第 60 回国語分科会国語課題小委員会（Web 開催）・議事録

令和５年７月２１日（金）
１５時 ００分 ～ １７時 ００分
文部科学省３階３Ｆ１特別会議室

〔出席者〕

（委員）森山主査、滝浦副主査、木村、齋藤、佐藤、中江、長岡、成川、古田、
前田、村上、山本（真）、山本（玲）各委員（計 13 名）
（ゲスト）大学共同利用機関法人人間文化研究機構 国立国語研究所
教授 小木曽智信氏
（文部科学省・文化庁）圓入国語課長、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
※ 森山主査と事務局は、文部科学省３Ｆ１特別会議室にて参加。

〔配布資料〕

１ 国語分科会国語課題小委員会（第 59 回）議事録（案）
２ ローマ字のつづり方に関する委員の意見（第 59 回まで）（案）
３ 現在（ならびに近い将来）の日本語を書き表すのに適したローマ字表記を考える（斎藤純男委員提出）
４ ローマ字のつづり方に関する実態調査の内容（案）
５ 現代語の書き言葉コーパスが果たす役割 ―『現代日本語書き言葉均衡コーパス』の意義と今後の課題―（大学共同利用機関法人 人間文化研究機構 国立国語研究所 小木曽智信氏提出）

〔参考資料〕

１ ローマ字のつづり方（昭和 29 年内閣告示第１号）
２ 外来語の表記（平成３年内閣告示第２号）
３ 国語課題小委員会（23 期）における審議の内容

〔経過概要〕

１ 事務局から配布資料の確認があった。
２ 前回の議事録（案）が確認された。
３ 斎藤委員から、配布資料３「現在（ならびに近い将来）の日本語を書き表すのに適したローマ字表記を考える（斎藤純男委員提出）」について説明があり、説明に対する質疑応答及び意見交換が行われた。
４ 事務局から、配布資料４「ローマ字のつづり方に関する実態調査の内容（案）」について説明があり、説明に対する質疑応答及び意見交換が行われた。
５ 小木曽智信氏から、配布資料５「現代語の書き言葉コーパスが果たす役割 ―『現代日本語書き言葉均衡コーパス』の意義と今後の課題―（大学共同利用機関法人 人間文化研究機構 国立国語研究所 小木曽智信氏提出）」について説明があり、説明

に対する質疑応答及び意見交換が行われた。
6　森山主査から主査打合せ会の設置について提案があり、了承された。主査打合せ会の委員については、参加希望者を募った上で森山主査が決定することとされた。
7　次回の国語課題小委員会について、令和５年９月 11 日（月）午後３時から５時まで、オンラインで開催する予定であることが確認された。
8　質疑応答及び意見交換における各委員の発言等は次のとおりである。

○森山主査

それでは定刻になりましたので、ただ今から第 60 回、今期３回目の国語課題小委員会を開会いたします。今回もオンライン上でのウェブ会議となりましたが、よろしくお願いいたします。

本日は、議事次第のとおり、（１）ローマ字のつづり方に関する検討、（２）今後検討すべき課題全般に関する調査等、（３）その他という内容で協議を行いたいと考えています。

事前に資料を御覧になっているとおり、本日は、斎藤純男委員にお願いをし、音声学・音韻論の観点から、ローマ字に関するヒアリングを実施することとしています。また、ローマ字に関する実態調査について、さらに、国立国語研究所の「現代日本語書き言葉均衡コーパス（ＢＣＣＷＪ）」について、同研究所の小木曽智信さんからお話を伺い、意見交換をしていただく予定です。

それでは、ローマ字のつづり方に関するヒアリングに入っていきたいと思います。本日は、斎藤純男委員にお話を準備していただいています。

斎藤委員は、音声学・音韻論を専門となさっていて、日本語以外にも、モンゴル語をはじめとする中央ユーラシアの言語、文献にもお詳しく、ローマ字による表記にも御関心が深いと伺っています。

今回は、現代の日本語における意味の区別に関わる音の違いに注目し、体系的にローマ字で表すための方法についてのお考えを話していただきます。

これまで国語課題小委員会では、主に訓令式とヘボン式という二つのローマ字のつづり方が教育や社会生活において用いられてきたということを前提としつつ議論を進めてきた側面がありました。したがって、この二つのつづり方を離れた議論はほとんど行っておりません。

本日は、今後の議論を深めていただくために、あえて今あるつづり方の体系にとらわれずに、現代の日本語で確認できる意味の違いに関わる音の対立に注目し、それらをできるだけ過不足なく表すために、どのようなローマ字表記の体系が可能であるのかという観点から整理されたお考えを準備してくださっています。

また、外来語をローマ字で書く場合にどうするかといったことも話題になってきましたが、その辺りのこともお話に含まれると思っております。

斎藤委員からは、配布資料３「現在（ならびに近い将来）の日本語を書き表すのに適したローマ字表記を考える（斎藤純男委員提出）」をお預かりしています。こちらを御覧になりながらお話を聞いていただきたいと思います。それでは、斎藤委員、よろしくお願いいたします。

○斎藤委員

斎藤です。よろしくお願いします。お手元に資料があると思いますが、画面共有もしながら説明させていただきます。

御紹介にありましたように、これまでのシステムとは少し違うものを考えてみようということで、今日は議論の手掛かりの一つとして、日本語の音の構造という面からローマ字表記をどう考えたらいいかということについてお話しいたします。

