参考資料２

今後における日本語のデジタル言語資源の整備・活用の在り方
（報告）

令和７年３月１７日

文化審議会国語分科会

今後における日本語のデジタル言語資源の整備・活用の在り方
（報告）

目　次

参考資料

今後における日本語のデジタル言語資源の整備・活用の在り方（報告）

はじめに

言語資源は、我が国の言語文化を伝える貴重な資料である。これらの利活用を進め、社会課題に対応するためには、デジタル言語資源を増やしていくとともに、各種の言語情報を付与した言語コーパスの整備を図っていく必要がある。

これまでの経緯に目を転じると、令和5年9月29日の国語分科会において、「日本語のデジタル言語資源の整備に関する国語分科会の見解」[参考資料4]がまとめられた。この中では、「安心して依拠できる確実な言語資源として、国語施策をはじめ、広い分野で更に活用されることが期待できる」こと、「時代ごとの日本語の姿を文化財として残していくという意義」を持つことから、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（以下「国立国語研究所」という。）の「現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese）」（以下「BCCWJ」という。）を将来にわたって定期的に更新していくことが求められている。さらに、「国民が安心して活用できる言語資源の整備を、国の事業・施策として位置付け、積極的に推進することを提案する。デジタル時代の新しい国語施策の一つとして、BCCWJを確実で信頼の置ける自然言語のデータベースとして整備・拡充するとともに、デジタル言語資源の運用の在り方までも視野に入れた検討を進めていくよう求めるものである」と述べている。

この見解を踏まえ、文化審議会国語分科会は、令和6年度、国語分科会の下に言語資源小委員会を設け、言語の果たす機能や我が国における言語コーパスに関する整備状況、海外の言語コーパスの現況を捉えた上で、今後における望ましいデジタル言語資源―とりわけ言語コーパスの整備や活用の在り方等に関して、発展的かつ俯瞰（ふかん）的に検討を行ってきた。

なお、本報告で用いる「言語資源」、「デジタル言語資源」及び「言語コーパス」の定義は、次のとおりである。

- **言語資源**：電子化されているものもされていないものも含む、書き言葉や話し言葉など多様な言語資料の総体。
- **デジタル言語資源**：言語資源のうち電子化されたもの
- **言語コーパス**：デジタル言語資源のうち、品詞名など各種の言語情報が個々に付与されたもの

以下は、国語分科会及び言語資源小委員会における議論をまとめたものである。

1 基本的な考え方

言語は、人間が思考する上で、また、他者と意思疎通を図る上で不可欠なものである。さらに、言語は、アイデンティティーに関わり、文化芸術の基盤を成し、「生きる力」の土台となるものであり、歴史と伝統の中で培われてきたものである。これまでの歴史の中で用いられてきた言語の在り方を振り返ることができれば、それぞれの時代における言語の変遷や往時の諸相を把握することが可能となる。

言語は、例えば、書籍や文献といった書き言葉から、演説や日常会話の中での話し言葉等に至るまで多様な形態で存在し、機能するものであるが、意図的に保存しない限り消え去っていく存在でもある。この点、辞典や全集物、録音・録画資料、言語コーパスといったものは、言語を意図的に保存した言語資源と言うことができる。

言語資源は、電子化されることにより、言語研究、言語政策、言語教育等においてより有用となるばかりでなく、近時、生成AIの基盤となる大規模言語モデルでの活用も期待されている。デジタル言語資源の果たすこのような役割に照らせば、我が国の言語資源を保存し、通覧・分析可能な状態にしておくことは、言語関係のみならず、社会課題の解決にも寄与し、社会の発展の上でも意義深いものである。こうしたことから、デジタル言語資源を「未来へ伝える文化遺産」と位置付け、我が国全体（産学官民）で多様な言語コーパスを構築し、活用していく土壌を整えていくことが必要であると考える。

上述した基本的な考え方の下、以下のとおり、「言語コーパスの意義・現況」、「言語コーパスの活用事例」、「生成AI等社会の変化・進展がもたらす影響」、「今後に向けた言語コーパスの整備・運用の方向性」等について示すこととする。

2 言語コーパスの現在

(1)言語コーパスの意義・現況

言語コーパスとは、実際の言語活動を体系的に記録して、電子的に検索可能な状態にして公開したデジタル言語資源の一つである。言語コーパスの要件としては、理想的には、次の7点が求められる。

① 電子化（電子的な検索が可能であること）
② 公開（囲い込まず誰もが利用できること）
③ 真正性（実際に生じた言語活動の記録であること）
④ 均衡性（様々な言語使用域を可能な限り均衡的に網羅していること）
⑤ 代表性（定められた母集団の縮図であること）
⑥ 規模（利用目的に応じたデータ量を提供できること）
⑦ 付加情報（電子的な検索に必要な情報が付加されていること）

言語は、多様であり、場面や状況に応じて使用されるものも変わってくる。そして、意味・用法が次第に変化するものでもある。言語コーパスは、文脈における個々の語の機能やその経年変化の解明を目指す言語研究において広く必要とされてきた。また、情報学においても、計算機で言語を処理するための学習用データとして、言語コーパスが注目されるようになってきた。

