第２回 国語分科会言語資源小委員会（Web開催）議事録

令 和　６ 年　８ 月　１ 日
１４時５５分～１７時００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　第１回国語分科会言語資源小委員会議事録（案）（委員限り）
２　言語資源の在り方に関する主な意見
及び言語資源の在り方に関する委員の意見（第１回分）（案）
３　辞書編集者から見た言語資源
－現代日本語書き言葉均衡コーパスについて－（神永委員御提出資料）
４　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）

〔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 事務局から配布資料２「言語資源の在り方に関する主な意見及び言語資源の在り方に関する委員の意見（第１回分）（案）」の説明があった。
５ 神永委員から配布資料３「辞書編集者から見た言語資源－現代日本語書き言葉

均衡コーパスについてー」に基づいて説明があり、質疑応答が行われた。

6　石川副主査から配布資料４「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－」に基づく説明があり、質疑応答が行われた。

7　二つの説明を踏まえ、コーパスの活用事例や活用の可能性についてを中心に意見交換が行われた。

8　次回の言語資源小委員会について、令和６年９月２６日（木）10時から12時まで、対面及びオンラインで開催する予定であること、今季後半の御予定伺いを発送する予定であることが確認された。

9　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

定刻までまだ少し時間はございますが、皆様おそろいですので、第２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。本日もハイブリッド形式で小委員会は開催しております。文部科学省の会議室に御出席いただいている委員の皆様、オンラインで御参加いただいている委員の皆様がいらっしゃいます。また、傍聴者の方々もオンラインでこの会議を御覧になっているということで、どうぞ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」、「有識者による事例等紹介」、「その他」という内容で御協議いただきたいと考えております。事前に資料を御覧いただいていますとおり、本日は、前回の審議を受けて、神永委員には辞書編集の現場でのコーパス活用のお話を、石川副主査には英語のコーパスとその活用についてのお話をしていただくこととなっております。お二人の委員のお話を踏まえまして、コーパスの活用という点を中心に御議論いただきますようお願い申し上げます。

前回の議事録（案）の確認をいたします。事前にお送りした資料を既にお読みいただいているところでございますので、大きな問題がなければこの案をお認めいただきたいと思いますが、よろしいでしょうか。

○前川委員

議事録を拝見しまして、実は、前回の私の発言の中にちょっと誤りがありまして、それに気付いたんですが、それは別に修正する必要もないでしょうか。

○鈴木国語調査官

事務局からお答えいたします。修正はまだこの後１週間受け付けさせていただきますので、修正の必要があると御判断なさった部分については、修正していただくことが可能でございます。よろしくお願いいたします。

○前川委員

具体的には、前回御紹介いただいた資料の中のドイツのコーパスとチェコのコーパスについてちょっと発言をしたんですが、その後、私、たまたまヨーロッパに出張する機会がありまして、両方とも実際に機関に行って確認してきたんですね。そしたら、ドイツの方はちょっと私の認識が間違っていまして…、チェコの方は合っていたんですけれども。そのことを、後ほど事務局に連絡すればよろしいでしょうか。

○鈴木国語調査官

はい。事務局にメールで御連絡いただければ、それに合わせた形で対応をいたします。

○前川委員

分かりました。よろしくお願いします。

○相澤主査

文言修正等あるいは訂正等は、１週間後、８月８日までに御連絡いただくということでお願いいたします。軽微な修正を超えて大きく変更あると事務局の方で判断される場合には、私に確認して、委員長一任ということでお願いさせていただければと思います。その後、公開されることとなりますので、どうぞよろしくお願いいたします。

それでは、議事に入りたいと思います。

まず初めに、前回の言語資源小委員会で出された意見を振り返るということで、前回出されました資料―本日の参考資料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）」につきまして、前回は「検討の観点について」という枠組みに関して皆様に御意見を賜りました。それにつきまして村瀬課長から御説明をよろしくお願いいたします。

○村瀬国語課長

お手元の配布資料２と参考資料１を御用意いただければと存じます。

審議事項につきましては、ただ今も相澤主査から御案内ございましたけれども、参考資料１に３点、お示ししているところでございます。

配布資料２は、この審議事項の柱を踏まえながら前回の御議論を整理したものでございますが、主な御意見につきまして御紹介したいと存じます。

本小委員会の審議事項につきましては、まず、「言語資源をいかに位置付け、その保存・活用の在り方をどのように考えていくか」という論点がありました。配布資料２の１ページ目、御覧のように、「コーパスはある時点の言語を記録した「言語資産」のようなものだということであり、どのような形で言語を保存していくのが望ましいのか」、そして、どのように「次の世代に受け渡していくのがよいのか」というお話がありました。

一つ飛びまして、「日本語コーパスの整備・発信について」ということでありますが、これについては、整備が着実に進み、「言語研究に占めるコーパスの位置は確立された」が、今後の課題として、多言語社会などを見据えた整備という話もあったところであります。

このほか、続いての○になりますが、技術的には著作権あるいは個人情報に係る対応についても御指摘があったところであります。

続きまして、２ページになります。見方を変えて、発信という側面になるわけでございますが、ややもすると一般の方からは「それほど身近な状態にはなっていない」ということもありますので、広くコーパスの魅力を発信していくとよいというお話もあったところです。

続いて、これからの日本語へ影響を与えていくものについて、御覧のような御意

見がございました。生成ＡＩモデルの中にあるバイアスというものが、逆に人間に影響を及ぼすおそれがあるといったような話、あるいは、続いての〇になりますが、こうしたＡＩ全体としてのインパクト以外にも、サイバー化、多文化・多言語化といった社会の変化にも留意すべきとする御意見があったところであります。

そして、次のテーマであります。「今後における望ましい言語資源の在り方について」に関しましては、審議事項の方にも、社会課題の解決に資するコーパスの在り方という形で示されておりますけれども、一つ目の〇の中で、後段の方になりますが、「社会のイノベーションにもしっかり貢献していける」ようなコーパスの応用可能性についても御指摘があったところでございます。

続いて３ページになります。このほか、利便性に資するインターフェースの改善であったり、あるいは、集めたデータの磨き上げであったりといったお話があったところでございます。

このように、前回、貴重な御意見を数々いただいたところでございますので、本日も、今後に向けた望ましい言語資源の在り方に向けまして、例えば社会課題の解決に資するコーパスの在り方も視野に入れながら御意見を賜れれば、誠に幸甚でございます。

〇相澤主査

ありがとうございました。それでは、ただ今の説明について御質問、御意見等ございましたら、よろしくお願いいたします。　　（ → 挙手なし。）

よろしいでしょうか。では、前回の議論や審議事項等を踏まえまして、本日も審議を進めてまいりたいと思います。

お手元の議事次第の２、「事例等紹介」ということで、既に御紹介いたしましたとおり、本日は神永委員と石川副主査に御発表をお願いしております。

まずは神永委員に、辞書編集の現場におけるコーパス活用の事例を御紹介いただきたいと存じます。よろしくお願いいたします。

〇神永委員

よろしくお願いします。

それでは、「辞書編集者から見た言語資源」ということでお話をさせていただき

ます。

私は辞書編集者ですので、辞書編集者としての立場からということで、コーパス活用の実態と、今後想定される活用の可能性、さらに、新たなコーパスへの要望についてお話をさせていただければと思っております。

最初に一言お断りしておきます。私は2017年まで小学館の辞書編集部に所属しておりまして、定年退職後の現在も辞書の編集は小学館のものしか関わっていません。したがいまして、ほかの辞書出版社がここ数年の間コーパスをどのように活用しているのかということにつきましては、一切分かりません。ですので、今回お話しすることは、小学館と私というかなり狭い範囲での事例だというふうにお考えください。

これからお話しすることなんですけれども、国語辞典としてのコーパスの活用の可能性、現行の辞書のコーパスの活用事例、コーパスへの要望という３点についてです。

国語辞典としてのコーパスの活用の可能性とは、現在考えられる可能性でして、今後多くの辞書でコーパスを活用するようになれば、様々な可能性が考えられるだろうと思います。

次の、現行辞書のコーパスの活用事例につきましては、実際に、僅かなんですけれども活用した事例がありますので、それを具体的にお話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、最後の、コーパスへの要望なんですが、これもまた、辞書編集者である私個人の考えであるということをお含みおきいただければと思います。

まず、国語辞典におけるコーパス活用の可能性、辞書編集としてコーパスはどのように活用できるかということなんですが、このことにつきましては、ここにも書いてありますように、「１．見出し語選定の参考になる」、「２．記述する意味の説明の参考になる」、「３．例文の参考になる」、「４．その語が使われる場面や文体、あるいは位相についての情報が得られる」、「５．表記に揺れのある語の表記をどのように考えたらよいのか、その根拠が得られる」、「６．ある語と連接しやすい語の特定が可能になる」。現時点では以上の六つの可能性が主に考えられると思います。

ただ、先ほども申し上げましたけれども、今後更に様々な可能性を見いだせるか

もしれません。それは飽くまでも辞書編集者の今後の課題ではあるんですけれども…。

まず、「１．見出し語選定の参考になる」ということなんですが、どういうことかと言いますと、「立項すべきかどうか検討中の語の使用頻度がわかる」、「一定以上（といってもその基準は編集部で決めるしかないが）、頻度数の高いものは、使用範囲が広く重要語だと考えられるので、見出し語として立項する候補にできる」、「従来、この項目選定の作業は一部編集者の勘に頼っていた部分だが、コーパスによって客観的な判断材料を得ることができる」といったようなことです。何が大事かと言いますと、見出し語選定のための客観的データが得られるということなんですね。もちろん、使用頻度が低くても重要と考えられる語もあります。一概には言えないんですが…。全てを使用頻度に頼るわけにはいかないということはあるんですけれども、特に新語などの場合には、どの程度定着してきたか、ある程度コーパスによって裏付けが取れるわけで、客観的な判断材料が得られるということはとても大きなことだと考えております。ただ、新語まで範囲に入れて考えるということになりますと、コーパスを常時更新させていかなければならないので、それなりに時間と手間が掛かるという問題も生じるかなという気はいたします。

