第9回 国語分科会言語資源小委員会（Web開催）議事録

令 和　7 年　9 月　9 日
15時 05分 ～ 17時 05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、武田、前川、森各委員（計7名）

（文部科学省・文化庁）山田国語課長、中田課長補佐、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

※ 相澤主査、石川副主査、神永、武田、森各委員及び事務局は、
5F5会議室にて参加。

〔配布資料〕

1 第8回国語分科会言語資源小委員会議事録（案）（委員限り）

2 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～
（神永委員御提出資料）

3 これからのデジタル言語資源の整備・活用を考える（石川副主査御提出資料）

〔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 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　文化審議会国語分科会）

〔経過概要〕

1 事務局から定足数の確認があった。

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 神永委員から配布資料2「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～」に基づいて説明があり、質疑応答が行われた。

5 石川副主査から配布資料3「これからのデジタル言語資源の整備・活用を考える」に基づいて説明があり、質疑応答が行われた。

6 二つの説明を踏まえ、コーパスやデジタル言語資源をめぐって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12月を予定しているが、具体的な日程については調整中であること、日程が決まったらすぐにお知らせす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9回、今期、3回目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に実際にお越しくださいまして、出席されている方々、オンラインで御参加の方々とがいらっしゃいます。また、傍聴者の方々はオンラインでこの会議に御参加されていますことを御承知おきください。

音声トラブルがありまして開催の時刻が遅れましたことをおわび申し上げます。

さて、本日は議事次第のとおり、「言語資源の整備と活用」について、「有識者による事例等紹介」、「その他」という項目で協議をいたします。事前に資料を御覧になっていらっしゃると思いますが、本日は、コーパスを活用する具体的な事例の御紹介ということで神永委員にお願いをしております。

また、石川副主査から、「これからのデジタル言語資源の整備・活用を考える」と題して、課題として見えてきたこと、進むべき方向性についてお話をいただく予定でおります。

それでは、議事に入ります。御紹介しましたとおり、本日は神永委員から「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～」という題目で御紹介いただきます。昨年度は、辞書編集においてコーパスをどのように活用し得るのか、一般論として整理いただいておりますが、今回はもう少し具体的な御紹介ということでどうぞよろしくお願いいたします。

○神永委員

よろしくお願いいたします。「国語辞典におけるコーパス利用について」ということで

お話をさせていただきます。

私、『日本国語大辞典』という辞典を編集しておりまして、その『日本国語大辞典』の事例を基にお話をさせていただきたいと思っております。ちょっと宣伝ぽくなってしまうかもしれませんけれども、お許しいただければと思います。

『日本国語大辞典』というのは、今の画面の左の方にあります緑色の表紙のもの、これが第1版でして、初版なんですけれども、それからその下のほうのちょっとオレンジ色の表紙のものがありますけれども、これが第2版です。私、この第2版の編集に関わってきました。そして第3版ということで、今改訂を進めているというようなわけです。

今日の話の内容につきましては、大まかに申し上げてこのようなことになります。まず一つ目は、辞書の用例について。辞書になぜ用例を載せるのかといったことを『日本国語大辞典』を中心にお話しさせていただきたいと思っております。

それからもう一つは、『日本国語大辞典』第3版、今編集に掛かっておりますけれども、これのコーパス活用についてということで話をさせていただきます。

皆さん、もう既に御存じだと思うんですけれども、『日本国大辞典』について簡単に歴史と言いますか、このような流れで編集が行われてきたんだということをお話しさせていただきたいと思います。

『日本国語大辞典』は、元になりました『大日本国語辞典』というのが－もう戦前の辞典なんですけれども－ございまして、それの改訂というよりは、本当に全面的な新しい辞典を作るということで、戦後になりまして編集を始めまして、1972年から73年の間に初版、全20巻の辞典を刊行いたしました。その後、第2版を2000年から2002年に刊行しております。つまり、30年ほどかけて刊行して、改訂の作業をしてきたというわけです。

初版は元々活版印刷であったんですけれども、第2版では初版のデータをOCRで読み取りまして、ある程度電子的に処理が可能になったということで編集を行いました。そして組版のほうも、当時としては最新の技術を駆使しまして行っております。

今回の第3版では、更に進歩したもので編集を進めていきたいというようなことを考えております。

一応大体今までの流れ、今後の流れと言いますか、今後どのような形になっていくかということをお話ししておきます。初版から先ほども言いましたけれども第2版までの間に約30年、それから第2版から第3版に掛けても30年というようなことで刊行を

考えております。 20年、30年というのが特に意味はないんですけれども、大体編集のノウハウと言いますか、そういった技術的なことを次の世代に伝えるには、ちょうど20年、まあ、30年たっちゃうとちょっと大分怪しくなってくるんですけれども、できれば少なくとも20年ぐらいで改訂をしていきたいなというようなことは考えて編集をしてきました。

さて、『日本国語大辞典』の場合には、用例主義ということで編集を進めてきたんですけれども、今回の第3版におきましても用例主義ということで編集を進めていきたいと考えております。

初版から用例主義を採用していますけれども、初版の「発刊の辞」を見ますと、国語大辞典の生命は、まずその用例文にあるとうたっているんですけれども、膨大な文献資料から語の使用例、用例を採集しまして、それを基に立項して、解説の－語釈ですけれども、執筆を行ったということです。初版や第2版と同じく第3版でも、先ほど申し上げましたように用例主義を継承していこうと考えております。

一般的な国語辞典ですと、作例と言えるものを掲出しております。作例というのは、辞書編集者が内省で作る例が多いんですけれども、そういったものを主に入れるようにしております。実例もあることはあるんですけれども、これはかなり少数ではないかと思います。時代とともに、こういった作例は、コーパスを使って、これを応用した形で作例を作るというようなこともやってまいりました。これに対しまして、『日本国語大辞典』では基本的には実例を掲出するというような方向でやっております。

そういった中で、やはり私たちの『日本国語大辞典』でやってきたものは、一つは古典資料、それから近現代の資料、さらにはデジタル化された資料、こういったことを検討したり活用したりして用例として掲出したいと考えております。膨大な用例をどのように採集するか。国語研のコーパスなどデジタル化された資料をどのように活用するかということが課題になるかなと考えております。

『日本国語大辞典』の用例というのはどのような形で引用されているかということをちょっと御説明しておきますと、まず、実際に小説、詩歌、随筆などの作品や、古辞書、歴史資料など、こういったもので使われた例を載せるようにしております。大体資料数としては3万点ぐらいなんですけれども、3万点の資料の中から100万用例を収録するというようなことをしております。一番古い例は聖徳太子の時代のものなんですけれども－これは実はちょっと日本語的にどうかなと私も思ってはいるんですけれども－聖徳太子の時代に書かれているお経の注釈書のようなものがあるんです。そういったものを

古い例として採っております。それから、一番新しいものとしましては、ちょうど第2版が刊行されたのが2000年だったものですから、昭和末年のものまでを典拠として示すということをしております。

そして、『日本国語大辞典』における用例採集の目安なんですけれども、どのようなことで用例を採ったかと言いますと、まず一番古いと思われるものということなんですが、いわゆる初出例を意識してはいるんですけれども、最古の例とは必ずしも言えないかなというようなことは考えております。それから、見出し語の意味を酌み取りやすいもの。取りあえずその語の存在証明として引用している用例もあるんですけれども、割と意味が取りやすいもの、同じ時代のものがあったとしたら、意味が取りやすいもの－酌み取りやすいものを中心に引用するようにしています。それから、主要作品の用例ということで考えているんです。主要作品って、じゃあ一体何というような問題ももしかしたらあるのかもしれませんけれども、一応できるだけ有名どころの作品、古典なんかもそうなんですけれども、有名どころの作品は載せられたら載せたいということを考えておりました。それから、その後の文字遣いの特殊なもの。これは例えば、堅固で丈夫なという意味の「がんじょう」とありますけれども、ここにありますいろんな表記があるんですが、できるだけこのような様々な表記の用例を取るようにするということも考えておりました。それから、その他の意味で適切なもの、つまり派生義なんですけれども、多義語の場合に、いろいろな展開をしていくわけなんですが、そういったものをいろいろと探しまして、掲出するということをしております。

これは第2版での用例採集の方法なんですけれども、一応当時は二通りの方式をやっておりました。一つは、いわゆるスカウト式用例採集と言っていましたけれども、資料を選びまして、この資料から用例を取るということを選定するわけなんです。その資料を基にして採集者の判断で用例を採取するというような方式でした。採取した用例は、カード化して、それを一つ一つ入力していったというようなことをやっております。それから、もう一つは、重点的に用例を採取すると決めた資料が幾つかありまして、これは編集部で独自にその本文を入力してテキストを作ってしまいました。ただ、残念だったんですけども、当時コーパスという発想がほとんどなかったものですから、それを活用し切れなかった部分もあるんですけれども。いずれにしましても、全ての文章を入力して、ここからここまでは用例部分として示しますよというような記号を付けながら、印を付けながら、見出し語の候補として示すといったようなことをやっており

ます。当然のことながら、それは読み仮名がなかった－ないものにつきましては、資料を見た人間が読み仮名、読みを付けるというようなことはやっております。そういったものは、いわゆるKWIC方式、検索方式、キーワードを前後の文脈つきで表示する方式ということで行っておりました。

これはこんなカードを作ったんだよというような例なんですけれども、第2版ではこういったようなカードを作りながらやりました。それから先ほども申し上げましたように、全文を入力して、その中からここの部分だけを取るというようなことを表示するということもやっておりました。

