言語資源小委(R8.2.26)

資 料 2

今期の言語資源小委員会における審議経過報告（案）

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。

令和6年度の言語資源小委員会においては、言語コーパスのことを中心にデジタル言語資源の現況を把握した上で、その整備や活用の在り方について検討を行い、「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・国語分科会）を取りまとめた。

令和7年度の言語資源小委員会においては、上記報告において示された方向性を踏まえ、補完的事項について審議を行った。具体的には、次の四つの課題を設定し、有識者による報告と意見交換を中心に審議を進めた。

○言語コーパスなどデジタル言語資源整備の取組の進捗状況

○言語コーパスなどデジタル言語資源を活用した社会課題解決のための取組

○地域の文化的基盤としての地域の言語資源の在り方

○外来語をめぐる課題と言語コーパスの活用　など

以下は、上記の課題設定の下、行われた、今期の審議経過を整理したものである。

1 言語コーパスなどデジタル言語資源整備の取組の進捗状況

前川委員（国立国語研究所所長）による「BCCWJ2の設計と構築」と題した報告を実施した。

・BCCWJ2の拡充作業は、比較的順調に進行

・2025年度末には、最初の部分的なデータの公開を予定

（現行のBCCWJ1とは別コーパスとして公開するが、串刺し検索には対応）

・2026年度から成果発表会開催を予定
・2028年度で構築を完了予定
・BCCWJ2は、2006年から2025年までの日本語データを追加するもので、追加されるデータは、
2006年～2025年出版書籍（サンプリング） ⇒ 1億語
3回の学習指導要領改訂に対応した、
小学校～高等学校の全教科の教科書1種ずつ（全文） ⇒ 2000万語
2024年～2026年SNS ⇒ 70万語
を予定
・SNSのデータは、これまでのコーパスに含まれていない、対話となっている書き言葉のデータであり、新しい価値、新しい使い道の可能性有
・検索において表示される語数をこれまでよりも少なくすることで、思想又は感情の享受を目的としない、軽微利用の範囲に収まるように対応

報告を踏まえた意見交換において、
・BCCWJ2は、国語辞典の編集に間違いなく活用でき、国語辞典の内容が大幅に変わることにつながる可能性を持っている。
・BCCWJ2は、人間だけが作っていた言葉の時代の最後の日本語のデータ集合になる可能性があり、大規模言語モデルの評価などで新しい価値が出てくるかもしれない。

という評価が示された。

2 言語コーパスなどデジタル言語資源を活用した社会課題解決のための取組

渡部泰明氏（国文学研究資料館館長）による「データ駆動による課題解決型人文学の創生～データ基盤の構築・活用による次世代型人文学研究の開拓～」と題した報告を実施した。

・2014年～2023年に30万点の古典籍をデジタル化し、データベースとして公開
・データ範囲の近代初期までの拡張と構造化テキストの作成
・公開したデータベースを利用した異分野融合によるデータ駆動型研究の推進
（文献観光資源学、典籍防災学 など）

報告を踏まえた意見交換において、

・多くの人の利用のためには、現代語訳や校訂本文、英訳もあるとよい。
・精度が100％でない段階のテキストデータも使わせてもらえることで研究が進む面もあろうから、公開について考えてほしい。
・広く国民に資する基盤的言語資源の構築には、国からの長期的支援が必要である。
・形態素分析の技術と国文学研究所が作成したテキストと結びつくことで、異分野融合、関連分野融合が進み、課題解決につながるものと思う。
・デジタル化された資料の公開が進むことは好ましいことである一方、オリジナル資料も保管されていることは日本文化の発信において重要なことである。

という意見が示された。

神永委員（国語辞典編集者）による「国語辞典におけるコーパス利用について～『日本国語大辞典』を中心に～」と題した報告を実施した。

・用例主義を採用している『』日本国語大辞典』の改訂作業が進行中
・用例掲出に当たり、デジタル化された資料をいかに活用するか
　頻度数による見出し語選定の参考
　記述する意味の参考
　見出し表記の参考
　コロケーションの参考
・辞書を入り口とした多種多様な知的情報への動線のデザイン

