

言語資源の在り方に関するこれまでの議論の整理

文化審議会国語分科会言語資源小委員会では、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、今後における望ましい言語資源の在り方等に関して検討を行ってきたところ、以下のとおり、これまでの議論を踏まえた現時点における言語資源の在り方に関する整理を示す。

1. 基本的な考え方

- 言語(言葉)は、人間が思考し、他者と意思疎通を図る上で不可欠なものである。また、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。
- このような言語の果たす意義に照らし、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で、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に至るまで多様な形態で存在し、機能するものであるが、体系的な言語資源として代表的なものは、辞典や言語コーパスといったものである。
- これらの言語資源は、言語研究、言語政策、言語教育等において有用であるばかりでなく、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。言語資源の果たす役割に照らせば、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、社会の発展の上でも意義深いものである。こうしたことから、言語資源を「未来へ伝える文化遺産(仮称)」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ます。
- なお、本委員会における議論は、昨年9月に国語分科会において取りまとめた「日本語デジタル言語資源の整備に関する国語分科会の見解」と軌を一にするものであり、より発展的かつ俯瞰的に将来の言語資源の在り方を示していこうとするものである。
- 上述した基本的な考え方のもと、以下のとおり、コーパスの意義・現況、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. コーパスの意義・現況等

(1) コーパスの意義・現況

- コーパスは、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言語資料である。コーパスの要件としては、電子化及び公開性のほか、真正性（実際に生じた言語活動の記録であること）、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）、代表性（定められた母集団の正確な縮図となるよう取り出されていること）が求められ、さらに一定レベルの規模を有している必要がある。
- そもそも言語は、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。コーパスは、文脈における言語機能やその経年的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、コーパスが注目されるようになってきた。
- 国立国語研究所が開発してきたコーパスは、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言語資源として、「現代日本語書き言葉均衡コーパス」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、同コーパスに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。
- コーパスは、ある時点の言語を記録したものであるが、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。コーパスは、言わば潜在的文化財としての価値がある。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と言える。
- 過去四半世紀で、国立国語研究所における日本語コーパスの整備は顕著に進み、言語研究に占める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は、SNSや非母語話者による日本語産出などである。
- 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

たが、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、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と考えられる。

(2) コーパスの活用事例

- 国語施策の観点からは、コーパスは、施策の検討に際しての語彙データ等言語を巡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。
- 辞書編集におけるコーパス活用の可能性としては、次の6点が挙げられる。第一に、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、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。第二に、記述する意味の説明の参考になること。コーパスによって、従来把握しきれていなかった意味・用法の広がり把握することが可能になること。第三に、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。第四に、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。第五に、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。第六に、ある語と接続しやすい語の特定が可能になること。
- 教育面でも、コーパスは、例えば、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うる。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。
- このほか、コーパスは、前述のとおり、言語学、情報学等の学術研究における活用が期待されている。

3. 今後に向けたコーパスの整備・運用等

(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、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一方で、すなわち、人間がモデルに言語データを登録する一方で、モデルが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。
- この点、元来、言語データにはバイアスが含まれている可能性がある。また、生成AIによる出力結果のバイアスをどの程度許容するかについては、それぞれのコミュニティにおける社会通念に基づいて判断されることとなると考えられる。
- 今後、学習データが不足していく中で、大規模言語モデルの訓練に、モデル自体が生成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自体が生成したテキスト、人間と生成AIがやりとりしたテキスト、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。
- 言語空間のモニタリングに際しては、人間と生成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) 今後に向けたコーパス整備・運用の方向性

- コーパスをよりよいものにしていくためには、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。

(取組の継続)

- 取組の継続に関して、一般に、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、言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。
- 言語データの収集態勢には、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(当初のコーパス開発者以外の第三者が当該コーパスの開発を承継する形式)があり得るが、実質的には、コーパス開発は、属人的で職人芸的なものと言える。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、組織化するための工夫が必要であり、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。

(コーパスの普及)

- 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、利便性を考慮した検索の在り方を研究することが必要である。
- 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ところである。
- さらに、ユーザーインターフェースの観点からは、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)をわ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、社会におけるコミュニケーションの手段として、人間関係の構築や社会のトラスト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つくものとなる。
- 上記を踏まえ、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、言語教育や医療、社会福祉分野への応用が考えられる。言語教材については、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、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の観点からは、子育てや合理的配慮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により生成される言語は、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。

4. 将来社会を見据えた考慮事項

- 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、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、長期的に見て社会に裨益す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り、今後に向けたコーパス整備の考慮事項としては、前述したとおり、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やその一層の普及等を図っていくことのほか、このようなコーパス開発やメンテナンスに対する意義の再確認や、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が挙げられる。
- 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、取り組まれてきたが、急速に社会に普及する生成AIに係るデータの収集等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。

5. 今後に向けて

- 言語資源小委員会においては、引き続き、将来の言語資源の在り方についての議論を深め、コーパスに関する諸外国における状況や社会における活用状況など、更なる調査・検討を深めていく予定である。

これまでの言語資源小委員会における審議状況等

○第1回(令和6年7月11日)

- ・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(2月に開催予定)