第2版での用例採集法の続きの話です。採取した用例データ、用例の採集の仕方は2種類あると申し上げましたけれども、これを初版の全データと言いますか、全文を－当時は活版印刷だったものですから－OCRで読み取りまして、読み取ったものとそれから用例データとをマッチングするというような作業をしました。そして、その中からマッチ項目とアンマッチ項目ということに分けて処理するというようなことを行っています。マッチ項目に関しましては、用例の追加候補になったわけですし、それからアンマッチ項目ということになりますと、これは新立項項目の候補になったというようなわけです。どちらかというと資料優先、用例優先というような形だったかなと思います。

第3版のお話を少し先にしてしまいますと、第3版では資料優先で用例を採取するものもありますけれども、多くのデジタル化資料を使って幅広く用例を採取する方法を採りたいと考えております。だから、かなりいろんな用例が集まってくるんじゃないかなというようなことを想定しております。デジタル資料に関してそれだけ期待も大きいんですけれども、あえて言うならば、これはどうしようかなと判断に迷うようなものが一杯出てきてしまう可能性もあるのではないかと思われます。

辞典に用例を付けることの効用と言いますか、何でそんなことをしているのかと言いますと、これは『日本国語大辞典』の初版と第2版の編集委員をされていた松井栄一先生がおっしゃっていたことなんですけれども、実際に辞書に用例を載せるとこんなことが分かるというような話をいたします。その語が確かに使われていたという証拠になるということ。それから、その語の使われていた時代が分かる。その語の意味の理解を助ける。その語の使い方の手掛かりが得られる。その語の読みに関する資料となる。その語の表記に関する資料となる。こういった六つの事柄を松井先生はおっしゃっています。

確かにそのとおりなんですけれども、実は私は個人的にもう一つ追加したいなと思っていることがございます。個人の感想みたいなものなんですけれども、ちょっと一言言わせていただきますと、いわゆる誤用ですね－従来なかったものの使い方、用法がどの程度広まっているのかというようなことがこの用例採集によって分かるというようなことが付け加えられるのではないかと思っております。

例えばどんなことかと言いますと、これがそうなんですけれども、「的を射る」、「的を得る」という問題がありまして、『日本国語大辞典』、割と早い時期と言いますか－これ第2版なんですけれども、「まとを得（え・う）る」、用例があったものですから、項目として載せたんですね。これは高橋和巳の『白く塗りたる墓』という小説に出てきたものですから、載せたんですけれども、実は「的を得る」に関しましては、辞書の中でもちょっと判断が割れているものなんです。

どういうことかと言いますと－具体的に申し上げますと、ちょっと辞書の具体名が出てしまいますけれども、例えば『明鏡国語辞典』第3版－『明鏡国語辞典』の最新版なんですが－「『当を得る』や『要領を得る』との混同から『的を得る』とするのは誤り」と言っております。ところが、『三省堂国語辞典』の第8版、これも『三省堂国語辞典』の最新版なんですが、「『的を射る』『的を得る』とも、戦後に広まった言い方。『得る』は『要領を得る』などと同様、『うまくとらえる』の意味。同義語『正鵠を〈射る／得る〉も二通りの言い方がある』と言っていまして、『三省堂国語辞典』の方は「的を得る」の方を認めるという立場になっています。

こういったような辞書によって判断が割れるものがあるんですけれども、『日本国語大辞典』の場合には、高橋和巳の用例がありましたから、載せているんですけれども、『三省堂国語辞典』の注記を基にして少し補いたいな－語釈を補いたいな、用例を補いたいなというようなことを考えております。例えば「的を得る」の表記で、「得」と書くか、平仮名で「える」と書くかなんですけれども、これをBCCWJで検索しますと、29例出てきます。作者名は表示しませんでしたけれども、1989年、92年、94年といったものが用例として見付かるというようなことですので、こういったものを使って補っていくことが可能なんじゃないかなと考えております。

「的を射る」の方を見ますと、『三省堂国語辞典』の方も、「射る」も「得る」も戦後に広まった言い方と言っています。実際に的に矢を当てるということに関しましてはもちろん古い例があるわけなんですけれども、的確にその要点を捉えるというもう一つの意味、

今話題にしている意味ですけれども、これの用例が『日本国語大辞典』も、「島崎藤村」という平野謙の評論にしか初出が採れませんでして、これをどのぐらい遡れるのか、できれば戦前まで遡りたいなというようなことを考えております。デジタル化資料などがありますから、そういったものを基にしてあるいは採れるのではないかなとは思っているんですけれども、ちょっと現時点ではまだ見付けられていないような状況です。

用例の話長くなるんですけれども、実は採取した用例というのはそのまま引用するわけではありませんで、出典検討というのを行っています。これ実はちょっと私の個人的なことを言わせていただきますと、『日本国語大辞典』で出典検討というのは私が担当していた部分でして、どんなことかと申しますと、前提としましては、用例と語釈は不即不離の関係にあるということですので、原稿の執筆者が採取した例を決められたテキストで当たり直す作業なんですね。これが出典検討です。

文字、仮名遣いを点検し適当な長さに引くという、ちょっと校正的な部分もあるんですけれども、それから、決められた注記、ジャンル名とか、それから小題、番号、歌なんか番号が付けられていますけれども、こういったものを入れたりとか、作者名を付記したりとかすることをしています。

それだけではありませんで、実はもっと大事な仕事がありまして、出てくる、採取された用例というのは、時代を問わず、ジャンルを問わず出てきますけれども、その意味をしっかりと読み取りまして、語釈と併せて考えて、その語釈の例として適切かどうかということを判断するような作業までしております。

そういった意味で、『日本国語大辞典』の場合には、用例を大事にするというのは、出典検討までをした上でのことだということを御理解いただければと思います。

そして、これはちょっと小さくて見づらいかもしれませんけれども、初版の時の原稿なんです。当然のことながら手書きで書かれているものなんですが、「あいおい」というものの原稿です。升目が書いてありますけれども、そこに解説が書いてありまして、そこの下のところにちょっと空白で欄が設けられています。よく見ると、「岩文」とか、「黙阿弥」とか何か書かれているんですが、これが実際に出典検討したよという証拠になっているわけです。つまり、用例を示すための決められた底本があったんですけれども、それをちゃんと見ているよということがこれで分かるわけです。

第2版になりますと、ゲラ上に初版の記述が表示されますから、それにこういった形で採取された新しい用例をこのようにどんどん付け足すというようなことをしております。

ここでは洒落(しゃれ)本の新用例が多いんですけども、洒落本の底本は主に洒落本大系なんですけれども、これを基にしてと言いますか、ちゃんとこれを見ているよということをここで分かるようにしております。

第3版になりますといろいろと変わってくるだろうとは思います。実は用例を採取するために編集部では資料をいろいろと集めておりまして、この資料を集めて、全部、書庫的なものが現在も神田のすずらん通りにあるビルに保存されております。ひょっとしますと、この第3版においてはこういった資料を集める必要がなくなるかもしれないなというようなことをちょっと考えております。デジタル化された資料が増えてきていますから、そういったものを使うことによって、こういった資料をいつまでも保存しなければいけないということはなくなるわけです。もちろん初版、第2版で使った資料は処分するわけにはいかないんですけれども。私はこれらの資料を保存するために絶対に捨てないでくれと会社にお願いして、あちこち引っ越しの度に一生懸命守ってきました。これがなくなることはないだろうと思いますが、大幅に増えることも今後はないかもしれないというようなことは考えております。

それから、右側の写真なんですけれども、これはちょっと見づらいんですが、『日本国語大辞典』の『平家物語』の用例を引用した底本です。岩波の古典文学大系の前の版の見返しなんですけれども、ここに手書きで一杯いろんなことが書かれています。これはどういうことかと言いますと、『平家物語』は御承知のようにいろんな本があるわけなんですが、『日本国語大辞典』の『平家物語』の第1底本というのが岩波の古典文学大系本なんです。読み本系のものなんですけれども、これで用例がもし採れなかった場合には違う本で探してくださいということ－指示がここに書かれています。見づらいんですけれども、第2底本としまして屋代本という本を表示しています。以下、ずっといろんな本が書かれていますけれども、2番目の底本にしているものがもしなかった場合には、さらにその先、駄目だったらまたさらにというようなことでやるというようなことをやっていきます。これもまた出典検討の仕事なわけです。

第2版の用例の課題と展望なんですけれども、実は全ての時代の用例を満遍なく採取しているわけではなかったんです。

それから、実は、用例主義と言いながら、見出し語を立てていながら、用例のないものが多少あります。ちょっと恥ずかしい話なんですけれども。

それから、初出例を載せるように心掛けていますけれども、実際には更に古い例が

見付かることも多々あります。今ももっと古い例あったよというようなことをいろんな方から御報告いただいているわけです。

それから、第2版までは紙版が中心だったんですけれども、今、デジタル版もあります。取りあえず最初は紙版でやっておりましたので、分量の制約があったんですが、そうしますと、やはり基礎語なんかになりますと、ちょっと近代例が手薄になっちゃっている場合があるんです。これは言ってみれば、あっ、何だこの語ってもしかしたら江戸ぐらいでしか使われないのかというようなことにもなりかねませんので、実際には明治、大正、昭和まで使われているんだよということを示さなきゃいけないわけですから、そういったものを何とか補いたいなということを考えております。

それから、もっと広く取りたいなということを考えておりまして、当時はやりませんでしたけども、漫画からの用例をどのように扱うかというようなことも検討したいと考えております。