二つ目の「記述する意味の説明の参考にできる」。従来、辞書の語釈は、「カードなどを使って集めた用例をもとに、帰納的に記述」してきました。ただ、カードなどで語の用例を採集しただけですと、どうしても「その語の多様な意味をじゅうぶんにカバーしきれないこともあった」ことも事実です。実際、新たにカードを使って採集し直してみると、従来の辞書では触れていなかった意味で使われた使用例が見付かるということがしばしばあります。「コーパスによって、そうした意味用法の広がりを確実に押さえることができる」のではないかと考えております。電子データであるコーパスは、ある語の検索結果として得られた用例を並べ替えたり、分類したりすることができるんですね。これによって、一つの語に意味が複数存在する多義語なんですけれども、意味記述の順序－ブランチ分けした際の順序ですが－使用頻度順に変更することができるというわけです。それらを参考にすることによって、使用実態に沿ったより使いやすい内容の辞書にすることができると思われます。

三つ目。「例文の参考になる」ということです。国語辞典における例文は、二つに

分けることができます。一つは、「日本国語大辞典」のような辞典なんですけれども、その語の意味による実際の使用例を、典拠を示して引用しているものです。このような実際の例は、「広辞苑」、「大辞林」、「大辞泉」といった中型の辞典でも、多くはありませんけれども引用しています。もう一つは、小型の国語辞典に多いんですけれども、実際の使用例ではなく、辞書編集者が、集まった使用例を基に作った、辞書独自の例文です。これは我々、「作例」などと呼んでいます。コーパスから、用例主義である「日本国語大辞典」のような辞書で引用する例文を見付けられないこともないんですけれども、現時点では、日本の文献を網羅的に収めたコーパスは残念ながら存在しませんので、コーパスからそういった用例を探し出すことは非常に難しいと思われます。今後の課題としまして、そのようなあらゆる日本の文献に対応したコーパスが構築されるのであれば、用例主義の辞書でも大いに活用できると考えられます。

もう一つの辞書の例文、つまり、辞書編集者が作る例文の方なんですけれども、こちらは、コーパスを参考にすることによって適切かつ典型的な例文を作成することができると思います。コロケーションに関しても、ある語がどういった語と共起しているのか、その実態を読み取ることができ、これも典型的な例文として辞書に載せることができるだろうと考えられます。実際の使用例である用例とは違いまして、こちらの方は非常に期待値が高いものだと考えられます。

四つ目なんですけれども、「その語が使われる場面や文体、あるいは位相についての情報が得られる」ということです。このことはコーパスへの要望にもなります。言ってみれば、俗語とか雅語とか文章語とか口語とか隠語とか幼児語、男女差とか地域差とかいった、位相に関するきめ細かな情報が欲しいということなんです。そのためには、その文章の詳細な書誌情報というのが欠かせませんし、「日本十進分類法（NDC）の情報はもとより、ＮＤＣに配された書籍をまんべんなく対象としたコーパスが望まれる」ということです。

五つ目に、「表記に揺れのある語の表記をどのように考えたらよいか、判断のための根拠が得られる」ということです。例えばなんですけれども、「悪事にカタンする」という「カタン」、これどういうふうに書くかということなんですが、本来の表記は「荷物」の「荷」を使った「荷担」なんですね。ところが、最近は、「加える」の「加」を使った「加担」も多く使われているということなんです。これについては

また、後で少し詳しくお話をいたします。

それから、外来語の表記もあるんですが、例えば、化粧の意味の「メイク」か「メーク」かというような揺れのあるもの。皆さんどっちを使っているか分かりませんけれども、これも分かるわけでして、これもＢＣＣＷＪでどのようになっているのか、後ほど御説明したいと思っております。

六つ目なんですけれども、「ある語と連接しやすい語の特定が可能になる」ということです。例えば、ある動詞について、それがどのような名詞や助詞を伴って使われるか、客観的な根拠を得ることができるということなわけです。例えば、「流す」という他動詞と「流れる」という自動詞の場合に、それぞれ助詞を伴った形を見ていきますと、「ヲ流す」の場合どのような語と使われることが多いかとか、「ガ流れる」の場合はどうかといったことが分かるわけです。それで、それらを辞書で示すことができるということです。実は、こういったことを従来の辞書では余り積極的に取り上げてきませんでした。ちょっと怠慢ではあったのかもしれませんけれども…、幾つかの国語辞典では、その語釈の前に「……を」とか「……に」とか格を示す情報を示しているものもありますけれども、ほとんどの辞書は例文の中でそれを示しただけというのが実情なわけです。

そういった形で作った例文なんですけれども、辞書編集者が作った例文は、規範性、可能性を考慮して作られてはいますが、実際には余り使われていないものだったということは絶対にないとは言えない気がするんですね。その逆に、コーパスでは、非常に頻度が高い用法なんだけれども辞書には載っていないというようなことも考えられます。コーパスを活用することによって、このような食い違いを阻止することができるだろうと思われるわけです。

以上、現時点で考えられるコーパスの活用法を示しましたけれども、先ほど申し上げましたように、活用法というのはこれだけではないと考えられます。辞書編集に関わる者として、どのようにコーパスを活用したらいいか、より良い辞書にできるのかということなんですけれども、今後の検討課題だろうと考えております。

画面に辞書の画像が出ていますが、現行の辞書のコーパス活用事例ということで、－今お話ししてきました六つのコーパス活用の可能性なんですけれども－僅かではあるんですが実際の活用事例がありますので、具体的に御紹介しておこうと思います。

その前に一言お断りしておきますと、事例というのは、主に小学館の「現代国語例解辞典」の第５版というので実際になされたものです。ちょっと宣伝っぽくなって恐縮なんですけれども、この辞典は1985年に初版を刊行しまして、2016年に改訂第５版を刊行したんですが、実は1985年の初版は私が担当しまして、別にどうでもいい情報なんですけれども、2016年の改訂第５版でコーパスを初めて活用しています。ほかの出版社の国語辞典は分かりませんけれども、恐らくこの「現代国語例解辞典（第５版）」が、国語辞典としては最初にコーパスを本格的に活用した辞典ではないかと思います。

では、「現代国語例解辞典（第５版）」ではどのようにコーパスを活用したかということですけれども、凡例でこのようなことを書いています。「日本語書き言葉均衡コーパス」を参考にしたと具体的に述べているわけなんです。それによりますと、見出し語形、表記欄、異形、補注などにおいて参考にしたと。言ってみれば、かなり限定的な参考の仕方なんですね。それから、コーパスを基にしましたコラムを250ほど掲載しています。それらの内容も、表記や語形、語の連接に関するものだけで、かなり限定的と言えるかもしれません。

このような限定的なことしかできなかったというのは、辞書の編纂にコーパスを活用するには、各語の用例の分析を本当にもう腰を据えてやる必要がありまして、それを行うだけの時間的な余裕とかノウハウがまだなかったといったことが最大の要因ではないかと思われるんです。これもまた今後の課題ではあるかなという気がいたします。

「現代国語例解辞典（第５版）」の活用事例についてなんですけれども、例えば見出し欄についてなんですが、「アイデア」と「アイディア」ということ。つまり、「アイデア」にするのか「アイディア」にするのかという、「ィ」の小さい表記が入るかどうかということなんですけれども、実は、「現代国語例解辞典」では、第４版－前の版なんですが－「アイディア」、小さい「ィ」入る形を本項目としていました。ところが、このコーパスを見ますと、小さい「ィ」が入るものの「アイディア」の頻度数というのが２０％台とかなり少数だったことが分かったんですね。そのために、コーパス（BCCWJ）を基にしまして、見出し語の形を「アイデア」という形に変えました。

それから、先ほど申し上げましたけれども、「メーク」と「メイク」なんですが、

これもＢＣＣＷＪでは「メイク」の方が優勢になったんです。そのため、見出し語の語形を、「メーク」だったものを「メイク」に改めています。さらに、余りここでは時間がないので省略しますけれども、「メイク」・「メーク」が使い分けなされているということも分かったので、これはこれで面白いなと思うんです。

あと、この「メイク」・「メーク」、「アイデア」・「アイディア」につきましては、ＢＣＣＷＪは話し言葉のデータは含まれていませんので、話し言葉のデータも含めた形で見られたら更によかったのかなと思います。これはコーパスへの要望なんです。

実は、このような見出し語の語形というのは従来、辞書編集者の主観によって決められていた部分が大きかったんですけれども、ＢＣＣＷＪによって客観的な判断材料が得られるようになったというわけです。

先ほど少し触れましたけれども、「カタン」についてです。コーパスでこの「カタン」というのを調べた結果、「現代国語例解辞典」の中で、補注とそれからコラムを作成することができました。補注の方は、「〈他人の荷物を担う意から「荷担」が本来の表記だが、現在は参加の意識から「加担」を用いることが多い〉」という内容です。今の画面の左側のところにその補注があります。

右側が、作成したコラムなんですけれども、前半は補注とほとんど同じ内容なんですが、さらに、「「思いやり」から「自分の意識」に主眼が移ったからなのか、「戦争に加担する」、「犯罪に加担する」など、ネガティブな単語と結びつくことが多い」と説明されています。