ローマ字は表音文字の一つです。表音文字というのは、その名称から、一般には発音を表すのが目的と考えられているかもしれません。しかし、言語学では、表音文字の究極の目的は、意味の単位である語を示すことであると考えられています。つまり、表音文字による表記というのは、相手に意味を持つ単位である語を思い起こさせるための手段、手掛かりとして単語の発音を利用しているというわけです。そのためには、発音の大体のところが分かれば良いので、正確な記述である必要はありません。これは、言語の基本的な機能は意味を伝えることであることを考えれば、当然のことと言えるかもしれません。したがって、子音しか表記しないアラビア文字のようなものや、発音と文字が大きく離れてしまった英語のつづりがそのまま使われているということがあっても、特に問題はないわけです。

そうは言っても、文字による表記を新たに作る場合は、意味の区別に影響する音の違いは表記に反映されていた方が便利だと言うことができます。したがって、効率的な表記を作るに当たっては、意味の区別に関わる音の対立を基盤として考えるのが良いと言えます。

音の対立という基準に基づいて論理的に考えていけば、これは誰が考えても似たような結果が得られると考えます。今日ここに提出させていただくものは、過去にローマ字表記として提案されたことがあるものや、音韻分析で示されたことのあるものとかなりの部分重なっていて、特別にオリジナルなものでも、変わったものでもありません。その分野の文献をお読みになったことがある方は、あれと同じかと思われるかもしれません。

現在の日本語では、かつて不可能であった子音と母音の組合せが現れていますので、ここでは、それを含めて考えていきたいと思います。

考えるに当たっては、幾つか気を付けないといけない点があります。まず、今ある音の新しい組合せが存在しなかった過去の時代において定められた表記法から出発するのではなくて、現在の日本語そのものを観察しながら考える必要があります。

次に、音の対立に基づいて一貫させる必要があって、仮名表記のローマ字による置き換えを混入したりしないということになります。仮名のローマ字による置き換えというのはどういうことかというのは、配布資料の２ページに参考として出しておきましたので、見ていただければと思います。そういう別の文字の置き換えを混入してはいけません。

それから、論理で考え通さなければなりません。これまでの表記法の慣れからくる感覚で考えたのでは、うまくいかなくなってしまうことがあるかもしれません。例えば平仮名でも、「本を読む」というときの「を」を「お」で書いたら、おかしいと感じるのは、その表記に慣れていないからであって、「お」を使っていけないという理由にはならないわけです。表記ではありませんが、例えば、ら抜き言葉なども、出てきた当初は非常におかしいと感じられましたが、今はそのような感じはしません。ファッシ

ョンなどでもそうです。出てきたときにはおかしな感じがすることがありますが、慣れてしまえば普通になってしまいます。例えば、最近はスーツにリュックサックを背負った方を時々見掛けます。最初に見た時にはおかしく感じましたが、今はおかしいという感じはしません。そういう慣れの問題がありますので、表記を考えるときに、過去の表記の慣れによる感覚から考えてはいけないということです。

また、五十音図は、古い時代の発音に基づいてできたものですので、現在の日本語の発音、音を考える際には、それをそのまま基とすることはできません。

さらに、現在行われているパソコンの入力システムなどに影響されることなく考える必要があります。ローマ字表記なども、仮に今の入力システムで入力を行うと面倒なことがあるとしても、それは後から変えればいいことであって、今それを気にする必要はありません。全て現在の日本語の音から考えいく、そういう必要があるということです。

実際にどのようにしたらいいかということを簡単にお話ししたいと思います。意味の区別に関わる音の対立から考えます。意味の区別に関わらなければ、それは書き分ける必要はないということになります。まず、短母音と音節の初めに来る子音について見ていきたいと思います。

配布資料の４ページ（１－１）で出したものは、恐らく、どなたがどのシステムで書いても異論が出ず、一致すると思われるものです。短母音の「あ・い・う・え・お」は、これ以外の方式を取るということは余り考えられません。

次に、ヤ行の子音です。ヤ行の子音は母音との組合せで言うと「や・ー・ゆ・（イェ）・よ」が可能です。「イェ」は、単語の数としては少ないですし、外来語に限られますが、今、現れていますし、発音できますので、ここに含めておきます。２番目の空欄になっている部分は、もう少し摩擦が伴うような音になると、例えば英語の year というときのようになりますが、日本語の場合は、そういう摩擦が伴う音ではありませんので、この部分は存在しません。音声記号で書けば、ドイツ語やデンマーク語といった言語で使われているように、［j］になりますが、ここでは英語式にｙを使って書いています。

それからワ行の子音ですが、これも、現在は外来語によって「わ・ウィ・ー・ウェ・ウォ」という組合せができています。真ん中の「ｕ」との組合せ「wu」は日本語では存在しないということで、ここは組合せがないということになります。

次です。いろいろな音の組合せを考えていかなければならなりませんが、案１と案２と二つお示しします。まず、実際の発音に近い形から見ていきます。サ行に関連するところですと、「さ」の子音と「しゃ」の子音の二つあります。例えば「傘」と「貨車」は、[s]という音と[ɕ]という音で区別されますから、この二つは区別しないといけないということになります。

後ろにどの母音が組み合わさるかということですが、今の日本語では、「す」の子音も「し」の子音も、どちらも全ての母音と組み合わさる。「さ・スィ・す・せ・そ」と、「しゃ・し・しゅ・シェ・しょ」、全部の組合せができますから、この二つは区別する必要があるということになります。