それらを解消するためにコーパスや外部のリポジトリを活用するということを考えているわけで、第3版へ向けてのコーパス利用という話になっていくわけなんです。ちょっと話を戻しますけれども、前回と言いますか、昨年、私、国語辞典においてBCCWJをコーパスはどのように利用できるかということをお話しさせていただきましたけれども、じゃあ、『日本国語大辞典』ではコーパス－BCCWJなどを中心とするコーパスをどのように活用できるかということを考えてみますと、一つは、頻度数によって見出し語選定の参考にできるということ。

それから、記述する意味の説明の参考になる。多義語の場合ですね。こういったものは非常に有効かなと思っております。

それから、もちろんですけれども、新立項項目の参考にもなります。それから、表記に揺れのある語の見出し表記を決める根拠が得られるということで、外来語なんかもこれ恐らく中心になっていくかなということを考えております。

それから、初出にはならないにしても、例ナシ項目－先ほど申し上げましたけれども、多くはないけれども、存在するということなんですが－これの用例を探すことができると思っております。

それから、ある語と連接しやすい語の特定が可能になるということで、典型的な使用例。コロケーション。これ実は『日本国語大辞典』、ちょっとコロケーションの典型的な使用例の用例が私から見ると手薄だったのかなというような気もしておりますので、

これも補えるのではないかなというようなことを考えております。ちょっと夢を語ったみたいですけれども。

『日本国語大辞典』第3版の改訂の要諦です。第2版は校正紙の段階からデータへと移行させまして、書籍版から電子版という形を取っていましたけれども、今度は、書籍版中心だった第2版と同じ形ではなくて、電子版中心にまずやって、それから書籍版という流れを考えております。ペーパーベースドからデジタルベースドに移行するというようなことで、編集作業も、提供の形態も、デジタル技術を駆使して進化させるということを考えております。

編集作業の進化ということで申し上げますと、まず、コーパスなどの外部リポジトリ、国語研のコーパスとか、それ以外のものもありますけれども、そういったものを基にしましてデータ収集を行います。

それらを駆使するために日国用のコーパスの検索システムを開発したいと考えております。国語研のコーパス以外も、ほかの文献資料からも用例を採集しまして、こういった語彙項目のデータベース、用例のデータベース、出典のデータベースといったデータベースを作ります。これは様々なデジタル化資料を活用することになるわけなんですけれども、必要な資料をコーパス化していきまして、独自の検索システムを作成した上で、編集支援システムへと連携させていくというような流れになります。資料から用例とその出典情報を採集しまして、編集支援システムに取り込むということです。

さらに、編集支援システムというのは、我々編集部の中でのデータ編集になるわけなんですけれども、用例を基にその語の意味や語釈、あるいは既存の語釈を新しく採集した用例を基に修正していくということです。それを組版プログラムに乗せていきまして、エクスポートしたXMLをプログラム処理で自動組版にしていくというようなことで、そこの中に先ほど申し上げております出典検討の作業も入ってきまして、校正システムに乗せていくというような流れになっていきます。

編集部の考えといたしましては、『日本国語大辞典』を知的情報のハブとして考えております。まず国語研のコーパス－BCCWJですね。それから、先日お話しいただきましたけれども、国文学研究資料館の様々な出典原本の影印もございますし、それから出典情報の書誌情報などもありますけれども、これを活用させていただく。それから、国会図書館のデジタル化された資料もありますので、これも活用していく。それ以外にも、ほかのデーターベース、他の機関のリポジトリなども活用しながら、これを

どのように運用するか、登載するかということなんです。これを、ジャパンナレッジというサイトがあるんですけれども、ここへ載せて、様々な形でリンクが貼れればいいかなというようなことを考えております。

『日本国語大辞典』で引用された用例を基にして、国語研、国会図書館、国文学研究資料館といったところでお持ちの資料に戻れるというようなことにできればいいかなと考えております。辞書を入り口とした多種多様な知的情報への動線をデザインしようというのが私どもの『日本国語大辞典』編集部の今の考えであるということです。

以上で私の説明は終わりにさせていただきます。どうもありがとうございました。

○相澤主査

神永委員、ありがとうございました。

それでは、ただ今の神永委員の説明につきまして何か御質問がありましたらお知らせください。挙手にて、あるいは御発言にしてお知らせいただければと思いますが、いかがでしょうか。

○石川副主査

大変貴重なお話をありがとうございました。言語資源を縦横に活用する日国第3版に大いに期待を持ったところでございます。それを踏まえまして三つ小さな質問をさせてください。

1点目として、神永委員は用例カード作りという伝統的なスタイルと言語資源を縦横に活用する新しいスタイル、両面を使って辞書作りをされてきたわけですが、これから業界に入ってくる若い辞書編集者の方々を考えましたときに、実際のお仕事においては、用例カードを手作業で作っていくという段階を飛ばして、いきなりデジタルで始まっているという実態があるわけです。こうした現状の中で辞書作りのノウハウの継承に関して、何か失われるものがあるとお考えでしょうか。また、もしそうであるならば、そこをどのように補完していくべきなのか、この辺りについてお考えがあれば御教示ください。

○神永委員

そうですね。カードを作るという作業は恐らくなくならないとは思っているんですよ。ただ、それをある程度やり方を理解している、それができる人にしかもうできなくなって

いるのかもしれませんので、新しい人たちにはそこまで求めるのはちょっと難しいかなとは思います。

だから、先ほど申し上げましたように、幅広く、この例を取りたいではなくて、ここからこれだけ用例として使えそうなやつが上がってきたよという、その中から選ぶという、今度は別の能力が必要になっていくんじゃないかなとは思います。

○石川副主査

ありがとうございます。2点目として、先に見出し語を決めておいて、それをコーパスで調べるだけでなく、コーパスから、これまでなかった見出し語を新たに発掘していくこともあるかと思うのですが、そういう場合はどのようにデータを使っていくことになるのでしょうか。

○神永委員

それは見出し語データが元々『日本国語大辞典』の場合50万あるんですけれども、それと、コーパスなどから採取した立項するための語彙の候補のデータができますから、データ同士をマッチングさせるんですね。マッチングさせますと、マッチしたものとアンマッチとできるわけなんですけれども、そのアンマッチのものが新立項の候補になると考えています。

それは実は第2版のときでもかなりそれに近い形でやっておりました。ただ相手が50万項目という膨大な数でも、アンマッチはものすごい数が出てきちゃうだろうと思いますので、それの選定をどうするかということも大事な仕事になっていくかなとは思っています。

○石川副主査

ありがとうございます。3点目、最後の質問ですが、歴史主義の辞書と伺いますと、日本語では日国、英語ではOED（Oxford English Dictionary）が思い浮かぶわけですが、日国の編集において、OEDをベンチマークされる、ないし参考にされる、というようなことはあるのでしょうか。

○神永委員

最初のときにはやはりOEDを目指してみたいなところはあったんですけれど、もう今の段階では、本当に私たち独自でやっていく。ちょっと私もOEDのことを詳しくは知りませんけれども、恐らく英語のものというのは、イギリスのものだけじゃないですよね、当然のことながら用例として取るものは。ただ、日本の場合には、飽くまでも国内で成り立ちますから、それほどOEDほどは大変ではないのかもしれないです。

○石川副主査

分かりました。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。ほかはいかがでしょうか。

私から一つ、最後から3枚目なんですけれども、編集作業の進化、DXという形で、編集のシステム自体も今回開発をされるとお伺いしまして、こういうシステムがあれば、30年に一度ではなく、定常的に辞書編集作業をしていくことも考えられるということでしょうか。

○神永委員

おっしゃるとおりです。

○相澤主査

継続は重要というような。

○神永委員

年に何回もというのは無理かもしれませんけれども、ある程度、1年に1回とか、2年に1回ぐらいの形の常時改訂というのは今後可能になっていくだろうなと思うんです。

その場合にちょっと問題もあるんですけれども、前のバージョンをどうするのかということがありまして、それから、その版を何と呼ぶのかという。3.1バージョンなのか、3.2バージョンなのかという、それもまた考えなきゃいけないですよね。

○相澤主査

分かりました。

○神永委員

ただ、おっしゃるとおり、かなり頻繁に改訂は可能になってきます。そうすると私たちも非常にやりやすいんですけれども、今年はここの部分だけやろうとか、そういうことが可能になるので、今年は本当に江戸時代だけ中心にやってみようとか、それから、今まで小説ばっかりだったけども、ちょっと随筆なんかも徹底的にやってみようとか、何かそんなことができるはずなので。

それから、新しいコーパスなりデータベースが、言語のデータベースができたときには、それを基にして使わせていただけるのであれば、それを参考にしてというか、そこの部分を取り込むこともできるかもしれないので。だから、私たちの作業は今後のコーパスの発展ともしかしたら不即不離の関係にいくのかもしれないんです。

○相澤主査

ありがとうございます。バージョン3.09.85みたいな、そんな形で、30年後にもう1回、大改訂、バージョン4になるみたいなイメージ－とても面白いですね。

○森委員

伺ってよろしいですか。改訂する場合は、更新なんですか。元のものが消えてしまうのかどうか。

○神永委員

『日本国語大辞典』の場合には、ほとんどこの項目は余り使われてないから消そうということはありません。小型の国語辞典ですと、やっぱり分量の問題もあるので、今使われなくなった項目は削除しようというので、話題になりますけれども、辞書から消された言葉みたいなのもありますけども、『日本国語大辞典』の場合には、基本的に項目は消しません。ただ、底本を変えるときがあるんです。結構古い時代に刊行された底本を使って用例を採っていたことがあるんですが、それが最近の研究成果で新しい本が出てきますと、それに当たり直したことによって、底本の方でそれを読み違えていたとか間違っていたことによって、そういう言葉があったと思っていたはずなのに、その言葉自体

