第1回 国語分科会言語資源小委員会（Web開催）議事録

令和 6 年 7 月 11 日
14時58分～16時55分
文部科学省 5階 5F5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）

（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 文化審議会国語分科会言語資源小委員会委員名簿
- 2 文化審議会国語分科会言語資源小委員会の会議の公開について（案）
- 3 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）
- 4 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
(文化審議会国語分科会 令和5年9月29日)
- 5 日本語の言語資源 特に言語コーパスについて（前川委員御提出資料）

〔参考資料〕

- 1 国立国語研究所現代日本語書き言葉均衡コーパス（BCCWJ）の概要
- 2 言語資源に係る関係記述（抜粋）
- 3 文化審議会国語分科会運営規則

〔経過概要〕

- 1 第1回の言語資源小委員会の開催に当たり、村瀬国語課長から挨拶があった。
- 2 事務局から配布資料の確認が行われた。
- 3 事務局から配布資料1に基づいて言語資源小委員会に所属することになった委員の紹介があった後、定足数の確認があり、参考資料3「文化審議会国語分科会運営規則」に基づき、委員の互選により、相澤委員が言語資源小委員会主査に

選出され、主査就任に当たっての挨拶があった。

- 4 参考資料3「文化審議会国語分科会運営規則」に基づき、相澤主査により、石川委員が副主査に指名され、副主査就任に当たっての挨拶があった。
- 5 事務局から配布資料2「文化審議会国語分科会言語資源小委員会の会議の公開について（案）」の説明があり、言語資源小委員会の会議を公開することが了承され、以降の議事が公開された。
- 6 事務局から配布資料3「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）」について説明があり、説明に対する質問及び意見を尋ねたが、特に発言はなかった。
- 7 事務局から配布資料4「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（（文化審議会国語分科会 令和5年9月29日）」について説明があり、説明に対する質疑応答が行われた。
- 8 前川委員から配布資料5「日本語の言語資源 特に言語コーパスについて」について説明があり、説明に対する質疑応答の後、意見交換が行われた。
- 9 次回の言語資源小委員会について、令和6年8月1日（木）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。
- 10 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。

○鈴木国語調査官

定刻にはちょっと早いですが、皆様お集まりでございますので、ただ今から第1回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。本日は今期最初の小委員会でございますので、後ほど主査を選出いただくまでの間、事務局で進行させていただきます。御了承をお願いいたします。

なお、会議は実際にお集まりいただきながらも、オンラインでの参加でもできるような形で今後も継続する予定でございます。

それでは初めに、国語課長の村瀬より御挨拶申し上げます。

○村瀬国語課長

まず初めに、委員の皆様におかれましては、国語施策の御審議に御協力を賜って

いることにつきまして、心より御礼申し上げたいと存じます。そしてまた、本日は御多忙の中、御出席を賜りまして、誠にありがとうございます。

本小委員会では、今年度より新たに言語資源の在り方をテーマに御審議いただくわけですが、いよいよ本日から本格的な検討のキックオフになります。言語資源をめぐるまは、言語の意義を振り返りながら、また近時は、生成A Iであったり、あるいは大規模言語モデルとの関係であったり、こうしたことも踏まえながら、望ましい在り方について御審議を深めていただきたいと考えているところでございます。正に温故知新と日進月歩が相半ばするテーマであろうかと存じます。どうぞ委員の皆様から、御指導、御知見を賜りながら進めてまいりたいと存じますので、よろしくお願い申し上げます。

以上となります。

【委員の互選により相澤委員を主査に選出】

○鈴木国語調査官

それでは、相澤主査から御挨拶を頂きたいと思います。よろしくお願いいたします。

○相澤主査

国立情報学研究所の相澤でございます。この度は、主査という大役を仰せつかりました。よろしくお願い申し上げます。

ただ今御紹介いただきましたとおり、私の専門は自然言語処理や情報検索という分野でありまして、皆様とはちょっと異なるかもしれないんですけども、計算機を通して人間の言語を理解していくという観点から、これまで研究をしてまいりました。

近年、生成A Iというものが出てきた後の自然言語処理の研究分野における変化は非常な勢いでありまして。とはいえ、研究分野の変化は、大きな社会全体の流れの中で言えばごくごく一部にとどまっています。より大きく社会の人間の生活の至るところで生成A Iというものが人間の言語活動に影響を及ぼしていると日頃感じております。コーパスという形で、生成A Iは人間の言語を学ぶわけですけども、

そのコーパスの中に埋め込まれた知識ですとか規範、倫理といったものを獲得していく、そういう過程がある一方で、生成モデルの中にあるバイアスとかが、逆に人間に影響を及ぼしていくという、そういう双方向的な関係にあると思っています。そのような観点から、いかに言語が変化していくかということを観察、捉えていくことが今後ますます重要になってくるであろうと考えております。

これから本小委員会において、現在の日本語コーパスの状況を踏まえながら、望ましいデジタル言語資源の在り方について審議を重ねていくと伺っております。私も一研究者として長年の間、コーパスには大変お世話になってきたこともありまして、今後に向けて微力ながら貢献させていただけると有り難いと考えております。

言語政策や言語研究、A I 等との関係なども視野に入れながら、その在り方について、皆様の御意見や御知見にお頼りしながら一緒に考えていければと存じますので、どうぞよろしく願いいたします。

【相澤主査の指名により石川委員が副主査に】

○鈴木国語調査官

それでは、石川副主査からも御挨拶を賜りたいと思います。よろしくお願いいたします。

○石川副主査

皆さん、こんにちは。神戸大学の石川と申します。副主査として、主査をサポートして、皆さんと一緒に議論を深めてまいりたいと存じております。私が考えておりますことを3点だけ申し上げます。

一つ目としては、コーパスの魅力を広く発信していく、そういう機会にできればよろしいのではないかと存じます。コーパスというのは、言語学者であるとか、あるいは自然言語処理の研究者にとっては非常になじみの深いものなのですが、一般の方から見て、コーパスがそれほど身近かという、恐らくそういう状態にはなっておりません。このギャップを埋めることも、恐らくは今回の会議の一つのミッションであろうと考えています。

二つ目は、コーパスの応用を考えていくということです。コーパスは、一義的に

は、言語学の研究、共時的あるいは通時的な言語の研究に大変有用な研究資料であるわけですが、同時に、コーパスは、より大きな社会変革ですとか、社会のイノベーションにもしっかり貢献していける素地を持つものと考えています。こうした可能性についてもしっかり見ていくことが必要であろうかと思えます。

最後に3点目として、コーパスにはいろいろな側面があるのですが、その一つは、ある時点の言語を記録した「言語遺産」のようなものだということでもあります。であるならば、どのような形で言語を保存していくのが望ましいのか、あるいは、これまでに保存されたものをどう保持し、次の世代に受け渡していくのがよいのか、いろいろ考えないといけないことは多くあるかと存じます。そうしたことを委員の皆さんと議論しながら、良い方向を見付けていくことができると考えております。

以上です。どうぞよろしくお願い申し上げます。

○鈴木国語調査官

ここからの議事につきましては相澤主査にお願い申し上げます。よろしくお願いいたします。

○相澤主査

それでは、お手元の配布資料2「文化審議会国語分科会言語資源小委員会の会議の公開について（案）」を御覧ください。こちらは6月3日に開催されました国語分科会で御了解を頂いた「文化審議会国語分科会の会議の公開について」を踏まえました案で、言語資源小委員会の会議について公開の在り方を定めようとするものでございます。事務局から資料について御説明をお願いいたします。

○鈴木国語調査官

配布資料2「文化審議会国語分科会言語資源小委員会の会議の公開について（案）」、こちらの御説明をいたします。

これは、今お話のあったとおり、去る6月3日の国語分科会で御確認いただきました「文化審議会国語分科会の会議の公開について」、そちらに倣ったものでございます。この言語資源小委員会の会議についても公開の在り方を定めようと、そうした内容でございます。

まず、「会議の公開」について。特別な場合—例えば主査を決めるなど人事に関することが例として挙げられるかと思いますが—そうした場合を除きまして、常に公開することになります。

続いて「会議の傍聴」について。事前に申し込み、登録された傍聴者が会議を傍聴することができます。コロナ禍前は会議室に傍聴席を設けておりましたが、現在はオンライン参加の委員もいらっしゃいますので、傍聴についても全てオンラインで配信して、それを御覧いただくという形を取っております。ですから、委員の皆様からは傍聴者の姿が見えません。それでも、傍聴されている方がいるということ意識していただければと思います。

そして、「会議資料の公開」でございます。これも原則として全て公開になります。

また、「議事録の公開」について。この場で委員の皆様がお話くださったことは議事録の形でまとめられ、原則として公開されます。小委員会が終わりましたら、次の小委員会の前までに議事録（案）を皆様にお送りし、そこで御確認いただいたものが小委員会での資料となり、小委員会において御承認いただいて初めて公開となるという、そのようなプロセスを取っての議事録の公開でございます。皆様の御確認、御承認を経っていない段階で公開することはございません。その点は補足させていただきます。

○相澤主査

ただ今の御説明につきまして御質問等ありましたらよろしくお願ひいたします。

○前川委員

今日の傍聴者の数はどれぐらいなんですか。

○鈴木国語調査官

今日はまだ傍聴については、公開はしておりませんが、傍聴の登録者数で申しますと20名でございます。

○前川委員

という方が多いとかいうのは分かりますか。

○鈴木国語調査官

一番多いのは、国立国語研究所の関係の方々でございます。あとは、マスコミ関係の各社の記者の方。元教員という方も若干いらっしゃいます。

○前川委員

分かりました。ありがとうございます。

○相澤主査

ほかはいかがでしょうか。よろしいでしょうか。 （→ 挙手なし。）

では、御異議がないようでしたら、「文化審議会国語分科会言語資源小委員会の会議の公開について（案）」、こちらについては合意いただいたことといたします。

それでは、これに従って、ここからの議事についてはオンライン上で公開することとなります。御了解をよろしく願いいたします。

【オンラインでの議事の公開開始】

○相澤主査

国語分科会言語資源小委員会、新しく設置されましたので、改めてよろしく願いいたします。

まず、主な審議事項についての確認から入ってまいります。配布資料3「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）」ということで、村瀬課長から御説明をお願いいたします。