タ行に関わる部分です。これも「た」の子音と「ち」の子音と「つ」の子音、全部種類が違います。かつては組合せに制限がありましたが、昔は組み合わさらなかった子音と母音が、今は組み合わされるようになりました。「た」の子音では、「た・ティ・

トゥ・て・と」と、全ての母音と組み合わさりますし、「ち」の子音も「ちゃ・ち・ちゅ・チェ・ちょ」、「つ」の子音も「ツァ・ツィ・つ・ツェ・ツォ」と組み合わせて発音することができますから、この三つの子音も区別することになります。

先ほどの「しゃ」の子音もそうですが、一般のローマ字では足らないので、補助記号を付けて区別しておくことにします。

ダ行やザ行に関わるところです。例えば、「だ・ぢ・づ・で・ど」の「ぢ」と「ざ・じ・ず・ぜ・ぞ」の「じ」は、かつては発音が違いましたが、今はほとんどの方言で同じになってしまっています。歴史的な変化によって変わってきたので、平仮名で書いた場合、文字と発音がずれているところがあります。これも、「だ」の子音、「ざ」の子音、「じゃ」の子音、三つ違うものがあって、それぞれが五つの母音と組み合わさります。「だ・ディ・ドゥ・で・ど」、「ざ・ズィ・ず・ぜ・ぞ」、「じゃ・じ・じゅ・ジェ・じょ」となるわけです。ここまでは全ての母音と組み合わさる子音です。

次に、母音との組合せに一部制限があるものです。「か・き・く・け・こ」の場合は、「か・く・け・こ」の子音と「きゃ・き・きゅ・キェ・きょ」の子音の２種類あります。例えばこの「か」を発音しようとして口を開く前に止めて口の形を観察します。それから「きゃ」のときも、「きゃ」と言おうとして口を開くのを止めて口の形を観察すると、違いがあるということが分かります。「き」の子音も同じようにしてみますと、「か」の子音よりも「きゃ」の子音と近いということが言えると思います。そういったことで、配布資料の５ページのような並びになっています。

ローマ字表記をするときに、「き」の子音は、実際に表面的に現れる音の種類としては、「きゃ」や「きゅ」の子音と近いわけですが、なぜそうなっているかというのを、後ろの母音のｉの影響で、この[k]というのが[kʲ]という音に変わっていると解釈して、「か・き・く・け・こ」がこの[k]の一つの子音で、「きゃ・きゅ・キェ・きょ」が、記号を付け加えた形で表すことができます。このように考えた方が全体として均整の取れたきれいな形になります。

ガ行、ナ行といったものも同様です。ガ行の場合は、いわゆるガ行鼻濁音というものがありますが、共通語の土台となっている東京の言葉ではもうほとんど消えているということと、もともと[g]と[ŋ]によって区別される単語というのはほぼないということで、意味の区別に関係がありませんし、それを持っていない人も多いということで、わざわざ書き表す必要はないということになります。

「な・に・ぬ・ね・の」と「にゃ・にゅ・ニェ・にょ」、「ば・び・ぶ・べ・ぼ」と「びゃ・びゅ・ビェ・びょ」、「ぱ・ぴ・ぷ・ぺ・ぽ」と「ぴゃ・ぴゅ・ピェ・ぴょ」、「ま・み・む・め・も」と「みゃ・みゅ・ミェ・みょ」、これらも同じです。

「ら・り・る・れ・ろ」と「りゃ・りゅ・リェ。りょ」も同じです。ラ行は、英語、フランス語、ドイツ語などの場合、このｒの文字で表せる子音が歴史的に大きな変化をしましたので、日本語のらりるれろと遠いような感じを受けますが、言語によって、この文字に対応する音は様々で、ポルトガル語やスペイン語などは日本語のラ行と似た響きを持っていますし、ｒで書いておいていいかと思います。

少し複雑になるのがハ行に関わるところで、これも子音が三つあります。種類としてはもう少しありますが、区別する必要があるのは三つです。「ひゃ・ひゅ・ヒェ・ひょ」の場合は、口の真ん中より少し前の方で出す[ç]というような音です。それから、「ファ・フィ・フェ・フォ」の音は、上下の唇を近付けて[ɸ]というような音になりま

す。「は・へ・ほ」の場合は、喉の奥から出す音と、それから口の上の奥の方から出す音で、喉の方から出すのは[h]となりますが、口の上の方の奥ですと[x]となります。どちらも現れます。「へ」も「ほ」も同様です。

「ひ」と「ふ」ですが、「ひ」の場合は、「ひゃ」の[ç]が現れることもありますし、こちらの「は・へ・ほ」と同じように[x]が現れることがあります。

それから「ふ」の場合も、丁寧に言えば、[ɸ]と唇を使うことがあるかもしれませんが、実際には[h]というように、もっと喉の奥の方で、唇は使わないで発音する音が普通に現れています。

「ひ」の子音が「ひゃ」のときと同じように発音される場合があるとしても、それは、後ろのｉの母音の影響だと見ます。「ふ」の子音が唇を使って「ファ」の子音のように発音されることもありますが、それは後ろに来る母音のｕの影響でそのような形になっていると考えると、このハ行に関わる部分はローマ字で、配布資料の６ページのようにすることができるだろうということになります。