がなくなっちゃったということがあります。そのときは消します。

○森委員

紙の辞書であった場合は、第2版と第3版を並べれば変わった部分が分かると思うんですけれども、データで更新作業がされていくようになると、3のときはこうだった、3.1でここが変わったというのが分からないかもしれなくて。

○神永委員

そうなんですよ。そこなんです。

それをどうやって残すのか。

○森委員

3のときはこうでしたが、3.1ではこう変えましたというようなことが注釈で出てくるといいんですけれど。

○神永委員

残せるかどうかなんですね。でも、それは膨大なデータ量になっちゃうので、どうしようかという問題があるんですね。

○森委員

そうですね。

○神永委員

もちろん、今、デジタルで展開している辞典もあって、あまりそういうこと気にせずにどんどん更新しているのもありますけれども、『日本国語大辞典』の使命としては、もしかしたらそれはこのように変わったんだということが分かるようにするのも大事かなとは思っています。

○森委員

紙の辞書は変わったときに割と大々的に改訂版が出ましたよというお知らせがある

んですが、電子版は、いつの間にか変わってる？ということがあって、変えたら変えた、その判断の理由も知りたいなあと利用者としては思います。

○神永委員

ありがとうございます。課題として。

○相澤主査

ありがとうございます。では、前川委員、よろしくお願いします。

○前川委員

どうも面白い話ありがとうございます。いろいろな面でDX化をお進めになるという話でしたが、一番中心的な課題である語釈とか語義の分類のところですよね。そこを例えば、今、AIがそこに手が届きつつあると思うんですけれども、今後そういった可能性は検討されるのかどうかということをちょっと伺いたいなと思いました。

と言いますのは、つい先日ですけども、国語研で青山学院の近藤先生に御講演いただきまして、その内容が正にそれだったんですよね。『源氏物語』のテキストを与えて、何だったかな、語釈を完全に自動でAIにやらせる。そうすると出てきたのは、かなり妥当なと言いますか、間然するところがないようなものが出てくるという実例。これは一つの事例ですけども、ともかくそこまで今来ているんだということをお示しになって、聴衆が大いに驚いた状態だったんです。今後、急激にそれが進むと思うんですよね。また、それが人間との関係ということもありますけれども、ただ、作業を大いに効率化するであろうということは期待できる。そこら辺、何かお考えになっていることがありましたらちょっと教えていただければ幸いです。

○神永委員

近藤先生は『日本国語大辞典』第3版の編集委員です。まだ編集部の方ではAIを使ってという話は具体的には出てきていませんけれども、恐らく近藤先生の方からそういう話が編集部の方に来ると思いますので、そのときには、そういったAIを活用した語釈を作るということももしかしたら検討されるのではないかと思っております。

ただ、現時点ではまだそこまでは編集部の中では考えていないと思います。

○前川委員

分かりました。ありがとうございました。

○相澤主査

ありがとうございました。

神永委員、大変貴重な御講演を頂き、どうもありがとうございました。

続きまして、配布資料3「これからのデジタル言語資源の整備・活用を考える」について、石川副主査からお話を頂きます。どうぞよろしくお願いいたします。

○石川副主査

それでは、始めさせていただきます。今期の当小委員会におきましては、「言語コーパスなどデジタル言語資源の取組の進捗状況」、「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」の4点が検討課題の例として挙げられております。本日はそれらを整理しまして、また、事前配布資料の順番を少し変更させていただきまして、論点1「言語資源と外来語」、論点2「言語資源の整備」、論点3「生成AI時代の言語資源」の3観点にまとめてお話をさせていただきたいと存じます。

論点1「言語資源と外来語」

まず、論点1「言語資源と外来語」についてであります。外来語を論じる際には、まずもって、2003年から2006年に行われましたいわゆる「外来語言い換え提案」が記憶に上るところでございます。

このプロジェクトは、高齢者の介護や福祉に関する広報誌の記事などに片仮名語が氾濫しており、肝心の、高齢者にとって非常に読みにくいものになっているという現状認識を出発点とするもので、特に高齢者とのコミュニケーションにおいては、外来語の使用を適切に考え直していく必要があるのではないかという重要な問題提起がなされました。外来語を適切に見直すことは、国語審議会が従前より申していた「平明で、的確で、美しく、豊かである」日本語というものにもつながっていくと考えます。

言い換えの趣旨でありますけれども、外来語を一切使ってはいけないと言っているわけではありません。例えば、行政文書であるとか、新聞であるとか、そういう公共性の高い文章においては、様々な読み手がいることを念頭に、様々な人々に伝わりやすい表現を考えていく必要があるわけで、そうした文脈の中で、外来語の適切な使用を考え直すということが目指されていたわけでございます。資料にもありますが、まずは、官公庁、報道機関、省庁、新聞などを対象にしたプロジェクトであったわけです。

調査方法については、「国語に関する世論調査」、つまり、国民の意識調査が行われました。調査は4回に分けて行われ、外来語を見てもらい、どのぐらい分かっているか、あるいは使われているかをアンケートで調べました。

その結果に基づき、各種の外来語を、理解している国民が全体の4分の3を超えるもの、4分の3程度のもの、半数程度のもの、4分の1未満のものに区分しました。このうち、理解している国民の数が4分の1に満たない外来語を取り出し、これらを公的場面で用いることに再考を促しました。現代から見ても、大変に価値のある、また、意義深いプロジェクトであったわけです。

2002年の時点におきましては、資料に挙げているような外来語が特に理解度が低いものとなっておりました。「インキュベーション」、「エンフォースメント」、「コンソーシアム」などが並んでおりますが、2025年時点で改めて見てみますと、かなり一般化したものもありますし、そうでないものもあります。ともあれ、2002年においては、これらが特に理解度が低かったわけです。

そして、国民全体でも、また、60代以上においてもともに理解度が低かったものを取り上げ、いわゆる言い換えの提案がなされていたわけでございます。対象になったのは、「アクセシビリティー」、「インフォームドコンセント」、「ネグレクト」、「ビオトープ」、「フリーランス」、「ワンストップ」などの語でした。

このように、理解度を直接アンケートで調べるという方法に加え、現代であれば、コーパスを使った調査も考えられるでしょう。2000年代前半においては使用できるコーパスが非常に限られていたわけですが、今は多様なコーパスができていますので、それらを組み合わせることでより詳細な調査ができるかもしれません。

さて、外来語を論じる際には、様々な論点があります。一つは表記揺れの問題です。資料にございます小椋先生の研究を御覧ください。表記揺れにも様々なものがあり、「カテゴリー」と棒引きで伸ばすのか「カテゴリ」と切るのか、「メール」と書くのか「メイル」

と書くのか、「スムース」とするか「スムーズ」とするか、幾つかのパターンに問題を分けることができます。この調査では、現代日本語の外来語の表記揺れを複数のパターンに分け、BCCWJ1のデータでどちらが多いか検証しています。ただ、語末長音の有無の問題だけを見ても単純には行かず、単語によって傾向はまちまちです。「モニター」、「ユーザー」、「ドライバー」などであれば長音が付きやすい一方、「ヘア」、「ブラウザ」、「セキュリティ」となると長音が減ってくることがコーパス調査で明らかになっています。

また、野崎氏による最近の論文では、先行研究で余り注目されていなかった新しい外来語に注目し、朝日新聞データベースを使った調査が行われています。

資料を御覧いただきますと、「レシピ」が2004年以降、「スマホ」が2013年以降、「ウェブサイト」と「ブログ」が2005年以降から増えていることなどが分かります。「ホームページ」と「ウェブサイト」を比較しますと、かつては圧倒的に「ホームページ」だったわけですが、徐々に「ウェブサイト」に置き換わっている実態も新聞調査から明らかにされています。

こうした大型の言語資源を使った調査というのは、今後、外来語の問題を考える際に、重要な基盤になるものであろうと考えます。

さて、外来語については、直近の2023年、2025年にも意識調査が行われており、その報告も上がっています。

例えば、一般と行政と比べますと、行政の方が外来語を好んで使うという傾向があります。恐らく何らかの新しい施策を行うときに、外来語で概念を目立たせたい、場合によっては、そのことによって予算取りを行いたいといった事情もあろうかと考えます。

以上を踏まえまして、今後の言語資源と外来語研究の関係について整理したく思います。先ほど御紹介した言い換え提案が2003年、2004年頃になされ、それから既に20年以上が経過していることを考えますと、今後は、信頼できる大規模言語資源を定期的に拡張しながら、より短いスパンで、外来語の使用実態の変化を定期的に観測していく必要があろうかと思います。国語分科会で検討していくべき事柄かもしれません。

そうした調査の結果として、新しい言い換え提案の方向に話が進んでいくかもしれませんが、私見では、そこまで行かずとも、この数年、どんな外来語が増えてどんな外来語が減っているのか、どんな外来語が新たに生まれているのか、こうしたことを信頼できる資料から明らかにし、定期的に社会に発信していくだけでも十分意味のあること

だと考えます。こうした情報発信は、国民の外来語問題に対する意識を高めていくきっかけになります。また、メディアや行政がそうした情報に触れる中で、外来語の安易な使い過ぎという現状に一定の見直しがなされる可能性もあるでしょう。

外来語問題の議論は、ややもすれば、外来語の全面否定といった極端な主張につながりがちですが、先ほども申し上げましたように、目立つ外来語によって新しい概念を打ち立て、新しい政策課題や社会課題としてハイライトしていくことは外来語の重要な機能です。一概に外来語が悪いというわけではなく、初期の言い換え提案にもありましたように、様々な国民各層に正確に伝わることが重視される場面において、使用に問題のない外来語や、使用時に検討が必要な外来語などの整理がなされていくとよろしいのではないかと考える次第です。

