第5回 国語分科会言語資源小委員会（Web開催）議事録

令和6年12月26日
10時00分～11時30分
文部科学省5階・5F4会議室

〔出席者〕

- （委員）相澤主査、石川副主査、植木、神永、齊藤（美）、武田、
前川各委員（計7名）
- （ゲスト）今道晴彦広島大学大学院人間社会科学研究科准教授
- （文部科学省・文化庁）村瀬国語課長、武田主任国語調査官、鈴木国語調査官、
町田国語調査官ほか関係官

〔配布資料〕

- 1 第4回国語分科会言語資源小委員会議事録（案）（委員限り）
- 2 ドイツ語コーパスの整備状況とその活用例（今道晴彦氏御提出資料）
- 3 言語資源の在り方に関するこれまでの議論の整理
(令和6年12月10日第88回国語分科会・配布資料3)

〔参考資料〕

- 1 言語資源小委員会における主な審議事項及び当面の予定（案）
- 2 継続・普及・実装 日本語言語資源の未来に求められる3つの方向
—英語コーパス研究からの提言—（石川副主査第2回御提出資料）
- 3 言語資源の社会実装—「集めた後」を見据えて—
(石川副主査第4回御提出資料)

〔経過概要〕

- 1 事務局から定足数の確認が行われた。
- 2 事務局から配布資料の確認が行われた。

- 3 前回の議事録（案）が確認された。
- 4 今道晴彦氏から配布資料2「ドイツ語コーパスの整備状況とその活用例」に基づいて説明があり、質疑応答が行われた。
- 5 事務局から配布資料3「言語資源の在り方に関するこれまでの議論の整理」及び参考資料1「言語資源小委員会における主な審議事項及び当面の予定（案）」を使って論点及び議論の現状について確認するための説明があり、質疑応答が行われた。
- 6 今期末の言語資源小委員会の議論のまとめに向け、意見交換が行われた。
- 7 次回の言語資源小委員会について、令和7年2月17日（月）10時30分から12時30分まで、対面及びオンラインで開催する予定であることが確認された。また、国語分科会については、令和7年3月11日（火）15時から17時まで、対面及びオンラインで開催する予定であることが確認された。
- 8 質疑応答及び意見交換における各委員の発言等は次のとおりである。

○相澤主査

定刻になりましたので、ただ今から第5回の文化審議会国語分科会言語資源小委員会を開催いたします。

本日も文部科学省の会議室で御参加の方々とオンラインで御参加の方々をお迎えしまして、ハイブリッドの開催となっております。また、傍聴者の方々もオンラインでの御参加となりまして、この会議を御覧になっていることを御承知おきください。

さて、本日は議事次第のとおり、「言語資源の在り方について」、「有識者による事例等紹介」及び「その他」という内容で協議を行ってまいります。事前に資料を御覧いただいているとおり、本日は、これまでの審議内容を踏まえまして具体的な情報を頂くということで、海外のコーパスの状況と活用事例について、広島大学の今道晴彦先生からドイツ語の状況についてお伺いさせていただきたいと思っております。その後、先日の国語分科会で説明いたしました議論の整理を見直しつつ、意見交換を通して議論のまとめに向け、肉付けしていきたいと思っておりますので、それぞれの委員の皆様のお立場から幅広く御発言を頂ければと思います。今道先生にお

かれましても、御議論にも御参加いただけましたらと思います。

それでは、早速ですが、議事次第の2に移ります。「有識者による事例等の紹介」ということで、本日は、広島大学の今道先生からお話を頂きたいと存じます。前回、フランス語でございましたが、今回はドイツ語コーパスの現状と活用事例ということで、ドイツ語のコーパスをめぐる状況について御紹介いただけると伺っております。

では、今道先生、よろしくお願ひいたします。

○今道発表者

よろしくお願ひします。では始めさせていただきます。

○相澤主査

よろしくお願ひいたします。

○今道発表者

本日の発表ですが、まず、「ドイツ語コーパスの整備状況」ということで、ドイツ語コーパス開発の歴史につきまして簡単にお話しし、その後、ドイツ及びオーストリアの代表的なコーパスを幾つか紹介いたします。次に、コーパスがどのように活用されているのか、特に社会への応用がどうなっているのかが重要な関心事となっているわけですが、私の知る限り、ドイツ語の主要コーパスは、ドイツ語研究への応用を主たる目的としておりまして、その活用例のほとんどが言語研究であること、個人の研究レベルでは、社会系分野や医療現場での応用も少し見受けられますものの、前回、石川副主査がお話しされましたランカスター大学によるプロジェクトのような他分野への広範囲に及ぶ応用といった動きは今のところ見られないということをお話しいたします。

しかし、言語学の枠組みにとどまることではありますが、国民的議論を引き起したドイツ語の正書法の改正に関して、ドイツ及びオーストリアの主要コーパスが判断時の経験的基盤として大変重要な役割を担っており、ドイツ語圏におきまして社会貢献を果たしているということをお話ししたいと思ひます。

それでは、「ドイツ語コーパスの整備状況」からお話しいたします。ドイツ語コーパス開発は1964年に遡ります。この年に西ドイツのマンハイムにドイツ語研究所（IDS）ードイツ語では「イーデーエス」と言いますーが設立されまして、その年にMannheimer Korpusプロジェクトが始まります。背景としましては、ドイツ語研究がナチズムの影響を受けて、価値中立的で客観性に重きを置く実証主義的な模範から逸脱してしまった過去を克服し、言語学を自然科学同様の厳密な科学として位置付けて、言語現象も法則として捉える啓蒙主義的な理想に回帰を目指す学問的風潮があったことが関係しているとされています。当時、東ドイツは学問研究を国家のイデオロギーに服従させようとしておりましたので、それと対峙する意味合いもあったようです。

また、この時期にアメリカではノーム・チョムスキー（Noam Chomsky）が言語を生物学的な現象として捉える動きがありました。しかし、中央ヨーロッパに位置し、多言語が行き交うドイツでは、言語を社会的現象として捉え、内省ではなくデータ収集によって、その本質に迫ろうとする風潮があったことも関係しているとされています。

また、こうした考え方は、電子計算機の普及によって、更に刺激されたようです。ただし、世界初の「ブラウンコーパス」が発表された事実を、当時、ドイツ語研究所の関係者がどれほど認識していたのかは不明とされています。

ちなみに、このドイツ語研究所というのは、日本の国立国語研究所に相当する研究機関でして、近世から現在にかけてのドイツ語を対象に、ドイツ語の記録と研究を主な目的とするドイツ語研究の拠点となる組織です。

このドイツ語圏初のMannheimer Korpusですが、後にVer. 2が出ましたため、Mannheimer Korpus 1 と呼ばれます。完成は1967年で、総語数は約220万語です。ノーベル文学賞を受賞しましたトーマス・マンや、後にノーベル文学賞を受賞しますハインリヒ・ベルやギュンター・グラスなどの著名な作家を含む文学作品や学術テキスト、新聞記事などによって構成されています。

まず、ドイツ語研究所でパンチカードに記録されまして、次に、近隣のーと申しましても、今でも電車で1時間弱掛かるのですがーダルムシュタットという町にあります計算機センターに持ち込んで機械可読化され、その結果をドイツ語研究所で修正するという過程を経て構築されたようです。ただし、完成後にすぐに代表性・

均衡性の問題が指摘されまして、Mannheimer Korpus 2 が構想されます。こちらは法令、使用説明書、娯楽小説といったVer. 1 では無視されたジャンルのテキストも含まれます。ただし、総語数は約30万語ということで、極めて小規模のコーパスでした。

