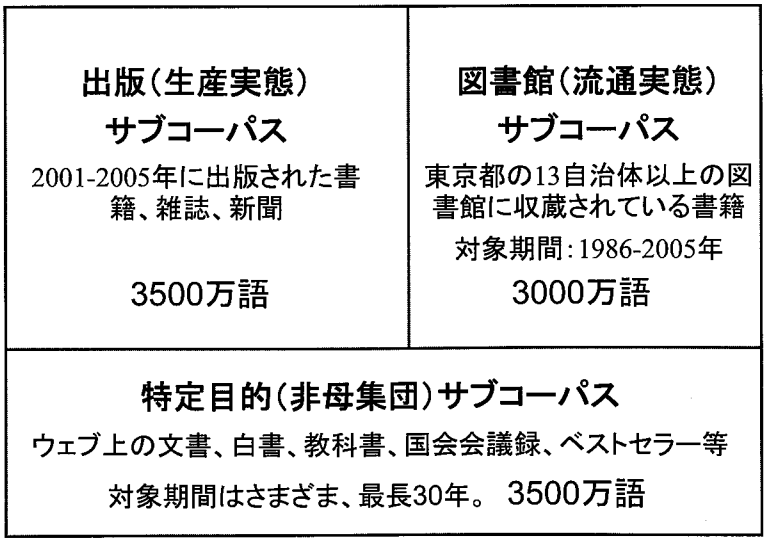

図２　『現代日本語書き言葉均衡コーパス』の構成

出版（生産実態）サブコーパスでは全体の74%を書籍のサンプルが占めている。この比率を異常に高いと感じる読者があるかもしれないが、これは誤りではない。書き言葉がどれだけ生産されたかという点に注目するとこういう結果が得られるのである。

ここで、それはおかしい、100万部のベストセラーのサンプルは1000部しか売れなかった本からのサンプルの1000倍の確率でコーパスに格納されるべきだと考えることも

できる。そのようなサンプリングは、書き言葉の受容の側面に注目したサンプリングと呼ぶことができるし、たしかに受容の実態に注目したコーパスが設計できるとよいのだが、残念なことに出版物の販売実績データは公表されていないか公表されていても著しく不正確であることが多く、母集団を定義することができない。

そこで、いわば次善の策として考案したのが「図書館（流通実態）」サブコーパスである。このコーパスの母集団は東京都下52自治体の公立図書館（１自治体＝１図書館とみなす）のうち13館以上に収蔵されている書籍で、ISBNが付与されており、1986年から2005年の期間に出版されたものの全体である。冊数にして約33.5万冊、文字数にして480億字（推定値）の母集団であり、出版サブコーパスのうち書籍部分の母集団（31.7万冊、485億字）とほぼ同一である。ここから出版（生産実態）サブコーパスの書籍と同数のサンプルを抽出する。

図書館（流通実態）サブコーパスは、単に出版されただけでなくある程度まで広い範囲に流通したことが確実な書籍を母集団としていること（したがって特殊な専門書や公序良俗に反する内容の書籍が排除されていること）、および、対象期間が20年間にわたっていることの二点において出版（流通実態）サブコーパス中の書籍部分と異なっている 。

最後に「特定目的」サブコーパスがある。これは国立国語研究所の研究活動のために必要とされるデータのうち、出版ないし図書館サブコーパスのサンプリングによっては十分な数が集まらないと考えられるものや、そもそも母集団を定義してサンプリングを実施することが不可能なものを格納するためのコーパスである。

特定目的サブコーパスには以下のようなデータが収録される。「白書」（500万語）、「国会会議録」（500万語）、「検定教科書」（200万語）、「ベストセラー」（300万語）、「Yahoo！知恵袋」（1000万語）、「Yahoo！ブログ」（1000万語）。

「白書」は過去30年間に政府が刊行した白書から無作為抽出したテキスト、「Yahoo！知恵袋」と「Yahoo！ブログ」はYahoo！Japan株式会社から提供されたデータから無作為抽出したテキスト、「国会会議録」は国会図書館がインターネットで公開している衆参両院の議事録のうち過去30年分を対象として無作為抽出したテキスト、「ベストセラー」は過去30年間にベストセラーリストに載った書籍951冊から無作為抽出したテキスト、「教科書」は小中高の検定教科書（各教科各学年につき1種を選定）から無作為抽出したテキストである。

これらのテキストが何故必要とされるかについての説明は不要であろうが、インターネット上のテキストを組織的に入手して公開できたことは、『現代日本語書き言葉均衡コーパス』の現代語コーパスとしての価値を高めたものと考えている。

４．開発の現状

『現代日本語書き言葉均衡コーパス』の開発費は約６億円と推定された。国立国語研究所の運営費交付金だけでは開発費をまかなうことができなかったので、2005年の秋に筆者が代表となって文科省科学研究費補助金特定領域研究を申請した。幸いこの申請はただちに採択され、2006年度から５年計画で特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築」（略称は「日本語コーパス」）を開始できた。

本稿執筆時点で『現代日本語書き言葉均衡コーパス』の構築は５年計画の第４年度が終了しようとしている。目標とする１億語に対して9000万語以上のサンプリングと電子化を終了しているが、そのうち著作権処理まで終了しているのは約5000万語分にとどまっている。

我が国もふくめて、近代的な著作権法をそなえる国であるかぎり、現代語コーパス開発における最大の障害は著作権処理である。『現代日本語書き言葉均衡コーパス』の開発では、著作権処理を専門にあつかつ開発グループを構成し、常勤１，２名、非常勤４，５名の体制で処理を進めてきた。「Yahoo！知恵袋」「Yahoo！ブログ」のように、提供元で一応の著作権処理が済んでいるデータや、国会会議録のように法律上は著作権が存在するが、国会の総意として著作権を主張しないという合意が形成されているデータもあるが、これらはあくまで例外である。