繰り返しになりますが、外来語のように、絶えず有りようが変化していく事象について、エビデンスド・ベースドで議論していく場合には、コーパスの継続的更新がやはり重要になります。外来語の定点観測のような事業を何らかの公的な形で位置付けることができれば、BCCWJのような大規模コーパスの継続的拡張に対して国の支援を要請する根拠にもなろうかと考える次第です。

さて、言い換えプロジェクトに関与された国立国語研究所の相澤先生に、今回の報告に先立ちましてお考えを伺ったところ、五つのことを御教示いただきました。

まず、当時は意識調査の形で行ったわけですけれども、現在もう一度この問題を扱うのであれば、言語コーパスの活用は不可欠になるだろうということでした。ただし、コーパスがあれば全て解決するというものではなく、コーパス調査と従前の意識調査をうまく融合させていく方法論に留意すべきだろうという御助言も頂きました。

また、外来語問題のように、社会の現実的な問題の解決を目指す研究においては、問題の所在の把握、優先すべき課題の見極め、対処方法の適切な選択を念頭に置くべきであるという貴重なコメントも頂きました。

さらに、調査結果を社会に発信する際に、社会にどう受け取られるかを考慮することの重要性についても御指導いただきました。いわゆる言い換え提案については、本来は「分かりにくい外来語を分かりやすくするための言葉遣いの工夫」であったものが、短く「外来語言い換え」と言ったために、この短い名称だけが広まってしまい、あらぬ誤解を招いた面もあったのではないかというコメントも頂いたところであります。

こうしたコメントとともに、相澤先生には、今後の外来語調査へのデジタル言語資源

の活用について期待をお示しいただいたところでございます。

論点2「言語資源の整備」

続きまして、論点2「言語資源の整備」という問題を扱いたいと存じます。ここでは、社会課題解決のための言語資源の在り方、あるいは地域の言語資源の活用という点も含めてお話をいたします。

言語資源については、幾つかのレベルがあることに留意する必要がありましょう。一つ目は、国語研などの機関による大規模公開コーパス、1億語を超えるような巨大コーパスで、公的な性質の強いものです。二つ目は、研究所や大学レベルで構築する中規模の公開コーパスです。三つ目は、個人研究者が個人の研究のために作っている小規模で、多くは非公開のコーパスです。コーパス言語資源ということを考える場合、第1階層、第2階層にばかり目が行きがちですが、同時に第3階層にも適切な目配りをしていく必要があろうかと思います。

では、第3階層のコーパスの実相はどうすれば調べられるのでしょうか。一つの方法は、科研費のデータベースを調査することです。これにより、科研費研究においてどのようなコーパスが作られているか網羅的に調べることができます。

今回の報告のために行った簡易調査の結果ですので、若干数字が変わるかもしれませんが、ざっと見たところ、科研データベースに入っている研究課題の中で、課題名に「コーパス」を含むものが約800件近くあり、2021年以降に限りますと126件程度となります。つまり、この5年間で、1年当たり平均25件程度のコーパスが作られているということでございます。

次に、126件の研究課題の「研究概要」をテキストデータとして保存し、平易なマイニングをいたしました。資料にありますように、当然ながら「コーパス」という語が大きく出ておりますが、そのほか、「言語」、「研究」、あるいは「日本・語」などがコーパスを使った研究のキーワードとして上がってきているところでございます。また、「発話」、「教材比較」、「文法」などの語も含まれ、コーパスを作る人や研究する人が何に関心を向けているかが分かります。

科研の費目ごとに見ますと、基盤Aが約10件、基盤Bが約10件あります。基盤Cがやはり多く、54件です。加えて、若手や萌芽（ほう）などでもコーパスに関連する課題が存在します。

簡易なものではありますが、これらの126件を内容別に整理してお示しすることで、個人研究者が構築している第3階層コーパスの実態を概観したいと存じます。ただ、この分類は私的・便宜的なもので、研究者御本人の意図と合っていない場合があることをあらかじめお断りしておきます。

全体を大きく、A：現代日本語系、B：中世・近世・近代日本語系、C：日本語教育系・情報工学系、D：非言語・医療福祉系、E：英語系、F：その他諸語系の六つに分けます。A（現代日本語系）については、異ジャンル別コーパス、コロケーション研究用のウェブコーパス、芥川賞作品コーパス、語義情報付きコーパスなどが入ります。やや特殊なものとしましては、話題別会話コーパスや、LINEコーパス、子供会話コーパス、テレビゲームコーパスなどがあり、更にユニークなものとして、選択的夫婦別姓に関する賛成派と反対派の言説を集めたコーパスや、地方議会コーパスなどがあります。また、今期の小委員会の検討トピック例にも挙げられています地方の言語資源に関係して、東北方言音声コーパスや日本語諸方言コーパスなども作られているようです。

続いて、B（中世・近世・近代日本語系）については、古辞書、抄物、江戸口語、洒落（しゃれ）本などのコーパス、崩し字研究用コーパス、明治・大正期の『赤い鳥』コーパス、帝国議会コーパスなどが作られています。

C（日本語教育系・情報工学系）については、学習者にとってのモデルになり得る資料が幅広く集められており、論文コーパス、授業動画コーパス、看護教科書コーパスなどがあります。また、日本語教師養成に関して、日本語教師発話コーパスなどもあります。さらには、学習者支援の目的の下、各種の学習者コーパスがあり、ユニークなところでは、日本語聴解テストの受験者への事後インタビューコーパスなどもあります。テストが終わった後で、なぜこの問題であなたはこう答えたのかという質問を投げ掛け、学習者の回答を録音して書き起こすことで、個々の判断に関して、学習者の頭の中で起こっていたことを事後的に再現することができるわけです。

国語やそのほかの教科教育への応用を考えたものとしては、外国ルーツ児童・生徒のための国語教科書コーパス、DDL（data driven learning：学習者にコーパスを検索させ、語の使用パターンを自発的に学ばせる教授法）用の国語教科書コーパス、中学校社会科コーパスなども作られています。

また、情報工学系のコーパスも各種作られております。前回の小委員会における相澤主査の御報告にもあったように、LLMの存在感は日増しに大きくなっており、

ユニークなところでは、自動音声認識技術の評価に用いる「なりすまし」データを集めたコーパスなどもございます。

ここまで、現代日本語、過去の日本語、日本語教育、国語教育等に関わるものを見てきましたが、加えて、D（非言語、医療福祉系）の中の非言語データを集めたものとして、手話コーパス、ジェスチャーコーパス、身振りのコーパス、舞踊動画コーパスなどが存在します。医療福祉への応用を考えたものとしては、自閉スペクトラム症（ASD）の方や、精神障害をお持ちの方、また、高齢者による発話を集めたコーパスなどもあります。これらを使って病状を自動診断する仕組みや、AIで健康相談を自動化するプロジェクトなども進んでいます。

このほか、E（英語系）やF（その他諸語系）に入るものも多数ございます。特に英語のコーパスは多く、英語のオリジナルテキストと邦訳を同時に収集するパラレルコーパス、英語教員採用試験問題コーパス、患者・医療従事者発話コーパス、通訳学習用コーパスなどがあり、学習者コーパスも各種存在します。学習者コーパスを調べることで、日本の英語学習者や、世界の英語学習者がどこに課題を抱えているか、問題を診断することが可能になります。その他の諸語についても多様なコーパスが作られています。

以上、駆け足で、科研で作られてきた、第3階層に位置する小規模言語資源を見てまいりましたが、それらは非常に多様でありまして、社会課題の解決を目指したコーパスや、地域言語資源の保存を目的とするコーパスなど、本小委員会で議論してきた、または、これから議論しようとしているタイプのデータも収集されている現状が見えてきました。

ただ、ここで述べておきたいのは、科研費で構築されたコーパスの大半は、最終的に非公開のままだということです。通常の研究においては、コーパスは、研究の目的ではなく、手段として位置付けられます。つまり、先に調べたい事柄があり、それをうまく調べるために、手段としてデータを集めるわけです。ゆえに、研究成果としてアウトプットされるのは、通例、コーパスではなく、コーパスを使って書いた論文のみになります。研究者の視点で考えると、科研費をもらった以上、論文公刊の必要はありますが、それに加えて、面倒な著作権処理やインターフェース開発を行って、自身が構築したコーパスをあえて世に公開することについては、義務もインセンティブもないわけです。このため、経費と労力を掛けて作られた価値ある言語資源の多くが、結果的に、死蔵されるという現状が生じています。

コーパスの公開には多くの研究上の意義があります。通例、研究者は自身のリサーチクエスチョンを解決するためにデータを作るわけですが、そうして出来上がったコーパスが全く違う関心を持っている人にとっての貴重な研究資源になる場合もあります。公開することで、一つのコーパスからより多くの研究が生まれ、さらには、関心を持った二次・三次利用者によって、より多くのコーパスが副次的・連鎖的に生み出されていくわけです。

ここで目配りをしておかないといけないのが近年のオープンサイエンスの潮流です。御承知のように、2025年度より、科研費の成果である研究論文を、電子版を有するジャーナルに掲載する場合は、ジャーナル購読者のみならず、誰もが無償で読めるよう、公的アーカイブにおいて同時公開することが研究者に義務付けられました。この考え方の基本は、公金でなされた研究の成果は広く国民に還元すべきだというところかと思います。しかし、この方針は、コーパスには適用されません。オープンサイエンスの時代に入り、コーパスを使って書いた論文は誰もが読めるようになりましたが、コーパスの多くはなお非公開のままとなりそうです。