その後、1970年代に、ブラウンコーパスをモデルとして、LIMAS-Korpusという書き言葉均衡コーパスが構築されます。ドイツ国立図書館で作成された書誌情報を基に、500のテキストを選定しまして、1テキスト当たり2,000語という制約を設けて作成されております。総語数は約100万語になります。

ただし、当時こうしたコーパスの利用はまだ限定的で、例えば、ある動詞が属格目的語を取るか、それとも前置詞目的語を取るかといった調査をする場合、可能な動詞と全ての語形をリストアップし、プログラムを書いてパンチカードに入力した上で、ダルムシュタットの計算機で解析をしまして、専門家による修正を経まして、研究者が手動で属格目的語と前置詞目的語を分類して、結果を分析する必要があったようです。

ちなみに、ドイツ語研究所では、1969年に1階フロアの全てを占めるほどの計算機が導入されますが、1980年代に至るまで十分な活用をされず、Teubert&elica2014は、データ分析というよりは専ら事例を取り出すための採石場として利用されることが多かったと述べています。

その後、ドイツ語研究所では、ボンの新聞コーパスをはじめ幾つかのコーパスが構築されます。これらのコーパスを新しいプラットフォームに統合しまして、書き言葉のアーカイブが構築されます。このアーカイブは後述しますドイツ語書き言葉参照コーパスのDeReKoとして継承されることになります。また、アーカイブ化に合わせまして、言語学的研究の経験的基盤を築くための新しい検索ソフトウェア(COSMAS)の開発も進められます。

このプロジェクトは1991年に始まるのですが、先ほど申しましたような事例抽出のような業務負担を軽減するため、独立した新規のプロジェクトとして進める決断がなされます。1993年に申請を行いまして、1994年より、短期間ですけれども、EUの助成を受けることが決定します。これまでのCOSMASとは組織としても別のCOSMAS 2というプロジェクトが始まります。2002年にWindows版、2005年にWeb版がリリースされまして、ドイツ語研究所のコーパスが広く普及する契機となり

ます。

なお、COSMAS 2は現在も現役でして、主な機能は、検索した単語の前後の文脈を表示しますコンコーダンス検索とコロケーション機能です。

ただ、1990年代に構築されたシステムということで、現在の多様なニーズに対応すべく、2011年よりその後継としてK o r A pという検索ソフトも開発されています。正規表現のほか、各種の検索クエリ言語にも対応しておりまして、スクリプト言語のP y t h o nからも利用可能です。統計解析言語のR言語も取り入れまして、分析結果を視覚化できるようにデザインされています。

次に、ドイツ語の主要コーパスとしてドイツのコーパスを紹介します。ドイツでも様々なコーパスが構築されておりまして、本日は先ほど少し触れましたドイツ語研究所の書き言葉コーパスのD e R e K oと、余り知られていないかもしれませんが、ベルリン・ブランデンブルグ科学アカデミーのDWDSという二つの巨大なコーパスを紹介いたします。以下、本発表では、ドイツ語の発音に従いまして、「デーベーデーエス」と呼びます。

先ほどお話ししましたCOSMASやK o r A pを介して利用できます、ドイツ語書き言葉参照コーパスのD e R e K oですが、公式ホームページによりますと、総語数は約576億語で世界最大規模をうたっています。このコーパスは、言語研究の経験的基盤として役立つことを主たる目的としておりまして、新聞記事、文学テキスト、学術テキスト、一般向けの科学テキストなど、多種多様なテキストを含んでいます。

ただ、その大半は新聞記事です。年代は近世からのデータを収録していますが、大半が新聞記事ですので、比較的最近のドイツ語ということになります。ちなみにD e R e K oの中には、日本人研究者が作成しましたトーマス・マンコーパスやグリムコーパスも含まれます。このコーパスは無償で利用できまして、毎年データが更新されます。1990年5月から2002年3月までは州の助成を受けています。

次に、もう一つ、ドイツを代表しますコーパスとして、ベルリン・ブランデンブルク科学アカデミーのDWDSを紹介いたします。DWDSというのは当初、1952年から1977年にドイツ科学アカデミーによって編纂さんされました『現代ドイツ語辞典』を電子化することを目指しておりました。その後、状況が変わりまして、DWDS - K e r n k o r p u sを作成し、20世紀一後に21世紀も含まれますが一の単一言

語辞書の経験的基盤として役立つことに目的を修正します。DWD Sは大規模であること、ジャンルのバランスが取れていること、重要な文学作品や作家を数多く含むことを要件としております。

この要件を踏まえまして、イギリスのブリティッシュ・ナショナル・コーパスを参考に、時代とジャンルにおいて均衡性が担保されたDWD S-K e r n k o r p u sを構築しました。1900年から1999年までをカバーし、文学テキスト、新聞、学術テキスト、ノンフィクション、そして、これは実際のDWD Sの検索では除外されますが、話し言葉から構成されます。

なお、DWD Sでは、DWD S-K e r n k o r p u sが中心的なコーパスとなるわけですが、その後、様々なコーパスが開発されまして、既存のコーパスや、アカデミーの別のテキストアーカイブも同じプラットフォームに入りまして、現在は五つのカテゴリーから成る巨大なコーパスとなっております。

全てを合算しますと、総語数は約705億語となります。ドイツ語研究所のD e R e K oを上回っておりますが、合算したコーパスがあるわけではありませんので、世界最大規模をうたっておりません。

なお、DWD Sはドイツ科学アカデミーの一部のプロジェクトドイツ語デジタル辞書学センターのプロジェクトでもあります。前者（ドイツ科学アカデミー）は連邦政府、州政府より、後者（ドイツ語デジタル辞書学センター）は連邦教育研究省より助成を受けています。

DWD Sコーパスは、オンライン上で無償利用できまして、様々な辞書が電子化されています。単語の頻度情報、地理的分布、時系列変化、コンコーダンス、コロケーション、類義語間のコロケーションの違いが調査可能で、処理速度も速く、使いやすく、最近はドイツ語研究所のD e R e K oよりもこちらのコーパスを好んで利用する人も少なくないように思います。

続いて、オーストリアのコーパスとして、Austrian Media Corpus—単にa m cと言います—という書き言葉コーパスと、二つの話し言葉コーパスを紹介いたします。

まず、a m cという書き言葉コーパスですが、こちらはオーストリア科学アカデミーの1組織であるオーストリアデジタルヒューマニティー・文化遺産センターによって構築されています。公式ホームページによりますと、総語数は約127億語で、オーストリア通信社より資金及びデータの提供を受けておりまして、毎年データが

追加更新されます。ドイツ語研究所のD e R e K o同様に、いわゆるモニターコーパスに分類されます。

こちらのグラフ(スライド24)は、a m cの収録記事数の推移を示すものですが、コーパスの収集が始まります1986年からしばらくは緩やかな滑り出しでしたが、データ収集技術が向上します1990年代以降は、急速に収録記事数を増やし、2000年以降は毎年150万以上の記事が収録されています。

a m cは、その名称からも予測がつかますように、ジャーナリズム散文を収集したコーパスです。ただし、オーストリア全土からデータが集められているという特徴があります。Webコーパスとして有名なSketch Engineを介して利用できます。

次に、発音しづらいのですが、DiÖというのがあります。DiÖは、オーストリアにおけるドイツ語を意味する話し言葉コーパスでウィーン大学、ザルツブルク大学、グラーツ大学、オーストリア科学アカデミーにより共同運営されています。連邦教育科学研究省のオーストリア科学基金の助成を受けておりまして、プロジェクトは2016年に始動いたします。

データの内容は、インタビュー、学校でのグループディスカッション、友人間の会話、朗読などの制御されたデータも含まれます。オーストリアの56地点で収録されておりまして、収録人数は約470名です。この中には、昨今の都市の実態を踏まえまして、L2話者も含まれるそうですが、人数や習熟等につきましては不明です。なお、当該コーパスは現在のところ、一般公開されておりません。