少し余談でもあるんですけれども、漢字の書き分けに関しては、コーパスは和語の異字同訓に関する情報を得る際に非常に威力を発揮しそうな気がいたします。細かな話なんですけれども、例えば「ノボル」の漢字表記について見てみますと、「ノボル」と書く漢字は、「上」、「登」、「昇」の三つが考えられますが、実際に2014年に発表されました文化審議会国語分科会の「異字同訓の漢字の使い分け例（報告）」にも、この三つの表記の書き分けについて触れられています。

この「ノボル」を、コーパスを使って実際調べて、「現代国語例解辞典」に反映させております。そうしますと、このようなことが分かりました。高いところへ移動する場合は「登る」、数量がある程度に達する場合は「上」、概念的な上昇の場合は「昇」と表されているようだというようなことです。かなりはっきりした使い分け

がなされているということが分かるわけです。

こういったコラムを「現代国語例解辞典」で作ったわけなんですけれども、この左側の円グラフを見ていただくとお分かりのように、「のぼる」と仮名表記のものがかなり多いことがコーパスで分かります。文化審議会の「異字同訓の漢字の使い分け例（報告）」では、飽くまでも異字同訓の漢字の使い分けについて説明するので、漢字をどのように書くかだろうと思うんですけれども、できれば、実際には仮名で書く人も多いんだということを何かしらの形で触れられているといいのではないかなという気がいたします。

最後に、今後拡充されるコーパスにつきまして辞書編集者としてどのような要望があるのか、簡単に述べさせていただきたいと思います。これも飽くまでも辞書編集者としてのことですので、それをお含み置きください。

一つ目、「データとして格納したサンプルは長めの文章が望ましい」と。「日本語書き言葉均衡コーパスの範囲というのは、書籍全般、雑誌全般、新聞、白書、ブログ、ネット掲示板、教科書、法律などのジャンルにまたがって、１億４３０万語のデータを格納しており、各ジャンルについて無作為にサンプルを抽出しています」としています。無作為にサンプルを抽出したというのは、著作権の問題もあってやむを得ないことだろうとは思うんですけれども、実際に辞書の編纂に使う資料としては不完全ですとしか言いようがありませんで、全文とは言わないまでも、もっと長めの例文が欲しいところです。それによって、意味の微妙な異なりや位相の違いなどが分かると思われるからなんですね。

それから２番目、「さまざまなジャンルの書籍のものをまんべんなく収録する」ということで、先ほども日本十進分類法に配されたその書籍を万遍なく対象としたコーパスが望まれると申し上げましたが、様々なジャンルの書籍が万遍なく収録されたものであってほしいと思います。

それから３番目、「書きことばに偏らずに、話しことば、また、マンガの吹き出しのせりふなども対象に」していただきたいと思います。いろんな様々な場面で使われているような文章が載っていると有り難いということです。

それからその次、「明治以降の書籍、文献を対象にする」ということです。収録対象の刊行年代については、「日本語書き言葉均衡コーパス」は、メインとなる書籍の場合は１９８６年から２００５年になるとしていますけれども、日本語の歴史を

考えた場合それでは不十分でありまして、少なくとも明治以降の書籍、文献は対象にしていただきたいと思います。

これは大変な作業になるということは分かっているんですけれども、特に私は現在も「日本国語大辞典」という用例主義の辞典に関わっていることもありまして、コーパスがそこまで対象にしていただけますととても有り難いと思います。

実は、このようなことを申し上げますのも、ちょうど１週間ほど前なんですが、「日本国語大辞典」の版元であります小学館が、「日本国語大辞典」の第３版の編集に着手することを発表したからなんです。刊行の予定は８年後で、2032年からなんですけれども、現在、収録数50万項目で100万用例の辞典なんですが、これをどこまで増補できるか。この、日本で最大の唯一の用例主義の辞典なんですけれども、１出版社のものとしてではなくて、全ての日本語を使う人々の辞典として、産学官一体となって育てていっていただければと思います。そのためにも、是非コーパスを活用して増補したいです。更に申し上げておきたいのは、実はこの「日本国語大辞典」もまた、言語資源の一つであるということなので、その点も皆さん、是非、心に留めておいていただければと思います。

さらに、次ですけれども、「より充実した検索ソフトの開発」ということで、コーパスを活用するためには、検索を行う検索ソフトの開発も欠かせないことだと思います。現在、ウェブアプリケーションである「少納言」、「中納言」とありますけれども、それが更に充実した内容になると有り難いと思います。

それから、「その他、ほしい情報として」ということで、いわゆる誤用のデータなんですね。これも欲しいんですね。というのは、誤用と言われる表現のデータの蓄積によって、誤用の広まり具合も分かりますし、それから、多分、文化庁で毎年やっています「国語に関する世論調査」のアンケート結果を裏付けるものもできるでしょうから、さらに、なぜそのような誤用が生じたのかというような解明する材料にもできると思いますので、是非これは残しておいてほしいなと考えております。あとは、分類語彙表の番号とのリンクとか、固有名詞の区別とか、詳細な読みの情報とか、絵文字とか、外字・画像で表示されたものも含めて検討していっていただければと思います。

ちょっと駆け足になってしまいましたけれども、以上、辞書編集者から見た言語資源ということでお話をさせていただきました。どうもありがとうございました。

○相澤主査

神永委員、どうもありがとうございました。そうしましたら、ただ今の御発表に関する御質問がございましたら、よろしくお願いいたします。意見交換の時間は別途、最後に設けさせていただければと思いますので、御質問ありましたらよろしくお願いいたします。

○石川副主査

大変啓発的なお話をありがとうございました。１点お伺いしたいのですけれども、出版社におかれましては、恐らくインハウスというんでしょうかね、社内で持っていらっしゃるある種の言語資源的なものもおありかと思うんですが、そうしたものを使う場合と、あるいはＢＣＣＷＪのようなものを使う場合、その辺りのあんばいというようなことはどのようになっておるんでしょうか。

○神永委員

出版社にもよるんだろうと思いますけれども、はっきり申し上げまして、小学館の中ではインハウスのコーパスで検索もできるような形のものは持っています。それを基にして見ることの方が私としては多いですね。

○石川副主査

ありがとうございます。

○相澤主査

私から１点、お伺いさせていただければと思います。誤用であるかどうかとか、あるいは、この言葉はもう死語であって使われていないとか、収録の紙面の制約から収録対象から外すべきであるといった、そういった判断が実は難しいのかもしれないと思ってお伺いしていたんですけれども、コーパスというのはそういったときに役に立つ、あるいは、どんなコーパスがあれば、「なくす」判断ができるのでしょうか。

○神永委員

そうですね、やっぱり、誤用かどうかの判断に関しましては、かなり古い時代のものからないとちょっと判断し切れないこともありますね。私は「誤用」という言葉を使わないようにしているんですが、今は誤用だと言われていますけれども、実は古い時代も、かなり古い時代に使われていたもので、それがやがて使われなくなって、誤用みたいに思われてしまっているというようなものも結構、古いところの例を見ていくとあるんですね。だから、やっぱり相当な年代のもの、長い範囲でのものがないと、判断はできないだろうと思います。

○相澤主査

ありがとうございます。

○齊藤（美）委員

御説明ありがとうございました。大変勉強になりました。コーパスを活用するとより質の高い辞書ができそうだというのも、お話を聞いていてよく分かったんですけれども、時間だったり労力だったり、あとはお金もだと思うんですが、コスト面としては、よりスムーズにできそうだとか、やはり大変なものは大変だとか、どういった感触でいらっしゃいますか。

○神永委員

そこなんですよね、問題が。先ほど申し上げましたけれども、限定的なことしかできなかったというのが、やはりそういったコストと時間との兼ね合いでして、本当にもう全ての項目について一つ一つやっていけばいいのかもしれませんけれども、なかなかそこまでやり切れないんですよね。そうすると、ある程度はピンポイントで、この語、この語みたいな感じでやらざるを得ない。だから、全ての語に対してちゃんとした客観的なデータを基にしてやっているかと言われると、ちょっと危ないかなというようなものもあります、実際に。

○齊藤（美）委員

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。　　（ → 挙手なし。）

ありがとうございます。続きは議論の時間に御意見を頂くことにいたしまして、続きまして、石川副主査に、英語のコーパスの整備状況とコーパスの活用事例ということで、配布資料４を使って御説明をいただくことといたします。よろしくお願いいたします。

○石川副主査

それでは、「継続・普及・実装」というテーマでお話をしてまいります。

最初に、お断りということなんですけれども、今回の話、飽くまでも個人による捉え方ということで御理解いただきたいと存じます。また、著作権の関係で、公開資料では図版を削除していることもお断り申し上げます。できるだけ誤りのないよう資料を作成したつもりですが、もし誤植・誤記が確認された場合は、何らかの形で修正を出したいと考えております。

さて、この小委員会では、日本語の言語資源について今後どういうことが必要か考えていくわけですが、私見では、恐らく三つのことが必要ではないかと思います。一つ目は、データを継続的に取っていくということですね。二つ目は、データを作るだけではなくて、それを普及させていくための工夫を考えていくということ。そして、三つ目は、最初の二つを出発点として、言語資源の社会実装の実現に向けた道筋を立てるということであります。