言語コーパスは、ある時点の言語を記録したものである。言語データは、時間の経過とともに消失しやすくなるものであるため、文化財的な価値は時間の経過とともに増していくこととなる。その意味で、言語コーパスは、言わば潜在的文化財としての価値がある。それだけに、どのような形で言語を保存するのか、次の世代にどのように引き継ぐのかという視点が必要である。

さらに、こうした言語コーパスに代表されるデジタル言語資源には、広く諸学への応用、言語を使ったあらゆる研究の下支えをするインフラとしての価値があると言える。

大学等の研究機関で様々な小規模な言語コーパスが単発で開発され、公開されているが、文化庁の「アイヌ語の保存・継承に必要なアーカイブ化事業」により、アイヌ語のデジタル言語資源化が進み、「アイヌ語アーカイブ」（公益財団法人アイヌ民族文化財団（国立アイヌ民族博物館））等の整備が行われるとともに、「消滅の危機にある方言の記録作成及び啓発事業」（「危機的な状況にある言語・方言のアーカイブ化を措定した実地調査研究」が前身事業）によって蓄積されたデータを活用して、沖縄県が「しまくとぅばアーカイブ」を作り、公開しているなど、消滅の危機にある言語・方言の言語コーパス化も進められているところである。国立国語研究所が開発してきた言語コーパスは、おおむね奈良時代から現代に至るまでの日本語の歴史的な変化をたどることが可能であり、書き言葉、話し言葉とも様々な言語使用域を把握することに資するものとなっている。とりわけ、現代日本語に関しては、国立国語研究所における代表的で基幹的なデジタル言語資源として、BCCWJが整備されている。これは、主に、1986年から2005年の間に用いられた約1億語の書き言葉を収めたデータベースである。現在、文化庁の事業（信頼できる言

語資源としての現代日本語の保存・活用のためのデジタル基盤整備事業）により、現行のBCCWJに2006年から2025年までのデータを追加し、全体として2億語規模に拡充する事業が進行中である。

過去四半世紀で、国立国語研究所における言語コーパスの整備は顕著に進み、言語研究に占める言語コーパスの位置は確立されたが、いまだ道半ばである。とりわけ、将来の日本語に強い影響を及ぼすと考えられるものとして注目される、SNS上の日本語や非母語話者による日本語産出などの言語コーパスの整備は十分ではない。

一方、言語コーパスを整備する上で、かつては著作権に係る権利処理が実務上の負担となっていた。しかし現在においては、デジタル化・ネットワーク化の進展に対応するために柔軟な権利制限規定が整備され、著作物に表現された思想又は感情の享受を目的としない利用については著作権法第30条の4に規定されたことなどから、今後は運用面において、著作権に配慮しつつ、実務の円滑化に資するよう関係者が事例を積み上げていくことが求められる。

（2）言語コーパスの活用事例

国内における言語コーパスの活用事例（可能性も含む。）としては、次のようなものが挙げられる。

国語施策の観点からは、言語コーパスは、施策の検討に際しての語彙データ等言語をめぐる実態の把握に資するものとして、また、施策実施後の反映状況の確認などフォローアップツールとして、重要な役割を果たしている。

辞書編集における言語コーパス活用の可能性としては、次の6点が挙げられる。

① 見出し語選定の参考になること（立項すべきかどうか検討中の語について、言語コーパスにより使用頻度が判明することで、経験則に基づく選定作業を客観的に補完するものとなること。）

② 記述する意味の説明の参考になること（言語コーパスによって、従来把握しきれていなかった意味・用法の広がりを把握することが可能になること。）

③ 例文の参考になること（ある語がどのような語と共起しているかが把握可能となり、例文作りが効率的となること。）

④ ある語が使用される場面や文体、位相についての情報が得られること（例えば、俗語と雅語、文章語と口語、男女差、地域差といった位相に関する情報を

得られる可能性があること。）

⑤ 表記に揺れのある語についての判断材料を得られること

⑥ ある語と連接しやすい語の特定が可能になること

教育面でも、言語コーパスは、例えば、国語科の教材や教育内容の改善に資するものであり、また、これらをアップデートするための基礎資料となり得るほか、日本語の初中級学習者にとっては、教科書を補完する生きた用例を提供するものとなり得る。さらに、上級学習者や日本語を母語としない研究者等にとっては、学術研究のための論文作成等の際に、辞書等を補完する目的で参照可能な専門的資料としても有用である。

言語コーパスの整備や活用ということでは、日本国内だけでなく、海外における事例も見ていく必要がある。以下、言語資源小委員会において事例報告していただいた、英語、フランス語、ドイツ語に係る言語コーパスについて整理しておく。