次に、こちらも発音が難しいのですが、WBÖコーパスです。こちらはオーストリアと南チロルにおける歴史的バイエルン方言辞典プロジェクトの一部として、1963年に出版が始まりました。WBÖという方言辞典のオンライン化に伴い構築されたコーパスです。データの内容は、その大半が、方言辞典が作成されました際に収集されました手書きのメモです。2016年11月から2038年12月という長期にわたりまして、オーストリア科学アカデミーより助成を受けています。なお、規模は不明で、こちらも一般公開されていません。

それでは、「コーパスの活用例」として、ドイツ語研究所のD e R e K oの事例を見たいと思います。ドイツ語研究所は、国内外のドイツ語研究の経験的基盤として役立つことをD e R e K oの主たる目的とした上で、想定する応用分野としまして、心理学、神経学、認知科学、言語療法、コミュニケーション科学、メディア学、

統計学を挙げています。同時に、D e R e K oの活用例・実践例の収集と公開も行っておりまして、公式ホームページでは、D e R e K oの活用例・実践例がリスト化されています。

ドイツ語研究を主たる目的としていることもありまして、そのリストの大半は学位論文を含むドイツ語研究論文と書籍です。その他の活用例としましては、辞書や授業資料などが挙げられておりまして、他分野での応用例は記載されておりませんし、社会への還元といった事例も確認できません。

しかしながら、冒頭でも触れましたように、D e R e K oは、国民的議論にも発展しましたドイツ語の正書法の改正において、重要な役割を担っています。正書法というのは、日本では余りなじみがありませんが、簡単に申しますと、20世紀初頭から続く、つづりに関する規則です。フランス語はつづりと発音が一致しないということが知られていますが、ドイツ語にもつづりと発音が一致しないケースや、様々なつづり方が見られまして、学校や官公庁では統一的な規則が必要とされるようになりました。

国民的議論に発展し、大きな社会問題となったのが、1996年にウィーンで取り決められました正書法の大幅改正でして、取決めでは、1998年に施行され、2005年までが移行期間とされました。しかし、大幅改正への批判が大きいことから、2004年にドイツ語正書法委員会が設置されます。ドイツ語研究所にその事務局が設置されまして、2006年より、改正された正書法の一部につきまして部分修正が施されます。

ドイツ語正書法委員会では、2004年の設立当初より、つづり方の実態を捉えるために書き言葉コーパスが利用されています。利用されているコーパスは、委員会の構成員である辞書出版社のD u d e nと、現在は廃刊しましたが、W a h r i gの辞書編纂用のコーパスと、先ほど紹介しましたD e R e K o及びa m cを土台としています。委員会構成員の全ての地域をカバーし、ジャンルのバランスを考慮して正書法用に構築されています。

現在の正書法のコーパスは、O K K—正書法中核コーパスと呼ばれまして、総語数は約145億語で、正書法の大幅改正が取り決められます1年前の1995年から現在までのドイツ語圏の書き言葉が収録されています。なお、Krome2020によりますと、コーパスはドイツ語正書法委員会設立以前から用いられており、当初はD u d e n及びW a h r i gの辞書編さん用のコーパスが重要な役割を果たしていたそうです。

こちらの図（スライド35）は、正書法中核コーパスのOKKがカバーします7地域のコーパスサイズを視覚化したもので、広範囲にわたって収録されているのが分かります。

先ほど申しましたように、OKKは1995年からデータを収録しておりまして、大幅改正から部分改正を経て、現在に掛けてのつづり方がどのように変わってきたのかを追跡することができます。こちらのグラフは、Krome2020が、「…にとって難しい」という意味の動詞 *s c h w e r f a l l e n* のつづり方の変化を調べたものです。大幅改正以前、*s c h w e r f a l l e n* は一語書きされておりましたが、当該動詞は形容詞 *s c h w e r* と動詞 *f a l l e n* から構成されますため、1996年の大幅改正では分かち書きをすることが決定されました。しかし、2006年の部分改正で、再び一語書きが認められまして、その後、旧形式の一語書きに戻る様子が見えてきます。

ちなみに、反義語の *l e i c h t f a l l e n*、「…にとって容易である」も同じ特徴を有していますが、*s c h w e r f a l l e n* ほど強い傾向を示さないことも報告されており、どのような傾向を示すかは単語によることを示唆しています。

最後に、オーストリアの活用例として、先ほど紹介しました三つのコーパスを使った事例を見ていきます。a m c も D e R e K o 同様、正書法の改正に利用されています。こちらのグラフは、名詞と動詞から構成される動詞を対象に、一語書きと分かち書きの両方の事例を、Dorn et al. 2020が調査したもので、大幅改正と部分改正後に、先ほどと同様の傾向を示すことが報告されています。

このようにコーパスは、つづり方の変化を観測していく上で、客観的な資料の一つとなり、正書法の改正の一翼を担っています。正書法は国民生活に影響を及ぼす重要な問題ですので、コーパスが社会において大きな貢献をしていると言えます。

先ほどお話ししましたが、a m c は主としてオーストリア全部のジャーナリズム散文によって構成されています。スライド39の上の図は、「学者」という意味のドイツ語の二つのバリエーションを調べたもので、下の図はオーストリアにおけるその分布を色分けして視覚化したものです。「学者」と表現する場合、オーストリアではドイツ同様に学問を意味します *W i s s e n s c h a f t* に *- e r* という接尾辞を付けるのが一般的ですが、東部・南東部では、*- e r* の前に *l* を入れた形式が好んで使用されることを示しています。

また、次の図（スライド40）は、合成語を構成します二つの形態素をつなぐ接続要素の分布を調べたものです。上二つが「ギフトバスケット」、下二つが「包装紙」を意味しまして、左二つが接合要素がない事例で、右二つが接合要素のsが挿入される事例です。G e s c h e n kが「プレゼント」を意味しまして、k o r bが「バスケット」、p a p i e rが「紙」を意味します。下の「包装紙」につきましては、全体的傾向として、接合要素の入らないG e s c h e n k p a p i e rが多いことがうかがえます。ただ、地域により程度差があることも伺えます。上の「ギフトバスケット」につきましては、地域の差がはっきりしていることがうかがえます。このように、オーストリアでは地域差、方言差を調べるためにもコーパスが活用されています。

次に、話し言葉の活用例を見てまいります。こちらの図（スライド41）は、話し言葉コーパスのDiÖをコーパスとして縮小語尾の分布を調査したものです。縮小語尾とは、簡単な例を挙げますと、世界的な食品メーカーのネスレ（Nestlé）のレに相当します。図を見ますと、オーストリアでは地域によって使用される縮小語尾が大きく異なるのがうかがえます。

最後に、方言コーパスのWBÖの活用例を見ます。こちらの図（スライド42）は、英語の従属接続詞i fに相当しますw e n nで導かれる文において、二人称単数主語のときの動詞の活用を調査したものです。大半は主語と一致しますが、南バイエル方言を中心に、主語と動詞の一致が見られない地域があることがうかがえます。ドイツも方言研究での応用がありますが、オーストリアでドイツ語が問題とされる場合と、ドイツでドイツ語が問題とされる場合では、求められるものが異なる部分もあり、ドイツでもスイスでもなく、オーストリアのドイツ語の特徴付けや方言や言語地理学的研究が多いように思われます。

最後にもう一度、本日の内容をまとめますと、ドイツ語の主要コーパスは、大規模なものが多く、商用目的でない限り、特にドイツでは無償利用が基本とされ、公的資金による継続的な支援によって成り立っているということでした。具体的な助成額につきましては公表されておりませんが、オーストリアでは、a m cのように、民間の外部資金によるプロジェクトもありました。