○村瀬国語課長

承知しました。委員の皆様、お手元に配布資料3を御用意いただけますでしょうか。御説明いたします。

主な審議事項といたしまして、まず前提でございますけれども、御覧のとおり、「言語・言語資源について」でございますが、言語、言葉というのは、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ございまして、アイデンティティーに関わり、生きる力の土台となるものでございます。そして、それは歴史と伝統の中で培われてきたと言えるかと存じます。したがって、これまでの歴史の中で用いられてきました言語の在り方を振り返ることで、言語の変遷であったり、あるいは往時の諸相であったりといったものを捕捉することが可能となる、そのようなものかと存じます。こうしたことから、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、言語研究や言語政策、言語教育、こうした分野はもちろんそうでございますけれども、社会全体としても意義深いもの、このように考えられないだろうかと思われるところでございます。

そういたしますと、このような意義を踏まえまして、言語資源を、例えば仮称でございまして、「未来へ伝える文化遺産」と位置付けた上で、言語資源の果たす役割と今後に向けた在り方を検討していくということがあり得るのではないだろうかと思っております。具体的には、多様な言語コーパスを言語資源として構築して活用していく、こうした土壌を整えていくことを視野に入れて、次に申し上げる三つの観点から御検討いただくこととしてはどうだろうかと思っております。

では、具体的な「検討の観点について」でございますけれども、大きく3点というところでございまして、まずそもそも論でございます。言語の果たす機能や本質をどのように考えていくのか。言語資源の位置付けであったり、その保存・活用の在り方であったりをどう考えていくのかということを広い視野で考えていくという観点が1点目。

2点目が、現在のフェーズという観点から、現在の整備状況といったものを確認して、その活用の在り方について、どのように考えていくのかという点でございますけれども、具体的には、先ほど申し上げました言語研究や言語政策等といった観点に加えまして、近時の課題でございます大規模言語モデル、こうしたものとの関係での整備や活用の在り方があろうかと思われま。

最後に、将来に目を向けてということになるわけですが、ただ今申し上げました既存の言語資源の整備状況を踏まえまして、今後における望ましい言語資源の在り方について考えてみるというのがあろうかと思っております。とりわけ、社会課題の解決に資するような観点から言語資源の在り方を考えていくというのがあろうかと思っております。そのためには、どのようなケース、分野にお

ける言語資源コーパスがあるとよいのか。実社会における課題はどのようなものがあるのか。そして、それを乗り越えるためのソリューションとしてのコーパスの活用可能性について、どのようなものが期待されるものとしてあり得るのだろうか。こうした事例やケースについて、もしよろしければ委員の皆様からお出しいただきながら議論を深めていただければと考えているところでございます。

以上が審議事項についてですが、続けて、「当面の予定」についても申し上げたいと存じます。

本日が検討のキックオフになるわけですが、ただ今申し上げた「検討の観点」に沿いながら御議論を進めていただきたいと思っております。当面はおおよそ月1回のペースということで、例えば、国内外におけるコーパスの状況や業界におけるコーパス活用の実情といったようなものを御紹介いただきながら御議論を深めていただければと考えております。そして、秋には国語分科会への御報告を挟みながら、後半への議論へと深めていただきたく存じます。事務的には以上のように考えているところでございますけれども、委員の皆様、どうぞよろしくお願いいたします。

以上でございます。

○相澤主査

ありがとうございます。では、ただ今の説明に対しまして、何か御質問、御意見などございましたらよろしく御発言をお願いいたします。主な審議事項として、言語資源の果たす役割と今後に向けた在り方の検討ということで検討の観点をお示しいただきました。予定としては、前半、後半の二つの期間がございまして、特に前半について詳細にお示しいただいたという形になっておりますが、いかがでしょうか。

(→ 挙手なし。)

では、特にならなければ、次に、お手元の配布資料4、前年度の国語分科会で取りまとめられた「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」の内容を確認してまいりたいと存じます。この報告がこの小委員会での議論の一つの出発点となると位置付けられます。では、配布資料4の説明を事務局、どうぞよろしくお願いいたします。

○鈴木国語調査官

配布資料4「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」—こちらの説明をさせていただきます。

まず、「1 社会や技術の変化と言葉の在り方」、その次、「2 信頼できるデジタル言語資源としてのコーパス」、ここまでの部分—ここの中では、世の中の目まぐるしい変化が言葉に及ぼす影響を踏まえ、いざというときに安心して頼りにできるような言語資源が必要である—そのような指摘をしている部分でございます。

中身としては、ウェブ上の巨大なデータを利用した生成AIがいろいろなところで広く使われるようになってきて、データの汚染ですとか一部企業による寡占状態、不透明性など、そうした課題が挙げられております。言葉に関する事柄として、国語施策の観点からも留意していくことが必要であるという認識を持つようになっていくという指摘、そして、ウェブデータを基にした大規模言語モデルとはまた別に、ウェブにはない書籍など紙媒体、その自然言語データを対象とした質の高いデータベースを、AI時代においても確保しておくことが重要である、このような指摘もされております。

このことに関しましては、実は参考資料2—いわゆる「骨太の方針」から引用している部分があるんですけども、こちらでも質のよい言語資源、言語データというものについて言及されていて、国としても重視していくという方針が示されているところと重なることを御紹介しておきたいと思っております。

続きまして、「3 BCCWJの優位性」、「4 BCCWJの課題」。この2か所では、国立国語研究所にございます「現代日本語書き言葉均衡コーパス（BCCWJ）」、こちらについて紹介しております。こちらは非常に信頼の置ける言語資源として考えられるところですが、課題も持っております。一番ネックになるところが、実は20年近く前のデータで止まってしまっているということでございます。これが追加されるということが必要だろうということで、今年度からその部分の追加の事業が動き出している状況でございます。

これからの国語を考える上で信頼できる言語資源としてどのようなものが必要であるか、それを考えたときに、現実的な選択肢として、国立国語研究所が平成23年に公開した「現代日本語書き言葉均衡コーパス（BCCWJ）」があります。このB

CCWJについては参考資料1で大まかなところは紹介しておりますが、後ほど前川委員からの御紹介もあるかと存じます。

このBCCWJは、作成されて以来、まだ本格的な拡充はされてきておりません。これを更に整備・拡充することによって、これからの時代においても、より確実で信頼の置ける言語資源を確保できるのではないかと考えている、このような指摘がなされているところでございます。

最後、「5 BCCWJの可能性」、「6 国語施策としてのデジタル言語資源の整備」。こちらでございますが、BCCWJが再整備された場合には、信頼の置ける言語資源としてどのようなことが期待できるものであろうかということ、それからデータの追加と整備の取組を今後国語施策として定期的に行っていくことが望まれるのではないかと、そのようなことが提案されている部分でございます。

BCCWJは、日本語の研究者、また日本語教育に関わる方々にとっては大変親しみのある欠かせないものになっていると言えるかと思います。さらに、出版社での辞書の編纂^{さん}やIT企業でのデータ活用など、民間でも広く用いられており、将来にわたってデータが追加され、整備されていけば、活用の範囲は更に広がるものと考えられています。加えて、その時々における日本語の姿を残す、ある種の文化財としての価値も生じるものと指摘されているところです。

さらに、最後の部分を見ますと、「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」という要望が述べられておりまして、これはちょうどこの言語資源小委員会が立てられる一つの契機となったと考えることができるかと思います。

以上が配布資料4の内容でございます。

○相澤主査

ただ今の御説明に対しまして御質問等ございましたら、挙手、あるいは御発言にてお知らせください。いかがでしょうか。

○前川委員

これはもう公開された文書ですので、今から指摘するのもあれなんですけれども、2か所、私としてはちょっと違うかなと思うところがございまして指摘させてい

たきます。

一つは、これは事実に関わることですけれども、4の最後のところで、各国語のコーパスのサイズがいろいろ紹介されていて、BCCWJが相対的に見て小さい、その事実自体はそのとおりなんです、取上げ方で、例えばチェコ語で10億語とありますが、これ、実は現在600億語ぐらいになっていて、それも実は中身の多くはウィキペディアを中心としたウェブのデータですよね。そういう意味で、ちょっと比べる内容と時期がばらばらになっているかなという印象を受けました。ただ、最終的な結論に異論があるわけではございません。

もう1点、4の頭のところで、BCCWJの開発が止まった原因の一つとして、共同利用機関になった時に、よりアカデミックな方向を目指したからということが書いてありまして、もちろんそういう要因もあるんですけど、一番大きな原因は、もともとBCCWJというのは外部資金で作っているんです。国語研のCSJもそうでしたけれども、全部補助金で頂いていて、研究所の予算でこんなものは作れない。それが、その後、大きな予算が取れなかったことが一つの大きな原因だろうと思っております。ですから、研究所の性格以外に、そういう外的な要因が大きかったと私は考えております。

○相澤主査

ありがとうございます。事務局から何かありますでしょうか。

○鈴木国語調査官

ありがとうございました。海外の事例については詰めが甘かったというところで反省しております。

あと、もう一つのBCCWJのその後のところについては、外部資金の問題があることは存じてはいたんですけども、この部分ではそこまで踏み込んで記述ができませんでしたので、前川委員にとっては不本意な記述になってしまっております。申し訳ありません。御指摘どうもありがとうございます。