英語では、イギリスのBritish National Corpus（BNC）とアメリカのCorpus of Contemporary American English（COCA）が代表的な大規模言語コーパスである。これらの言語コーパスは、単に文脈における語の使われ方を調べるインターフェイスだけでなく、様々な分析結果を示せるインターフェイスが準備されていて、言語研究者や辞書編集者だけでない、幅広い層がユーザーとして想定されている。また、特定分野に特化した言語コーパスも多く作られている。活用参考資料5に関しては、辞書をはじめとして言語教材や評価手法の開発のほか、言語コーパスやコーパス分析技術を言語そのものの研究の枠組みを超えて、広く社会課題の解決に生かすという「コーパスアプローチ」の観点から、ヘルスコミュニケーションの研究（「肥満症」などがどのように表象されているかの分析）、患者へのフィードバックの研究、ワクチンや終末期医療に関する言説研究、企業コミュニケーションの言語研究、ヘイトスピーチ分析、国内政治や国際問題の関する言説研究などに言語コーパスが貢献している。社会課題の解決に貢献する存在として言語コーパスは認識されているのである。

フランス語では、書き言葉の言語コーパスも話し言葉の言語コーパスも作られているが、大規模な言語コーパスは少なく、小中規模の言語コーパスが多い。それらの言語コーパスの多くは、政府の予算でORTOLANGというネットワークインフラで一元化されていて、成果として、

①質が保証されたリソースをプールすることで、言語の分析、モデリング、自動処理に関する研究を国際的な最高レベルまで促進すること

②公的研究機関内に設置されたリソースやツールの利用を促進し、産業パートナーと共有すること

③公的研究機関が獲得した専門知識を共有することで、フランス語とフランスの地方言語を促進すること

という3点が期待されている。

こうした言語コーパスは、辞書作成、学習教材作成、フランス語学習者のライティングの自動評価システム作成、フランスの地方言語の研究、言語政策で活用されている。

ドイツ語では、内省ではなく、データ収集によって本質に迫ろうとするドイツ語研究への応用を主たる目的とし、ドイツやオーストリアにおいて、一部民間資金のものもあるが、主に公的資金で作られている。ドイツでは書き言葉の大規模な言語コーパスが複数あり、オーストリアでは、書き言葉に加え話し言葉の言語コーパスもある。両国の言語コーパスは、正書法の改正の際、つづり方の変化を観測する客観的な資料として活用され、国民生活に影響を及ぼす重要な問題の解決に寄与するという社会貢献を果たしている。また、ドイツでは、ドイツ語そのものを明らかにする言語研究での活用事例が多く、オーストリアでは、ドイツ語の地域特性や方言の研究での活用事例が多い。

3 言語コーパスの今後

(1)生成AI等社会の変化・進展がもたらす影響

今日、変化の激しい社会にあって、言語が変化する速度は一層増している。日本語がこれから変化していく要因としては、サイバー化、高齢化、多文化・多言語化、そして生成AIによる影響などがあり、これらは、社会の変化・進展に応じて生じてくるものと考えられる。とりわけ、生成AIに関しては、大規模言語モデルが、言語データの中に埋め込まれた知識等を獲得していく(人間がモデルに言語データを登録する)一方で、生成する言語の中に生じ得るバイアスといったものが、逆に人間社会に影響を及ぼすという双方向的な関係にあると考えられる。ただ、そもそもの言語データにバイアスが含まれている可能性も否定できない。生成AIによる出力結果のバイアスについては、それをどの程度許容するかは、それぞれのコミュニティーにおける社会通念に基づいて判断される問題であろう。

現在、大規模言語モデル学習用のテキストデータの不足が指摘される中で、大規模言語モデルの訓練に、生成AI自体が産出したテキストが活用されることが予想される。言語空間には、これまで参照されてきた出版物等のほか、生成AI自体が産出したテキスト、人間と生成AIがやり取りしたテキスト、生成AIの影響を受けて人間が作成したテキストが混在していくことが考えられ、こうした中でいかにテキストの品質を保持していくかが課題となる。それだけに、信頼できる言語データは大いに求められているのである。

言語空間のモニタリングに際しては、人間が作り出したテキストと生成AIが作り出したテキストをそれぞれ区別するなど、混在するテキストをどのように見極めていくかという点にも留意が必要である。

(2)今後に向けた言語コーパス整備・運用の方向性

言語コーパスをより良いものにし、社会から認知されていくためには、BCCWJのような大規模言語コーパスの整備拡充はもとより、各種の小規模言語コーパスにも視野を広げ、これらの開発促進や、言語コーパス開発後も、取組の継続、普及及び社会実装を意識することが重要である。

(言語コーパスの整備・拡充)

BCCWJのような基盤的言語コーパスについては、長期にわたる拡充が確実になされるよう支援する必要がある。

さらに、書き言葉としての言語コーパスとともに、話し言葉としての言語コーパスに関しても、例えば、方言が有する地域コミュニティーを支える意義や消滅の危機にある現況に鑑み、一般国民がアクセスしやすく、未来へ伝える地域資源として、普及・啓発活動に有用なものとなるよう、方言コーパス等の機能の充実を図っていくことが必要である。