ドイツ語の主要コーパスは、ドイツ語研究や教育への応用を主たる目的としており、ドイツ語研究にコーパスを利用するケースが大半を占めました。ただし、国は

その重要性を認め、長期にわたり公的資金を投じています。

オーストリアでは、先ほど見ましたように、ドイツ以上に地域特性や方言、または、言語地理学に関する研究も多く、コーパスの利用目的がドイツとは異なる部分があることをお話ししました。

ドイツ語コーパスの社会貢献としては、ドイツ語正書法委員会においての貢献が挙げられることをお話ししました。

発表は以上になります。どうもありがとうございました。

○相澤主査

今道先生、ありがとうございました。では、ただ今の御講演につきまして、御質問がありましたら、どうぞよろしくお願いたします。（→ 挙手なし）

では私から。細かいところになるのですが、ドイツ語においてコーパスの主要部分を新聞記事が占めるというお話がありました。日本だと新聞記事を使うに当たっては、それなりの制約が起きてくると思いますので大変印象的だったんです。ドイツあるいはオーストリアにおいては、新聞記事は公共性の高い言語資源として、少なくとも研究用途には最新のものがコーパスの中に取り込まれているという形で理解しても大丈夫でしょうか。

○今道発表者

そうですね。著作権の問題がありますが、公共性の高いデータと考えられているように思われます。ドイツでは、学生割引を利用しますと、比較的安価で新聞を購読することができます。複数の新聞を定期購読している学生も少なくありませんし、若い人たちも日常的に新聞を読んでいるという意味でも、公共性が高い言語資源であると理解されていると思われます。

○相澤主査

ありがとうございます。ほかはいかがでしょうか。

○齊藤（美）委員

オンラインの齊藤ですが、よろしいでしょうか。

○相澤主査

よろしくお願ひいたします。

○齊藤（美）委員

御発表ありがとうございました。私も細かい点ですが、どういった内容なのか、もう少し伺いたいところがありましてお願ひ申し上げます。

DWDS コーパスについて、スライド19で、1)、2)、3)、4)、5)とコーパスの内容を教えていただいたんですけど、そのうちの五つ目の「特殊コーパス」のところにもいろいろ挙げていただいている中で、「映画字幕コーパス」と、あと「チャットコーパス」というのがどういった内容のものか、もしよろしければ伺いたいです。映画字幕というのは、ドイツ語のせりふを文字にしているのか、それとも、せりふだけじゃなくて、例えば「今、雷の音が鳴っている」とか、そういったもの一、視覚障害者の方のための字幕のようなものなのかということについて。

あとチャットと言いますと、これは公共空間にあるものではなくて、個人のものかなとは思ったんですけども、どのようなものなのか、少し伺えたら幸いです。お願ひいたします。

○今道発表者

まず、「映画字幕コーパス」ですが、無料字幕データベースのopensubtitles.orgを利用して構築されています。正確なことは分からないのですが、これまで当該コーパスを利用しました限りでは、「今、雷の音が鳴っている」といったような説明や注は見たことがありませんので、恐らく含まれていないと思います。

次に、「チャットコーパス」についてですが、こちらのコーパスは20年ほど前に構築されたもので、ドイツ語圏初のオンラインコミュニケーションを対象としたコーパスです。コーパスの中身ですが、例えば、ユーザーからポスティングされたチャットデータがありますし、Eラーニングや専門家のインタビューも含まれます。また、サポートなどの実務的な場面でなされたチャットのやり取りや、テレビ番組などで実施されたチャットのやり取りなど、非常に多岐にわたる種類のデータが含まれています。このコーパスは、当時ドルトムント大学に所属されていた研究者らが

構築したコーパスで、後にDWDSコーパスで利用できるようになりました。

DWDSコーパスには、登録なしに利用できるデータと登録が求められるデータがありまして、こちらの「チャットコーパス」は登録なしでは利用できません。

また、DWDSコーパスに統合される過程で、作者などの情報が追加で付与されているのですが、「チャットコーパス」を利用する際は「不明」と出てきます。ですので、個人が特定されることのないよう配慮がなされているものと思われま

○齊藤（美）委員

ありがとうございます。細かいところを伺いました。大変ありがとうございます。

○相澤主査

ありがとうございます。ほかはいかがでしょうか。

○前川委員

よろしいでしょうか。

○相澤主査

よろしく願いいたします。

○前川委員

国立国語研究所の前川と申します。先ほどのお話の、相澤先生のお尋ねの中で、新聞のことがあったのですが、全くこれはたまたまですけど、私は今年の夏にIDSを訪問いたしまして、所長さんとか、ヴィット（Witt）さんというコーパスを作っておられる担当の方と長い時間お話しさせていただいて、そこで、私も全く同じ疑問を抱いております、新聞の著作権処理はどうしているんですかということは伺ったんですね。

そうしたら、私はやっていないのかと思っていたら、やっているんだそうです。それで、大変でしょうという話もしたら、以前は新聞社が非常に協力的で、すぐOKを出してくれたんですけど、だんだんせちがらくなってきて、最近は余り…一代償

を要求されたりとかということだと思いますけども、難しくなってきましたというお話でした。これはたまたまそういう話をしましたので、横から補足させていただければということです。

それに関連して、今道先生にお尋ねしたいんですけども、ほかのコーパスで、特にすごく大きいのがありましたよね。七百何億語とかいうような、旧東のほうで開発されたものだと思いますけど、あれも著作権処理をどうしているかということ、もし御存じでしたら教えていただきたいんですけど。

○今道発表者

先ほど前川委員もおっしゃったように、著作権処理は難しい問題であると思います。最近はこの新聞社も過去の記事を集めて独自にデータベースを構築して、販売しておりますので、なかなか無償で提供することは難しいということだと思います。

ベルリン・ブランデンブルク科学アカデミーは、ドイツ語研究所と比べますと、規模も少し小さくなりますので、その意味でも大変な苦勞をされていると思われる。なかなか表には出てくる内容ではありませんので、実際の著作権処理がどのようになされているのかは、一般の利用者には分からないのが実情です。

ただ、恐らく、ベルリン・ブランデンブルク科学アカデミーもドイツ語研究所と同様、法的手順を踏んでいると思います。ダウンロードできるのは、検索語を含む一文ないし、前後の文に限るなど、利用に当たって、データ提供者との間で何らかの取り決めもしていると思います。その制約がドイツ語研究所とDWDSで異なるという部分もあるのかもしれませんが、公表されていない情報ですので、直接お話をお伺いしない限り、推察するしかない状況です。

○前川委員

どうもありがとうございました。

○相澤主査

ありがとうございます。では、ほかはいかがでしょうか。

○石川副主査

石川です。今道先生、大変貴重な話をありがとうございました。非常に勉強になりました。お話を伺って、私が特に重要であったと思うところを5点ほど簡潔に申し述べたいと思います。

1点目は、ドイツ語のコーパス研究が、ナチの時代のイデオロギーに影響された言語研究から距離を置くためという目的を持っていたということです。これは、私は不勉強にして存じませんで、大変参考になったところです。

2点目は、ドイツでは、正書法など、国語に関する国家的な決定を行う際に、コーパスのデータを使って、エビデンスに基づいた議論をしていこうという姿勢が貫かれているというところです。これも大変参考になりましたし、本邦で行われているローマ字表記の試論にも参考になるお話であったかと思います。

3点目は、そういう分析を可能にするコーパスがしっかり構築・維持されていることです。1995年から現在まで、ある程度同じ基準でデータを取りためているコーパスが多数存在し、そのサイズが100億語を超えていること、さらには地理的な多様性といった要素も組み込まれていることは大変すばらしいと思いましたが、これは日本にも参考になることだと感じた次第です。