○前川委員

この場でそういうことまで含めて検討していただければ幸いです。

○鈴木国語調査官

ありがとうございます。

○村瀬国語課長

前川委員から貴重な御指摘を頂いたわけですが、この小委員会の場で、前年度に作成したペーパーを、ある意味土台にしてと言いますか、礎にして、またブラッシュアップを図っていくことになりますので、是非より良いものをとということで、委員の皆様、どうぞよろしくお願いいたします。

○相澤主査

前川委員の御指摘の点については、公開資料の一番最後の「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」というところにも、気持ちとしては反映されていると思います。ほかはいかがでしょう。オンラインで御参加の委員の皆様もよろしいでしょうか。

(→ 挙手なし。) ありがとうございます。

続きまして、いよいよと言うか、日本語のコーパスや日本の言語資源について、委員の皆様それぞれの捉え方について詳細に御意見を賜るということで、その議論のための共通認識のために、本日は国立国語研究所所長の前川委員に、「日本語の言語資源 特に言語コーパスについて」ということで御紹介いただけることになっております。よろしくお願いいたします。

○前川委員

よろしくお願いいたします。それでは、早速始めたいと思いますが、スライドが30枚ぐらいありますので、30分—できればもう少し早く切り上げたい—議論の時間を取れるようにしたいと思います。それでは、始めさせていただきます。

今御紹介いただきました国立国語研究所の前川と申します。今日はこういうタイトルでお話をさせていただきます。内容ですけれども、今お示ししているような三つのことについてお話ししようかなと思います。

最初に、そもそも「言語資源」という言葉が何を意味しているかということを通

単に御説明し、それから日本語のコーパスの整備について御紹介して、その問題点とか今後の課題について考えるという順番です。

最初にお断りしておきたいんですけども、私、国立国語研究所に所属しているんですが、これからお話しすることは、基本的に私の個人的見解として申し上げるということで、研究所としての公式な見解を述べるものではないということを御承知ください。それから、私、7年ほど前までは開発の現場で働いておりましたが、その後離れましたので、必ずしも最新の知見を反映していない可能性はあると思います。もし問題にお気づきでしたら、是非この後、御指摘いただければと思います。

それでは最初に、「言語資源とはそもそも何か」ということで、ちょっと余計なスライドかもしれませんが、ウィキペディアに「言語資源」というページがありましたので、それを取ってきました。これをネタにちょっと話をしようと思います。この定義、よくまとまっているんですが、幾つか問題があります。まず、英語で「Language resource」となっているのは、普通、複数形で「resources」ではないかと思いますが、そのことはともかくとして、一つの問題は、ここでは言語資源というものが自然言語処理の技術のためのリソースだと位置付けられているのですけれども、元々そうではない。そして、この小委員会、分科会としてもそれだけを考えるわけではないということがあると思います。それから、作成に非常に手間が掛かるために、その数がいまだに少なくとても高価であるということがあるのですけれども、高価かどうかよりも、現状における問題は、一部の特権的な組織とか、会社とか、そういうところしか特にウェブ上のデータを大量に備蓄することができないこと—それがむしろ問題であろうかと思いません。アクセシビリティに極端な差があるということです。そういう意味でも、公的な組織の関与というものが何かの形で期待されるのではないかと考えております。

二つ目の段落のところに、ウェブから取ってきたデータに関して、一般的にノイズが多くて価値のある情報が云々^{うんぬん}ということが書いてあります。これもそのとおりとさえいえばそのとおりでいいんですけども、現実には非常に多くのものがウェブのデータを基に開発されているわけです。それに対して、例えば先ほどから御紹介いただいていますBCCWJはどのように利用されているかということ、大規模な言語モデルが作り出すものに対するフィルター、BCCWJを一つのがみとして、ある程度ちゃんとしたものができているかどうかというものを評価する、そういう形で使

われているということが現実でございます。

最後の段落には著作権の問題が書いてありまして、これは確かに非常に重要な問題です。今日の私のこの後の話では触れることをいたしません、そういったものがあることは改めて指摘、強調しておきたいと思えます。

以上を前置きにいたしまして、では、言語資源の具体例としてはどんなものが挙げられるかということがこのスライドでございます。御覧のように、いろいろなものがございまして、今日はその中で一番上に挙げておりますコーパス、ランゲージコーパスに関して説明をしたいと思えます。

コーパスというのは何かというと、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ということです。次の次ぐらいのスライドでもうちょっと詳しく説明しようと思えます。ほかの言語資源については、時間の関係もありますので今日は省かせていただこうと思えます。

今御覧いただいている図は何かと言いますと、グーグルのN g r a m V i e w e rを使って、緑色のがL a n g u a g e r e s o u r c e s、あとの二つの点がL a n g u a g e c o r p u s と c o r p o r a、単数形と複数形、その時代の変化を見たわけですけれども、御覧のように、90年代の半ばぐらいから急激に伸びてくる—ここ数十年の間に急激に普及した概念であり、世の中でそれだけ重要性が増えているものだというをお分かりいただけるのではないかなと思えます。この線が関連していることから、言語コーパスを論じることによって、ある程度言語資源全般を論じることにも通じることを示しているのかもしれない。

このスライドは、言語コーパスとは具体的にどういうものか、どういう要件があるかということをもとめたものです。これは10年ほど前に私の書いた論文の中で指摘したことです。七つポイントを挙げております。ここは大事ですので、ちょっと丁寧に見ていこうかなと思えます。まず真正性—A u t h e n t i c i t yという言葉が私が勝手に訳したんですけれども、実際に生じた言語活動の記録である—作られたものじゃないということですね。それが非常に大事なポイントだと思えます。

それから均衡性。バランスですけれども、これは日本語なら日本語、英語なら英語という一つの言語にいろんな側面があります。どういう場面で使われるかということによって言語は変化してくるわけですけれども、それを我々は「レジスター」と呼んでおりますが、様々なレジスターをできるだけ広くカバーしているというこ

とです。

それから代表性。これは対象によりますけれども、例えば母集団が決められるのであれば、その母集団をできるだけ正確な縮図として取り出してきているかどうかということです。例えば、話し言葉なんかは典型ですけれども、母集団そのものが決められないことがありますので、その場合は必ずしも代表性を論じることはできないということになります。

規模ですが、これはどっちかという目的志向の概念で、どういう目的に使うかが決まったときに、その目的のために十分な量が確保されているということです。

電子化、これは現代では当然の要請としてあります。ただ、検索する際に、単なる文字列検索しかできないんじゃないわけじゃないわけで、もうちょっと高度な検索ができるようにいろいろな付加価値を付けることをいたします。御存じのとおり、例えば日本語ですと、そもそも分かち書きがされていませんで、単語を調べようと思っても、その場合にはまず連続したテキストを単語に分割するという作業が必要になってまいります。そういった付加情報。それから、話し言葉であれば、例えばイントネーションの情報なんかを取ろうと思えば、それを何かの形で検索できるような形に記号化してやるというような作業が発生してきます。非常にいろいろな側面の付加情報がございます。

最後、公開ですけれども、これは誰もが利用したい人が利用できるということです。ただ、この最後に関して一言申し上げておきたいのは、最近国の方で「公開」という言葉を完全な無償公開という意味に限定してお使いになっているということがあるようなんですけれども、私はそれにはちょっと違和感を覚えておりまして、要するに、使いたい人が適正な価格で使えればいいという意味で「公開」という言葉を使っております。無償であることを必ずしも意味しないと考えております。

今七つ要件を述べましたが、これを全部満たすコーパスは実はそう多くはない。もうちょっと条件を緩めるとどうなるかというのがこの次のスライドでございまして、真正性、規模、電子化、公開ぐらいが満たされていけば、コーパスという最低限の条件ではないかと思えます。ただ、この今赤くなったところの条件ぐらいでデータを見ますと、それは実際上アーカイブと呼ばれるものであつたりするものと余り区別できない形になるかもしれません。これについては、後でまたちょっと触れます。

このスライドは、それじゃ、言語コーパスにどんな種類があるかということをもとめたものです。皆さん既に御存じのことかと思いますが、一応触れておきますと、均衡コーパスというのはたくさんのレジスターをカバーしているということ。サンプルコーパスと全文コーパス。全文コーパスというのは対象の全体を入れてしまうということです。サンプルは対象が非常に大きい場合に、その中からの確な方法でサンプルを取って全体が分かるように取ったものということです。モニターコーパスというのは、あるコーパスを作って、そのまましておくのではなくて、例えば毎年更新していくとか、5年、10年に一度更新していくという、そういうタイプのものをモニターコーパスと言っております。言語を継続的に監視していくという、そういう意味から「モニター」という言葉が使われているのではないかと思います。それから共時、通時、これは言語学の基本概念ですが、共時というのはある一つの時代だけを対象としたコーパス、通時というのはそういったものを幾つか並べて比較できるようにしたものということです。相同コーパス、これはコンパラブルコーパスという言葉を読んだんですけども、同じ設計で複数の言語のコーパスを作る。非常に有名なのは、アメリカ英語のブラウンコーパスというのがありますけれども、その設計。例えば、その中の何%は新聞で、何%は書籍で、その書籍の中では、例えば宗教に関する領域がどれぐらいでという、そういうサンプルの割合を決めたリストがありまして、それに従っていろいろな英語の方言だとかほかの言語のコーパスを作る。そうすると比較がしやすいというようなことで考え、昔よくこういうものが作られました。それから、対訳コーパス、これはある言語のコーパスがあったときに、それを幾つかのほかの言語に翻訳するという形のものでした。日本語に関しては、中国で中日対訳コーパスというのがかつて作られたことがあります。内容は多くは文芸作品でした。学習者コーパス、これはその言語を母語としない人たちー学習者の方が使った言語サンプルのコーパスということです。