報告を踏まえた意見交換において、

・紙の辞書では改訂で変わった部分が分かるが、データ更新の場合、変わった部分が分からなくなるのではないか。
・語釈や語義の分類でAIを活用するという可能性はないのか。

という意見が示された。

3　地域の文化的基盤としての地域の言語資源の在り方

大野眞男氏（岩手大学名誉教授）による「消滅危機方言の記録作成及び啓発事業の意義―日本語の多様性のためにできること―」と題した報告を実施した。

・危機言語・方言に係るユネスコや文化庁の動向は、各地の方言の価値を見直し、活性化させる取組の後押しする力

・言語には、意思疎通の道具としての機能とアイデンティティー表出の機能とがあり、後者の機能に注目して、取り組む必要
・消滅危機方言の記録作成のための新たな調査と埋もれている既存データ（例：各地方言収集緊急調査）を掘り起こしての言語資源の整備の必要性
・各地の方言データの社会還元は、医療介護分野など高齢者の場や教育現場の方言学習材開発の場に必要
・方言の価値の見直し、方言を通じた地域への愛着の深化は、地域文化の活性化に寄与

また、畑中悠樹氏（東洋美術印刷株式会社）による「方言を活かしたブランディングの取り組みについて」と題した報告を実施した。

・都市の人口集中と過疎化に対抗するための街づくりの一環として方言を使った施策が、地方の自治体や企業の間で見られる。
・現在、方言は地方の温かみやオリジナリティーといった肯定的なイメージを持ってもらえるもの
・メディアでは標準語が使われているので、方言はインパクトが強く話題になりやすい。
・方言をむやみに使うと不自然になり、敬遠されてしまうので、わざとらしさがなく、地元の人が使っているようにする必要
・方言の使用は、県外観光客向けに地域性をアピールするものだけでなく、現地の人のふるさとへの帰属意識や愛郷心を刺激するものもある。

これらの報告を踏まえた意見交換においては、

・方言は多様であるが、データ化して一律に検索できるようにするためには画一化する部分も必要となるので、データ化やコーパス作成は難しいであろう。
・消滅危機言語の記録作成は非常に大きな課題であり、どのようにデータベース化して将来のために残すかを考える必要
・各地で方言資料が作成されているので、今後、そうした資料をいかに取り込んでいくか考える必要がある。特に、方言研究者が方言調査に使った磁気テープは将来消失するおそれがあり、緊急に対策を講ずべき貴重な資料である。
・方言ブランディングにおいても方言データベースがあれば活用される。
・目の前の短期的な目標に心が傾きがちな中、方言の継承への動機づけを強める工夫が必要であろう。

という意見が示された。

4 外来語をめぐる課題と言語コーパスの活用

小椋秀樹氏（立命館大学教授）による「コーパスを活用した外来語表記のゆれに関する調査」と題した報告を実施した。

・外来語の表記について、BCCWJを用いた調査と通時コーパスを用いた調査を実施
・表記と発音の関係を見るため、日本語話し言葉コーパス、日本語日常会話コーパスを用いた調査を実施
・国語施策においては、実態調査と意識調査から日本語の実像を把握し、必要な改定を進め、受け入れられていると考える。この実態調査の部分で、継続的に拡充されたコーパスが有益である。

金愛蘭氏（日本大学准教授）による「新聞経年コーパスを用いた外来語の基本語化の研究」と題した報告を実施した。

・書き言葉である新聞語彙において、抽象的な意味を表す外来語が基本語化し
・自作した、20世紀後半の大規模な新聞経年コーパスに基づいて調査

これらの報告を踏まえた意見交換においては、

・記事において、同じ語が続くのを避けるために、外来語を含む類義語を使用
・外来語の問題を科学的に議論していくには、経年コーパスが必須であり、各新聞社が持っている自社の記事データを言語学的にアノテーションし、共通のプラットフォームに載せて、言語研究に使えるようにするといったイニシアチブが出てくると、外来語のみならず、日本語研究の更なる発展に有用
・辞書の中では外来語の扱いが、対応する日本語の意味に置き換えるだけになっているが、外来語が既存の類義語の一部の用法に進出していることもあり、扱いを検討する必要
・明治期の文献を読むと、表記は様々で、読みづらい気もしながら、これはこれでよいという思いも抱き、表記の統一がそんなに重要なのかという気もする。