4点目に関して、私はコーパスの分析結果を分かりやすく人に見せることも大事だと考えているのですが、K o r A pの様々な工夫、さらにはDWDSの新しい検索システムなどが非常に興味深いと思えました。ボタンを押すとグラフが出てきたり、マップ表示で場所が分かたりする出力もすばらしいと思えました。

最後に、5点目は、そうしたいろいろなコーパスが、国や公的機関の長年にわたる財政支援によって成立しているということです。日本との比較で強く感じましたことは、ドイツの場合はコーパスを支援するためのお金の出どころが非常に多いですね。お話を伺っているだけでも、連邦だったり省だったり、教育省だったり、ときには国防省だったり、国を越えてEUだったり、さらにはデジタルドイツ語辞書学センターや新聞社などの協力もあるとのことで、幅広いセクターがコーパスの価値を理解し、その構築・維持に資金を投入しておられることが、質の高いコーパスが長期にわたって作り続けられていることの背景にあるのだと強く感じた次第です。価値あるお話をありがとうございました。

最後に、短い質問を一ついたしますが、ドイツにいらした御経験を踏まえ、実際

の現場で、コーパスのようなものに対して、研究者の方々がどのような意識、態度を持っておられたか、よろしければ御教示いただけませんかでしょうか。

○今道発表者

マンハイム大学には留学したことがあります、ドイツ語研究所での滞在歴はありません。ただ、これまで何度か研究所の担当者には話しを伺う機会がありましたが、こちらのコーパスは次から次へとデータを追加更新していくいわゆる「モニターコーパス」ですので、著作権処理の問題もそうですし、新たなデータ提供先の開拓も、重要な問題であるようです。また、話を伺いまして、研究所のコーパスを様々なところで活用してもらうことにも注力されているという印象を持っていたのですが、コーパスを用いたドイツ語分析の具体例やデータベースなどが以前にも増して公開されるようになったと思います。また、このところ分析結果を視覚化する試みをされておられるのですが、これも、近年のコーパス言語学の動向や、利用者の利便性なども重視されていることが背景にあるように思います。

○石川副主査

どうもありがとうございます。

○相澤主査

ありがとうございます。ほかはいかがでしょうか。（→ 挙手なし）

私からもう1点だけ。細かな質問となりますが、29ページにWBÖコーパスのプロジェクト期間を御紹介いただいています、2038年と書いてあったのは非常に驚きました。これは本当に、22年間もサポートを受けるプロジェクトになっているのでしょうか。

○今道発表者

そうですね。私も驚きましたが、期間には間違いがないようです。

○石川副主査

助成期間が3年や5年というのではなく、20年間余りというのは、コーパス構築

のような時間の掛かるプロジェクトから見ると、ある種の理想の形かもしれないですね。

○相澤主査

そうですね。ありがとうございます。よろしいでしょうか。

○村瀬国語課長

今道先生、どうも御発表ありがとうございました。

社会背景として、ドイツの場合は、東西ドイツが統一されたのが確か1990年頃だったと存じますけれども、拝見しますと、1990年以降のプロジェクトといったようなものが目立つようにお見受けしました。

また、先ほど地理的な言語分布の在り方についても御発表の中にあっただかと存じます。

この点、ドイツの場合は、歴史的には地方分権といったものを戦後推進してきた部分もあろうかと存じますが、こうした東西ドイツの統一、あるいは地方分権、そういう社会制度や社会背景を受けての言語コーパスの構築といったものへの影響につきまして、もし関係がありましたら、御教示賜りたいと存じます。

○今道発表者

ありがとうございます。東西統一や地方分権といった社会背景のコーパス構築への影響はないように思います。ただ、御指摘のように、古くから地方分権という歴史的背景がありますし、ドイツ語は方言学的に北と南に二分され、さらに様々な方言がありまして、方言研究への応用ということで、各地域のドイツ語を集めたコーパスも構築されています。また、東と西に分かれていたということもありまして、東西のドイツ語を集めたコーパスや、東西統一前後のドイツ語を集めた転換期のドイツ語コーパスなどもDWD Sのアーカイブに入っています。その他、ドイツ、オーストリア、スイス、イタリアのドイツ語圏である南チロル地方のドイツ語を集めた、ドイツ語圏のドイツ語コーパスプロジェクトもありました。ただ、国を越えた国際プロジェクトということもあるせいでしょうか、当初の計画どおりに進まず、公開はされていますが、未完成で終了しました。そういう意味では、地域性は決し

て軽視されているわけではないように思います。

○村瀬国語課長

どうもありがとうございました。大変参考になりました。

○相澤主査

ありがとうございます。よろしいでしょうか。

では、改めまして、今道先生、大変貴重な御講演を頂き、ありがとうございました。

○今道発表者

ありがとうございました。

○相澤主査

続きまして、村瀬課長に小委員会のまとめに向けた議論の枠組みとして、配布資料3と参考資料1についての御説明をお願いいたします。石川副主査の参考資料2、3もございますね。

○村瀬国語課長

配布資料3を御用意いただけますでしょうか。また、相澤主査からもお話がございましたが、参考資料としまして、参考資料1が検討の観点をまとめたもの、参考資料2と3が、石川副主査による先日のプレゼンテーションの資料でございます。

それでは、配布資料3でございますけども、これに沿って話をしてまいりたいと存じます。この資料は、先日の国語分科会で、これまでの議論の整理として、中間的におまとめいただいたものでございます。振り返りとして主な点について申し上げたいと存じます。

まず、「1 基本的な考え方」のところでございますが、ここは言語の意義について触れているところでございます。特に上から三つ目でございますが、こちらのところで、先ほども、御発表の中でもございましたが、言語資源は、言語研究、言語政策に有用であるということ、それから、それだけでなく、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るということ、そしてまた、その下でございますが、この言語資源というのを「未来へ伝える文化遺産」と位置付けているというところがまずあるわけでございます。

続いて、次のページの「2 コーパスの意義・現況」というところになりますが、こちらの三つ目の○でございます。こちらで国語研によるコーパス事業、すなわち「現代日本語書き言葉均衡コーパス」について言及がございます。

そしてまた、その次のページでございます。活用事例ということで、(2)のところでございますけれども、順に辞書編集、教育面、そして情報学等における取組についても触れられているところでございます。

続いて、「3 今後に向けたコーパスの整備・運用」についてですが、このうち(1)において、社会の変化・進展がもたらす影響について記載がございます。こちらの3ページの末尾のところ、主だったものとして、例えば、サイバー化であったり、生成AIによる影響であったりなどがあるとされる中、次のページに移りますけれども、三つ目と四つ目の○になろうかと存じますが、言語データにはバイアスがあるということを前提とした上で、四つ目の○のところでございますが、大規模言語モデルの訓練に係る課題ということで、今後学習データが不足すると見込まれる中、生成AI自体が作成したテキスト、あるいは人間と生成AIがやり取りしたテキストなどの活用をどう考えていくのかという、そういった課題意識に裏打ちされたお話もあったところございました。

そして続いて(2)でございますが、ここで、コーパスの整備・運用について、4ページから5ページにわたっておまとめくださっています。(2)のリード文のところに、開発後の取組の継続、普及、社会実装について言及がございますが、これら3点が重要であるという御認識を示してくださいました。例えば、普及につきましては、5ページになろうかと存じますが、この上段の方におきまして、一般ユーザーの利用を意識すること。その上で、利便性を考慮した検索の在り方やインターフェース、こういうことにつきまして、普及に関しては考慮していくということ。続いてその下、下段でございますが、社会実装に関する記載部分におきましては、具体的に挙げてくださったわけでございますが、言語教材、あるいは医療面での活用可能性、福祉面もそうでございますけれども、そういった分野での活用可能性について御紹介いただいております。