最後に、括弧に入れて「Reference corpus」というのを挙げておきました。これは何と訳せばいいのか、まだ考えがまとまってないんですけども、いわゆる標準性、もうちょっと言えば規範性を備えた人々によって認知されているようなサンプルを集めたコーパスということです。これは言語学的にはなかなか難しい概念なんですけれども、現実には世の中には幾つかの、ドイツ語のDeReKoなんていうのは、正にレビューとかレファレンスなんですけれども、

そういったところでは、例えば新聞の記事とか文芸作品とかいったものを中心にしたものをレファレンスと呼んでいるようです。

それから、コーパスの類似概念として、アーカイブズと言語データベースを挙げておきました。アーカイブズというのは、典型的に言うと国会会議録のようなもので、はっきり限定された対象を全て保管していくというのがアーカイブズでございます。余り付加情報は付けなくて、メタデータ、誰がどこで言ったぐらいのことは付けるんですけども、それ以上の検索用の情報は余り付けられないことが多いようです。典型的には、国会会議録のほかに青空文庫ですとか、最近、国立国会図書館が盛んに出しておられますデジタルコレクション、昔の書籍、そういったものが入るかと思います。

言語データベースは、必ずしも真正性を前提としないコーパス的なものであるとか、あるいは存在するコーパス、それを加工して2次的に派生されたデータといったものが言語データベースと呼ばれていると思います。典型的な例を挙げますと、例えばあるコーパスにどういう漢字が何回出てくるかとか、どういう単語が何回出てくるかとか、共起しやすい単語のペアはこれであるとかいったものを派生させたようなもの、そういったものがデータベース、少なくとも我々はそう呼んでおります。以上がコーパスに関する説明でしたけれども、何のためにコーパスが必要とされるのかということを経つかの側面からお話ししようと思います。

最初の理屈として言語研究がございます。言語研究、これを述べると長くなりますので簡単に済ませたいと思いますが、まず言語というのは非常に多様なものであって、その全体を我々が考えるだけで一頭の中で内省するだけで捉えることはほとんど不可能だということが出発点としてございます。そこに例が幾つか挙げてあります。「山がそびえる」とか「撤回しないべきだ」とか「気付かなそうな」、これ、いずれも日本の文法書の中で誤用であるとはっきり書いてある。しかも、かなり有名な文法書の中でそう書いてある例なんですけれども、実際には、「山がそびえている」「撤回するべきでない」「気付くそうにない」というのが正しいんだということになるわけなんですけれども、実際にはここに出ているような例は一杯出てまいります。それを全部、本当に間違いなんですかという、そういう話です。それを研究する必要があるということです。

2番目は、先ほどから「レジスター」という言葉を何遍か使っておりますが、言

語というのは、それはどういう状況で使われるかで初めて使われる形が決まってくるという面があります。典型的には、日本語の敬語、待遇表現なんかがそうだと思いますが、そういうことを状況とともに言語を把握するというような意味でもコーパスは必要になってまいります。

3番目は、言語というのは変化していったって、なおかつ人間が意図的に新しく創造する、クリエイティビティーを発揮することもできるということです。新語だとか新表現だとか、あるいは外来語から言葉を取り入れる、集団の中でしか通じない言葉をあえて作り出す、伝統的な表現が変わっていく—これは国語課が毎年調査されているような表現の意味の変化みたいなこと、ああいったことがあります。そういった変化を捉えるためには、やはりコーパス的なものが必要になってくるということです。

最後に「Phraseology」と書きましたが、これはコーパス言語学の中で時々使われる言葉です。伝統的な言語学では単語が寄り集まってセンテンスができると考えるんですけども、その中間にフレーズのレベルがあって、そこが非常に重要であると。チャンクとかフレーズとか単語のある程度の固まりで文末満のようなものを重視しようという、そういう見方がございます。

次に、情報科学の世界ですけれども、これは私なんかは申し上げることでもないんです。ちょっと口幅ったいことになりますけれども、大きく分けて、昔から人工知能の研究は行われてきているわけです。1950年ぐらいから始まって、80年代の半ばぐらいまでは大体、いわゆる記号ベースの研究が中心であったと思います。それは実は言語学と相性がいいような内容なんですけれども、80年代の終わりぐらいから、いわゆる統計ベース（学習ベース）の研究が主流になってまいりまして、その学習用のデータとしてコーパスというものが非常に注目を浴びるようになってきたということです。

皆さん御存じのとおり、目覚ましい成果が出ている音声認識や合成は完全に実用化されましたし、現在は生成AI、大規模言語モデルが日々注目を浴びているわけですけれども、こういったものは全て学習ベースになってからの成果と考えるべきです。音声認識や合成は昔からやっていますけれども、いわゆる実用化のレベルに達したのは、最終的には学習ベースになってからだだと思います。

このスライドは、潜在的文化財としての価値ということで、私、20年以上前から

ずっと言っているんですけど、最近ようやくいろいろと取り上げていただけるようになってうれしく考えております。言語データは放っておくとどんどん消えていく。典型的には話し言葉がどんどん消えていくわけですがけれども、書き言葉も実は放っておくと記録は残っていない。BCCWJを作りましたときに分かったんですけれども、例えばある種の雑誌とかといったものはほんの数年前、四、五年前に出版されたものでももう手に入らないんですよ。出版社も持っていませんし、国会図書館に入っているかというとなぜしも入っているとは限らない。そういったことはよくありまして、実はどんどん消えていくというものです。そして、最後に書いておきますのは、時間とともに文化財的な価値がどんどん増していくと。これは例えば、今、平安時代の人と話していた日本語が書き言葉としては残っておりますが、それが録音されて残っていたら、日本語の研究はまるっきり変わってくるんですよ。そういう意味で、日本語の研究に限らず、文化財的な価値は時間とともにどんどん増していくということでございます。

最後に、科学的な言語政策の礎としてのデータとしてのコーパスも考えられると思います。これはこの後また論じることになりますけれども、私の認識として、これから日本語は非常に大きく変化していくんじゃないかと思います。その原因はやっぱり社会の変化にあるわけで、社会の変化をもたらす幾つかの要因というのは今そこに挙げておりますーサイバー化、高齢化、多文化・多言語化、そしてAIのインパクトといったようなものが挙げられると思います。これについては後ほどまた触れますが、大事なのは、そういった変化が起きたときに、例えば教育とか文化政策に当然影響が及ぶわけですよ。それに対する何かの施策を行うときに、科学的なデータに基づいてできればそれに越したことはないわけです。これまではそういうデータが必ずしもない状態で御苦労されてきたと思うんですけども、これからは、そういったデータが十分ではなくてもある程度整備されていて、その分析に基づいて施策が決められていくというような形が理想的ではないかと考えております。

大体20分ぐらい過ぎましたので、あと少しスピードを上げたいと思います。ここからは国語研究所を中心に、これまでコーパスがどのように整備されてきたかという話をさせていただこうと思います。

まず国語研究所の話ですけれども、これから幾つか図をお見せします。この図の

見方ですが、真ん中にありますのが時間軸だと考えて、歴史の軸です。こちらが最初に文献が現れる奈良時代、こっちが現在、ここら辺に昭和とか平成があるということです。最初にCSJというのを構築いたしまして、それは話し言葉で、独話で平成の辺りのものを捉えているということです。これが開発のきっかけになりました。その後、2006年から2010年に掛けて、先ほどから時々話題になってまいりますBCCWJというコーパスを構築いたしました。これは2000年前後のデータを中心といたしまして、書き言葉のうち、均衡コーパスですので、比較的幅広いレジスターをカバーしております。同じ時期に、明治から昭和の初めぐらいにかけての近代語の太陽コーパスというのも構築いたしました。

この後、我々、共同利用研究機関になりまして、第2期中期計画期間になりますと、今度は歴史コーパスの構築をいたしました。その後、方言のコーパスですとか学習者のコーパスにも着手をいたしました。と同時に、この時期にウェブコーパスを作りまして、これはかなり規模がでかくて、ウェブを全体的にクロールしてきて、その中から日本語を取ってきて、283億語ぐらいの規模のものを作っております。

その次の第3期中期計画期間になりますと、今度は歴史コーパスが非常に大きく拡張されたということと、また、話し言葉の方で日常会話コーパスのような、本当にモノログじゃなくて日常の会話とか、3名での対話とか多人数での会話とか非常に自発性の高いデータを映像とともに記録したものを作りました。そのほか、幾つかの既に存在しているコーパスを加工したようなものもございます。

現在何をやっているかというのがこの最後のスライドで、現在は学習者の作文の横断コーパス。時間とともに追跡するようなコーパスを作っていたり、それから先ほどの日常会話コーパスには子供の発話が入っておりませんので、そこをカバーしたようなコーパスを作ったり、先ほど御紹介がありましたけれども、BCCWJの後継コーパス、拡張版、1億語のものを作る事業が今年から始まっております。国語研究所ではこういうことをやってまいりました。その結果、細い線ですけれども、奈良から現代まで言葉の歴史的な変化をたどることもできるようになりましたし、近代における進行中の変化を見ることもできる、それから書き言葉と話し言葉の両方でいろいろなレジスターを把握するというようなことが、十分ではありませんけれども、何とかできるようになってきたかな、これが我々の二十数年間の成果かと考

えております。

次のスライドでは、国語研究所外にどのようなコーパスがあるかということを一覧にしました。実は非常にたくさんございまして、ここでは代表的なごく一部のものを示しただけです。下の方に挙がっております二つのものは国外で開発されているものです。その中に一部として日本語が入っているということでございます。詳しい紹介は省略させていただきます。

こういったたくさんのコーパスがどこで入手できるかと言いますと、大きな配布機関として四つございます。今ここに挙がっているものです。お手元のスライドには、三つ目のNII-IDRというのは抜けていたかと思えます。私、うっかりして入るのを忘れていましたが、ここで紹介させていただきます。