〈継続的データ収集の必要性〉

まずは、一つ目、継続的データ収集から考えていきたいと思います。

なぜ継続的収集が必要なのかということですが、一般に、コーパスというのは、言語母集団に対する統計学的な標本であるわけです。つまり、母集団が厳密に確定されて初めて資料のサンプリングが行えます。この意味で、日本語全体や英語全体を母集団にすることは事実上できないんですね。そういった大きな母集団だと、データの全体像、つまり、輪郭が固定しませんので。サンプリング基準が決められないわけです。そこで、例えば1980年代の英語や2010年の日本語というように、年代や年号の枠をはめて母集団を定義していく必要があるわけでございます。

こうしたコーパスを、均衡コーパス、標本コーパス、スナップショットコーパスなどと呼びます。時間の流れの中で言葉は変化していくわけですが、ある瞬間を決め、その瞬間をスナップショットで撮影するように固定し、そのときの言語の在りよう、例えばどんなジャンルがどのぐらいの比率で存在するのかなどを調べ、それがうまく再現できるようにコーパスを作っていくということになるわけです。

したがいまして、一般的なコーパスは、どれも、ある瞬間に限定されたもので、言語変化を追跡していくためには、コーパスを一度作って終わりとせず、データをずっと取り続けていかないといけないということになります。

ここで、継続的にデータを集める必要性をもう少し考えてみたいと思います。英語の場合、ＣＯＣＡというコーパスがありますが、これは1990年から2019年まで、毎年、同じサンプリング基準でデータを取り続けている面白いコーパスなんです。このコーパスで、例えば、助動詞の「ｍｕｓｔ」の頻度を５年刻みで調べてみると、1990年から2019年まで線形的に頻度が下がっているように見えます。数年程度の比較だと、本当に増えているのか減っているのか、あるいは偶然の変化なのか分からないわけですが、ＣＯＣＡのように、数十年のスパンにわたって同じサンプリング基準でデータを集めた資料があれば、様々な語や文法項目の使用度や用法がどう変わっているのかが見えてきます。もちろん、こうした判断を行うためには、データが真に信頼できる形で継続収集されている必要があるわけですが、そのようなものがございますと、個々の語や表現の使用度や用法の変化を科学的に検証することも可能になってまいります。

では、コーパス開発において、どのようにすればデータを継続的に取っていくことができるのでしょうか。英語の場合で考えてみたいと思います。

まず、二つの方向性がありまして、一つ目は、同一母体型、つまり、同一の機関ないし個人が継続的に取っていくという方向ですね。もう一つは、拡張分散型で、当初の開発者以外の第三者が自発的に参画してきて、それぞれの関心をベースとして、結果として、データが継続的に取られていくという方向です。順に御説明申し上げます。

まずは、同一母体型のうち、機関が中心になるタイプについてであります。英語のコーパスの中で、非常に大きく、影響力も大きいものとして、Ｂｒｉｔｉｓｈ　Ｎａｔｉｏｎａｌ　Ｃｏｒｐｕｓ（ＢＮＣ）というものがございます。これは1994年

に完成したものなんですが、ランカスター大学のほか、学習者向け辞書の編纂(さん)を手掛けるオックスフォード大学出版局やロングマン社などが協力して構築した、当時としては最大のコーパスで、コーパス言語学全体の歴史の中においても記念碑的なコーパスということになっております。このコーパスを使って世界中で多くの研究論文が書かれたほか、各種の英語辞書の改訂にも使われました。

このコーパスは1994年の完成後、幾つか微修正やデータフォーマットの更新などがなされますが、基本的にデータは変わっておりませんので、コーパスに入っているものは全て1994年より古いデータということでございます。均衡コーパスの宿命は、完成した瞬間から古くなるということで、ＢＮＣも1994年より新しいデータがないわけですので、ＢＮＣには「スマホ」も「アイフォン」も入っていませんし、windowsと言えばほぼ全て窓だということになるわけですね。

このＢＮＣですが、膨大なコストを掛けてようやく完成したものですから、同じ枠組みで新しいデータを取ることはもう二度と無理だろうと思われていました。しかし、最近になって、ＢＮＣ2014というものがリリースされました。初版の1994年から見て20年後のデータということになります。担当者は別ですが、初版を作ったランカスター大学が引き続き関与しました。さて、この新しいＢＮＣ2014でwindowsを調べますと、大文字のWindowsが多く出ており、きちんと新しいデータが入っていることが分かります。ＢＮＣ2014のデータは、後で御紹介するＢＮＣlabというサイトやＣＱＰwebという検索プラットフォーム上で誰でも検索できます。さらに、書き言葉と話し言葉のデータ全体がLancsBoxというコーパス検索プログラムにバンドルされておりまして、これも無償でダウンロードして活用できるようになっております。

以上、同一母体の中でも、大学等の機関が中心になってデータの継続収集がなされた事例を紹介しました。次は、同一の個人研究者が継続的にデータを収集している希(け)有な例を御紹介したいと存じます。

先ほどもお話ししたＣＯＣＡですが、正確にはＣｏｒｐｕｓ　ｏｆ　Ｃｏｎｔｅｍｐｏｒａｒｙ　Ａｍｅｒｉｃａｎ　Ｅｎｇｌｉｓｈと言います。これは米国のブリガムヤング大学の教授であったマーク・デービス氏が2008年頃から構築を開始したもので、基本的には個人研究として、以後ずっとデータを集めてこられたわけですね。

このコーパスの作り方は、ウェブ上にあるアメリカ英語の書き言葉・話し言葉を母集団とみなし、同じサンプリングフレームで毎年データを取っていくということでございます。テレビと映画、話し言葉、小説、雑誌、新聞、学術という６分野から、毎年、等量のデータを集めいています。現実の言語分布を考えると、単純に等量でよいのかという批判はあり得るわけですが、そもそも、ジャンルごとのデータ量は毎年変わっていくもので、その意味で、長期的に継続してデータを集めようとする場合には、便宜的に、等量、つまり同比率にするというのは、理解できる考え方ではあります。こうして、６分野から各400万語、合計で毎年2,400万語のデータを取っています。

これまでに、1990年から2019年までの30年間で、６分野合わせまして7.5億語を収集しています。それとは別に、2012年の時点で収集したブログやウェブデータがあり、これを加えまして、現在、10億語のデータとして、無償で公開されております。

以上で、同一母体が継続的にデータ収集を行ったものとして、同一機関が関与したＢＮＣ、同一個人が関与したＣＯＣＡという二つの例を御紹介申し上げました。

さて、英語コーパスの継続的データ収集にはもう一つ別の方向がございまして、それが先ほど申し上げた拡張分散型です。当初のプロジェクトに関わっていなかった第三者が自発的に参画してきて、それぞれ、元のコーパスを自分なりに受け継いで、自身の関心でデータを集めていくという、言わば、分散型のデータ収集ですね。

この分散型システムの基になっていますのが、アメリカのブラウン大学で1964年に作られたＢｒｏｗｎコーパスです。これは1961年のアメリカ英語の書き言葉をバランスよく集めた100万語のコーパスです。１サンプルは2000語、これを500本集めて合計100万語になっています。

このサンプリングフレームが非常にシンプルで明瞭であったため、いろいろな人が同じサンプリングフレームを使って別のデータを集め始めました。これにより、Ｂｒｏｗｎと比較できるデータが各種作られることになりました。例えば、Ｂｒｏｗｎのデータ年である1961年を起点として、時代変化を見ようということで、1931年や1991年のデータが集められました。さらに、Ｂｒｏｗｎが集めたのはアメリカ英語ですが、地域差も見ようということで、イギリス英語、インド英語、オーストラリア英語、ニュージーランド英語なども集められました。こうしたコーパスが散発的・分散的に作られたのですが、元となるサンプリングフレームは同じですので、

これらは結果的に相互比較なものとなり、これらを総称してＢｒｏｗｎ　Ｃｏｒｐｕｓ　Ｆａｍｉｌｙなどと呼んだりしております。このサンプリングフレームは他の言語にも適用可能で、私自身、Ｂｒｏｗｎの小説セクションをまねて日本語小説のコーパスを作ってみたこともあります。以上、拡散分散型の継続収集の例としてＢｒｏｗｎとその関連コーパスを御紹介しました。

本節のまとめでありますが、１点目は、「コーパス作りは、実質的には、機関でなく個人が担う場合が多い」ということです。２点目は、「多くの場合、当該教員が転出・退職すれば、機関としてのコーパス開発はそこで終わる」ということです。ＣＯＣＡがその例で、当初はブリガムヤング大学という機関の枠組みで開発が行われていましたが、当該教員の退職後は個人による運営に切り替わっています。実際、大学や研究所と言っても、実際にコーパス構築の作業を担うのは、一人ないし少数名の研究者です。コーパス作りというのは職人芸的なもので、全体の設計、データの収集法の決定等、個人研究者の専門的判断に依存する部分が多く、これを機関における業務として組織化していくのは、体系的な支援が必要です。機関に対する長期的な財政支援がないと、当該研究者のノウハウを他の研究者が継承できず、結局、中心人物が辞めてしまうと、ノウハウごと失われてしまうということが起こりかねません。３点目は「教員が大学等を退職後に、個人として収集を継続しようとしても、研究資金（収集費、開発費、サーバー等維持費）が不足したり、コーパスの所有権を主張できなかったりして、開発が止まる場合もある」ということです。幸い，ＣＯＣＡはそうはなりませんでしたが、英語では、中心的に構築されていた先生が大学を退職後、データ自体が非公開になってしまった例もあります。４点目は、「同一機関における継続収集が可能になるのは、当該機関において先行者を継承する研究者がうまく出現した場合に限られる」ということです。ランカスター大学がＢＮＣ1994とＢＮＣ2014に継続して関与し得たのは正にこのためですが、これは例外的と考えておく方がよいでしょう。最後に５点目は「コーパスのサイズを合理的に圧縮し、かつ、踏襲可能なサンプリングフレームを提供することで、多様な第三者の参画を促す方向もあり得る」ということです。１億語、10億語といった巨大コーパスであれば、機関が時間と予算を掛けてやるべきものでありましょうが、それとは別に、もう少しサイズを落として、かつ作りやすいサンプリングフレームを公開することで、Ｂｒｏｗｎのような形で、幅広い人が自由に参画をし、自分の関心に応