結びに「4 将来社会を見据えた考慮事項」ということで、これは6ページのところになりますけれども、こちらの一つ目の○でございますが、コーパスが我が国の社会課題の解決に資する役割があると期待されるということも踏まえて、国民共有の資産として構築していくということについて記載がございます。そして二つ目の○、それから四つ目の○にわたってになりますけれども、先ほどもございましたが、二つ目の○のところでは主要な研究機関への継続的な支援ということ、そして四つ目の○でございますが、多文化・多言語社会の到来といったものを見据えて、そして中期的な研究課題として日本語母語話者でない方々が話す日本語コーパスの整備であったり、あるいは先ほどもありました生成AIによる日本語への影響であったりといったものについての評価の在り方といったことも御審議いただいたところでございました。

このレポートの概略はただ今申し上げたとおりでございますけれども、この取りまとめ、御議論に当たりましては、参考資料1に、検討の観点ということでお示ししております。こちらは参考資料1を御覧いただきますと、その下段のところには3点ほどこの観点を記載していたわけでございますけれども、それぞれ改めて申し上げていきたいと存じます。

一つ目が言語の果たす機能をどう考えるのかと。その上で言語資源をいかに位置付けて保存・活用の在り方をどう考えていくのか。そして実際の話として、2番目として、現在におけるコーパスに関する整備状況はどうであったかという確認を行った上で、将来に向けた話としまして、今後における望ましい言語資源の在り方についてどう考えていくのか。とりわけ社会課題の解決といったことを意識すると、どういったことがあり得るのかということにつきまして御検討を賜ってきたわけでございますが、今一度、この議論の取りまとめの充実に向けまして、御議論を深めて頂ければと存じます。

雑駁^{ばく}でございましたけれども、以上でございます。よろしくお願いたします。

○相澤主査

村瀬課長、ありがとうございます。では、ただ今の御説明につきまして御質問等がありましたら、よろしくお願いたします。よろしいでしょうか。

○石川副主査

石川です。質問というよりも、大変手際よく議論をおまとめいただいております。事務局の御尽力に感謝申し上げたいと思います。この問題は非常に多角的なところがございしますが、そういったところにも目配りの効いたまとめになっているなと拝読をして感じた次第です。ありがとうございました。

○相澤主査

ありがとうございます。では、これまで、本日で5回目の議論の場となりますが、これまでの議論を踏まえまして、また、前回と本日お話のありました海外のコーパスの状況や、コーパスを活用した社会課題の事例、ただ今の御説明等、お手元の配布資料3、あるいは参考資料1、社会課題の解決の実例につきましては参考資料2、3等も御参照いただきまして、今後、小委員会の議論のまとめを充実していくに当たっての意見交換を行いたいと思います。

皆様の御専門分野等も踏まえまして、御意見賜りたく存じますので、御自由に御発言を頂ければと思います。今道先生もよろしく願いいたします。

では、まず石川副主査から、この参考資料2、3に合わせていただけてということにつきまして、まとめにおいて検討すべき要素というものを挙げていただければと存じます。

○石川副主査

ありがとうございます。小委員会で2回にわたって報告の機会をお与えいただきましたことを、まず感謝申し上げます。その中でも一貫して申し上げてきたわけですが、1990年代から2000年代、2010年代辺りというのは、「コーパスをまずは作ってみる」という時代でございましたが、今は、「既に一通りのものが出来上がった、その後」という新しい段階に来ているのだらうと思われまます。その中で、新しい見取り、パースペクティブ (perspective) のようなものが言語資源の研究活用にも求められてくるわけです。それを私は「継続」と「普及」と「社会実装」というキーワードで御説明したところであります。これが全てではないと思いますけれども、こうした観点をより強調していきながら、コーパスを含む言語資源の価値とその有用性を社会に広く伝えていく、それによって、国全体で言語資源の研究を後押しして

いただく機運を醸成し、関連の研究者が長いスパンで言語資源の開発・管理・維持に当たれる体制を作り上げていくことができると考えておるところです。

○相澤主査

ありがとうございます。そうしましたら続きまして、これまで事例等の紹介をお願いすることのできなかつた委員の皆様、これまでのヒアリングや意見交換も踏まえまして、まとめに向けまして、こうした要素を加えたらどうか、または、ただ今石川副主査からお話のありました継続、普及、社会実装、この三つの観点についてのアイデアですとか、そういったことがございましたら、お考えを伺えればと存じます。

こちらは、あいうえお順となってしまうと恐縮ですけれども、植木委員からということで、どうぞよろしく願いいたします。

○植木委員

ありがとうございます。もう既におまとめいただいたところで重複してしまうのですが、私が個人的に重要だと感じている三つの観点から、本当に簡単な感想のようなものを、申し上げたいと思います。

一つは、書き言葉と話し言葉という点です。私は古典を勉強しているので、当然ながら書かれた言葉を収集している歴史コーパスを非常に有り難いものとして重要なデータとして活用させていただいているんですけども、一方で、私事で恐縮ですが、私自身は、メロディーに乗せて歌われた歌謡を対象としているところもあって、非常に口語表現が多かったり、貴族階級が記した文献には出てこないような俗語も多く現れ、用例がないので解釈を確定するのが難しいということに直面する場合が多く、そういう経験から、話し言葉の収集というのも非常に重要と感じています。石川副主査もおっしゃっていましたが、ずっと継続していけば、今の現代語も将来には古語になっていくわけなので、今の口語が将来分からなくなってしまうということがないように、今ならできる口語の蓄積というのも重要かと考えます。

2点目は、既に申し上げたのですが、ダイバーシティ推進の観点からです。石川副主査、川原先生に、医療や子育ての際の有効な声掛けの話をお聞きしてきましたが、合理的配慮が義務化される中で、医療従事者や保育者だけでなく、社会の

広い範囲で一特に大学などの教育現場でも、発達に特性のある人たちに対して分かりやすい話し方というのが求められているので、そのようなことにも役立つという説明は社会に訴求する力があるんじゃないかと思いました。

また、社会の構成員として日本語を母語としない人々も増えてきて、「やさしい日本語」というのが注目されており、「至急避難してください」ではなく、「急いで逃げてください」というような、自治体でも災害の際の言い換えのマニュアルなども作っていると思うのですけれども、そういう表現の蓄積も社会的に意義があるのではないかと思いました。

3点目は、コーパスを作っていく上での様々な連携のことです。私事で恐縮ながら、私は現在ある通信社の記事について検証する委員会の委員を務めているのですが、通信社や新聞社の記者も、言葉、特に時事的な言葉には強い関心を持っておられて、社内的に辞書を作っておられたりします。連携することで、コーパスを作る際に有用な知見を得られるかもしれないし、虫がよすぎるかもしれないのですが、資金面でも協力していただければ、と。今日の今道先生のお話では、オーストリアでは、民間の通信社から資金が出ているということでしたので、そういうことが可能になるといいなと感じました。

以上です。

○相澤主査

どうもありがとうございます。大変含蓄のある御意見、ありがとうございます。では、続きまして齊藤（美）委員、よろしく願いいたします。

○齊藤（美）委員

ありがとうございます。先ほどの村瀬国語課長からの御説明の参考資料1の「主な審議事項及び当面の予定」の最後の部分にある「社会課題の解決等に資するような専門的分野特化型コーパス等の在り方」というところで主に考えてみました。

私の専門である翻訳学の立場から、コーパス活用の可能性について思い付いたことです。また、望まれるインターフェースの機能—先ほどもお話がありました、それについても—これはもう既に多く研究されていると思いますが—言語教育に関わる観点から申し上げたいと思います。