こういった言語資源がどれぐらい使われているかということなんですけれども、いろんなメジャーを考えることができますけれども、ここでは私どもが公開しております「中納言」というシステムのユーザー数と利用件数をグラフにしてみました。御覧のように、時間とともに非常に単調にどんどん伸びている状態です。ユーザー数にして、現在ですと5万人を超えていて、年間検索数で250万件を超えて、もう300万に迫るような勢いで伸びております。今年は多分300万に達するのではないかと思います。とにかくユーザーはたくさんいるということです。

それから、これはBCCWJじゃなくて、その前に作った日本語話し言葉コーパスなんですけれども、どれぐらい論文が書かれているかというグラフです。引用件数です。それを工学系と言語系に分けて示してみました。面白いのは、両方とも1期と2期があって、最初はリリースしたときにばーっと使われるわけなんですけれども、もう一つの波が2010年に来ているんです。これは明らかに深層学習のテクニックが実用化された時期と重なっております、データに対する需要が非常に増えたんだということがよく分かります。

それから、これは論文が英語で書かれているか日本語で書かれているかの違いなんですけれども、御覧のように、1期は日本語の方が多いいんですけれども、2期になると英語が圧倒的に多くなるという形で、それだけ世界的に今は利用されているんだということを表しているかと思えます。

ここから、現状の問題点と今後の課題ということでお話しします。現状の問題点をまとめたのがこのスライドです。様々なコーパスが開発されていますが、大規模

なものも余りなくて、小規模なものが多いということです。それから、継続的な開発が行われていない。それから、書き言葉の場合ですけれども、ウェブを使ったものか、あるいはBCCWJを加工したものが多くて、それ以外のものが意外と少ないということです。それから、デジタル・ヒューマニティーズなんかでよく言われますFAIR原則というのがありますよね。FAIRってFindableで、Accessibleで、Interoperableで、Reusableであることが望ましいということですが、特にそのうちのインターオペラビリティ（相互利用性）ということが余り確立されていない。要するに、仕様が必ずしも標準化されてないことが多いということです。それから、歴史コーパスの場合は、現在はまだ古いところは文学作品が中心になっている。それから、権利処理がいまだに大きな問題であり、そして、まだまだ整備されてない、足りないレジスターというものが一杯あるということです。

それについてまとめたのがこのスライドです。これは国語研の同僚にも何人か話を聞きまして、どんなコーパスが必要ですか、今後必要になりますかということをもとめてみたものです。意見はみんなほとんど一致しておりまして、ここに挙がっているようなものが今後必要になるだろうということでございます。一々触れていくと時間が大変ですので、ちょっと飛ばしますが、数点触れると、やっぱり社会の多言語化によって、日本語以外の方が話している日本語のコーパスはどうしても必要だろう。それから、AIの影響なんかを評価できるようなデータもやっぱり必要だろうと思います。後の質疑のところでも御質問を頂ければ説明しようと思います。

このスライドは、先ほどのうちSNSとか、Computer-Mediated Communication、CMCの中にどんなものがあり得るかということ、最近国語研に来ました落合さんという方が発表されたものがありましたので、そのスライドをお借りして示したものです。一つのCMCの中でもこれだけいろんなものがあるということです。

それから、これは私の頭の中でずっと考えていることなんですけれども、先ほどから何遍か申しました、これから日本語が大きな変化期に入っていくだろうと。その際に、どういう言語問題が生じて、先ほど申しましたように、有効な言語政策とかを考えるためにどんなコーパスが必要になるかということをもとめた概念図です。これは説明すると非常に時間が掛かりますので省かせていただきますが、次のスラ

イドで主要なものを取り上げております。

日本語がこれから変化していく要因として、大きなものとしてこの三つ。先ほどもちよつと触れましたね。インターネットの影響、AIの影響、そして日本社会の多言語化といったものが考えられるかと思えます。ごく簡単にそれぞれ触れたいと思いますが、今出てきたスライドはネットの中です。私が漫画かなんかを読んでいたときに出てきた広告なんですけれども、「どの国籍の女の子を育みたい？」という、これ、ネットの広告で出てくるんですよ。この日本語は何だろうということで、これがおかしいわけなんですけれども、おかしいと言いたいんじゃないで、どうしてこういったものがパブリックな面に出てくるのかという、その原因が面白いと思うんです。これ、機械翻訳の間違いではないと思います。多分人間がこれを書いたんだろうと私は思っております。機械翻訳ではこう間違えないと思うんです。そこから辺、ちよつと御意見を伺えればと思えますが、こういったものが世の中に結構たくさんこれから出てくるわけです。そうすると、単純接触効果と言いますが、触れていると自然に思えてくるんですよ。それで言語が変わっていく、こういう影響は決して無視できないと私は考えております。

これはセンサーデータの例で、これは私がやっていることなんですけれども、例えばこういうデータを取ることができます。

(映像上映)

これは何をやっているかという、要するにしゃべっているときの口の中の動きをセンサー、MRIですけれども、使って捉える。こういったものと言語の関係を捉えるようなコーパスも現実に作ることができる。私、既に一部公開しております。

まとめですが、こういう形になります。ちょうど30分になりましたので、ここでやめようと思えますが、一応読みますと、過去四半世紀で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ました。しかし、いまだ道半ばであり、今後データ整備が望まれる多くのレジスターが残されております。

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CMC関係や日本社会の多言語に関する部分が弱いと考えられます。一方で、情報処理領域での需要も著しく増大しています。国費を組織的に言語データの整備に投入するのであれば、整備の順序について超領域的な熟議が必要だろうと思えます。また、諸外国における政策の比較研究なども重要ではなかろうかと思えます。

以上でございます。

○相澤主査

前川委員、大変ありがとうございました。大変情報量の多い示唆に富んだ御発表を頂きました。

それでは、本日、資料の中で提示された観点等も踏まえまして、委員の皆様から御意見や、御専門の立場から御発言を頂ければと思います。よろしくお願いいたします。

○齊藤（美）委員

齊藤です。よろしいでしょうか。

○相澤主査

よろしくお願いいたします。

○齊藤（美）委員

前川委員のお話を伺っていて、大変勉強になりました。ありがとうございました。いろいろ伺いたいこととかいうか、気になることとかあるんですけど、まず一つ、御発表の中で、今後構築が望まれるコーパスのところで、漫画、アニメ、ゲーム、世界にとっての日本語というのを一つ挙げていただいたと思うんですけども、これはもう少し詳しくお話を伺えたらうれしいなと思います。これは、翻訳される前提がある日本語という意味なんでしょうか、それとも、日本語が母語話者じゃないけれども、日本語を楽しみながら、勉強というか、生活の中で遊びとして使うみたいな、そういった要素なんでしょうか。よろしくお願いいたします。

○前川委員

その両面があると思うんですけども、特に意識しておりましたのは後の方でしょうかね。私、日本語教育は専門でも何でもないんですけども、ただ、一橋大学とかで、そっちに関連するようなところでしばらく連携教授みたいなことをやっておりまして、そこでつくづく思うんですけども、動機としてそういう形で日本語

に興味を持ったというのが圧倒的に多いということ、それから、人によりますけれども、実は字幕で流れていて、そういう形で日本語を聞いて、それで実際上かなり覚えてしまった—日本語を教師に付かずに、学校に行かずに、そういった形で覚えてしまったという人もいますよね。翻訳されたものに関してまで入れてしまえば、もう圧倒的な日本の今、最も影響力のあるメディアだろうと思いますので、そういう幾つかの面を総合して、外国に一番流れている日本語はここだろうと考えて書いた次第です。

○齊藤（美）委員

ありがとうございます。よく分かりました。アニメなんかでネット上で公開されているものに違法なものもあるんですけど、ファンの人が字幕を付けて、それで、ネット上に公開してしまうみたいなものもあるわけです。その中で日本語を勉強したいというか、元のアニメのこと、作品をよく知りたいという動機が強いんだと思うんです。日本語の要素も説明を注で付けたりしながら、字幕でやりながら流すのが喜ばれるというようなこともあるので、気になって伺いました。ありがとうございます。

○前川委員

ありがとうございます。

○相澤主査

では、ほかはいかがでしょうか。

○神永委員

大変分かりやすい御説明ありがとうございました。辞書編集者の立場として伺いたいですけれども、私、小学館で国語辞典の編集を長い間やってきた者として、辞書で、特に気になることの一つに外字の問題があるんです。それをコーパスとしてどのように扱うのか。要するに、辞書の中では外字を残しておきたいんですけれども、コーパスを活用した場合、残しておきたいんですけれども、それを検索のためになくしてしまうとちょっと困ってしまうことがあります…。外字だけじゃな

くて、画像だけで表現されているものとかも当然あると思うんですけども、そういったものもどうしているのか、どのように考えていらっしゃるのかということ伺いたいと思ったのですが、いかがでしょうか。

○前川委員

外字の問題は、今御指摘にあったとおり、検索のしやすさと、それから言語資料としての価値との間にトレードの関係がどうしても生じてしまうわけですね。一つ考えられる方策としては、検索のためには代表的なもので置き換えておいて、そこに元々の字はこれであったというようなことを何かのメタデータ、埋め込みデータとして残しておくというようなものが現実的な解ではないかなと思います。

○神永委員

そうですね。やはりそういう形で残しておいていただければ我々としても有り難いですし、活用できると思います。ありがとうございます。

○相澤主査

ほかはいかがでしょう。 (→ 挙手なし)