じて様々なデータをアジャイル的に集めていく、そういう動きが本邦で広がることも期待したいです。つまりは、機関と個人、同一母体型と分散型、ツートラックでのデータの継続的収集が望ましいのではないかと考える次第です。

以上で一つ目のセクション、継続的データ収集についてお話ししました。

〈コーパス普及を促進する取組の必要性〉

続きまして、本日のお話の二つ目のポイント、普及のための工夫についてお話をします。

コーパスでは様々な検索の手法があるのですけれども、最も代表的な手法は、ｋｅｙ　ｗｏｒｄ　ｉｎ　ｃｏｎｔｅｘｔ、略してＫＷＩＣと呼ばれる方式です。大層な名前が付いているんですが、発想は非常にシンプルで、中心語を中央に、その文脈を前後に並べて表示することで、「文脈における語や表現の使われ方」を調べるというものです。中心語の左側や右側で並び替えをすることで、キーとなる語や句が、他のどういう語と共起しやすいかが見えてきます。

さて、用例からパターンを抽出するトレーニングを積んでいる言語学者や辞書編集者にはＫＷＩＣは非常に使いやすい出力形式なのですが、それ以外の方にとっては、ＫＷＩＣ画面だけ見て、そこにどんなパターンが隠れているか言ってみろというのは、非常に難しい作業なんですね。つまり、幅広い層に広く使ってもらえるよう、コーパスの普及を考えるならば、コーパス検索結果についても、ＫＷＩＣ以外のアウトプット方法を考えていく必要があるでしょう。

コーパスを普及させるための分析結果の見せ方に関して、英語の世界でどういうことがなされているのか、幾つか御紹介します。まず１点目は、「共起語の一括表示」です。ＣＯＣＡの検索画面で、ある語を入力すると、その後の近傍に出現しやすい語が、頻度を示す棒グラフと共に一覧で出てきます。例えば、「big」と「large」のおのおのについて共起語を調べると違いがすぐ分かります。

２点目として、語と語の結び付き、つまり、コロケーションをもっと見やすくできないかという工夫もいろいろなされています。ランカスター大学で作っているLancsBoxというソフトでは、ある語を入力すると、その語と関係の深い語の全体が、円形のコロケーションネットワークとして、ネットワークグラフの形で出力されます。この辺りは日本語のコーパス研究でも是非取り入れたいところですね。

３点目は、ジャンルの変化や時代別の変化を直感的なグラフで示すものですね。

ＣＯＣＡの検索画面などでこれが可能です。

４点目は、同じくＣＯＣＡで実現されているものですが、類義語や反義語などをペアで同時に入力すると、片方に出やすい語を統計的に選び、自動的にハイライトして表示してくれる機能です。２語の意味や用法の違いを探る上で便利です。

５点目は、関連情報の一括集約表示というもので、ＣＯＣＡですと、ある語を入力すると、その後に関してコーパスから得られるあらゆる情報を一つの画面で一挙に表示してくれます。例えば「large」を調べると、その後が英語の各ジャンルでどのくらい使われるのか、どのような定義があるのか、類義語には何があるのか、largeを含む連語―クラスターと呼びますがーにはどのようなものがあるのか、largeが出やすいトピック・話題は何か、largeの共起語にはどんなものが多いか、どんな単語がlargeと語形的に関連しているのか、さらにはlargeが具体的な文脈の中でどのように使われているのか、これら全てがコーパスから自動で引っ張り出され、一つの画面に一度に出てくるわけですね。辞書の項目記述がオンラインで勝手に作られていると言ってもいいような状況が既に実現しているわけなんですね。

最後に６点目はＢＮＣ2014の話し言葉データを検索するＢＮＣｌａｂという検索サイトの機能ですが、ここでは、入力された語に関して、話者の性別、年齢、居住地、社会階層などで頻度がどう違うか、また、ＢＮＣ1994とＢＮＣ2014の20年間で頻度がどう変わったか、さらには同じＢＮＣの書き言葉と話し言葉で頻度がどう違うか、これらが全て、美しいグラフで出力されます。もちろん、こういう比較では誤差の可能性に注意しないといけないわけですが、このシステムでは裏で仮説検定が走っていて、様々な要因の中で統計的に有意な要因だけがきちんと分かるようになっているんですね。

本節のまとめとしては、「コーパスを普及させようと思うなら、検索システムの工夫も必要で、その際、ＫＷＩＣだけでは十分ではない」ということです。例えば、中学校や高校の先生方が、授業の準備のときにコーパスをちょっと使ってみようと思ってくださったとしても、結果がＫＷＩＣでしか見られないのであれば、やはり、コーパスの本当の価値というところまでは実感していただきにくいのではないかと思います。この点については、言語学者・辞書編集者に限らず、幅広い一般国民を想定ユーザーと考えた場合、どんな形であれば使いやすいか、どういう形であればもっと使ってもらえるか、産官学の視点も含め、いろいろ議論していく必要もある

でしょう。

〈社会実装に向けた取組の必要性〉

以上で、一つ目の「継続」、二つ目の「普及」と話してまいりました。最後に三つ目の「社会実装の実現」について４点お話しします。ただ、もう時間がほぼ来ていますので、本日は見出し項目のみでお許しを頂いて、また機会があれば、もう少し詳しくお話ししたいと存じます。

まず１点目、外国人向けの第二言語としての英語(English as a second language: ＥＳＬ）辞書への応用ということで、英語におきましては、完全なコーパス準拠の辞書というのが既に普及しています。1987年のCollins Cobuildの辞書がこの種の試みの嚆矢（こうし）とされています。この辞書の初期版の表紙には、Helping learners with real Englishと記載されており、realには下線まで引っ張ってあるんですね。「本物の英語」で学習者の手助けをする辞書が遂に完成したという著者の気持ちが伝わります。日本語の場合、なかなか全面コーパス準拠の辞書というのは出ていないわけですが、この背景には、国語辞書の主なユーザーが母語話者だということもあるように思います。やはり、母語話者と学習者では、辞書に望むべきところが違いまして、コーパスから分かる細かい用法パターンや各種の用例というのは、もちろん母語話者が読んでも面白いんですが、そうした情報をより求めているのは、やはり母語の直観を持たない学習者なんですね。英語は、全世界に圧倒的な数の学習者がおりますので、その人たちのニーズが追い風になって、コーパス準拠辞書が広まったという部分がございます。

２点目として、コーパスが社会の言語規範にも大きな影響を及ぼしています。これはＢＮＣ2014の話し言葉版がリリースされて数年後にＴｈｅ　Ｔｅｌｅｇｒａｐｈという新聞に載った記事ですが、ＢＮＣ2014のデータが公開されたことで、イギリス英語の古い文法のルールの見直しの機運が高まったことが報道されています。コーパスができたことで新聞記事が書かれ、それを契機として社会的な議論が起こり、やがては人々の言語観が更新されていくというのはコーパスのもたらす大きな社会的インパクトと言えるでしょう。

３点目は、言語教材の変化です。ケンブリッジ大学出版局ですが、ケンブリッジのコーパスを全面的に使った英会話教材を開発しています。会話の教本というと、決まり文句的なやり取りを暗唱するイメージがありますが、この教材では、様々な

表現について、口語においてはどういう用法が何％ぐらい用いられ、高頻度の用法トップ３は何かといったことが、非常に細かく紹介されています。母語に直接触れることができない海外の学習者にとって、こういう情報は有り難いものになると思われます。

最後に４点目として、コーパス言語学の社会的応用についてです。たとえば、本年、2024年には、*Corpus Linguistics for Health Communication*（健康関連コミュニケーション研究のためのコーパス言語学）というような本も出ています。社会における健康増進や社会福祉の改善におけるコーパスの貢献がテーマになった本です。例えば、肥満（obesity）というのは、英国に限らず多くの国で健康上の課題の一つですが、新聞コーパスを用いてobesityの共起語や文脈を検討することで、それが社会的にどう捉えられているかが解明でき、それを基にすることで、その治療をどうやって進めていけばよいかのヒントが得られます。この本には様々な研究が紹介されていますが、その一つでは、肥満症の患者さんとお医者さんの対話集会での録音データが分析されており、医師との意見交換を経て患者さんの意識が段階的に変化していく過程が調査されています。また、別の研究では、医師と患者の会話、看護師と患者の会話、一般の人の会話を比較し、医療者たちが無意識のうちに使っている特有の言い方を特定しています。こういうことが分かってくると、医師や看護師の養成教育において、どういう話し方をすると患者と良い関係が築けるのか、効果的に指導していくことも可能になりそうです。

本節のまとめとしましては、コーパスは、社会の問題の解決に様々な形で貢献できるということです。言語教育や、社会的な言語意識はもとより、医療や公衆衛生の改善にも有用です。こうした可能性は、日本語でも今後大いに模索されていくべきでしょう。

〈全体のまとめ〉

さて、本日のお話全体をまとめますと、コーパスは作りっ放しにして終わりにするのではなく、継続・普及・社会実装の３点を考えていくべきだということになります。この方向を考える場合、四つのことが重要になろうかなと思います。