例えば書籍データの場合、25000件以上のサンプルについて、著作権者の連絡先を調査して実際に連絡し、コーパスでの利用許諾をもらうという作業を実施している。書籍の場合、これまでに約７割超の権利者に連絡をとり、その７割以上から許諾をもらっているので、全体としては５割以上のデータについて著作権処理が完了している。

残る５割はどうなっているかというと、明示的に利用を拒否されたのは全体の５％程度であり、大半は連絡先が判明しないか、連絡しても反応がないケースである。これらのサンプルをコーパス公開時にどのように扱うかは、これからの１年で決定することになるが、明示的に拒否されたサンプル以外はできれば公開したいと考えている。

５．行政改革の空騒ぎ

KOTONOHA計画とその嚆矢としての『現代日本語書き言葉均衡コーパス』の開発は、日本語研究への貢献と国民へのサービスを両立させるという点で、独立行政法人国立国語研究所の存在意義を端的に示す研究開発プロジェクトであった。ここで国民へのサービスとは、『現代日本語書き言葉均衡コーパス』のような言語資源が狭義の日本語学研究だけにとどまらず、日本語情報処理研究やIT産業、また国語辞典、日本語辞書の開発や日本語能力測定などでの応用的利用に供されることを意味している。

コーパスにこのような応用的価値が存在することは、広く識者の認めるところであった。上述の科研費特定領域研究の審査の過程でも応用の可能性が高く評価されていたし、国立国語研究所の外部評価においても『現代日本語書き言葉均衡コーパス』関連事業はA＋評価を獲得した。

国語研のコーパス開発に世の中が注目してくれていることには、識者の評価に頼らずとも、多くの状況証拠が挙げられる。新聞、雑誌、ITメディアなどで、過去５年間に15回以上の報道があり、そのなかには全国紙の夕刊１面のトップ記事が含まれている。学術の領域でも人工知能学会が2009年秋に特定領域研究「日本語コーパス」の特集を組んだほか、日本語学会、日本言語学会、英語コーパス学会、情報処理学会などが『現代日本語書き言葉均衡コーパス』を含む言語資源に関する特集記事やシンポジウムを企画した。

このようにKOTONOHA計画は順風満帆の様相をみせていたのであるが、2007年末にいたって予想だにしなかった事態が発生した。同年11月初旬、当事の福田康夫内閣のも

とで、独立行政法人問題を検討していた行政減量・効率化有識者会議」（当事の担当閣僚は渡辺喜美行政改革担当大臣）が日本語教育事業の廃止などとならんで、国立国語研究所における言語資源開発についても「日本語コーパス事業については、民間でも行われていることから廃止、もしくは民営化を検討されたい」という判断を表明した。

この判断が一体どのような根拠のうえになされているのかは正直なところ理解が困難だった。そこで文化庁国語課を通して有識者会議に「民間でも行われている」実例を示してほしいと要望したところ、「小学館コーパスネットワーク」いう回答がかえってきた。御存知の方も少なくないだろうが、これは英国で開発されたBritish National Corpusを小学館が日本でネット配信している事業である。当然日本語とは関係がない。

早速「我々は日本語のコーパスを開発している。事実誤認でないか」という反論文書を提出したのだが、この文書は有識者会議から受け取りを拒否されてしまった。行政にはそのような奥の手があることを初めて知ったが、そのとき文化庁の某課長補佐は「受けとりを拒否した以上は先方が負けをみとめたも同然で、これでコーパス事業は安心」としたり顔で説明してくれたものである。しかし同年12月24日の閣議では、独立行政法人国立国語研究所の廃止と大学共同利用機関法人への移管が決定されたことはご存知のとおりである。役人の説明もあてにはならないことを学ぶはめにおちいった。

閣議決定を受けて文科省は科学技術学術審議会に国語研究の将来像の検討を委託し、2008年４月には『「国語に関する学術研究の推進について」報告（案)』がパブリックコメントに付された。この報告は幸い言語資源整備の必要性を認める立場で執筆されており、胸をなでおろしたことを覚えている。

パブコメ終了後、言語学会と日本語学会の重鎮による新所長候補者の選考と研究員の選考人事を経て新研究所の組織が固まり、国立国語研究所を大学共同利用機関法人に移管するための法律が2009年３月に国会に提出された。この法律は可決されたが、衆参両院で日本語教育も含めて旧国語研が実施していた事業をすべからく継承すべしという付帯決議がついたこともまたご存知のとおりである。閣議決定から付帯決議まで、much ado about nothingの15か月間であった。

６．今後の展望

国立国語研究所は2009年10月１日をもって大学共同利用機関法人に移管された。新研究所には４つの研究系が設置され、そのひとつは言語資源研究系である。その他にコーパス開発センターも併置され、このふたつの組織で従来の言語資源の研究開発を継承することになった。コーパスの開発は主にコーパス開発センターが担い、コーパスを利用した日本語研究は言語資源研究系が担う。

ただしすべてが従来どおりというわけではない。上述の『「国語に関する学術研究の推進について」報告（案)』には、新国語研の守備範囲として日本語の歴史的研究が明記されており、言語資源開発においても歴史的コーパス（過去の日本語のコーパス）を対象に含めることになっている。現在は外部の専門家の助力を得ながら歴史的コーパスの設計に着手したところである。

新国語研の船出は政治の荒波にもまれることになったが、言語資源研究系は、大学を含めて我が国の国立研究機関に初めて設置された言語資源専門の研究組織である。これ

を健全に発展させることが、新国語研に課せられた大切な指名のひとつと考えて、今後とも言語資源関連の研究に邁進したいと考えている。