それでは、私から1点。先ほどの構築が望まれるコーパスを拝見しますと、マルチモーダルと言うか、テキスト、音声にとどまらず、画像ですとかセンサーデータ等も含まれた多様なコーパスが含まれています。一方で、テキストを中心とするコーパスについても、最近は「文書A I」という形で、文書のレイアウトですとか、例えばどの部分が太字になっていて、どの部分が箇条書になっていてという情報が注目されています。このように、人間が目で見ると文書のレイアウトそのものを見て理解するようなことをA Iもできるようになっていくという流れが重要になっていくと思っています。また、レイアウトそのものも文化だと思いますので、そういったところに着目したテキストコーパスの作り方もあるかなと思いながら拝見しておりました。この辺りは、いかがでしょうか。

○前川委員

実は、B C C W Jはそういったことをある程度考慮しているんですよ。段落の構

成だとか。レイアウトを完全に再現できるような情報は提供できていないと思うんですけれども、部分的に再現するための情報はかなり苦勞して付けたんですよね。ただ、実際の利用例を見ていると、それを活用した研究って、私の知る限りないんじゃないかな…、少なくとも言語研究の側で言うと、ほとんどないみたいに思います。相澤主査がおっしゃったように、例えば実際の新聞をぱっと見てどこを見るかとか、雑誌を見て開いたときに、どこから読むかとかいうようなのは、実際もちろんレイアウトは影響するだろうと思うんですよね。そういう意味では、ある程度自動化できるのであれば、是非残しておきたいデータではありますね。

○相澤主査

ありがとうございます。ほかはいかがでございましょうか。

○石川副主査

前川委員のお話を伺いまして、今後どういうことを考えていったらいいのか、個人の意見として、5点ほど申し上げます。

まず一つ目は、やはり定期的にデータを取り続けていくことの重要性です。このたび、BCCWJのVersion 2というものが計画されており、これは非常に有り難いことではあるのですが、Version 2が出て、そこで打ち止めになってしまっただけでは困るわけです。これからずっと続いていく時間軸の中で、将来にわたって定期的にデータを取り続けていく、また、場合によっては、前だけでなく後ろにも遡り、同じサンプリングフレームで過去のデータも集めていく…今のBCCWJを起点として、前後両方向で、収集範囲を広げていくような取組が望ましいだろうと考えています。これが1点目です。

2点目は、今、前川委員の御説明にもありましたように、BCCWJは非常に綿密に作られています。そこに入っていないジャンルというものもたくさんございます。そうした新しいジャンルについても、継続的な時間軸の中でデータを集めていくことが必要になろうかと思えます。

3点目ですけれども、私自身も含め、コーパスを作る人は「作る」ことにのみエネルギーを注ぎがちです。しかし、作ったものを広く使っていただくためには、例えば、検索のためのインターフェースの在り方なども考えていく必要があるかと思

います。

4点目は、社会実装ですね。海外のコーパス研究、特に外国語学習者コーパスの研究分野では、「コーパス研究の成果で、外国語教育にもうすぐ大きな変革が起こるだろう」というようなことが古くから言われてきたのですが、実際には、目に見える変革というのは余り起こっていないようにも思えます。同じく、日本語コーパスの研究成果についても、日本語教育や国語教育、あるいは、社会の様々な層にどのくらい具体的に影響を及ぼしてきたのか、と考えますと、本当はもっとあっていはずのものが、まだそこまで至っていないのではないかと、とも思われます。この意味において、コーパスを作るだけでなく、その社会実装を考えていくこともまた重要な課題だろうと思います。

最後に、5点目として、今申し上げたようなことを可能にしていくためには、やはり継続的な財政支援と、コーパス開発を継続的にサポートしていく国の仕組みが必要です。それなしで、その都度その都度の競争的資金だけでやっていくということでは、長いスパンでものを考えられません。そうした体制が整備されるように、我々としても声を上げていくことが必要ではないかなと個人的に考える次第です。

前川委員のお話を伺いながら、言語資源の将来の展望について、こうしたことを考えました。

以上です。

○相澤主査

ありがとうございます。ただ今の御意見に関係してでも結構でございますし、そのほかの点でも結構でございますので、いかがでございましょうか。

○石川副主査

この機会に前川委員にお伺いしたいのですが、先ほど申し上げたように、ずっとデータを取り続けていくということが、もし日本において実現するとすれば、どういう要件が必要だとお考えでいらっしゃいますでしょうか。

○前川委員

端的に言うと、資金と人だと思いますけど。

○石川副主査

なるほど、ありがとうございます。

○前川委員

あと著作権ですよ。現代語のコーパスはもうそこから逃れられません。30条の4というので大分楽になったわけですがけれども、それを具体的にどう解釈するかということではまだ見解が分かれるところがあるように思います。本当にそのところを、一種の事例の積み上げでもいいですから、ここまではオーケーだ、ここはオーケーでというようなガイドラインがある程度はつきりするようになってくればやりやすくなりますよね。

○石川副主査

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。

今後という意味では、30ページ、言語問題の俯瞰図^{ふかん}は非常に含蓄に富んだスライドであるようにお見受けいたしましたけれども、御説明の時間がとても足りないということだったので、もし補足や追加で何か御説明いただけるようなことがありましたら…。

○前川委員

Roadmap2023というのを一応説明しておきますと、昨年度、文科省が募集した科学技術関係の大規模な、最大で100億円、数百億の規模まで可能というような10年間の研究プロジェクトの募集、それに私どもが応募いたしましたして、そこそこいいところに行ったんですけど、最後に落とされてしまいまして、残念ながら取れなかったんです。その申請書で用いた概念図です。真ん中に書いてありますように、先ほど私の話の中でも申しましたように、日本語は今、非常に大きな変化期に入る時期にあると思っています。日本語というのは過去に何遍か、例えば奈良時代から平安に

掛けてとか、あるいは室町時代とか明治とか大きな変化期があったと言われているわけですが、それに並ぶような時代にこれから入ってくるんじゃないかなと私は考えております。

その過程は、言語を研究している人間としては、是非緻密に研究したいことですが、一方、言語の変化は必ず社会の変化と同期しているわけですので、そこに生じるであろう多くの言語問題に対して、我々が何かの形で貢献できればそれに越したことはないわけです。それを表現しているのがこの図でございまして、真ん中に問題が発生するという、左側はそれを科学的に研究するために必要なデータとして現在どんなものがあるかということを示してあります。ただ、やっぱり足りない、だから、この計画で国費を投入していろんなデータを作りましょうというのが、右側のいろんな色で書いてあるところです。

細かいので見えにくいかもしれませんが、上の方はB C C W Jの拡張というようなことが書いてあります。これは実は文化庁のプロジェクトとは全然独立に考えたことですので、結果的には2億超えの構築、倍にするということを書いてあります。

そのほかに、時代的な変化を、統計的なモデリングを行うために、ある短い期間だけは飛躍的にデータを増やしたいということがございましたので、特に明治以降ぐらいから現代に至るぐらいに関しては、更にぐっと多いデータを取りたいというようなこと。

あとは下の方になってまいりますと、障害のある方の日本語だとか学習者のデータというようなものも取りたい。そして、これはこの分科会の話とは少し変わってまいります、下の緑色のところになってきますと、これはいわゆる、在日の外国人の方の言語の問題がありますので、それをきちんと把握したい。つまり、今日本でどれだけの外国語がしゃべられているか、実はデータが分からないんですね。一番いいのは、国勢調査にその質問を入れる一かなり多くの国で、国勢調査には「あなたはどの言語をしゃべりますか」、「家庭では何をしゃべりますか」というようなことは聞くんですけども、日本は聞かないものですから、本当はそうやってもらうのが一番いいんですけども、それが無いのでやりたいというようなことが、その緑のところを書いてあります。それから、日本人自体が今、どれぐらい読み書き能力があるのかということの調査もやりたい。そういった総体的なデータを取ってきて、それを科学的に分析するために、一番右側にありますような、我々は勝

手に「E3P-Linguistics」と呼んでいるんですけども、そういったものを開発したいというのがこの図です。E3Pの部分は幸い別途予算が認められまして、今、人間を募集しているところで、ちょうど近く、結果が出るところです。

最終的には、右下に「新しい言語観」と書いてありますのは、要するに、多文化、多言語の日本の社会の中において、日本語を—誰もが同じ言語を知っているという社会じゃなくなってくるわけですね。そういう多様性をはらんだ言語観のようなもの、先ほどお見せした、「どの国籍の子を育ちたい」みたいなもの—ものすごく拒絶感があると思うんですけども—そういったものを少し望ましい方向に変えていくというか、どういふのかを具体的に述べるのは難しいですけども、より許容性に富んだ、柔軟性に富んだ言語観が日本の社会の中で生まれるような、そういう方向に長期的なインパクトとしては持っていきたいと、そういう提案を表した図です。

○相澤主査

ありがとうございます。この御提案の中で、先ほど石川副主査がおっしゃったような継続性、つまり、10年というのは十分長期ではありますけれども、それを超えて、20年、30年とモニタリングしていくようなファクターも重要であるということ、結論としては持っておられるということですね。

○前川委員

この申請書の中では、余りそのことは書きませんでした。ただ、重要であることはもちろんです、当然のことです。

○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

再びよろしいでしょうか、齊藤（美）です。

○相澤主査

よろしく申し上げます。

○齊藤（美）委員

新しい言語観というところ、大変興味深いなと思って伺いました。というのは、今日、この会に参加しながら、日本語の定義と言いますか、どこまで含めるのかと言うか、コーパスとすることで、例えば私が今日参加する前までは、地域方言なんかどこまで含めるのかとか、地域方言だけのコーパスがあるのも存じ上げてはいたんですけども、例えば宮古語は含めるのかとか、あるいは独立したものとして別途作るのかとか気になっていました。辞書編纂者の方もいらっしゃるので、やはり言語の研究は規範を求めがちということもないと思うんですが、コーパスは、先ほど文法書の中では誤用となっているものも実際使われているとおっしゃっていて、そのとおりだと思うんですけど、どこまでは使われているからOKなのかとか。一般社会の中には拒否反応を示すような方もいらっしゃるのかなと思ひまして、正に先ほど挙げてくださった「どの国籍の女の子を育ちたい」というのなんかもそうだと思うんですけど、こういうのがたくさん使われるようになったら、じゃ、OKとして、いずれ文法書にも載るようになるのかとか。その辺りが難しいなと思って伺っていたんです。コーパスの御研究を専門になさっている方としては、そういうところはどのようにお考えになるのでしょうか。伺えたら幸いです。