一つ目は、コーパス開発を手掛ける主要な研究機関への継続的支援であります。本日のお話で、コーパス作りというのはかなりのところまで属人的であると申しました。数年間の競争的資金しかなければ、資金が尽きればその作業を担った人が解

散してしまい、ノウハウが失われてしまいます。この意味で、10年、20年といったスパンで息の長い基金を用意し、それを用いて、研究機関が次の世代の作業を担う人材を育てていけるようにすることが重要だと考えます。

二つ目は、前川委員の前回のお話にもあったんですが、言語資源の開発やメンテナンスが研究としてなかなか評価されにくい状況がございます。この点については、言語資源の開発・整備そのものを一つの研究分野として積極的に認めていくことが重要ですし、また、そのような方向に進むよう、関係者が尽力する必要もあるでしょう。例えば、科学研究費の研究分野の中に、「言語資源開発整備」を立てるというようなことも一つの方向性として考えられます。こうした枠があると、科研を得てコーパスを構築し、さらには別の科研を得て、過去に作ったコーパスを継続的に開発・拡充していくことがやりやすくなるのではないかと思われます。また、コーパス検索ツールの開発もこうした枠組みの中で取り組めるでしょう。なお、言語資源の共有財としての性質に鑑み、今後、科研で構築された言語資源については一般公開を義務付けるといったことも考えられます。

三つ目として、今日のお話で、コーパス普及のためには、一般の人が使いやすいような検索インターフェースの開発が必要だと申しました。これについては、例えば10人が10通りのインターフェースを作るのもよいでしょうが、加えて、みんなで使えるものを１個作っておいて、それをオープンプラットフォームにする。そこにみんなが自分のデータを放り込めば、同じインターフェースで様々に高度な検索が簡単にできるようになるといった方向も考えたいところです。英語の場合だと、国ではこれがＣＱＰｗｅｂや、Ｓｋｅｔｃｈ　Ｅｎｇｉｎｅといったシステムで、こうしたことがある程度実現しています。日本語においても、共有型の高性能コーパス検索インターフェースが作れないかといった夢も広がるところでございます。

四つ目は、言語資源の素材となり得るデータへのアクセス許可を国や公共団体にお願いしたいということです。医療、保健、工学、ビジネス、政治経済分野における発話データや書き言葉のデータは、コーパス言語学の手法を用いて分析することで、社会が抱える様々な課題の解決に大きく役立てることができるものです。しかし、こうしたデータへのアクセスは、現状では、個人研究者には極めて困難です。例えば、特別な研修等を経て、個人研究者にクリアランスを与え、匿名化されたデータへのアクセスを認めるといった公的な仕組みの整備もまた重要であろうかと考え

ます。

最後に、日本語言語資源の価値が広く社会に認知され、収集したデータがもっと広く活用されていくために、三つのことが重要であるということを改めて申し上げて、本日のお話を閉じさせていただきます。

１点目は、データは１回取って終わりとするのではなく、長いスパンにわたって継続的に取っていく、それによって初めて言語の変化が見えてくるということです。

２点目は、コーパスは言語データを集めるだけで終わりではなく、集めたデータから様々な情報が簡単に引き出せるような検索システムを工夫することもまた重要だということです。

３点目は、コーパスを言語学者や辞書編者のための道具にとどめず、社会実装を目指していくべきだということです。そのためには、まだまだ眠っているデータがありますから、そういうデータを利活用する仕組みも必要であろうということです。

以上で私からの報告を終えます。どうもありがとうございました。

○相澤主査

石川副主査、ありがとうございました。では、ただ今の御説明について御質問、ございましたら、よろしくお願いいたします。

○神永委員

一つよろしいですか。ちょっと辞書のことでお聞きしたいんですけれども、先ほどＣｏｌｌｉｎｓ　ＣＯＢＵＩＬＤ　Ｅｎｇｌｉｓｈ　Ｄｉｃｔｉｏｎａｒｙのお話をされていまして、コーパスを全面的に活用した辞書だったということで、英語学習者向けの辞典だったというお話でしたけれども、実際にその内容が大きく今までと違うような内容になったときに、従来のネーティブ向けの辞書との問題というか、その齟齬（そご）というのは、大きな問題にならなかったんでしょうか。

と言いますのは、実は私も前々から考えていて、それが実現せずに終わってしまったというか、もう出版社にいないものですから無理なんですけれども、日本語学習者向けの辞典を前々からやりたいと思っていまして、それができないんですけれども、もしこのコーパスを活用してやったときに、やっぱりそういう問題はひょっとすると出てしまうんじゃないかなという気がするものですから、お聞きしている

んです。

○石川副主査

そうですね、英語の場合は、日本語に比べると圧倒的に学習者の数が多いですよね。全世界に英語学習者がいますので、マーケットとして圧倒的にそちらの方が大きいわけですね。ですから、そのマーケットのニーズに即して辞書が作られる、そのために辞書が自然とコーパス準拠の方向に変わってきたというように理解しています。

○神永委員

それはもしかしたらあるかもしれない。やっぱり日本語学習者の数は、増やそうとはしていますけれども限られているんで、結局そのネックになったのが我々もそこで、日本語学習者向けの辞典ができなかったというのもそこなんですね。

○石川副主査

よく分かります。先ほど少し申し上げたのですが、母語話者が辞書に求めるものと学習者が辞書に求めるものは若干違う可能性があると思います。

○神永委員

そうかもしれませんね。

○石川副主査

うろ覚えで恐縮なのですが、例えば、「右」という語をどう定義するかというときに、国語辞典では、「心臓があるほうの反対側」とか、「時計の文字盤で０から６のある方」とか、こういう気の利いた語釈をそれぞれ考えますよね。母語話者の読者はそこを読んでにやりとする……。

こういうのは非常に母語話者的な辞書の読み方ですけれども、学習者はそうではありませんで、もし、日本語学習者が「右」を辞書で調べるとすれば、「右」は漢字で書く場合と仮名で書く場合、どちらが多いのか、「右」の前後に来る単語はどういうものがあるのか、「右」を含むよく出る表現にはどういうものがあるのか、何

よりも、自分自身が日本語で「右」を使うときにまねできるようなお手本が欲しいと彼らは考えるでしょう。こうしたニーズの違いというのが恐らくあるのだろうと思います。

ただ、日本語でも、将来、コーパス準拠の本格的な学習者辞書が出てくると、母語話者から見ても面白いなとなり、やがては母語話者にもそうした辞書が受け入れられていくという可能性はあるだろうと考えています。

○神永委員

ありますね。分かりました。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

では。よろしくお願いいたします。

○齊藤（美）委員

大変詳しい、たくさん御講義をありがとうございました。勉強になりました。

もし聞き逃していたら申し訳ないんですけれども、アメリカンイングリッシュとかブリティッシュイングリッシュと言うときのアメリカンとかブリティッシュというのは、例えばＣＯＣＡで言えば、アメリカで発行されている雑誌、アメリカで発行されている新聞、アメリカの学術誌というようなことで、アメリカンだというふうに判定されるんでしょうか。著者が誰かというところは、先ほどイギリスの方で地域差とかの話も出していただいたので、出身地はアメリカじゃないけれどもアメリカで活動している人が書いた記事とかも含まれるという可能性はあるということでしょうか。

○石川副主査

そうですね、御指摘のように、ある話者やある言語種の範囲をどう定めるのかというのは、英語のコーパス研究でも古くからある問題です。現状について言うと、

書き言葉につきましては、その国において出版された、その国の言葉で書かれたものという捉えをしている場合が多いかと思います。アメリカ英語の書き言葉を母集団にするBrownコーパスの場合も、アメリカの出版目録のようなものから標本を決めており、その意味で言うと、やはり、書き手の出身国うんぬんではなく、アメリカで出版されたもので、英語で書かれていれば、対象にみなすというふうなことになっていると思います。

話し言葉の収集につきましても、基本的にはその国に暮らして英語を使っている人が対象になります。全般的に見て、生まれという意味でのネーティブ性に過度にこだわらないというのが、英語のコーパス構築の基本的な方向かなと思います。この点に関して、応用言語学では、排外的なニュアンスを帯びがちな「ネーティブスピーカー」という概念を批判的に捉え直す動きが広がっており、そういったことも関係しているのではと考えています。

○齊藤（美）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。

私からも一つ。非常に基礎の質問となってしまい恐縮ですけれども、２番目の、普及のための工夫のところで、参考例として、いろいろなコーパスの見方、コーパスからの知識の得方について御紹介いただいたことに関する質問です。この背後には非常なアノテーションの労力があるのではないかと感じまして、その辺りのコストについてどうお考えかということと、今、例えば私の分野で言えば計算機、ＡＩを使うと、ある程度アノテーションは自動化はできるけれども、それほど正確ではないかもしれないという状況が出てきたとして、そういうものは果たして役に立つのかという点について、御質問させていただけますでしょうか。

○石川副主査

ありがとうございます。ＢＮＣの話し言葉データの収集では、話者一人一人について、性別、年齢のみならず、社会階層なども細かく調査して、これらをアノテー

ションしてテキストとひも付けた分析もできるようになっています。

このほか、品詞のアノテーションは一般的に普及しておりますし、構文のアノテーションや、意味論的アノテーションがなされているものもあります。意味論のアノテーションについては、意味タグ（セマンティックタグ）の付与を半自動で行うシステムもあるのですが、基本は辞書ベースでして、ある語に対して、辞書の定義を根拠として、「体に関する語」や「色に関する語」のようなタグを貼っていくわけです。一方で、文脈の中での意味変化などにもうまく自動で対応できるシステムというのは、知る限りでは、まだ一般的になっていないように思います。おっしゃるように、コストの掛かるところではありますが、あるところまでは機械化し、その後は、人手で修正するというふうなことになってくるのかなと思います。もちろん、AIにも期待したいところです。