7月の前川委員の御発表で、構築が望まれるコーパスの中に、在日・渡日外国人会話を挙げられていたことと、それから石川副主査の8月の御発表において、ヘルスコミュニケーション、例えば医療者と患者との間のコミュニケーションのためのコーパスの活用方法を御紹介いただいたことも思い出しながら考えました。

本日の配布資料3の方に、これからの日本語に非母語話者による日本語産出の影響が考えられること、それから今後、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についての言及がありまして、先ほども御説明があったところですが、この状況においては、翻訳・通訳、そして言語教育がこれまでも増して必要な場面、そして必要な人が増えてくることが考えられます。

そのような中、翻訳・通訳の研究者の間で、ユーザー教育、つまり翻訳・通訳サービスを利用する側ー翻訳される、あるいは通訳される言葉を発する側に翻訳・通訳への理解を深めてもらう機会の必要性が認識されて、議論されています。誰もが同じ言語を知っているという社会でなければ、特定の職業や立場にある人だけでなく、一般の人の書いたものや言ったことが翻訳・通訳される機会が増えることが考えられるからです。そこで、翻訳あるいは通訳しやすい書き方、話し方の検討のためにコーパスが活用できるのではないかと考えました。翻訳者や通訳者がどのように訳すかを検討するためのコーパスではなく、自分の書いた、あるいは話した言葉が翻訳・通訳される人が、どのように言葉を発すればよいかを検討する上で役立つコーパスを想定してお話ししています。

文学作品や映画など、訳しやすさが求められないジャンルもありますが、法廷を含む司法、それから医療、教育、役所、地域コミュニティーにおける翻訳・通訳においては、訳される前の文章や発話を翻訳・通訳しやすいように書き、話すことが重要だと考えられています。先ほど植木委員のお話にもありました「やさしい日本語」という種類の日本語がありますが、それとは異なり、難易度の低い語彙を用いることは翻訳・通訳ユーザーの場合、必ずしも必要ではありません。例えば法廷での通訳場面であれば、法律に関わる専門用語が使用されることもありますし、それは訳出先の言語でも対応する語が通常あるため、そのような難易度の高い用語を使用することに問題はありません。

では、何が翻訳・通訳しづらいかというと、例えば多義的であって解釈が複数できる文章、発話、あるいは文化的知識がないと理解できない、理解しづらい発話、

慣用句などです。プロの翻訳者や通訳者であれば、扱う二言語の言語的な知識のほか、文化的知識も持っていますので、理解はできますが、訳す際に困難が生じる場合があります。これまでの研究で、翻訳・通訳ユーザー側、例えば法律家の方などは、そのまま訳してください、何も足さず何も引かないでくださいと翻訳者や通訳者に頼むことがよくあるとされているのですが、そのまま訳すことが不可能な場合が多々あるわけです。つまり、そのまま訳したら—説明を追加せずに訳したら理解されない場合があるということです。こういった齟齬^{そご}を防ぐために、どのように書けば、あるいは話せば、そのまま訳すことができるのかを、コーパス上の事例を基に検討し、その結果として訳しやすい書き方、話し方を提示する指南書であるとか、動画などを作成することが行えるのではないかと考えました。

そのためのコーパスが理想的には、翻訳・通訳を介したコミュニケーションのデータのコーパス、つまり二言語、あるいは複数言語対応であると訳しづらい要素について検討しやすいため、よいと思います。ただ、必ずしも二言語対応コーパスでなくても検討は十分に可能で、日本語だけであっても、翻訳・通訳が必要な場面のデータがあれば、翻訳・通訳の専門家及びその分野の専門家、例えば法律家であるとか医療者、自治体職員、教員などが協働して分析できるのではないかと思います。

法廷など、録音・録画ができない場面では、オーセンティック (authentic) な本物のデータが入手できないことも多々あると思います。そのような場合は、例えば模擬法廷でのロールプレイであるとか、ドラマなどのフィクションをデータの収集対象にすることもできるのではないかと考えました。以上が翻訳・通訳に関わるアイデアです。

もう1点、言語教育関連についても申し上げます。本日の配布資料3の方です。これまでの議論の整理にあるユーザーインターフェースのところ、特定の語や語句を入力することで得られる情報として、用法や用例も挙げられていますが、その点が充実すると、異文化間の誤解や摩擦を減らすことにつながるのではないかと思います。

コミュニケーションは相互行為—インターアクションであることが多いので、2者間のやり取り、会話のデータがあるとよいと思います。例えば謝罪、謝り方に関するやり取りはどのように行われているのかが学べるようなコーパスです。謝罪

の言葉・表現として、「ごめんなさい」や「申し訳ない」などがあることは、辞書や日本語学習のための教科書などで十分に確認できるのですが、コミュニケーションとしての謝罪の表現はどのように行われているのかというのを、「ごめんなさい」や「申し訳ない」の前後にあるやり取りを見ることで学べる素材がたくさんあると、日本語運用能力を高めやすいのではないかと思います。

例えば、謝罪のやり取りはどのように始まるか。「ごめんなさい」、「申し訳ございません」から始まるのか、あるいは謝罪の対象となる出来事が生じた理由の説明から始まるのかなど。それから、それ以降の2言目以降、何と言うか。相手に負荷を掛けたことを認識している「大変なお手間を取らせてしまいました」などと伝えるのか、あるいは、「もう二度と同じことはしない」と宣言するのか、「以後気を付けます」などと宣言するのかとか。

そして、そのやり取りがどのように終わるかです。謝罪を受ける側が、「もう気にしないでください」などと言って、そして話題を変えたら、それ以降は謝らなくていいなど、「ごめんなさい」、「申し訳ない」と言うだけではなく、このような一連の流れの中で謝罪という行為は実際のところ成立しています。

この謝罪なり、ほかにも感謝であるとか提案であるとか、いろいろな行為が成立するためのやり取りを、コーパスを検索して複数見ることができると、日本語以外の言語が、第一言語の人たちが日本語によるコミュニケーションを学ぶ上で役立つ可能性があると思いますし、それに加え、日本語が第一言語であっても、何らかの理由でコミュニケーションに難しさを覚えている人たち、先ほどの植木委員の2点目にもありましたが、そういった方たちが自身のコミュニケーション方法について検討する上で役立つ可能性があると思います。

以上です。

○相澤主査

ありがとうございました。翻訳・通訳の立場、あるいは言語教育の立場からということで、非常に重要な御指摘を頂いたと思います。

では、続きまして武田委員、よろしくお願いいたします。

○武田委員

本小委員会におきまして、日本・日本語におけるコーパス開発、構築、そしてその現状ということを踏まえ、また、ヒアリングにおいて諸外国・諸言語におけるコーパスの開発、構築、社会実装、そういった事例を拝見いたしまして、改めて日本・日本語においてもコーパスの継続的な整備、今後に向けての新たな整備が必要であるということがよく分かってまいりました。

それを踏まえまして、今後、普及して社会実装ということになっていくわけですが、その普及において、書籍出版においては国語施策を広く国民に普及させるということが一つの責務であろうと考えております。なかでも、コーパスに最も関係するのは、辞書を中心とした書籍ということになろうかと思えます。

事例研究でも御報告を頂戴したとおり、従来、辞書編纂においては一定程度コーパス利用が進んできております。けれども、まだまだその利用の余地というのはあるかと思われまじし、また、継続的に開発されるコーパスを、継続的に活用していく必要もあろうかと思われまじし。そのことによって、間接的にはありますけれども、広くそのコーパスの意義を普及させることにつながるであろうと一そこからまた、直接一般のユーザーが使う道も開かれ、社会実装に広げていくことにもつながっていくだろうとも考えております。このまとめにおきましても、そういったところまで、若干でも触れられればよろしいかと考えているところでございます。