これらに加え、古典籍を含む経年的言語コーパス、またSNSのような新しい媒体を含む社会の幅広い言語使用に対応した言語コーパス、さらには、特定の社会課題に直結した各種の専門的言語コーパスなどについても、その機能の充実を図っていくことが重要である。

(取組の継続)

取組の継続に関して、一般に、言語コーパスは、ある期間に限定した言語データを収集したものであるから、言語変化を追跡していくためには、継続的に言語データを収集する必要がある。その際、同じサンプリング基準でデータを収集することにより、語や文法

項目の使用状況や用法に係る変化の推移を把握することが可能となる。

言語データの収集形式には、同一母体型（同一の機関が継続的にデータを収集する形式）又は拡張分散型（当初の言語コーパス開発者以外の第三者が当該言語コーパスのサンプリング基準を用いて開発を承継する形式）があり得るが、実質的には、言語コーパス開発は、属人的で職人芸的なものであったと言える。このため、全体の設計、データの収集方法等専門的なノウハウを承継し、組織化するための人材育成等における工夫が必要であり、言語コーパス開発を担う主要な研究機関への継続的支援が必要である。

海外の事例に見られるように、基幹となる大規模言語コーパスと同時に、特定分野に特化した小規模言語コーパスの開発も有用であり、それらの言語コーパスを横断的に扱えるプラットフォームの構築もインフラ機能を強化する一環として位置付けられる。

（言語コーパスの普及）

言語コーパスのより一層の普及に関しては、言語研究者、辞書編集者に限らず、一般ユーザーの利用も視野に入れる必要がある。このため、コーパスの価値を実感しやすい工夫が必要であり、以下のとおり、言語コーパスの活用を進め、その意義を発信するほか、利便性を考慮した検索の在り方を含むインターフェイスについて研究することが必要である。

・辞書における言語コーパスの活用を進め、その意義を発信することで、言語コーパスに触れるきっかけを作るとともに、言語コーパスの意義を理解してもらう取組が考えられる。

・検索に関しては、一般的には、キーワード検索というものがあり、文脈における当該キーワードの使われ方が示される。この点、例えば、ある語を入力すると、当該語の近傍に出現しやすい語の一覧を頻度とともに示すことや、コロケーションの観点から、ある語を入力するとその語と関係の深い語の全体像が、図解で示されるといった工夫などが考えられる。

・ユーザーインターフェイスの観点からは、ユーザーが特定の語や語句を入力するだけで、それに関する幅広い情報（類義語、多義語、意味、用法、用例、頻度等）を分かりやすい一覧の形で示すといったシステムを研究していくという視点もあると考えられる。

（社会実装）

言語は、社会における意思疎通の手段として、人間関係の構築や社会の信頼感の形成などに深く関わっている。言語コーパスの構築や分析、それに関わる手法の共有は、円

滑なコミュニケーション基盤の構築を必要とする様々な現場での社会実装に結び付くものとなる。

国民生活に密着する分野における社会実装に関しては、例えば、言語教育や通訳、医療、社会福祉分野への応用が考えられる。言語教育については、言語コーパスに収集されたデータの調査結果を踏まえ、例えば、口語においてどのような用法がどの程度用いられるのかに関する具体的な情報を提示し、高頻度の用法を手厚く解説した会話教材など、従前のものに比べ、より実践的でアップデートされた教材の開発が期待される。このほか、通訳の面では、法廷などの通訳場面で通訳しやすい日本語表現を分析することでスムーズなやり取りが期待される。医療面では、医療従事者と患者の会話をコーパス化し、当該コーパスを分析することにより、当事者間における信頼関係の構築に寄与する具体的な言葉のありようを解明し、さらには、得られた知見を医療人材向けの教育に応用していく可能性などもあると考えられる。また、児童福祉や障害者福祉などの社会福祉の観点からは、子育てや合理的配慮、日本語を母語としない人に対する伝え方への参考となるような会話を蓄積し、関係者へ提供することで、効果的な話し方や言葉の選択、ダイバーシティー推進に資する、包摂的社会の実現の一助としていく視点もあると考えられる。この点、生成AIにより産出される表現は、自然言語と異なり、感情や身振り手振りを伴って発せられるものではないことから、上記のような視点を持ち合わせておくことは、これからのAI時代においては有用であると考えられる。同様に、感情や身振り手振りを伴う感謝や謝罪の場面のやり取りのデータから、言語運用面の分析を行うことで、感謝や謝罪といった行為に資するコミュニケーション方法を導くことも期待される。

4 将来社会を見据えた考慮事項

言語コーパスは、第一義的には、共時的又は通時的な言語研究に有用なものであり、我が国の言語政策、ひいては文化政策の基盤的な部分に関わるものと言える。同時に、言語コーパスは、我が国の社会課題の解決に資する役割も期待されるところであり、長期的に見て社会の助けとなるような国民共有の資産を構築する視点を持ち合わせておくことも重要である。このような言語コーパスの意義に照らせば、我が国全体で多様な言語コーパスを整備していく機運を醸成していくことが必要である。