これが音節の初めに現れる子音の体系と、それに対応させたローマ字表記ということになりますが、この場合ですと、表記がたくさん、文字がたくさんになってしまいます。音韻的な解釈としても、できるだけ要素は少なく簡潔にした方がいいという考え方があります。実際にローマ字表記として使う場合も、少なくできるなら、その方が便利であろうということで、音の区別は保ったまま、また整然とした表記の体系も保ったまま、文字の数を減らすということを考えてみたいと思います。例えば「しゃ」「きゃ」「みゃ」というようなときは、口の構えがｉという母音を発音するようなときの構えになっています。それを、ｉという母音と発音的にはほぼ等しいヤ行の子音が後ろに付いている影響だと解釈します。そうすると、先ほどは、例えばサ行でしたら、「さ」と「しゃ」で別の子音なので、それぞれの子音に対応した別の文字をあてがっていたわけですが、「しゃ」の方の子音が、「さ」の子音の後ろに「や」の子音が来て、その連続が実際に現れたときに、お互いに影響し合って「し」という音になると解釈すれば、７ページの案２のような表記になります。

「ツァ」と「チャ」も同じです。「チャ」は、「ツァ」の子音の後ろにヤ行の子音が来ています。それが表に出てくるときに、お互いに影響して[tɕ]という音に変わると解釈すると、このような体系になります。

「ざ」と「ジャ」も同様です。音としては違いますが、それぞれに一つ一つの文字を与えるのではなくて、片方は二つの文字の組合せと解釈するということです。

これはａ・ｉ・ｕ・ｅ・ｏ全部と組み合わせることができる子音です。一部制限があるものに関しては、「きゃ」の子音を同じように「か」の子音と「や」の子音が並んでいると解釈すると、このようになります。

先ほど「き」は、音声としては下の方のグループだが、上の方のグループにした方が体系として均整が取れると言いました。配布資料の７ページ右側に参考として書いておきましたが、ア行とヤ行はそれぞれａ・ｉ・ｕ・ｅ・ｏと ya・yu・ye・yo です。ヤ行の２番目のｉのところが欠けていますが、カ行の場合も、ア行、ヤ行と全く並行的な分布を示しているわけで、整然とした体系となっています。

ガ行、ナ行、バ行、パ行、マ行、ラ行も同様です。

ハ行ですが、「ファ」は、唇の音の特徴は、ワ行の子音が後ろに来ていると解釈すれば、「hwa」のようになるわけですが、そうすると、これも ya・yu・ye・yo とａ・ｉ・

ｕ・ｅ・ｏと wa・wi・we・wo の分布と並行的できれいな体系になります。このように考えれば、はじめに、全ての音の違いを表したものについて、記号の数を少なくして同じ機能を果たすことができます。子音が並ぶというところだけは多少複雑になりますが、区別できるという点では、最初の案１と同じで、それがもっと記号の数からいうと簡略になっているということです。こういった方式だと、いろいろな外来語などで現れる音の組合せも十分に表すことはできるということになります。

「テュ」「デュ」などは、一般的な言葉には余り現れないかもしれませんが、歴史上の人名といったものでも見たことがありますし、実際に発音できないということはないだろうと思います。そういったものにも簡単に対応することができます。

ここまでは音節の初めの子音でしたが、音節末の子音に関して次にお話しします。いわゆる 撥(はつ) 音と促音です。

まず撥音です。撥音の場合は、実際に観察すると、９ページのように「さんばい」「さんだい」「さんにん」「さんがい」「さん。」「さんを」と、いろいろな音が出てきます。記号は有限ですから、これぐらいしか表せませんが、実際には母音の種類などにも影響されて様々にあります。こういう音の違いが現れるのは、その後ろに来る音などに影響されていて、後ろが唇を使う音でしたら、唇を使う[m:]になりますし、歯と歯茎の辺りで発音する「だ」「な」といった場合でしたら、歯茎の辺りで発音する[n:]になりますし、後ろの方で発音される「が」のような音の場合には、もう少し後ろの方の音になります。ここで共通しているのは、鼻音性を持っているということです。鼻音性を持つ何らかの音が現れればいいわけです。発音しやすいように、後ろの音と発音する場所が似てくるということです。それは場所によって決まっていますので、同じところに違う音が現れて意味を区別するということはありません。例えば「さんだい」というときには、上の歯とか歯茎の辺りに舌先がくっつく発音はありますが、唇を閉じて「さｍだい」というような言い方はしません。表面的にはいろいろな音が現れますが、音の違いによって意味を区別することはありません。これは人間で言えば、同じ一人の人物が、行く場所によって服装を変えるのと同じで、服装が変わっているからといって別人物とはしないわけです。それと同じように、これは表面的にはいろいろな発音で現れますが、この細かい音の違いは意味の区別には関係しないので、一つの音だと解釈します。

解説書などでは、必ずこうなる、といったニュアンスで書いてあるものがありますが、実際は話すスピードなどによって変わってきます。例えば「さんばい」は、必ずしも、最初から唇が閉じているというわけではありません。これは区別する必要がない、一つの音だと考えておけばいいということになります。

鼻音の要素がありますから、これを代表的な「ｎ」の文字で表記するといいかと考えます。しかし、そのまま表記すると、ナ行子音の音節の初めの「ｎ」と混同してしまいますので、何でもいいのですが、例えばここでは上に点を付けるような文字を導入しています。このようにすれば、例えば「hoṅ」「hoṅmo」「hoṅno」「hoṅo」というように区別ができます。