それから、これはコーパスの内容に関わるところで、著作権の問題があったかと思えます。著作権法の整備によって、そのこのところも従来より円滑に進むようになる方向であるということは御説明を伺いましたが、くれぐれも著作権に関しては権利侵害が起きることのないように、従来どおり慎重に進めていただければと、その必要があろうかと考えております。

簡単ですけれども、以上でございます。

○相澤主査

ありがとうございます。社会実装において辞書とコーパスというものが相互に連携すべきものであるということで、まとめにおいても言及という重要な御指摘を頂きました。また著作権についても、慎重な検討が必要であるというのも記載が必要な御指摘の検討できる項目であると感じます。

それでは、さらに自由な意見交換ということで、御意見のある委員の皆様がいら

っしやいましたら、御発言よろしくお願ひいたします。

辞書ということで、神永委員。

○神永委員

武田委員のお話と重なるのかもしれませんが、いろいろと各国のコーパスの状況なども伺いながら、私は辞書編集者としての、今コーパスをこのように使っているんだということを発表させていただきました。改めまして、コーパスの活用ということを考えてみたんですが、どの辞典でもそうですけれども、辞典は用例が命です。初めに用例ありきだと思うんです。

何でそんなに用例にこだわるのかと言いますと、用例によって分かることがあるわけですし。私の以前の発表でも報告させていただきましたけれども、その語が確かに使われていたという証拠にまずなります。それから、その語の意味とか用法の手掛かりが得られて、理解を助けることができるということになると思うんです。

あと、語の読みや表記の資料にもなります。

辞書に関わる者としては、そうした情報をできるだけたくさん集めたいという願望がありまして、もちろん辞書を出版している各社はそうした独自のコーパスを持ってはいるんですけれども、それを拡充させていきたいと考えているわけです。それによって、日本語の全体像がより深く理解できるわけで、辞書の記述なんかも更に充実したものになっていくだろうと思うんです。

コーパスというのは辞書にとっては当然のことながら表には出ない資料ですけれども、それを辞書編集の際にどのように活用しているのかということを発信するのも必要じゃないかなといろいろと考えまして、辞書編集者として、コーパスの活用を伝えていくことが大切だと思ひまして、例えばですけれども、裏データとしてこういうコーパスーコーパスではこのようになっていたんだよということを示せるといいのかなと。今、辞書はだんだんデジタル化されていますから、それによって可能になるんじゃないかと思うんです。紙の辞書ですとなかなかそれは難しい。実際にやったこともあるんですけれども、そこで終わってしまうような気がするんで、デジタル化で辞書を示すことによって、そういった裏データとしてあるコーパスの実際がこうだよということが示せると、更に日本語に対する理解度が上がるんじゃないかなと考えています。

いろいろとお話を伺いながら、辞書編集者のある意味妄想かもしれませんが、そんなことを考えました。

○相澤主査

ありがとうございました。では前川委員、よろしく願いいたします。何か全体を通して御発言等がありましたらよろしく願いします。

○前川委員

余りもうないんですけども、ずっと関わってきた一実際に作ってきた立場からいたしますと、いろいろな問題というのは非常に身近に感じるわけです。その大きなものとしては権利処理の問題が非常に大きいわけで。もし今後この委員会の答申とかそういったものを通じて、国として、国の施策として言語資源というものをきちんとやっていこうというような形になるんだとしたら、そこも含めて対策を最初から考えていかないといけないのかなということは強く感じます。

それから、これも余計な話かもしれないですけど、これまで作ってきた、そして今も作りつつあるので、その内情をちょっと御紹介いたしますと、一言で言うと自転車操業に近い状況です。元々外部資金で始めたということは以前にお話ししたと思います。今年度から、国語課さんからの資金を導入することができまして、資金の面では助かっているんですけど、人間の面で、それを担ってきた人間が、ここ20年間全然うちの中では変わってなくて、知識の継承ができない。だんだん高齢化が進んで。要するにこれは言語資源の問題ではなくて、国立の大学を含めた研究機関全体の問題ですけども、要するにお金がないから、若い人を育てることができないという深刻な問題が起きています。言語資源のようなのはノウハウを継承していかないとうまく長期的な取組というのはできないですよ。ですから、少しこの小委員会の範囲を超えるかもしれませんが、何か答申を最終的にまとめるようなときには、そういったところにも配慮をしていただけると、長年現場にいた人間としては非常にうれしいから、多分どこがそういう、こういう仕事をやっていくにしても完全に仕事を自動化するということは、今後とも当面の間は不可能だろうと思いますので、人間に対する配慮、安定的に雇用できるような環境というか、そういったものが現場では非常に求められているということを申し上げておきたい

と思います。

以上です。

○相澤主査

ありがとうございます。権利処理と人材の関係について、是非このメッセージを発信すべきという御意見だと思います。

ほかはいかがでしょうか、今道先生、追加でもし何かコメントがございましたら。

○今道発表者

様々な委員の先生方のお話を拝聴いたしまして、単にコーパスを使うというだけでなく様々な視点があるということ、今回いろいろ勉強させていただきました。どうもありがとうございます。

○相澤主査

ありがとうございます。ほかはいかがでしょうか。

私自身も非常に、皆様のお話をお伺いして勉強になるところがありました。特に本日、何人かの委員の方から御指摘が出ましたとおり、社会実装という観点を議論の俎上に挙げることは大切であると実感をいたしておりまして、例えば翻訳ですか、コミュニケーションといったところも、説得力のある実例として挙げていただいたと感じています。

逆にそういう重要な社会実装から出発して、ではそのために何が必要かということで、権利処理について真剣に向き合いましょうとか、人材育成に向けてはどうしたらよいかといった議論も深めていくという流れが、このまとめでメッセージとして伝わるとよいのではないかと感じました。

ほかはいかがでしょうか。石川副主査、よろしいですか。

○石川副主査

ありがとうございます。前回の小委員会が閉じた後の雑談で、コーパスの社会実装に関して、例えば、異なる分野の研究者が集まって何かコーパスを使った異分野連携研究のようなことができないものかという話になりました。どうやったらこう

したことが実現するのかこれからも考えていきたいですね。やはり一番大事なのは最初の一步だと思います。言語資源の研究者と、これまで余り接点がなかったような分野の研究者や、ビジネス分野の方々が、何らかの形で出会い、互いに触発し合って、正に「価値創発」と言いましょうか、これまでなかった、新しい社会問題解決の視点が出てきたというようなことから始まり、小さな成功体験が幾つか積み重なってくると、これは大きな可能性へとつながっていきます。また、その成果に社会の目が向くという好循環も期待されます。こうした仕掛け作りの方法を、また、この小委員会などを舞台にして考えていくことが必要かもしれないなと感じております。ありがとうございました。

○相澤主査

どうもありがとうございます。そうしましたら、ほかは御意見いかがでしょうか。よろしいでしょうか。

では、以上で意見交換を終了いたしまして、まとめに入りたいと思います。前回のフランス語、英語、今回のドイツ語のコーパスについて、その活用も含めて把握できまして、日本語のコーパスについても、社会課題解決の一助とするという観点や、一般の理解を得るための普及の工夫など、求められることを、小委員会として整理して発信していく必要があるということで、本日のまとめとさせていただければと思います。

今後については、この小委員会につきましては、議論のまとめの仕上げに向けまして、事務局に案を準備していくということになります。もし何かありましたら、追加で御発言がありましたら、事務局へお知らせ願います。

では、本日の言語資源小委員会はこれにて閉会といたします。どうもありがとうございました。