○相澤主査

どうもありがとうございました。ほかは、御質問等いかがでございましょうか。

（ → 挙手なし。）

そうしましたら、改めて石川副主査、御発表ありがとうございました。

○石川副主査

ありがとうございました。

○相澤主査

それでは、本日御説明のありました辞書編集の現場でのコーパス活用事例や英語のコーパス事例を踏まえて、自由な意見交換を行いたいと思います。検討の観点の中で、日本語コーパスの活用の在り方やニーズに即したコーパスの在り方といった観点も踏まえつつ、自由に、委員の皆様の専門分野におけるコーパスの活用事例や可能性、あるいは御知見を踏まえて、御発言を頂ければと思います。どうぞ御自由に、皆様、１回ずつは御発言いただく心積もりで、よろしくお願いいたします。これまでの質問の続きでも結構でございます。いかがでしょうか。

○齊藤（美）委員

事前に、コーパスをどんなふうに活用した経験があるかというのがもしあればというメールを頂いていたので、ちょっと考えてみたんですけれども、私自身が研究においてコーパスという名称が付いたものを使ったことは残念ながらないんですけれども、言語データベースを活用した経験はちょっとあるので、少しお話しさせていただければと思います。

私、翻訳の研究をしておりまして、特に明治時代を対象に研究しているんですけれども、翻訳によって作られた言葉、あるいは翻訳によって新しい意味が与えられた言葉を「翻訳語」と呼びます。「国民」や「民族」というのは翻訳語なんですけれども、それがどのように広まっていったかというのを研究したときに、国立国会図書館の当時は「近代デジタルライブラリー」と呼ばれていた、今の「デジタルコレクション」だと思うんですが、2015年辺りに研究したことなんですけれども、そこでタイトルや目次を調べまして、明治の何年代に発行されているものに「国民」や「民族」がどれぐらい出てくるというようなことを、数を調べたらだんだん増えていますねというようなことを―明治30年代から「民族」が増えてきて、それよりもうちょっと早い時代から「国民」は増えているとか、そういったことを調べたことがありました。

そのとき共同研究者だった、坪井睦子先生という方が書いた、同じ研究テーマでやっていた単著の論文なんですが、そこでは国立国語研究所編の「太陽コーパス」をお使いになっていて、やはりタイトルと本文において「国民」、「民族」がどれぐらい出現するかというのを調べていました。出てきたものに関しては、記事の内容も読んで、どういう文脈で使われているかということ、また、どういう意味を与えられているかということを分析なさっていました。

最近私が書いた佐藤美希先生という方との共著論文では、また同じ「太陽」という雑誌を研究対象にしたんですけれども、そこでは、翻訳についてどういうことが語られているかというー「翻訳言説」と呼んでいますがーそれを調べたくて、明治時代の発行分だけ扱って「太陽」を調べたんです。そのときには、「太陽コーパス」の方は、限られた年の分しかコーパスが入っていなかったので、インターネット上のデータベースで「ジャパンナレッジ」というのがあると思いますが、その中のコンテンツの一つである日本近代文学館を使用しまして、そうすると「太陽」の全ての号が閲覧可能で、キーワードで記事の検索ができますので、それで「翻訳」と入

れて選ばれた、ヒットした記事を全部分析して読んでいったということをしました。

こういった研究をしていますので、「国民」、「民族」以外の翻訳語も含めて、普及を調べる上であったり、あるいは翻訳言説、翻訳についてどんなことが語られているかというのを調べたりする文書を探す上でも、もしさらに充実度の高いコーパスがあれば、私の研究には必ず活用できるだろうなと思っています。今あるコーパスでもちろん活用できるんですけれども…。

あと、ちょっと関連しまして、大学で教員もしていますので、教育面においてどのように使えるかなと考えてみますと、卒業論文の指導、レポートの執筆とか、プレゼンテーション作成などに当たって、辞書をもちろん見なさいということは言えるんですけれども、辞書に続くよりどころとして、ふだん日本語で会話するけれども、ちょっと学術的な文章だったり学術的な話をしたりするときー言葉ということで、日本語が第１言語ではない学生や院生もいますので、そういう人たちを含めてー先ほども用例の話がありましたけれども、用例を調べる目的で結構使えるんじゃないかなと思いました。使いなさいと勧められるんじゃないかなと思いました。

この活用法においては、やっぱり、学術文書や研究書というジャンル、レジスターのデータがあるといいだろうなと思いました。

あとは、私の個人的な言語生活とかを考えますと、だんだん辞書でも、元々誤用だったけれども今は使われるようになっているというような説明がよく載っていると思うんですが、例えば、私が学内で見た文書で「後ろ倒し」と書いてあって、「後ろ倒し」は間違いというか、ちょっと良くないかなとか思いながら、特に何も言わずに読んでいたんですけれども、でも、大手新聞社の記事を読んでいましたら、結構出てくるものだなと気付きまして、もういいのかなと何となく思ったんです。けれども、その何となくなところをコーパスでちゃんと調べられれば、先ほど石川副主査も、だんだん減っていますとかいうのを「ｍｕｓｔ」なんかで見せていただきましたけれども、より納得度が高く言語使用できるんじゃないかなと思っております。

○相澤主査

どうもありがとうございました。では、ただ今の御発言に対してでも結構ですし、全く新しい御意見でも結構でございますので、何かありましたらお願いいたします。

○神永委員

今お話のありました国会図書館の「デジタルコレクション」に関しましては、私も非常に活用させてもらっていまして、「日本国語大辞典」で用例がないものがあるんですよ、いまだに。それから、もっと古い例が絶対にこれはあるはずなのにそうでもないというのがありまして、それがかなりなものがそこから取れるというのがありまして、今、本当に時間があるときにのぞいて、本当にもうピンポイントでその言葉だけ探しているというだけなんですけれども、ただ、調べるのがとても大変なんですね。もう本当にバーッと出てきちゃうんで、それをどうしようかと思いながら、その意味もそれぞれ振り分けていかないと……。

特に、意味が変わっていくものが非常に問題でして、例えば「勉強」なんていう言葉がありますけれども、「勉強」も、その本来の意味の、無理に何かさせるみたいな意味の「勉強」が元々の意味なんですけれども、それが、学問する「勉強」にどうも明治の初めぐらいなのかな、多分幕末ぐらいだと思うんですけれども、変わっていくんですね。その変わるところを探したくて調べたことがあるんですけれども、そうすると、かなり「デジタルコレクション」は有効なんですけれども、いろんな「勉強」が出てきちゃいますから、振り分けていかなきゃいけないという、ちょっと大変なこともあります。

そのような苦労をしながらですけれども、でも、私はとても国会図書館の「デジタルコレクション」というのは有り難いデータだと感じています。

○相澤主査

ありがとうございます。近代デジタルライブラリーということですけれども、古典文学を御専門にされていらっしゃいます植木委員、いかがでございましょうか。古典文学という観点からコーパス、何か御意見ございましたら、よろしくお願いします。

○植木委員

ありがとうございます。私自身も歴史コーパスは使ったことがあるので、研究者としても、学生たちももうそれはフルに活用して用例の調査などはしていて、非常

に有り難いものだと思っています。今日も言葉の誤用のことがいろいろ話題になりましたが、古語を現代に使う場合として、すごく細かい例を申し上げますと、結婚式で女性の司会者の方が、「○○さんはやんごとなき御用事で今日は欠席されます」とおっしゃったんです。

それは、多分、「よんどころない」とおっしゃりたかったんだろうなと思うんです。「高貴な用事」というのは余り考えられないと思うので。私は大変驚いて、おかしいと思ったんですが、皆さんは全く普通に聞いていて、それは、「やんごとなき」という言葉の意味が理解されていないからなのか、皆さんの頭の中で「よんどころない」に置き換えられていたのか、ちょっと分からず、突飛な例と言うか、珍しい例かもしれないんですが、これは古語であるのか、現代語としてまだ生きているのか……。前回も、言葉を取る範囲というようなことが話題になったんですけれども、古語と現代語の切れ目と言いますか、そういうことをどう考えたらいいのかということを、素人ながら思ったことがありました。

皆さんが使わなければ、言葉として見出しとして出てこないとは思うんですけれども、また一方で、その言葉を、例えば女性が多く使うとか、そのような使われ方を調べていく中で、古そうな言葉を使うのにふさわしい、多分、結婚式の披露宴というのは少し重々しい場所ですので、そういうところで、ちょっといつもと違う言葉を使おうとして古語が出てきたのかもしれないので、使われる場面と古語との何らかのつながりがあるのかもしれません。どう整理していいかは全然私も分からないんですけれども、現代においても古語が使われる場面や誤用をどういう形で整理したらいいのかというようなことを、個人的な疑問、感想として持っております。

○相澤主査

ありがとうございます。多様性という観点もあるかなと思いながら伺っていました。私などは飽くまで理系的な発想ではありますけれども、言葉を使うときに和語由来か漢語由来かちゃんと考えなさいなどを、教わりながら勉強したんですけれども。武田委員、いかがでございましょうか。

○武田委員

そうですね、今、ちょうど古語のお話も出ましたので、例えばですけれども、少

し気になるのは、「すべからく」という表現がありまして、一般的には漢文訓読に由来するということで、「すべからく……べし」という受け方だと思うんですけれども、近年は「ほぼ全て」の意味で使うようになってきまして、「すべからく……である」というような言い方も見られるようになってきました。