点がないと、後ろに母音が来たときに、「な・に・ぬ・ね・の」と区別できませんが、上に点が付いているのを見れば、後ろで切れるということが分かりますから、慣れればすぐに発音の区別ができるようになります。

促音も同じようにいろいろな音が出てきます。これも同じように、後ろに来る子音

の種類によって変わってきますが、その音の違いは、意味の区別に関係しません。したがって、一つの文字で表せばいいということになります。一般にはなじみがないかもしれませんが、言語学や日本語学では―大文字にすることが多いんですが―「q」を使うことが多いですし、ほかの音に使われていないので、ここではこの文字を使って表します。

ここまでは子音のお話でしたが、次に母音です。短母音は、先ほどのように、恐らく異論はないと思います。

日本語の場合、二重母音があるかどうかという問題がありますが、母音連続の場合も、恐らく問題はないと考えます。

長音に関しては、日本語の場合は母音の長短で意味の区別がありますので、表記上も区別する必要があります。世界の言語のどのぐらいが区別しているかというのは、何百かの言語の統計を取ったものを昔見たことがありますが、２割ぐらいの言語が長短を区別していると書いてあったと思います。現在もローマ字で 11 ページのような記号を使うことが提案されています。ほかにも、例えばハンガリー語は、アクサンテギュ（アクセント記号の一種）を使って長音を表すという方法を取っていますが、こういう方法を取ってもいいのかもしれません。

仮名表記で「えい」とされるものがありますが、母音連続の場合と長母音の場合があります。これは仮名表記に影響されず、実際の音の区別ということで、母音の連続の場合は「ei」、長母音の場合は「ê」などと表さなければならないと考えます。

長母音も、日本語の場合、母語話者の意識の上では二つの短母音に分けてしまうこともできますが、母音連続と長母音との区別をする必要があります。例えば「歯痕（はあと）」と「ハート」の区別ができるようにするためには、長母音の表記を用意しておいた方がいいと考えます。

補足ですが、今、ユニコード（Unicode）が確立していますので、補助記号を使った文字の使用に困難はありません。案２の方は、それほど補助記号を使った文字を入れていませんが、そういった記号については、今はもう心配する必要はないと考えます。

形態論と言うか、意味の単位についてですが、同じ意味の単位は同じ書き方で書くということにすると、例えば「matanai（待たない）」「matimasu（待ちます）」「matu（待つ）」というのは、全部「待つ」の語幹が文字の上でそろうわけですが、そうすると、「買う」のときはどうするのかという問題も出てきます。また、例えば外来語以外はそうするとした場合、一貫しなくなってしまいます。そのような形態論的な表記にせずに、音の区別で見分けた方がいいだろうと思います。

形態論的な表記を取っている言語も結構ありますが、日本語の場合は部分的なところになりますので、音の交代、形の交代として見ていけばいいだろうと考えます。例えばトルコ語の場合、「行く」というのは「git-」と「gid-」という二つの形があって、同じ形態素ですが、発音の方を表記しています。ほかにも幾つか例を挙げています。ロシア語のローマ字表記などにも、そのような不統一があります。

また、原語のつづりと外来語のつづりというのは別です。外来語は日本語ですから、日本語としてつづらなければいけないと考えます。実際に例えばトルコ語では、エレベーターはフランス語の ascenseur から来ているのですが、asansör とつづります。それから、フランス語の office は、トルコ語では ofis と書きます。日本語でも例えば外来語でエレベーター、ドライバー、オフィスは、上に提案した書き方で言えば 12 ペ

ージのようになります。フランス語のルポルタージュは ruporutāzyu というように書かなければならないだろうと考えます。

こういったシステムでうまくいくと思いますが、現実の観点からどうなのかということがあります。正書法として定める場合とそうでない場合とでは状況が異なるわけです。正書法として定める場合は、新しい文字を導入したり、新しいつづり方を導入したりして、例外的な表記があっても、全ての出版物がそれに従って書かれ、教育によって徹底できますから、問題は起こらないと思います。ところが、日本語のローマ字表記の場合は、正書法にするわけではありません。何のためにあるかというと、外国との交流において、自分の名前や住所をローマ字で書く必要があるだとか、漢字や仮名が読めない外国人が日本に来たときに、駅名などにローマ字表記があれば分かりやすくて便利だろうといったくらいではないかと思います。そうすると、新たなシステムを提案しても、それをどこまで徹底できるかということが問題かと思います。

例えば、「富士」は、今 Fuji というつづりで定着していますが、それを、先ほど提案したつづりで Huzyi のように変更できるのか、こういったことを全て変更できるのかということです。仮に変更できるとしても、実際には、人名、地名を表すぐらいで大して使わないのに、大きな努力、労力を払う必要、メリットがあるのかという問題があると思います。

また、現在ヘボン式に準じた表記が世の中に広まっていますので、例えば、現在の日本語の音の区別に対応できて、なおかつ、整然とした文字表記の体系を持つ、上に提案させていただいたようなものを基盤として定めておきながら、実際の一般の使用としては、一部の文字や文字連続をヘボン式に近い表記に置き換えて行っても良いとしておくのが現実的かと思われます。