こうした現状を変えるためのアイデアについても少し考えてみたいと存じます。全くの私見でありますけれども、例えば、文化庁などが中心になって、現状の科研費とは別に、「社会課題解決のための小中規模日本語コーパス整備事業」のような新しい助成制度を作るといったことが考えられるかもしれません。この制度では、研究論文ではなく－論文を出すことを妨げるものではありませんが－コーパスを構築し、公開することが応募の要件となります。また、本制度で作られたコーパスは、共通のプラットフォームに集約し、一般に公開します。プラットフォーム上で、本制度で作られた様々なコーパスが単一の検索系で網羅的に調べられるようになれば、本邦の言語関連研究の推進に大きな力となるでしょう。

また、仮にそのような新しい助成精度を作る場合、例えば、「基礎研究部門」と「社会課題解決部門」を立て、後者に明確なウエートを置くような仕掛けも有効でしょう。これにより、例えば、SNS上のヘイトスピーチ対策、電話やSNSを用いた各種の詐欺行為の防止策、また、多様な母語背景を持つ多人数会話の同時通訳システムの開発など、直接・間接に言語と関わりを持つ各種の社会課題の解決につながる言語資源を体系的に集めていくことができるかもしれません。また、一定数の言語資源が集まれば、続いて、公開されたコーパスを利活用して社会問題の解決提案を募るハッカソン

（hackathon）的な助成事業を企画するのも面白そうです。共通に使える多様な言語資源のセットがあれば、国民各層がそこから様々な情報を引き出し、従来にないアイデアやイノベーションの種が生まれてくるのではないかと期待されるところです。

以上、個人的な思い付きについて長く語ってしまい、失礼いたしました。

論点3「生成AI時代の言語資源」

最後に、論点3「生成AI時代の言語資源」についてお話します。ここまで言語資源の問題について多角的に見てまいりましたが、やはり生成AIのビフォーとアフターで、言語資源に対する見方も随分と変わってきたように思います。

先ほど前川委員からも御紹介がございましたが、私も、8月29日に開催された「言語資源ワークショップ2025」に参加して、青山学院大学の近藤泰弘先生による「AIによる言語資源の利用法－辞書データを中心に－」という御講演に衝撃を受けた者の一人です。近藤先生の御講演では、自身で集めた用例セットをAIに投げて、指定した語について、幾つの語義を立てるべきか、なぜそのように分けるべきなのか、分けた語義にそれぞれどのようなラベルを付けるべきかをAIに回答させる実験の結果が報告されていました。詳細は別として、全体として見れば、プロが作る辞書に近いような、極めて妥当なものが出来上がってきたということでございました。

私も、早速、近藤先生の作られたChatGPTのプロンプトを借用して、あるコーパスから取ってきた、動詞「行く」を含む25の用例をAIに投げて同じことをやらせてみました。その結果、資料にあるように、言語学者から見てもおおむね同意できるような分類がなされたことを、驚きを持って確認しました。

英語コーパスの研究でも同じようなことが進んでおります。9月2日に「英語コーパス学会」の大会が行われましたが、そこでは、ChatGPTによる自動分析機能をあらかじめ組み込んだコーパス検索ツールの紹介がありました。このツールでは、調べたい語を含む用例を一覧で書き出すことができるのですが、ボタン一つで、そうして抽出されたデータをChatGPTに分析させ、語義分類や用法パターンを回答させることができるとのことです。

昨年の私の報告で紹介した、「COCA」という大規模なアメリカ英語コーパスにも同様の動きがあります。同じく9月2日付けになるのですが、COCAの検索プラットフォームに、ついにLLMが統合されました。コーパス検索を行って必要なデータを抽出

しさえすれば、後は、LLMが共起語情報から分析対象語の語義を分類・整理し、個々の語義を書き出してくれます。また、そうして得られた結果を30言語に即時翻訳したり－つまり英語辞書の日本語版・中国語版・韓国語版などが一度に作れる－言語学者・翻訳者・教師など、指定したユーザータイプに合わせて出力結果を自動で最適化したりすることも可能になりました。どちらの例においても、これまで、長年経験を積んだ研究者にしかできないとされてきたデータ解釈の大部分を機械に行わせるわけで、コーパス研究にもある種の地殻変動が起こりつつあることを感じます。

もっとも、COCAを開発しているMark Davies氏は、LLMに人間の仕事を完全に代替させるのではなく、補完させることが重要だと強調しています。Davies氏は、ChatGPTなどのベースになっている言語データがブラックボックスであり、必ずしも信用できないことを踏まえ、人間が質の高い言語資料を収集し、その解釈をLLMに手伝わせるという分業ないし協業により、これまでの言語研究が飛躍的に進化する可能性を見ておられるようです。

まとめ

最後に、本日の報告をまとめます。

まず、論点1「言語資源と外来語」につきましては、継続的な外来語の使用実態調査が必要ではないか、その基盤として、大型コーパスの継続的拡充の必要性を打ち出していくことを考えてよいのではないかということを申しました。

論点2「言語資源の整備」については、大型コーパスだけではなくて、各種の小規模コーパスにも目を向けていく必要があるのではないか、特に、小規模コーパスの多くが非公開のままになりがちである現状に対して、何らかの打開策を考えるべきではないかということを申しました。

最後に、論点3「生成AI時代の言語資源」に関しては、LLMの言語データのブラックボックス性を踏まえ、信頼できるコーパスを人間が構築し、その解釈にLLMを活用する方向も考えられるのではないかといったことを申しました。この点に関して、生成AIには得手不得手があり、例えば、テキストデータを与えて語の頻度を数えるような作業をさせると、かなりのミスが出るといったことも報告されています。従来、言語研究における生成AIの活用については、頻度調査などの単純作業を機械にやらせて、データ解釈の部分を人間がやるというのが一般的でしたが、これを逆転させる方向も考えてみる

べきなのかもしれません。

以上で私からの報告を終わります。御清聴ありがとうございました。

○相澤主査

石川副主査、大変興味深いお話ありがとうございました。

それでは、ただ今の御発表につきまして御質問がありましたらよろしくお願いいたします。　　（ 挙手なし。 ）

では、まず私から。三つの大変重要な論点を御教示ただきまして、感銘を受けたところでございます。

社会課題というキーワードについては、前期からこの小委員会でも強く意識されているところでありますけれども、1番目の論点で「基盤的」と言うときには、この基盤は、どちらかというとコミュニケーションの、人対人のコミュニケーションの基盤という意味合いが強いように感じました。3番目の論点では、むしろ言語そのものがターゲットになっていて、想定されているのは言語に対するアノテーションであるのに対して、1番目の場合は、言語を読む人が分かるかとか、言語を介してコミュニケートしている人たちがどういうリアクションをするかということがターゲットになりますので、アノテーション対象は言語でなく人間になるとの印象でございました。

なので、コーパスを構築する際にも、アノテートする対象として、文字あるいは音声のあるスパンだけではなくて、読み手あるいは聞き手、あるいは言語を介して何かを理解したりエモーションを持ったりする人間も考慮するのが、社会課題としてのコーパス基盤に期待される拡張でもあるのかもしれないと思いながらお伺いしました。その辺り、御意見ありましたら是非お願いします。

○石川副主査

ありがとうございます。コーパスは、付与するアノテーションのタイプや対象によっても区分可能で、社会課題解決型コーパスを考える場合には、言語学的アノテーションだけでなく、対人コミュニケーション関連のアノテーションも重要ではないかという御指摘であったかと理解をいたしました。

おっしゃるとおりだと思います。これまで我々がコーパスと言いますときには、どちらかというと前者のみを前提にする傾向がありましたが、今後は、後者のタイプ、つまり、

実際のコミュニケーションを支える非言語的要因や、感情的要因などの情報を組み込んだコーパス整備も必要であろうという点、私も同じ思いを持っております。ありがとうございます。

○相澤主査

ありがとうございます。とてもすばらしい言葉で説明いただきました。

○石川副主査

いえいえ、そんな。

○相澤主査

では、ほかに御質問等いかがでしょうか。　（ 挙手なし。 ）

それでは、御発表に対する質問に限定せず、自由に御討論いただく時間も兼ねてということで、石川副主査の御発表を踏まえて、何か今後のコーパスの在り方、言語資源の在り方について引き続き質問や御意見などがありましたらお願いをさせていただければと思います。いつもどおり皆さんに御発言いただければと思っておりますが、まずは、石川副主査からもし何かこの点更に強調したいということがありましたら。

○石川副主査

既にたくさん話させていただきましたので、ほかの皆様からの質問、コメントを承りたいと思います。

○相澤主査

分かりました。ありがとうございます。

では、神永委員からお願いしてよろしいでしょうか。

○神永委員

やっぱり私の場合には国語辞典との関係ということでどうしても考えてしまうんですが、外来語の調査のことに関しまして、外来語調査は、継続的に定点観測的な調査が必要だし、発信が必要だというようなことをおっしゃってくださっていましたけども、正に

そのとおりだと思います。コーパスを利用して、私もどのように活用できるかー辞書でどのように活用できるかということを申し上げましたが、頻度数によって見出し語選定の参考にできるなんていうことを申し上げていますけれども、実はこれ、そうは言ってもなかなか難しい問題がありまして、ある程度、何かこういうのでやっているぞというのを見せていただくととても参考にしやすいんですよね。本当にやみくもに検索するというので何かが分かるわけではないので、やっぱりいろんなお力を借りながら、ここのところはこのように変わってきているんだというものが分かりますと、それによって見出し語の形を決めやすいとか、それから異形としてこれは必要なのかなとか、そんなことを考えられますので、そういった意味で是非お願いできればなということを考えました。