○前川委員

それに対する答えは一通りじゃなくて、いろいろあり得ると思うんです。だけど、私の実際に取った行動をお話いたしますと、コーパスを作っているときに、実はランダムサンプリングで雑誌とか書籍から我々は取ったんです。これは恐らく世界でも我々だけだと思うんですけども、そういうことをやると、実は日本の書籍とかのかなりの部分はポルノグラフィティ的なものなんですよ、驚くぐらいの比率がそうなんです。そうすると、そのサンプルをテキストとして入力して分析するというような過程で、こんなのやりたくないみたいな話が当然出てくるわけですよ。実際それが生じまして、そういったものは省くべきだというような意見も出てきました。

ただ、そのとき私がーこれが唯一の正解だとして申し上げるわけではありません

が一私自身は、「悪いけどやってくれ、省かずに、サンプルに引っ掛かったものはみんな入れる。それは、日本語全体のゆがみのないデータを作ることが我々の目的だから、我々が恣意的に省いてしまったんじゃ、そこが台無しになってしまう。だから、少なくともBCCWJにおいてはそうするんだ。」ということで、そこは押し切ったというか、納得してもらって、それで今のような形になっております。だから、BCCWJは引くと、そういうものが一杯出てまいります。公開してからですけれども、このコーパスを引くとそういった表現が一杯出てくるので、これはちゃんと作ってあると、そういう評価を頂いたこともあります。

ただ、それは目的次第ですね。先ほどちょっと申しましたリファレンスコーパスというようなものを作るのであれば、そこは何かそういうものを、外的な基準から導入する必要があるということは間違いないです。学校教育でBCCWJを使えないというのはよく言われる話です。引くと子供が喜んじゃうという話です。

○齊藤（美）委員

よく分かりました。ありがとうございます。

○相澤主査

ありがとうございます。先ほどのお話をお伺いして、書籍等は商用価値のあるテキストなのかなということを考えていて、一方でスライドの中で「打ち言葉」という新たなジャンルが出てきて、なるほど、いい表現だなと思いました。ここで、打ち言葉とかスマートフォンに話す、スマートスピーカーに話す場合の言葉は商用価値がある言葉として流通してはいないと思いますが、その二つを区別する必要性は感じておられますでしょうか。

○前川委員

御質問の意味は理解できたと思うんですけども、売り物になるような言葉ということですね。

○相澤主査

はい、そうです。

○前川委員

余りそういう観点から今まで考えたことはなかったんですけども、言われてみると確かにそのとおりですよ。書籍とか雑誌とか新聞もそうかもしれませんが、有償で世の中に出てきたものとそれ以外のものというのは、レジスターの性格がすごく違うかもしれませんよね。目的によりますけれども、これまでのコーパスというのは、基本的には有償のものを中心に多分作られてきていたんだろうと思うんですけども、それ以外の日常会話、我々が作ってきたもので言うと日常会話コーパスとか、ああいったものは言わば、それ自体に商品価値は余り認められないものを大量に集めた。だけど、恐らく伝統的に言語学が概念的に第一義的な対象としてきたのはそういったものだろうと思うんですよ。本当に言語学がそれをきちんと分析しているかどうかは別の話ですよ。概念的には、やっぱり話し言葉で、余り規範性にとらわれない日常に使われている言葉が、本当は言語学の対象であるというのが結構昔からの言語研究のテーゼになっていると思います。ただ、実際は必ずしもそういうものを分析していなかったと思いますけど。

○相澤主査

ありがとうございます。著作権がコーパスを作る上で非常に問題だということをお伺いして、著作権は著者の権利を守っているものなので、著作物に価値がある場合に特に問題になるのかと思うんですけども、一方でウェブ情報を扱ったりするときに、最近個人情報も非常に問題になってきて、法律という意味では、著作権に加えて、個人情報の方も大きな障害として考えていかねばならなくなっているという事は日頃感じております。やっぱり両方とも踏まえて言語資源というものを考えていかねばならないというものかなと、思いながらお伺いしました。

○前川委員

それは私、確かに著作権のことばかり言ってしまって、個人情報の保護ということをお述べませんでした。実はBCCWJのときもそれは幾つか問題になりました。まだそんなにスマートフォンとかも出てない。BCCWJに「スマホ」という言葉は入っていないんですよ。そういう時代だったんです。新聞社からデータを提供

いただいたんですけれども、犯罪記事で犯人の名前が実名で出ている、けれど、その後、服役して社会更生している可能性があるのも、それは消してくれと言われて、それは消したりということはしました。

それから、yahoo!に御協力いただいて、ブログのデータをかなり大量に入れているんですけれども、その中には個人情報結構入っているんですね。それは、実はNII（国立情報学研究所）にも御協力いただいたりして、可能な限り検索して消してあります。ただ、完璧ではありません。

○相澤主査

匿名化も必要ということになりますね。

○前川委員

はい。

○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。

○村瀬国語課長

先ほど前川委員から構築が望まれるコーパスということで、多言語社会を一つのコンセプトとして御提案いただいたかと受け止めた次第でございますが、今後に向けたイノベーションと関係させようとしたときに、コーパスをいかに実社会における課題解決につなげていくかという視点からは、委員が今お考えになられているコーパスについて、どのような期待可能性と言いますか、応用可能性があり得ますでしょうか。御示唆を賜ればと存じます。

○前川委員

ちょっと御質問の意図が理解できていないかもしれませんが、一応お答えします。

例えば外国の方が日本語で日本の情報サービスを利用するような場合というのは当然今後考えられるわけですね。音声認識が典型だろうと思いますけれども、あれは今のやり方ですと、音声のデータベースがあって、それと辞書との掛け算で確

率を計算して、確率が一番高いものを認識するというやり方ですので、そうすると、外国の方が話された日本語の音声データがある程度ないと利用できなくなっちゃう、ヒットする確率が下がっちゃうとかという、そういう非常に具体的な面でのメリットというものも考えられます。

それから、ほかにはプレーンランゲージという話をちょっとしたかと思うんですけども、「やさしい日本語」的なものですよ。日本語は必ずしも十分でない方に向けて、どういう日本語を書いて、行政なんかがどのように言葉を使えばいいかというようなことですね。それに関しては世の中にいろいろな試みがありますから、そういったものを集積して行って、望ましくは、その中の本当に分かりやすいものと必ずしも分かりやすくなってないようなものに対する評価値を付けるようなことをしたコーパスみたいなものを構築できれば、それは自動化機械みたいなことも含めて、かなり役に立つのではないかと考えております。ドイツなんかはそれは法律で要請されるんですよ。必ずそういう「やさしいドイツ語」で書きなさい一書かなきゃいけないんですね。そういったふうにいずれ日本も、強制するかどうかは別として、そっちの方にもう既に移っていきつつありますから、そういったものを支えるコーパスというようなものもあり得ると思います。ほかにも一杯あると思いますよ。日本語教育では絶対必要にされますしね。どの範囲の漢字を教えるかとか、どういう順番で教えるかとか、どういう単語を先に教えるかとか、全部やっぱり言語資源コーパスがないとできないことだろうと思います。

○村瀬国語課長

音声認識なり、自動翻訳なり、それなりにいろいろ普及が進んできたところかと思いますが、そうした中であって、更に進めていくとした場合に、コーパスとの関係でどの部分を意識していけばよろしいでしょうか。それから、先ほど評価値を付けるというお話がありまして、評価値とコーパスの関係というのはにわかにイメージしかねたのですが、もう少し掘り下げてお話をしてくださると大変有り難く存じますが、よろしゅうございますでしょうか。

○前川委員

評価値の話は簡単で、実はCSJの一部にはそういうものが付いているんですけ

れども、例えば「やさしい日本語」的に、日本語が母語でない人のためにやさしく書き直しましたという例が、今地方自治体中心に一杯ありますよね。それをその人たちが読んで、分かりましたか／分かりませんかという、そういう評定とか、そういうものが1/0でもいいから付いていれば、目的を達成しているかどうか分かるわけですよね。それができてきたら、あとはAIはかなり頑張れると思いますので、そういう非常に単純なものでもいいから、そこは2次的に付けばいいわけです。データとは別に評定者を雇って、その評定者にいろんな言語の人が入っていて付ける－それは日本に滞在している数の多い順に言語を選べばいいと思いますけれども－そういったデータベースは有益かと思います。

それから、最初の御質問は何でしたっけ。

○村瀬国語課長

より多言語社会を見据えてやっていくとした場合、「やさしい日本語」というのはその一つの例だと思いますから、今後に向けて課題解決によりつながるといふ部分の中にはあろうかと思しますので、今どの部分が不足しているのか、あるいはこの部分については上乘せが必要だろうといった部分があれば、コーパスの関係で、どこの穴を埋めていったらいいのかというそういう課題意識でございました。

○前川委員

分かりました。そういう問題は確かに存在するはずで。それで、音声認識－相澤主査の方が御存じかもしれませんが－機械翻訳にしても、ある程度データがカバーしてないと対応しないわけですよね。ですから、それをどこが一番抜けているかといったら、今直ちに私、お答えできませんけれども、どこが抜けているかを調べることはできるはずで、どういう部分に対して、例えば生成AIの反応が悪いか、それから、音声認識であればもっと単純で、多分ある種の単語が欠けているか、音声の面でのバラエティーが欠けているか、あるいは発話速度に対する対応が欠けているとか、そういったどれかですから、そっちは比較的簡単に、その気になってやれば、組織的に調べれば見付かると思います。