少し急ぎましたが、このようなことで議論の手掛かりとなることができればと思っております。ありがとうございました。

○森山主査

斎藤委員、ありがとうございました。外来音のことも含めて、音の区別、いわゆる音韻の対立ということに対応して、整然とした体系のローマ字の書き方についてお話しいただいたかと思います。

ただ今のお話の内容について、疑問点あるいは確認しておきたいことなど、御質問があれば伺っておきたいと思いますが、いかがでしょうか。

（ → 挙手なし。）

それでは、お話を踏まえての意見交換に入りたいと思います。今回は、専門的な内容でもありましたので、必要があれば、論点を絞ってまいりたいと思いますが、その前に、まずは感想等も結構ですので自由に御発言いただき、様々に議論していきたいと思います。いかがでしょうか。

○木村委員

貴重なお話をありがとうございました。

御提案くださったローマ字表記は、お話にあったように、一見ヘボン式のようなところもありますが、「ひ」や「ふ」を区別しない点、また、タ行の書き分けや拗音の扱いなど、音韻的に整合性が取れているものと拝察いたします。

配布資料の 12 ページ（４）補足の一つ目で、現状、補助記号を使うことがお話にあったところに関して少し考えたことがあります。長音に関して、「国語に関する世論調査」の「大阪」「神戸」の調査によって、いろいろな表記があることがよく分かっているところです。促音を「ｑ」を使っていらっしゃいましたが、長音には例えば「ｒ」を使うのはどうなのかといったことを考えた次第です。

○斎藤委員

例えば、中国語のローマ字システムもこれまで幾つか提案されたものはありますが、一つのシステムでは、声調を文字の上や下に補助記号として書くのではなくて、同列に並べて示すという方法がありました。長音を補助記号ではなく表すというのも一つの方法かと思いますが、「ｒ」を使った場合はどうなのか、ちょっと分かりません。どの文字を使うかにもよりますが、そういった方法も一つの案として可能かと思います。

○森山主査

ありがとうございます。ほかに、いかがでしょうか。

（ → 挙手なし。）

それでは、論点を絞ってまいりたいと思います。今回の御発表は、音韻論的な考え方に基づいて、できる限り破綻のない体系を追求した内容であると受け止めております。言い換えれば、いわゆる訓令式やヘボン式に欠けているところに着目されてお考えになったものということになるかとも思います。

それで、改めて斎藤委員にお伺いしたいのですが、訓令式やヘボン式の問題点に関して、どのようにお考えになっているのか、お話しいただければと思います。

○斎藤委員

その点は準備していなかったのですが、訓令式の場合は、現在できている音の組合せをうまく区別して表すことができないというところが一番大きな問題かと思います。ヘボン式に関しては、音の体系と文字の体系は一致しなくてもいいのですが、その点に関して、バランスが良くないように思います。

すみません、細かくチェックしていなかったので、今申し上げられることはそのぐらいです。

○森山主査

ありがとうございます。特にヘボン式などでは、長音の表記がないというのは本当に問題ですし、いわゆる撥音の「ん」の表記で、日本人の音の区別から外れたような「ｍ」と「ｎ」の区別があるといった問題なども含めて、検討は必要かと思います。

今お話しいただいた中で、外来語の音をどう考えるかというところがあります。訓令式の場合、特にそれがよく見えてしまうのですが、外来語について、現行の「ローマ字のつづり方」（昭和 29 年内閣告示第 1 号）で、例えば「ティ」は、つづりようがないわけです。そういう外来語にローマ字を用いる場合のことに「ローマ字のつづり方」では全然触れられていませんが、本日の斎藤委員のお話の中では、外来語に用いられる表記、音にも対応するということが示されています。ローマ字を外来語の表記に対応させるということになりますと、かなり大きな変更が必要になるということも考え

なければなりません。

ここで、参考資料２「外来語の表記（平成３年内閣告示第２号）」の２ページ目の表を御覧ください。右側に「シェ」「チェ」「ツァ」など、それから右側の下の方に「イェ」「ウィ」「クァ」などの表記があります。これらの表記に関しては、基本的に現在の「ローマ字のつづり方」では対応していません。今後の調査で明らかになればいいと思いますが、例えば、現実の社会生活の中で、外来語をローマ字表記する場合があるのかどうか、あるいは、その場合にどういうことが考えられるのかということに関して、外来語のつづり方とローマ字のつづり方の関係をよく考えていく必要があると思います。

今後、具体的に「ローマ字のつづり方」を検討していくに当たり、外来語にまで対応するのかどうか、その点、御意見を伺いたいと思います。もちろん、今後検討していく中で、更にその議論を深めていくべきことですので、今ここで決定する必要はないと思います。そういう点で、気楽に意見交換ができればと思っております。いかがでしょうか。

○斎藤委員

特別、外来語とか和語とか漢語とか、表記の仕方の面では区別する必要はないのではないかと思います。もちろん、外来語にしか現れない音の連続といったものはあったとしても、日本語として、日本語の音のシステムに従っているわけです。外国語から多少影響を受けて組合せができたりしたものはありますが、元の起源がどこであろうと、それは問題にしなくてもいいのではないかと私は思います。

○成川委員

外来語ということで思い出したことがあります。市町村名だと南アルプス市ぐらいしかありませんが、住所表記になると、アルカディア（山形県米沢市）などありますので、ローマ字で書けた方がいいかと思いました。そういうところがどのようにローマ字で書いてあるのか調べたことはありませんが、いわゆる市町村の下の地名表記の部分だと少しはあるかと思います。