今後に向けた言語コーパス整備及び活用における考慮事項としては、次のように整理できる。

①言語コーパス開発に取り組む主要な研究機関への継続的支援

②言語コーパスの価値を広く知ってもらうための普及策の工夫

③社会課題を解決する一助となる「コーパスアプローチ」の実践

さらに、

④言語コーパス開発やメンテナンスに対する意義の再確認

⑤素材となり得る言語データの収集が、円滑になされるようにすること

⑥書き言葉だけでなく話し言葉など多様な言葉を蓄積すること

⑦ノウハウの継承のための人材育成

もとより言語資源は、出版物等が創造的に生み出され、多様に存在することを通じて、重厚なものとなり、また活性化されるものである。このような活動は、これまで長い歴史の中で育まれ、取り組まれてきた。しかし、急速に社会に普及する生成AIに係るデータの収集等の取扱いの在り方を含め、今後も継続的かつ多様に行われ、国民が安定的にこれらから生み出される効用を享受できる機会や環境が維持されるよう、社会の動向やその在り方について注視する必要があると考えられる。

変化の激しい混迷の時代にあって、我が国は今後、多文化・多言語社会の到来により、必ずしも誰もが同じ言語を知っているという社会ではなくなってくる可能性がある。このような多様性を内包した社会においては、狭量な社会通念に縛られることなく柔軟性のある言語観の在り方を意識しておくことも必要な視点となる。このような視点に立てば、中長期的な課題として、日本語母語話者ではない人々が話す日本語の言語コーパスの整備の在り方や生成AIによる日本語への影響についての評価の在り方が、今後、研究対象になってくると考えられる。

おわりに

これまで述べてきたとおり、言語の果たす機能や我が国における言語コーパスに関する整備状況、海外の言語コーパスの現況を捉えた上で、今後における望ましいデジタル言語資源―とりわけ言語コーパスの整備や活用の在り方等に関して、発展的かつ俯瞰（ふかん）的に検討を行ってきた。

近年、例えば、デジタルヒューマニティーズ（情報人文学）の分野で、古文書、絵巻物なども含め人文学分野の様々な資料の電子化が行われ、公開されるようになってきた。こうした資料もデジタル言語資源と捉えることが可能であり、デジタル言語資源を活用した

社会課題への取組も出てきている。このような動きの中で、言語コーパスも、生成AIとの関わりだけでなく、様々な社会課題の解決に寄与するため、幅広い分野と結び付きを持ち、「コーパスアプローチ」という手法を共有していくことが望まれる時代にあるという共通認識を持つ必要があると考える。

結びに、今日、これまで行われてきた言語コーパスの整備という局面から、一歩進んで、言語コーパスを活用した新しい見取り図を示していく局面へと進化してきている。言語コーパスを用いた価値創造の有用性に照らし、言語コーパスを活用した社会実装に向け、社会全体の一層の理解増進と連携協力を望むものである。

参 考 資 料

参考資料1

文化審議会国語分科会委員名簿

（敬称略・五十音順　　◎分科会長　○副会長）

○	相　澤　彰　子	情報・システム研究機構国立情報学研究所教授、副所長
	石　川　慎一郎	神戸大学教授
	植　木　朝　子	同志社大学文学部教授
	大　島　中　正	同志社女子大学教授、元日本ローマ字会代表理事
	神　永　　暁	元小学館辞典編集部編集長
	川　口　敦　子	三重大学人文学部教授
	川　瀬　眞由美	株式会社テレビ朝日アスク取締役
	川　辺　章　絵	江東区立毛利小学校校長
	木　村　　一	東洋大学教授
	齊　藤　美　野	順天堂大学国際教養学部准教授
	斎　藤　純　男	拓殖大学外国語学部教授
	滝　浦　真　人	放送大学教授
	武　田　　京	一般社団法人日本書籍出版協会国語問題委員会副委員長、株式会社三省堂出版局辞書出版部次長
	棚　橋　尚　子	奈良教育大学国語教育講座教授
	常　盤　智　子	白百合女子大学教授
	中　江　有　里	俳優、作家、歌手
	長　岡　由　記	滋賀大学教育学部准教授
	成　川　祐　一	共同通信社用語委員長
	古　田　徹　也	東京大学大学院人文社会系研究科准教授
	前　川　喜久雄	国立国語研究所所長
	前　田　直　子	学習院大学文学部教授
	村　上　政　彦	公益社団法人日本文藝家協会常務理事、作家
◎	森　山　卓　郎	早稲田大学文学学術院教授
	山　本　真　吾	東京女子大学現代教養学部教授
	山　本　玲　子	京都外国語大学・短期大学キャリア英語科教授

参考資料2

言語資源小委員会委員名簿

（敬称略・五十音順　◎主査　○副主査）

◎相　澤　彰　子	情報・システム研究機構国立情報学研究所教授、副所長
○石　川　慎一郎	神戸大学教授
植　木　朝　子	同志社大学文学部教授
神　永　　暁	元小学館辞典編集部編集長
齊　藤　美　野	順天堂大学国際教養学部准教授
武　田　　京	一般社団法人日本書籍出版協会国語問題委員会副委員長、株式会社三省堂出版局辞書出版部次長
前　川　喜久雄	国立国語研究所所長