また、石川副主査（神戸大学教授）による「これからのデジタル言語資源の整備・活用を考

える」と題した報告の中で「言語資源と外来語」への言及が行われた。

・外来語の使用実態を考えれば、年次・隔年次での定点観測調査と結果報告が継続的に行われることで、国民の意識伸長や公的機関での一定の抑制効果を期待

・新しい概念を際立たせる効果がある一方で、受け手に正しく伝わるための工夫は必須

・上記調査のためにも日本語大型コーパスの継続整備が必要

5 生成AI時代の言語資源

相澤主査（国立情報学研究所教授）による「大規模言語モデルと言語資源」と題した報告を実施した。

・大規模言語モデルの構築に用いるコーパスの品質は、知識問題を解かせた正答率で調べる。

・良質のコーパスを作成するには、指示タグを取り除き、テキストを抽出し、言語種等によるフィルタリング、重複除去といったクリーニングを行う必要がある。

・大規模言語モデルは、事前学習、中間学習、仕上げ学習の3段階の学習が行われる。

・大規模言語モデルにはモデルごとに出力テキストに「口癖」がある。

・大規模言語モデルが生成する言語が人間の言語に影響を与えている現状がある。

・大規模言語モデルの生成する言語を解析するためには、モデル自体の透明性が求められるが、オープンなモデルは数えるほどしか存在しない。

報告を踏まえた意見交換においては、

・学生が生成AIで作ってきたものが非常に自然になり、内容も正しくなってきて、教育の現場でそれをどのようにしていったらいいのかということが大きな課題になってくる。

・BCCWJ1、BCCWJ2は、生成AIが関与していない状態の日本語である点に価値があり、大規模言語モデルを評価するときに利用できるのではないか。

という意見が示された。

また、石川副主査（神戸大学教授）による「これからのデジタル言語資源の整備・活用を考える」と題した報告の中で「生成AI時代の言語資源研究」への言及が行われた。

・生成AIの基礎データはブラックボックスなので、出自のはっきりしたコーパスから適切な用例を抽出し、生成AIにデータ解釈させる方向に進む可能性

・研究において、質の高いコーパスの開発と適切な用例抽出の部分が人間のリソースを

割く対象となる可能性

6 これからのデジタル言語資源の整備・活用における課題

石川副主査（神戸大学教授）による「これからのデジタル言語資源の整備・活用を考える」と題した報告を実施した。

・多種多様なコーパスが作成されているが、多くは公開されていない問題
（コーパス作成が目的ではなく、手段であるため）
・作成されたコーパスが、他の研究者の他の関心事に有益である可能性
・中小規模のコーパス作成とコーパスを活用した社会課題解決のための研究への支援

報告を踏まえた意見交換において、

・コーパスの作成に当たり、公開を前提とするのは少し酷ではないか。公開を前提に作成されるものを「コーパス」と呼び、そうでないものはデータとかデータベースと呼ぶということにしたらよいと考える。

という意見が示された。

おわりに

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はこれからも共有していく必要がある。その上で、今後も、デジタル言語資源の整備と活用について実態を捉え、中小規模のコーパスの存在も意識しつつ周辺の関連領域まで視野に入れて課題を整理していくことで、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の一歩としたい。

令和7年度 文化審議会国語分科会言語資源小委員会 開催の記録

●第7回　令和7年5月29日（木） 10:00～11:35

主査・副主査の選出

会議公開の確認

主な審議事項の確認

渡部泰明国文学研究資料館館長よりヒアリング

●第8回　令和7年8月4日（月） 10:00～12:00

前川委員よりヒアリング

相澤主査よりヒアリング

●第9回　令和7年9月9日（火） 15:05～17:05

神永委員よりヒアリング

石川副主査よりヒアリング

●第10回　令和7年12月23日（火） 15:00～17:00

小椋秀樹立命館大学教授よりヒアリング

金愛蘭日本大学准教授よりヒアリング

●第11回　令和8年1月27日（火） 15:00～17:00

大野眞男岩手大学名誉教授よりヒアリング

畑中悠樹氏（東洋美術印刷株式会社）よりヒアリング

●第12回　令和8年2月26日（木）

今期の審議経過報告（案）の審議