こういったものを辞書でどう対応するかということが、正に、日々、辞書編集においては問題になっているわけです。辞書を幾つか調べてみましても、対応しているものもあれば対応していないものもあり、そういったところで、こういうコーパスなどがあって実情を捉えることができれば、単に規範的になるだけではなくて、実情を踏まえた解説も検討していくことができるのではないかと思われております。

本日、神永委員より、辞書に関する事例ということで御紹介いただきましたので、若干それを補足させていただくような形で、私も出版社に所属しておりまして、辞書の関係の業務についておりますので、その点で補足になればと思います。

神永委員からは特定の事例という形で御紹介いただきましたけれども、私どもでも幾つかの国時辞典を刊行しておりますが、いずれもコーパスは既に利用しておりまして、コーパスがなければ、なかなかいい記述ができないという状態になっております。

辞書もかなり数が出ておりますし、また、読者対象も様々ですから、どこまで適用できるかということは分かりませんが、ただ、私どもで複数刊行している辞典では、既にコーパスも利用しながら、例えば今のような、「すべからく」をどう記述するかとか、そういったような事例も検討しているところかと思います。そういった中では、神永委員のお話と大体重なりますけれども、項目の採否の問題ですとか、あるいは適切な用例－用例と言いますか、「日本国語大辞典」のような用例を挙げるのではなくて、作例ですね－作例を検討したりする際にも、そのコーパスを基にして、実際に共起などを確認しながら、適切な例を編集委員会として考えていくというような形で編集しておりますので、そういう形でもコーパスを使っております。

また、例えば、項目の採用もありますけれども、今後廃項していくということも、辞書によっては、項目を採録すれば、一方では廃止していく、辞書から取り下げていくようなケースもありまして、そういったところの判断にもコーパスを使うということはあり得ると思います。

という形で、かなり同じような形で国語辞典においても広くコーパスを使うよう

になってきていると思われています。

ただ、石川副主査がおっしゃったように、それを社会実装としてどれだけ実現できているか、読者のところにどれだけお届けできているかというところは、これからの課題ではあると思いますので、先ほど、神永委員の企画のような、あれも１例だと思いますけれども、コーパスの良さを広く社会全般に共有していくような形に、辞書としても関わる必要があるのではないかと思われています。

また、若干観点がずれるかもしれませんが、英語の辞典、日本における英和辞典においては、先ほどのＣＯＢＵＩＬＤのお話に恐らく関連すると思いますが、今現在は基本的にはコーパスを使った辞典が主流になっていると思います。これは英和辞典でございますけれども、2003年に私どもの方でーもう20年前ですけれどもーそういったコーパスを活用した辞典を刊行いたしまして、それが大変御評価も頂いたということもあったのか、それ以降、他の英和辞典においても基本的にはコーパスを多かれ少なかれ活用するようになってきているようです。そういう形で、英和辞典、石川副主査のお話の中のＥＳＬと同じような位置付けなんでしょうか、外国語として英語を学ぶ、そういう立場において有益な情報はコーパスから取るというようなことが、かなり浸透してきているようでございます。それも踏まえて、国語辞典の方も考えなければというところがあるかと思います。

また、私は、漢和辞典を長く担当しておりまして、漢和辞典は余りコーパスと関係がないようにも思われますけれども、実は、20年以上前に既にコーパスと言いますかー厳密な意味ではコーパスではないかもしれませんがー台湾の中央研究院が漢籍電子文献を公開しておりまして、1990年代の終わりにそれを全面的に活用させていただきまして新刊の漢和辞典を刊行したという経験がございました。台湾のそのサイトで、中央研究院が公開してくださらなかったら、なかなかできなかった企画であろうと考えております。

そういう経緯もございまして、今回検討するべき日本語のコーパスにおいても、既に海外からの利用も多いと聞いておりますので、国の内外に向けて幅広くコーパスの有用性ということを考え、実際に構築していくべきではないかと考えているところでございます。

○相澤主査

どうもありがとうございました。では、本日、石川副主査のお話の中で継続的データ収集というのが第１番に出てまいりましたけれども、前川委員、日本語コーパスを長年構築されてきた立場で、何か御意見ございましたら、よろしくお願いいたします。

○前川委員

前川でございます。石川副主査の御指摘にあったことは、もうみんな非常にそのとおりでーそのとおりでと言うか、特に継続的にできれば、厳密に同じフォーマットで採るのは実際上不可能だと思うけれども、しかし、できるだけ同じような形でずっと採っていくということは、言語の変化とかを調べることには本当に必要なことだろうと思います。

ただ、何か言い訳めきますけれども、なぜ、例えば国語研究所でそういうことをこれまでやってきてないかと言いますと、ほかにやることが一杯あるんですよね。現代語だけを対象としているわけではない。それから、書き言葉だけを対象にしているわけではない。といういろいろなことがございまして、取りあえずインフラ全体をまず立ち上げよう。前回お話ししたこととは、重なるわけですけれども、歴史的なもの、それから様々なレジスターといったものを全体的にまず立ち上げようということで、それはここ二十数年で大体できてきたかなと思っておりますので、これからそういう、現在ＢＣＣＷＪとして始まっているそのアップデートの話が、行うことの方針になるーつまり、最初の例になっていくのではないかなと考えております。

そんなところなんですけれども、一つ、ちょっと、逆に私からお尋ねしてよろしいでしょうか。

○相澤主査

お願いいたします。

○前川委員

神永委員のお話を伺っていて、その中で、ＢＣＣＷＪのことかと思うんですけれども、文脈が短いということを御指摘になったように記憶しているんです。

○神永委員

はい、申し上げました。

○前川委員

具体的にはどれぐらいの長さが欲しいというか、どれぐらいの長さの状態で見ておられるんですか。

○神永委員

そうですね、特に何文字とか何語とかそういうことはないんですけれども、やっぱり、もうちょっとあったら分かるのにみたいなのが時々ありまして、本当にもうアバウトな感覚なんですけれども…。

○前川委員

ＢＣＣＷＪというコーパスは著作権処理をやっておりますので、実は相当長い範囲、入っているんですよね。あと、ビューワーでそれを見るときに、文脈を制限、長さを選べるようになっておりまして、デフォルトだと例えば20語とか30語ぐらいで表示するんですけれども、中納言－あれ確か最大500ぐらいまで増やせるんですよ。500ぐらいあれば、そこまで見えますので。

○神永委員

そうですね。それぐらいあればもう全く、問題ないです。

○前川委員

それから、ついでにもう一つですけれども、できたら話し言葉のデータもということもおっしゃっておられたと思いますが、確かにＢＣＣＷＪは、その名前の中のWが「ｗｒｉｔｔｅｎ」ですから、書き言葉だけなんですけれども、今日お話しになった「メーク」の例とか、あれは、例えば、ＣＥＪＣという日常会話コーパスだとか、それから話し言葉コーパスを中納言で検索していただくと、一杯例が出てきて、しかも、その発音まで転記していますから、「メイク」と発音しているのか、「メー

ク」と長母音で発音しているのかまで含めて、お知りになることはできます。

○神永委員

そうですか。すみません、そこまでは…。

○前川委員

一応そこまでやっていますので、ＢＣＣＷＪだけじゃなくて、ほかのは……。

○神永委員

統合はされてないわけですね。

○前川委員

そうです。ただ、実は、部分的には串刺し検索もできるような、「ＫＯＴＯＮＯＨＡ」というプラットフォームも提供しております。ただ、詳しく見ようと思うと、個々のコーパスに当たっていただく必要がある。

○神永委員

要するに、労力を惜しんでいるだけなんですけれども、串刺し検索ができるととても有り難いなという、そういうことですよね。

○前川委員

そこら辺は、石川副主査のお話にもあったように、インターフェースをいろいろと使いやすいものに変えていく必要はあるなと、そういう自覚はございます。

○神永委員

ありがとうございます。

○前川委員

一般の人というよりも、プロ向けにももうちょっとインターフェースを使いやすいものにしていかなきゃいけないなと。

○神永委員

そうしていただけると、とても有り難いです。

○前川委員

あとは、環境の中にある種の分析機能みたいなものも埋め込めるといいんですけどね。

○神永委員

そうですね。

○前川委員

例えば、今日、意味の話が何遍か出てきましたよね、誤用の話とか、それから類義語、あるいは多義語ですね。あれなんかも、一つのやり方としては、ユーザーが幾つかの検索した例に対して、これはこっちの意味、これはこっちの意味というような判定を幾つか行っていくと、いわゆる半教師あり学習というやつで、システムが裏で動いて、ほかのに対して、こっちの意味です、こっちの意味ですというのを自動的にやると。ここら辺、神永委員、御専門になるんじゃないかと思いますけれども、そういうシステムは、今は、できるだろうと私は思うんですよ。そういったものを将来的に組み込めるといいなと考えております。

○神永委員

ありがとうございます。

○相澤主査

どうもありがとうございました。では、時間となりましたので、本日の議事は以上とさせていただきます。本日は、日本語と英語のコーパスの活用について、社会課題の解決も含めて幅広く意見交換いただきました。本日いただいた御意見は、分類整理して、本日配布の資料２の更新という形でお示しをしたいと考えております。引き続きどうぞよろしくお願いいたします。また次回以降も、日本語の言語資源を

考える上で有益と思われることに関しまして、他の有識者の方、委員の皆様からお話を伺うことができればと存じます。

　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。どうもありがとうございました。