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（案）

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。