○森山主査

ありがとうございました。アルカディアという例が出ましたが、そういう地名表記などの中には外来語が出てくる場合も若干あるということですね。

ほかに、いかがでしょうか。

○古田委員

意見ではありませんが、思い出したことがあります。例えばセントレアというのが空港名で用いられていると思いますが、これは造語です。中部のセントラルとエアを付けて、さらにローマ字表記するとどうなるのかというと、難問のように思いました。

単なる外来語ではなく、英語やほかの外国語を用いた造語は、ほかにも例えばＪリーグのサッカーチーム名などにもあります。英語などの外国語がそのまま日本語化しているというだけではなく、造語のような形で複数の語が付いて創造しているケースもあると思うので、その点も考えないといけないかと思いました。

○森山主査

ありがとうございます。そういう点では、例えば、「バイオリン」を片仮名で書く場合でも、「バイオリン」と発音する人と「ヴァイオリン」と発音する人がいるのに合わせて、片仮名で書き分ける場合があるなど、外来語の発音にどこまで文字を合わせていくのかということに関する様々な揺れの問題というのもあるかもしれませんね。

ほかはいかがでしょうか。

○山本（玲）委員

外来語を表記するときには、日本が非常に外来語に対して懐の深い文化があるということもあって、例えばお店などでカフェオレとかモンブランとかは、フランス語のつづりになっています。一般の人や子供たちまでもが、英語に限らずいろいろな外国語のオリジナルのスペルを見ているということが結構起こっているという 稀(け) 有な国だと思っています。そういう中で、元のつづりが存在する言語を表すために、例えばカフェオレをローマ字でわざわざつづるということは、余り現実的に起こらないのではないかと想像します。

外来語を片仮名表記するときに、片仮名も従来五十音しかなかったものですから、それまで存在しなかった、ウに濁点の片仮名などを使うことが先行して始まってしまっているための混乱もあると思います。

そういう意味では、先ほどの斎藤委員が示してくださった資料を拝見すると、ローマ字からは離れますが、五十音を外来語に対応できるように整理してくださったところに、例えば「た・ティ・トゥ・て・と」という段があるというような整理をしてくださったことが、まず一つ大きな意味を持つという印象を受けました。

少なくとも外来語を表記する片仮名については、あの「た・ティ・トゥ・て・と」等の 50 以上の音のある表を使ってよいという整理がされるだけでも、外来語の片仮名表記について、きちんと整理が進む可能性があると感じました。

以上です。

○森山主査

ありがとうございます。この辺り、まだ様々に議論を深めていきたいところではありますが、斎藤委員の配布資料の最終ページに、ローマ字表記は何のために必要かという非常に重要な問い掛けをしていただいています。言わば交流の段階で問題になることであるということですね。それから、お話の中でも、最終的にはヘボン式のやり方にある程度合わせた形で考えていくというようなことも書いていただいていますが、その辺りも含めて、まず目的と、それから現実的にどういう書き方がいいのかということに関して、もし御意見等がありましたら、意見交換をしたいと思います。いかがでしょうか。

○滝浦副主査

斎藤先生、どうもありがとうございました。大変根源的な議論と御提案を拝見できて良かったと思います。

この国語課題小委員会では、今まで、特に前提もなしに、日本式・訓令式というもの

と、ヘボン式を比較するような感じで検討してきました。この国語課題小委員会に対して、ローマ字のことをどうにかしなさいということは多分期待されているのですが、どうしなさいということは全く言われていないので、どのレベルで何をすればいいのかということ自体が、実は私たちも分かっているわけではありません。今までは何となく既存の二つについて検討してきましたが、理論的には、全く新しくローマ字の体系を考案するということもあり得ることになります。そうした意味では、今まで一度もしてこなかったことですが、既存のものに左右されない形で、新しい体系を考案するとしたらどうなるかということの一つの究極の姿をお見せいただいたということは、この国語課題小委員会にとっても非常に大きなものになるのではないかと思いました。

今後どうなるかというのは、結局、では、どのレベルで何をやっていくのかという話を考えていくことになります。それが分からないところではありますが、大変参考にさせていただけるのではないかと思いました。

○森山主査

ありがとうございます。何のために必要であるかというローマ字の必要性という観点で申しますと、先ほど斎藤委員がおっしゃってくださったように、いわゆる広い意味での交流の場面での日本語の地名・人名の表し方だとか、あるいは日本語を知らない方が日本語の発音を勉強する段階でのローマ字を利用した読みであるとか、そういったことが確かに重要だと思います。さらにもう１点、教育的な観点で申しますと、小学校の３年生で学習するローマ字が、日本語の音の仕組みの学習など、前回の山本（玲）委員のお話にあったように、日本語というものの音の仕組みをどう考え、また、外国語、英語などの学習をするときに、その音の仕組みをどのように捉えていったらいいのかといった学校教育の中での位置付けというのも非常に重要ではないかと思います。

そういったことを考えますと、ローマ字の表記というのは、音韻論的な観点と同時に、今までのなされてきた教育的な在り方といったことと関連付けて考えることも必要かとも思います。もしその辺りのことでお考えがありましたら伺いたいのですが、いかがでしょうか。