機械翻訳なんかの方になってくると、私が申し上げるといいかげんな答えになりそうな気がするんですけども、できたら相澤主査にお答えいただければと思いま

す。

○相澤主査

村瀬課長、いかがでしょうか。

○村瀬国語課長

イメージをつかむことができました。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。十分に議論をする時間を取っていただきましたので、もう一言ずつ委員の皆様から頂く時間がございます。それで、やや強引な提案でございますけれども、委員名簿の順で一言ずつ、順番に皆様から頂く形で締めにさせていただきますと存じます。私、アなので、責任を取って最初に発言させていただきますので、その間に御準備をお願いいたします。

前川委員、どうもありがとうございました。大変勉強になりました。私、コーパスに関して、長年自分でもコーパスを使ってきたと思っていたんですけども、今初めて新しい見方ができたと思ったのは、コーパスというのは大きな日本語の基盤であって、その上に様々なアノテーションを載せていくことによって社会実装がされていくというクリアなイメージです。それは大規模言語モデルが非常な汎用モデルであって、その一つのモデルがいろんなタスクをこなしていくのとちょうど同じイメージで、網羅性がある日本語をカバーすることによって、初めていろいろな社会実証、先ほどの翻訳の例で言えば、カバレッジを広くすることによって性能評価をしっかりとしていくといったような実装が開けていくという、そういった見方でコーパスを見ていくということを学んだように思っています。今後引き続き議論していく上で、いろいろとまた勉強させていただければと思っておりますので、よろしくをお願いいたします。ありがとうございました。

では、石川副主査、よろしくをお願いいたします。

○石川副主査

いろいろやりたいことと言いますか、我々が考えないといけないことはあると思

うのですが、やはり、良い言語資源を作っていくということ、作ったものは皆さんに使っていただけるようにしていくということ、さらには、作り放しにしないこと、アノテーションのこともそうですが、いろいろな人が関わって、集めたデータの質を磨いていくということですね。

最後にもう一つ大事なことを申すと、こうしたことを長く維持していくこと、維持し、継続していくこと、そのための体制の必要性を一緒に考えてまいりたいと思った次第です。ありがとうございます。

○相澤主査

では、植木委員、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。私は日本の古典を勉強しているので、この場にふさわしくないような感じはするのですが、前川委員のお話で大変勉強になりました。齊藤（美）委員と前川委員のやり取りの中で、規範性と言いますか、文法的に間違っていない言葉を選んでいく必要があるのかどうかというようなところを興味深くお聞きしました。私は歴史コーパスは使わせていただいているのですが、古典の場合は書写されて残ってきたという点で、一定の基準をクリアしていると言いますか、そこまで言葉の乱れや、庶民的な俗語というんでしょうか、そういうものがない、ある程度洗練されて^{とう}淘汰されたものが残っているのです。取る範囲というところに余り迷うことがなく、それが現代のものと大分違うんだなということを改めて感じました。

今日、「打ち言葉」という面白い用語を、なるほど、そのとおりでなと思ってお聞きしたのですが、例えば笑うことをアルファベットの「w」で表す例や、同僚の国語学の先生が、卒論で絵文字を研究したいという学生がいて大変困ったというようなことをおっしゃっていたのを思い出しました。本当に素人のお伺いなのですが、そういう絵文字的なものの扱いはどのように考えたらいいのだろうかと思ひながらお聞きしておりました。本当にいろいろ学ばせていただきました。ありがとうございます。

○相澤主査

ありがとうございます。

前川委員、順番は最後なので絵文字については、その時にコメントを頂ければと思います。

では、神永委員、よろしく願いいたします。

○神永委員

私、辞書の編集をやってきたものですから、ちょっとまた違う視点からと言いますか、お聞きしたいこともまだあったんですけども、例えば国立国語研究所で分類語彙表ってとてもすばらしいものがあるんですけども、それとのひも付けというんでしょうか、そんなことはお考えでしょうか。アノテーションに関わるようなことなんだろうとは思うんですけども。実は辞書編集者としては、そういったものの情報が付いていると、それを活用しながら検索ができると非常に有り難いなどという思いがありまして、ちょっとお聞きしているんですが、いかがでしょうか。

○前川委員

簡単に一言で。やっています。

○神永委員

そうですね。ありがとうございます。是非それは進めていただきたいと思います。分類語彙表自体はまだまだ語彙数が足りない、これからも増やして行っていただきたいなと思うんですけども、それとのうまい形で連携していただければ、私たちとしてもとても有り難いと思いますので、是非それは進めてください。よろしく願いします。

○前川委員

基本の分類語彙表といろいろな形態素解析したものとの対応が取れるようにはなりつつあります。

○神永委員

そうですか。それは是非よろしく申し上げます。

○相澤主査

ありがとうございます。

ついでに、先ほどの絵文字のお返事も。

○前川委員

絵文字は、BCCWJでもやろうかなという話になったんですけど、ものすごいのが出てくるんですよ。何百字と使ったような、アスキーアートと言うんですか、それでちょっと挫折しました。だけど、日本語コミュニケーションのある特徴、ある一部でもありますから、余力があればやる価値があると思います。

○相澤主査

ありがとうございます。

では、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。今日は大変勉強になりまして、ありがとうございました。ある意味当然のことかもしれないですけど、今日改めて気付かせていただいたのが、国の予算を使って何か事業を行うというとき、今回ではコーパスを作るとか整備する、発展させていくというようなときに、やっぱり研究者のためになるというだけじゃなくて、石川副主査が何度かおっしゃっていたように、使いやすさとか、一般の方に魅力が伝わるような取組をしていくのが確かにとても大事だなと思ひまして、改めて研究者の端くれとしてーコーパス研究はしてないんですけどもー自身が引き締まる思いでした。

先ほど御質問させていただいた、規範と外れるようなデータを示すようなときって、辞書に書かれていること、文法書に書かれていることが正しいんだと思ひているような方々にとってはちょっと拒否反応みたいなこともあるのかと思ひますので、その辺りもしっかり説明していくような、今日の前川委員のお話などを聞いてくだ

されば分かることだと思うんですけど、そういうことも必要なのかなと思いながら参加させていただきました。

以上です。ありがとうございます。

○相澤主査

ありがとうございます。

では、武田委員、よろしく願いいたします。

○武田委員

前川委員をはじめ委員の皆様のお話、大変興味深く、勉強させていただきました。

質問ということではないんですけども、現状の報告と申しますか、コーパスと辞書編纂^{きん}の関係も幾つか言及していただきましたけれども、実際に辞書の編集において、コーパスを利用する機会は非常にございます。頻度ですとか傾向ですとか、あるいはコロケーション、言葉の結び付き、そういったところを確認しながら編纂しているところのございます。そういう形で有効に活用することができていますので、特にBCCWJは、ある一時期のデータで止まっているということのございますから、新たにまた構築される、石川副主査がおっしゃったような継続性ということが非常に重要だろうと思います。言葉自体がどんどん変わってまいります。辞書の編纂も、改訂のペースは思ったより早いと言いますか、数年に1回改訂するケースもあり、そういった中では、新語ですとか新語義ですとか、新しいそういった情報に対応していくということがございますので、そういう意味では、コーパスも常時改訂をするというような体制を取っていただければ辞書編纂にも役立つと思います。

辞書と申しましても、規範性ばかりを追求するものではございませぬので、現状を認識しながら現状をきちんと捉えていく一助として、コーパスの必要性は非常にあるかと考えております。また、書き言葉中心ということではございますけれども、話し言葉なども、既にそういう形のコーパスも構築中であるというお話も伺いましたけれども、そういったところにも配慮が必要かなとも思われております。

簡単ですけれども、以上のございます。

○相澤主査

どうもありがとうございました。

では、前川委員、最後に一言よろしくお願いします。

○前川委員

言いたいことは先ほど大体話したんですけれども、この小委員会で今後、議論をどういう方向に持っていくのか、そこは主査、副主査、あるいは事務局の皆さんの思惑ももちろんあると思います。ただ、やはり日本の言語政策、文化政策にとって基盤的な部分に関わることだと思います。それで、昔にはできなかったことが今、情報処理技術の発達によって、許容されるコストで実行できるようになってきたわけですね。ですので、いい時期にこういう小委員会ができて一本当は10年、20年前にできていると私はもうちょっと楽できたんですけど一ちょうど今がいい潮時なんだろうと思いますので、長期的に見て日本の社会に裨益^ひするようなそういう知財、新しい文化財、社会にすぐに役立つことも含めた文化財みたいな、そういった新しい国民財産をこれから作っていく、その出発点をいい方向に定められればいいなと思います。

ただ、その中には、権利関係のを中心、かなり生臭い話もどうしても出てくるんですよ。生臭いの代表は権利関係と、先ほどの規範性の話の二つだろうと思いますけれども、そこら辺を私自身もはっきりした、こうすればいいというアイデアがあるわけではありませんので、いろいろ皆さん、この場で話し合っ、いいアイデアをまとめられればと思います。

○相澤主査

ありがとうございます。では、大体お時間となりましたので、以上で議論の時間を終了いたしたいと思います。言語の果たす役割、日本の言語資源の位置付け、コーパスの在り方、活用について様々な御意見を頂きまして、ありがとうございました。本日頂いた御意見は、やや大変かと思いますが、内容ごとに分類、整理して、次回資料としてお示しするという考えております。

また、議論の基盤を共有するためにも、事務局と相談をしまして、有益と思われる事柄について、委員の皆様を含めまして、有識者の方からお話を伺うことも次回

以降、また検討していきたいと存じます。

では、本日の言語資源小委員会はこちらで閉会いたします。ありがとうございました。