参考資料3

審 議 経 過

国語分科会

○第87回（令和6年6月3日）

・国語分科会長・副分科会長の選出

・今期の検討課題確認

・ローマ字小委員会及び言語資源小委員会の設置

○第88回（令和6年12月10日）

・ローマ字小委員会における審議経過報告

・言語資源小委員会における審議経過報告

○第89回（令和7年3月17日）

・ローマ字小委員会の報告

・言語資源小委員会の報告

言語資源小委員会

○第1回（令和6年7月11日）

・主査・副主査の選出

・主な審議事項の確認

・前川委員よりヒアリング（国立国語研究所における言語資源等について）

○第2回（令和6年8月1日）

・石川副主査よりヒアリング（英語コーパスを踏まえた言語資源の在り方について）

・神永委員よりヒアリング（国語辞典におけるコーパスの活用可能性について）

○第3回（令和6年9月26日）

・相澤主査よりヒアリング（大規模言語モデルとテキストコーパスについて）

・川原繁人発表者（慶應義塾大学言語文化研究所教授）より
ヒアリング（生成AIにおける課題について）

○第4回（令和6年11月28日）

・石川副主査よりヒアリング（英語コーパスを活用した社会課題への取組について）

・杉山香織発表者（西南学院大学外国語学部教授）より
ヒアリング（フランス語コーパスの現状について）

○第5回（令和6年12月26日）

・今道晴彦発表者（広島大学大学院人間社会科学研究科准教授）より
ヒアリング（ドイツ語コーパスの現状について）

○第6回（令和7年2月17日）

・審議報告（素案）の審議

参考資料4

文化審議会国語分科会（令和5年9月29日）
日本語のデジタル言語資源の整備に関する国語分科会の見解

1　社会や技術の変化と言葉の在り方

● 社会の変化は言葉の在り方にも影響する。これは歴史上ずっと続いていることであるとも言えるが、インターネットが情報交換の中心となったデジタル時代において、言葉の変化する速度は更に増している。例えば最近ではよく用いられる「…していただくことは可能でしょうか。」といった表現は、20年ほど前にはほとんど見られることがなかった。国語施策においては、日本語を用いた円滑なコミュニケーションの実現を目指す上で、それら言葉の移り変わりや定着の様子を適切に捕捉していくことが必要である。

● また、新たな技術の進展によっても言葉の在り方は影響を受ける。例えばいわゆる大規模言語モデルとそれによる生成型AIが話題になっている。特にテキスト生成型のAIが実際に様々な分野で活用されつつある現状は、言葉の問題を考えていく上で看過できない変化の一つであろう。国語施策においても、今後、様々な課題に取り組むに当たっては、新しいテクノロジーの信頼性を慎重に判断しつつ、安全かつ有効に活用していくことが求められる。

2　信頼できるデジタル言語資源としてのコーパス

● 特に、喫緊の課題について検討する際に施策の根拠とすべき調査等を行うに当たっては、言葉をめぐる現状を適切に判断する上で、できる限り規模が大きく、バランスのとれた、信頼の置ける調査対象を確保することが不可欠である。そのためには、日本語に関する精度の高いデータを収集・保存し、将来にわたって安心して参照できるデジタル言語資源として整備していくことが望ましい。

● 生成型AI等に関わる大規模言語モデルは、ウェブから得られる数千億、更にはそれ以上の語を集めたものもある巨大なデータである。ただし、ウェブ上のデータは比較的容易に収集で

きる一方で、情報の信頼性に難があり、内容にも偏りが生じやすい。加えて、場合によっては著作権等の知的財産権や個人情報の保護に関わる問題が生じることとなる。最近においてはAIが生成したテキストをAIが重ねて学習に使い新たなテキストを生み出すといったデータの汚染や、悪意ある改ざんの危険も指摘されている。さらに、大規模言語モデルの多くは一部の世界的な企業によって寡占されており、その運用は必ずしも透明性の高いものではなく、また、一般の人々が自由に利用できない場合が少なくない。

● 一方、ウェブによる大規模言語モデルとは別に、日本を含む各国において、以前から、書籍や新聞など、ウェブ以外の分野までを含んだ自然言語のデータが収集され、確かな出典に基づくデータベースが構築されてきた。このような自然言語データベースは、「コーパス（Corpus）」と呼ばれる。

● 日本国内においても、例えば、国語研日本語ウェブコーパス、昭和話し言葉コーパス、日本語諸方言コーパス、日本語日常会話コーパス、名大会話コーパスなど、国や各学術機関等によって幾つものコーパスが構築されてきた。そのうち、最も代表的で基幹をなすものとしては、平成18（2006）年から独立行政法人国立国語研究所が構築を開始し、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への移管後、平成23（2011）年に公開した「現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese）」（以下「BCCWJ」という。）が挙げられる。