それからもう一つ、生成AI使った辞典の語釈の件なんですけど、語義分類に関しまして、実はかなり私たちは恣意的なのかもしれないー恣意的と言ってしまったらちょっと言い過ぎかもしれませんけれども、各辞書によって語義分類というのはかなり違っているものですから、それをやっぱりどこか、別に生成AIが示してきたものが正しいかどうかはちょっと検討しなきゃならないだろうとは思うんですけれども、一つそういうような判断材料として見るということも今後必要になってくるのかなということは考えました。

ただ、語義分類を本当に生成AIがーちょっと私も生成AIのことはよく分かりませんけれども、何か参考にするものがあって、例えば辞書データとか、そういったものを参考にしてもしやっているのであれば、何だかぐるぐる回っているようなことにもなりかねませんので、その辺のところは検討する必要はあるかもしれませんけれども、ただ、特に辞書の中で語釈が非常に長い、語義分類もかなりな数になっているというのがありまして、そういったものも一つ整理するような材料にできるのかなというようなことは思いまして、先ほど前川委員からも日国は考えているのかというようなことをおっしゃっていただきましたけども、今後の課題として考えるべきことなのだろうなということは思いました。ありがとうございます。

○相澤主査

ありがとうございます。日頃言語を扱うAIについて研究している私たちも、辞書の語義とAIの語義について考えることもあるんですけれども、例えばOCRと呼ばれる文字認識は、画像としての文字を読んで、その文字が何かを当てるんですけれども、書き順は分かっていないんですね。ちょうど生成AIも、言葉の意味は言い当てるんです

けれども、出典というか、意味の変遷のようなことが分かっているわけではないと感じます。例えば文字の書き順は、美しさとか文化に関係がありますし、語源とか語義の変遷というのは、やはり文化とか、恐らく何らかの形の文章の美しさに関係あるかもしれないと個人的には思ったりもしております。

では、武田委員、よろしくお願いいたします。

○武田委員

まず、前半の神永委員の御報告、非常に参考になりました。『日本国語大辞典』の改訂、時代に合わせた改訂に進んでいらっしゃるということで、完成が大変待ち遠しく思われております。コーパスの利用という点では、辞書の編集には非常に有用だと考えられますので、そのことによって、用例ではなく作例で解説する辞書も多くある中で、コーパスを活用して改訂する用例中心の『日本国語大辞典』の意義は従来に増すものと思われます。楽しみに待たせていただきたいと思っております。

また、石川副主査のお話も大変参考になり、示唆に富んでいるものと思われます。具体的な御提言もあり、できる限り対応していくことができればと思っております。

今回は、生成AIを使った辞書の語義分類や語釈の話がございましたけれども、辞書における語義分類・語釈については、既に示されておりますとおり、これは従来、辞書によってそれぞれでありまして、若干違うところもありますけれども、本来それぞれの辞書の考え方、方針に照らして解説するところがあるかと思われています。その意味で、生成AIなどを使って自動的にどこまで汎用性のある語義分類・語釈が実現するのかというところは非常に気になるところです。とはいえ、全て画一的な語釈のみになってしまって、正に辞書は一つあれば事足りるということになるということはないと思われますので、その意味では生成AI活用の余地はまだまだあるのかと思います。その上で、生成AIを使った語釈もここまで的確な形でできるという例も示されているということでありますから、今後に注目したいと思っております。

簡単ですけれども、以上でございます。

○相澤主査

どうもありがとうございます。では、続きまして、森委員、よろしくお願いいたします。

○森委員

神永委員には先ほど質問させていただきましたので、石川副主査の発表についてお伺いしたいところが幾つかあるんです。外来語の言い換えのところで、官公庁、報道機関など公共性の高い組織では言い換えていくというので、我々テレビ局はもちろん言い換える立場にあります。十分に浸透して、もう言い換えなくてもいいんじゃないかという判断は、手探りというか、肌感覚で行っています。もうみんな使っているよね、通じるよね、この番組の視聴者層年齢層であれば伝わるけれど、こちらの番組の時間帯は子供からおじいちゃんおばあちゃんまで見ているので、分かりやすく言い換えた方がいいよねというふうにも考えるのですが、官公庁に当たる文化庁の皆さんは、この判断はどんな感じになさっているんですか。

○石川副主査

課長からコメントが頂けると有り難いです。

○山田国語課長

石川副主査の御発表にもあったとおり、やはり役所の方が時々、片仮名の新しい言葉を打ち出して、よく分からないというような御指摘を受けるところもあると思うんです。私も昔関わっていたところで、「インキュベーション」という言葉を、ベンチャーの関係で出始めた時に使っておりまして、その時、私も自分自身よく分からないと思いながら使っていたところもありますが、やはり10年、20年とたつと、それは多分今の時代ではインキュベーションというのはそれほど説明を要しなくても通じる場面は増えてきたとは思います。

ですので、必ずしも、広く国民に通用しているものだけを使うということでは、役所の場合は、もちろん原則としてそういうことがありながら、場面によってはそういう新しい概念を打ち出す、これまでとは違う新しい政策を打ち出したいというときには、やはりちょっとそういう新しさということを伝えるための手法として新しい外来語を使うということはあるのかなと個人的には思っています。

○森委員

2002年の調査のようなことを今後またされる予定はあるんですか。

○鈴木国語調査官

2002年にやって、その後、何年だかたって、またやってはいるんです、同じような調査というのは。

それから、今課長から申し上げたことのちょっと補足的なことで、国語調査官のところには役所で外に出す報告書などの文章を見てほしいという依頼が時々来るんですね。そうすると、やっぱり外来語はすごく気になるので、チェックしますけども、まず新しい概念について注釈も何も付いていないときは、最低限注釈は付けてほしいという指示を出します。概念に関するところの考え方は、課長から説明のあったとおりなんですけれども、実はもう一つ、すぐに分かる－例えば「オープンする」とか、「スタートする」とか－意味はすぐ分かるんですが、別に「始める」などの言い方で何でいけないんだろうと思うことも多く、そういう部分は、意味は分かるし、全然難しくはないけれども、なぜわざわざ外来語をそこに使わなければいけないんだろうかという使用の必然性という観点でも、文章のチェックはしているということはございます。

○森委員

もう一つ。大学とか研究のことが分かってないので。社会課題解決のためのコーパス集中整備事業のところで、このぐらいの予算を付けて、こうすれば研究のためのコーパスの公開につながるんじゃないかというお話ですよね。この予算を付けてもらったら、その研究をしたコーパスを作った人たちは著作権の確認とかをしなきゃいけなくなるんですか。

○石川副主査

そうですね。そこについては、前回の前川委員の御報告で、BCCWJ2の構築に関して、新しく出ている著作権関連法に基づいて、コーパスを作る側のフリーハンドを広げていくという御提案もあったところでございます。その方針で臨んでいくとすると、10年前にやっていたほど大変なことにはならないのではないかと思っております。

○森委員

前回も思ったんですけれども、この事業は、著作権を持っている人に、その権利を強

く守らなきゃいけないと思わせるものなんだろうかと。文芸作品にせよ、報道等の記事にせよ、その内容が全て出るわけではないので。

○石川副主査

ユーザーが読むのは、基本的に、1行程度の文の断片ですからね。

○森委員

そうなんです。そのぐらいよくないですかって思っちゃうんですけど、なかなかそのハードルというのは崩せない。

○石川副主査

その点は、恐らく後ほど前川委員からいろいろと教えていただけるのではと思います。

○森委員

辞書の用例なんかでも新しいものって使えないんですか。

○神永委員

一応用例だから大丈夫だという判断で今までやってきたんですけれども、今後変わるだろうと思っていまして、どういう形で引用ができるのかはちょっと検討しなければいけないことになっています。やはり弁護士さんなんかも、いろいろと聞いてみますと、二つに分かれちゃうんですよ。大丈夫だという人とやっぱりかなりグレーですという人といまして、だからその辺のところをどうやって私たちの辞書で引用する－特に『日本国語大辞典』なんかは新しい小説なんかも引用しますから、たとえ短いものであっても、やっぱりそれは駄目だとなってしまったら、どうしようかなと－だからなるべくそれは回避してやらなければいけないのかもしれないということは今考えています。

○森委員

ちょっと話がずれますが、例えばテレビで何とか辞書によるとこんなふうに載っていると紹介するのは、それは著作権に当たらないんですか。

○神永委員

当たります。

○森委員

それが辞書の会社に連絡をして、しかるべきものを結んで使わせていただくという……。

○神永委員

はい。最近は－昔も割とテレビで使うような場合でも－使わせてくださいはありましたから、黙って使うというのはよほどのことがない限りはなかっただろうなと思います。

○石川副主査

そこは難しいところですよね。研究の世界で、引用、つまり、サイテーション（citation）を行う場合は、誰が何年に書いたどの文献かを明示しておくことで、著者からの引用許諾の取得を省略することが普通かと思います。

○森委員

論文の最後に一杯出てくる……。

○石川副主査

おっしゃるとおりです。そうしたリストに掲載された著書や論文の著者から、引用してよいかどうか、その都度、許諾を取っているということは普通ないので、著作権の扱いも、目的、つまりは引用の使い方に応じて変わってくるかと思います。コーパス構築における著作権の扱いについても、恐らくは、その目的に即した形で、独自のスタイルがあってしかるべきかと考える次第です。