3 BCCWJの優位性

● BCCWJは、世界的にもまれな高品質のコーパスとして知られる。全てのデータについて著作権と個人情報に関する処理が施されているとともに、検索用の文法情報が付されており、専門的知識を持つ人材の目によって一つ一つ確認が行われている。安定的な管理運営の下に確保された信頼度の高いデジタル資源であり、主に昭和61（1986）年から平成17（2005）年の間に用いられた約1億語の書き言葉を収めたデータベースとして広く各分野で活用されてきた。

● 特筆すべき点として、その均衡性が挙げられる。ウェブ上のデータ（掲示版、ブログ等）のほか、書籍、雑誌、新聞、白書、教科書、法令などを含む多種多様な媒体を広く対象とし、統計学的な手法に基づいてデータを収集することによって、現代の日本語における書き言葉の全体を偏りなく反映し、相似的にバランスよく縮小したモデルとなっている。日本語に関する専門的

な研究に用いられるのはもちろん、辞書の編さん、心理学・認知科学、自然言語処理等に幅広く活用されている。特に、日本語の研究における貢献度は非常に高く、例えば現在の日本語文法に関する研究成果の半数以上がBCCWJを活用しているという調査結果もある。

● また、BCCWJは一般に広く公開され、誰でも無料で使用することが可能である。情報の透明性も高く、資料の出典とコーパスの設計や構築の過程までが公表されている。詳細なデータは商業的な利用にも供されており、国際的なIT企業をはじめとする約70の企業と有償の利用契約が結ばれている。そのほか、平成22年の常用漢字表の改定においては、一般公開に先立って漢字使用の実態把握のためにデータが提供されるなど、国語施策・政策にも寄与してきた。

● 現在、国語分科会では、ローマ字のつづり方や外来語の表記に関する検討を行おうとしており、審議を支えるための適切な調査を必要としている。現代における言語使用の実態を捉えるに当たっては、様々な媒体における信頼のできる言語データを集めて分析する必要がある。この点で、あらかじめ各分野のテキストを広くバランスよく収集してあるBCCWJのような高品質のデータベースは最適であるとも言える。

4 BCCWJの課題

● BCCWJは、データの質をはじめ設計や機能は十分であるのに対し、収納されたデータは公開以後更新がなされていない。これは、国立国語研究所が大学共同利用機関法人人間文化研究機構に移管され、より学術的な研究機関としての性格を強めていることにも関係している。当初は最先端の研究としての側面を持って構築されたものであるが、既に稼働しているコーパスについて、それを拡大し整備することについては、新規性のある学術研究とはみなされないため、新たなデータを追加することは困難となっていた。

● 社会変化や技術の発展によって、急速に言葉の在り方が変化している現代において、国語施策をはじめ各分野で有効に活用するには、最新のデータを含むような更新を重ねていくことが不可欠である。BCCWJは上述のような事情により、平成17（2005）年から現在に至るまでの20年に近い期間のデータが収められておらず、現時点の日本語の姿を捉えるために用いることは難しい状況にある。

● また、日本のBCCWJが約1億語の規模にとどまる一方、諸外国のコーパスの多くは、定期的にデータの追加・更新がなされ、現在も整備・拡張されている。アメリカのCOCA（Corpus

of Contemporary American English）は約10億語、ドイツのＤｅＲｅＫｏ（Das Deutsche Referenzkorpus）は約550億語規模である。そのほかの国々においても、スペイン語で20億語、ロシア語で4億5,000万語、ポーランド語で10億語規模のコーパスがある。また、比較的話者の少ない言語においても、言語文化の保存や振興の目的も含め、チェコ語で10億語、スロベニア語で10億語の規模でコーパス構築が行われている。

5 BCCWJの可能性

● 今後、BCCWJのデータ更新・追加が行われていけば、安心して依拠できる確実な言語資源として、国語施策をはじめ、広い分野で更に活用されることが期待できる。また、将来にわたって定期的に更新を行うことによって、国語施策の成果がどのように現代の日本語に反映されているのか、その評価指標としても用いることが可能である。

● さらに、BCCWJのような精密に構築されたコーパスは、億単位の語による規模であっても、今後の大規模言語モデルの向上に貢献できる可能性がある。例えば大規模言語モデルは、AIによる再学習によって随時精度を調整する必要があり、この再学習においては、高精度なコーパスを言語モデルの規範として活用することができる。また、高品質のコーパスは、サイズが比較的小さな場合であっても、それ自体が生成型AI等を支える言語モデルとして、高い性能を引き出すといった研究もある。

6 国語施策としてのデジタル言語資源の整備

● ウェブ上の数千億以上の語を集めた大規模言語モデルを背景とする生成型AIのような新たなテクノロジーが登場し、期待を集めている。このような急激な変化を踏まえつつ、日本語の実態を正確に反映した、確かな出典によるより安全で信頼度の高い言語資源の重要性を、改めて見直すべきである。しかも、我が国においては、BCCWJによって既にその土台が整えられている。これを国語施策の観点から再整備し、適切にデータを追加していくことによって、これまで以上に有用な日本語のデジタル言語資源として、用途や利用者が広がっていくことが期待される。

● 再整備に当たってはこれまでの経緯を踏まえ、最新データの追加を定期的に行うような道筋を付けることが不可欠である。その時々の国語の在り方を映し出すことが将来にわたって可能になれば、時代ごとの日本語の姿を文化財として残していくという意義も見いだせよう。

● 文化審議会国語分科会は、こうした国民が安心して活用できる言語資源の整備を、国の事業・施策として位置付け、積極的に推進することを提案する。デジタル時代の新しい国語施策の一つとして、BCCWJを確実で信頼の置ける自然言語のデータベースとして整備・拡充するとともに、デジタル言語資源の運用の在り方までも視野に入れた検討を進めていくよう求めるものである。

参考資料5

英語における言語コーパスの社会実装例
（言語資源小委員会における報告及び審議に基づくまとめ）

1　言語資源の社会実装

- 言語資源は一義的には言語研究に資するものであるが、その枠組みを超えて社会の諸問題の解決に言語資源を生かすという方向も考えられる。

2　英国CASSにおける研究

- 一例として、英国では、ランカスター大学にESRC（経済社会研究評議会） Centre for Corpus Approaches to Social Science（CASS）という研究センターが設置されており、コーパス研究手法の社会科学への適用や、コーパス分析技術を修得した次世代の社会科学研究者の育成が実施されている。
- 扱う分野は、公衆衛生系、経済・ビジネス系、社会心理・社会思想系、国際・国内政治系、文学・言語学・言語教育系など多彩である。
- 公衆衛生系では、ヘルスコミュニケーションという観点から、メディアの発信などをコーパス化して、不安症・精神症・肥満症・ワクチン接種・マスク着用・終末期医療などがどのように描写され、どのように社会的に捉えられているかを分析する研究や、医療現場での発話をコーパス化して医療従事者向けの望ましい言語使用法を分析する研究などがなされている。
- 経済・ビジネス系では、金融関連テキストの言語処理研究、企業が発行する株主向けレターや財務情報などをコーパス化して企業コミュニケーションの在り方やコーポレートガバナンス改善の方策を探る研究などがなされている。
- 社会心理・社会思想系では、オンライン上での女性に対する差別的発話や各種のヘイトスピーチ等をコーパス化して分析し、その効果的な抑止策を探る研究や、地域別コーパスを用いた世論の分析、時代別コーパスを用いた各種概念（貧困など）に対する意識変容の研究などがなされている。
- 国際・国内政治系では、社会的課題（宗教問題、都市部の暴力事案、地球温暖化、移

民など）を取り上げ、メディアの発信やそれに対する読者のコメントなどをコーパス化することで、これらの課題に対する社会的意識の実態やその時系列的変化を解明する研究などがなされている。

- 文学・言語学・言語教育系では、文学テキストとGIS地理データを統合した空間人文科学の研究や、原典と翻訳を収集した対訳コーパスに基づく研究、言語能力テストにおける受験者発話を収集した大規模コーパスの構築研究、関連資料に基づく教育政策の批判的談話分析研究などが行われている。

3 そのほかの言語資源基盤型の社会実装系研究

- CASSの枠組みのほかでも、コーパスを基盤として、社会課題の解決を目指そうとする各種の研究がなされている。
- 医療系では、自閉スペクトラム症（ASD）や抑鬱症などの診断手法の開発研究、肥満症などの治療プログラムの改善研究などがなされている。
- 医療教育の分野では、医療従事者の言語トレーニング関連の研究が広く実践されており、一例として、非母語話者看護職員向けの英語の発話トレーニング法などの開発も行われている。
- 外国語教育の分野では、コーパスデータから重要語を抽出するなどして学習語彙表を開発する研究や、単語と習熟度の関係をモデル化し、特定語の使用度を手掛かりとして書き手や話し手の習熟度を簡易推定する手法の開発などもなされている。

4 英語コーパスを用いた社会実装型研究から学べること

- BCCWJのような大規模汎用コーパスは重要だが、それだけで社会実装型の研究を行うことは難しく、大規模汎用コーパスを核として、様々な社会課題に特化した多様な小規模専門コーパスが多数開発されることが望ましい。
- コーパスやコーパス言語学の枠組みを超え、「corpus approach」という視点の下、コーパス言語研究者と他分野研究者の協働を促す仕組みが望ましい
- 日本でも問題となっている各種の社会的課題（差別や偏見の解消、教育改革、行政システムの効率化など）を取り上げ、それらの解決を目指すために必要なデータを精査

し、目的別に、多様な小規模コーパスを開発していくことが望ましい。

- コーパス研究者は、こうした他分野のコーパス構築を直接的・間接的に支援し、収集したコーパスデータの分析手法に関して言語学研究で蓄積されてきた各種の知見を提供できる。