○森委員

ありがとうございます。

○石川副主査

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、前川委員には最後にフォローいただくことにして、あいうえお順で植木委員からよろしくお願いします。

○植木委員

今日も勉強させていただきまして、ありがとうございました。神永委員の話については、森委員がおっしゃっていたんですけれども、辞書は版が変わるときに、書き直したものだけが次に残っていくので、書き直したところが分からなくなってしまいます。実用的というより、研究の立場からになってしまうのかもしれないんですけど、やはり残されていくということが重要なのではないかなという気もしました。そのこと自体が言葉の歴史になるというか……、なので、どこまで残すのかというのは本当に難しい問題だとは思うんですけれども、残していただければ本当に有り難いと感じながらお聞きしていました。

同じようなことが、辞書を作るときのカード的なものにも言えるのかなと思います。私も今、本当に小さな辞書の作業に関わっているんですけども、前の版の見直しをした人たちの記録が－それは紙ではなく、エクセルの表の中に書いてあるんですけれども－いろんな人がこう思うというのを、例えば、字の色を変えたりしながら積み重ねてあって、私はそれを見て、最終的にどうするかを判断します。考察してきた過程、いろんな人たちが見直して確認してきたものが残っているのは非常に貴重だなと感じているところではあります。

今、例えば作家も原稿をパソコンで書くようになると、昔のように、第1校、第2校と書き直した推敲の跡も分からなくなってしまって、かつての推敲の跡をたどるような文学研究が成り立たないと言われますけれども、人の思考の跡が残るというのも非常に大事なことなんじゃないかなということを感想として持ちつつ思ってお聞きしていました。ありがとうございます。

石川副主査のお話はいつも非常に広い範囲で教えていただくことが多いんですけれども、個人的な興味としては、社会課題の解決というところで、非言語とか、医療福祉系のところ、面白く伺ったんですけれど、これはちょっと難しいかもしれないんですが、例えば場面緘黙(かん)というんですかね、ある場面で言葉が出なくなる、そういうのは、言葉がないんですから、取りようがないからどうにもならないのか。あるいは吃(きつ)音なんかだと、どういう言葉が言いにくいのかというような、それがある程度データ化されれば、改善に何か貢献できるところがあるのかなと。言いよどんだりしてしまう、そういう言葉についての何かデータみたいなものが取れると、貢献できる部分が－そういう困難を抱えている人たちに対して貢献できるところがあるんだろうかというようなことを思ったんですけれども、石川副主査、御存じのことがあったら教えていただきたいと思いました。

それと社会思想分析、先ほどのスライドの社会思想等分析支援というところで夫婦別姓言説コーパスというのがあって、本当にこれも個人的な興味で恐縮なんですが、これはどういうもので、どのようにそれを活用できるのかというのを個人的に是非知りたいと思いましたので、お願いします。

外来語の言い換えのお話をしてくださったんですけれども、日本語も、今、例えば海外の方、災害のときなんかに「避難してください」を「逃げてください」と言い直すというような、これは森委員もお詳しいと思うんですけれども、そういう日本語から日本語に言い換えるというようなこともしかするとそういう一覧というか、対応みたいなものがデータ化されていれば、いろんなところで普遍的に使えるのかなと思ったりします。

ほぼ感想のようなことなんですけど、3点、もしお考えなどありましたらお伺いしたいと思います。

以上です。

○石川副主査

植木委員、ありがとうございました。

3点御質問を頂きました。まず1点目は、医療分野への活用ということをもう少し柔軟に考えていってはどうかということで、私も同感であります。例えば、私の専門ではありませんが、吃音などについても、それが起こりやすい音素や、あるいは、それが起こりやすい文脈などがあると伺ったことがあります。こうしたことをデータ化していくことで、

吃音の起こりやすい条件を明示化し、そこに特化したトレーニングなどを考えることも可能かもしれません。

場面緘黙につきましても、それが起こる前後の環境的特徴、つまりは、緘黙を引き起こすトリガーのようなものを言語的に特定できないか研究していく余地はあろうかと思います。御指摘の点は非常に可能性のある分野であると、素人ではありますが、植木委員と同じように感じた次第であります。

2点目は、夫婦別姓コーパスに関する御質問であったかと存じます。こうした研究では、一般的には、賛成派と反対派の言説をそれぞれ集めて、コーパス分析の手法で比較していくことになります。賛成派の言説によく出てくる言葉や、反対派の言説によく出てくる言葉を統計的に特定することで、賛成派の人たちは何をもって賛成しているのか、反対派の人は何をもって反対しているのかといったことを、分析者の主観バイアスを抑えつつ、科学的に議論していくことが可能になります。昨年度に報告させていただいた海外の研究例ですと、移民や、セクシャルマイノリティーなど、社会における少数派の権利擁護に関して、賛否の言説を比較し、社会的意識の底流を探っていくような研究が増えているようです。

3点目につきましては、外来語の言い換えで、単に日本人向けに外来語を減らすというだけでなく、海外の方向けに日本語を分かりやすくするような言い換えも考えるべきではないかという御指摘を賜りました。これにつきましても、御指摘のとおりかと存じます。実際、日本語教育では、「やさしい日本語」のようなプロジェクトも進められているところで、そうした動きに言語資源を活用していくことも十分できるのではと考える次第です。

ありがとうございました。

○植木委員

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、前川委員、よろしくお願いいたします。

○前川委員

何点か私、指名もあったようなので、思い出せる範囲で申し上げたいと思います。

最初が、外来語の私どもの昔の研究を取り上げていただきまして－仕事を取り上げていただきまして、ありがとうございました。

この外来語の研究をやっていたのは、BCCWJなんかを作っているのと並行した時期だったと思います。ちょうど我々が独法に移管されて、独法というのは国民に対するサービスをやる機関だということですので、何がサービスになるんだろうということをいろいろと試行錯誤していた時代の一つのああいう試みもあったということです。

確かにそういうサービスの側面としては、私は価値のあるものだと思いますので、今後もし文化庁の方でそういったお仕事を継続されるのであれば、それは大変結構な意味のあることだろうと思います。

ただ、私個人として申し上げますと、分かりにくさということを構成している要素は、外来語も確かにあるし目立つんですけども、もう一つ、私個人的に非常に思っておりますのは、新しく作られた漢語ですね。特に長い漢語というのが大変私は分かりにくいと思っております。例えば我々の所属しております人間文化研究機構というところのプロジェクトの実例なんですが、「基幹研究」というのと「機関基盤型研究」というのがあるんですね。これ聞いて分かるかというと、私、実は「キカン」のところの字が違うんですよ。基幹研究というのは、基礎の基に幹の「基幹」。機関基盤型の「キカン」というのは研究機関の「機関」。こういうのはやめてくれって私いつも言っているんですけども、決まって中期計画に書き込まれているのでしようがないので使っておりますけども、そういうのが実は非常に多くて、聞いて全然分からないですよね。そういうのの方が私は実は実害が大きいんじゃないかという気がしないでもないです。これは個人としての感想です。ですから、もし分かりにくさの解消ということをやるなら、そういうところにも目を配るべきではないかなと思っております。これが第1点。

それから、公開されないコーパスが多いので、何かの手段がという御提案がありましたけども、私はこれはもちろんそういうのはあってもいいんですけども、多くのものにコーパスの公開を前提としてというのはある意味ちょっと酷かなという気がしないでもありません。つまり、100万円でやる研究と1,000万でやる研究と1億でやる研究はやっぱりおのずと違ってきたものですし、そこに公開ということの可能性を無理に要求すると作りたいものも作れなくなってしまうということがあるかなあと思いますので、

一律にやるのはちょっとどうかなと思います。むしろ問題は、大量の資金、例えば何千万とか何億というような資金を提供して行われるような研究で作ったものも公開されない、これは大変問題なので、むしろ審査のときにそういうところをきちんとやるべきではないかということを私は思っております。

それから、この小委員会の去年のごく最初の頃に私は話をいたしまして、コーパスというものの定義を述べたことがあるんですけども、その要件の一つに、コーパスとは公開されているものであるという定義を入れたと思います。それは実は今言ったような話で、公開を前提として作らないものはそもそもコーパスと呼ばない方がいいという私の元々の考え方－それはデータとかデータベースと呼べばいいのであって、コーパスという言葉は公開を前提として使いたいなというのが私の提案です。

それから最後に、著作権処理の問題で、引用で処理できないかというお話が出ていたと思います。これは本当に昔からそれはそれでできるといいと思っているんですが、BCCWJを作り終わる頃に文化審議会の著作権分科会というのに呼ばれまして、そういった話を今の著作権解放につながるような動きが始まったときですけども、そういう話をしたわけです。それで、著作権処理の現状の話をいたしまして、連絡がついて依頼すると95％の方は長さにかかわらず許諾を下さる。だから、ちょっと今の著作権法は厳し過ぎるんじゃないですかみたいな話をしたんですけども、その時に参加されていた法曹関係の方だと思いますけど、いや、その5％が問題なんですと強烈に反撃してこられるので、現実は－世の中はそういうものであって、法曹関係者の意識というのは何かそういうもんかなあということを強く感じたことを今も思い出します。これはそれがどうこう、両方、立場はあると思うんですけども、そういう立場もあるということはここで申し上げておこうかなと思いました。

以上です。

○相澤主査

ありがとうございました。では、お時間となりましたので、本日大変貴重な御講演と、有意義な議論いただきまして、前期報告のフォローアップとしても方向が示されたものと感じております。

それでは、本日の言語資源小委員会は、これにて閉会いたします。ありがとうございました。