○前田委員

本日の御発表、大変興味深く伺いました。本当に勉強になりました。ありがとうございます。

今の教育ということと関係があると思ったので、二つ、お聞きしたいことがあります。例えば、「たちつてと」を書く場合の書き方が４ページに案１としてあります。「ti」と書くと、「ティ」の音になるというところです。現在日本語には「ティ」という音があるので、「ti」は「ティ」のために使うという御説明はとてもよく分かるのですが、そうすると、例えばコンピューターで「ティ」という音を打ち出そうとするときに、「ti」と打ってしまうと、「ち」が出てきてしまいます。「ティ」という音を、片仮名で「テ」プラス「ィ」と書く以上、ローマ字で「ti」と表記するのと、片仮名表記とが、ずれてしまうということになるのではないかと思います。そこが教育上気になるというのが１点目です。

もう一つは、配布資料の７ページにある案２の、サ行とツァ行とザ行が並んでいるところです。「さ・す・せ・そ」の子音は「ｓ」でいいが、「し」だけは「sy」と書くというのは、音韻的には本当にこのとおりなのですが、五十音図というものが今後もずっと使われていくとすると、「し」だけが別の書き方になるというのが、現実的に少し分かりにくいかと思います。日本人が「し」の子音だけ違うということを理解しないといけないということが少し難しいかと思っています。例えば、「し」は「si」のままにしておいて、「スィ」という新しい音だけ何か別の書き方をするというような可能性はあるかと思いましたので、お伺いしたいと思いました。よろしくお願いいたします。

○斎藤委員

教育に関してですが、日本語のローマ字表記というのは、私は、音の教育をするためと考える必要はないし、そうするのもおかしいのではないかと思っています。もし日本語の音の構造を教育するのであれば、例えば「たちつてと」というのは、古くは「た・てぃ・とぅ・て・と」であったが、ｉの前で[t]が[tɕ]になって、ｕの前で[t]が[ts]になったということを教えれば、その方が子供に構造を理解してもらえることになるのではないかと思います。

入力のシステムについては、今のシステムで言うと、大人はある程度対応できるかもしれませんが、子供の場合、混乱するということは、確かに考えられるかと思います。以前のいろいろなシステムに基づいて入力システムが作られていますので、現実的に現在どうするかというのは別として、ローマ字表記とは何かを決定して、学校でも混乱がないようにするためには、入力システムの方を変えてもらうようにしなければならないかと思います。これは頭で考えれば簡単でも、実際にやるのは大変だと思いますが、そうしないと、いつまでも混乱は続くのではないかと思います。もちろん、今の入力システムを残しておいてもいいわけです。ほかの言語でも入力システムは複数ありますから、今のも残しておいて、新しい入力システムを作って、子供には最初、混乱のない方法で行うといったことも考えられるのではないかと思います。

○前田委員

「スィ」を別な表記にするという、そういう可能性についてはいかがでしょうか。

○斎藤委員

そうした場合、そこだけが例外的な表記になります。そうすると、またそこで混乱が起こる可能性があるのではないかと思います。

○前田委員

おっしゃること、よく分かります。

例えば、新しく日本語に入ってきた音の表記は何か別の形で書くようにして、例えば「ティ」を「ti」ではなくて「tyi」のようにして「ti」は「ち」に残しておく、「si」は「し」に残しておくと、音声的には少しおかしいだろうと思いますが、五十音図との対応ということだと、そういうことも考えられるかと思いました。斎藤委員のお話は大変よく分かります。入力システムの方を変えていくというのもあり得ると思いました。どうもありがとうございました。

○森山主査

ありがとうございます。本日の議論については、よく整理して、今後の検討に生かしていきたいと思います。斎藤委員に改めてお礼を申し上げます。ありがとうございました。

それでは続いて次の議題、（２）今後検討すべき課題全般に関する調査等について、意見交換をしていただこうと思います。前回の国語課題小委員会では、外来語の表記に関する実態調査について御意見を頂きました。既に調査に向けた入札への手続が進んでいると伺っています。

本日は、ローマ字のつづり方に関する実態調査について、事務局で作成した、たたき台を基に御意見を頂きたいと思います。御意見は、内容が現実的に可能であり、予算にも収まるようであれば、できる限り調査に反映していただきたいと思っています。

それでは、配布資料４「ローマ字のつづり方に関する実態調査の内容（案）」について、事務局から説明をお願いします。

○武田主任国語調査官

配布資料４を御覧ください。時間の都合もありますので、簡単に説明いたします。

前回、外来語の調査の内容について御相談しました。今回は、まだ少し時間がありますので、今日お話をして、ここで御意見を頂き、またその後、何か良い案があれば、更に事務局に御意見を頂ければと思っております。

では、事務局で現在考えている内容について簡単に説明します。

調査の趣旨ですが、現状のローマ字のつづり方の実態を把握したいということです。現在の内閣告示である「ローマ字のつづり方」の第１表、第２表の表記とどのように合っているか、合っていないか、あるいは、その外にあるような書き方がないか、特に長音記号の有無などについても把握したいと思っています。

調査対象は大きく分けて二つあります。一つは、国内におけるローマ字表記です。国の官公庁でもいろいろなルールがあります。これに関しては、調査するまでもなく、ルールをまとめてどこかの段階でお示しできると思いますので、それ以外のところの調査を考えています。一つは、自治体等でどうなっているか、次に、各交通機関でどうなっているか、三つ目に、企業でどうなっているかです。企業に関しては、海外向けのウェブサイトなどを調査できるのではないかと思っています。四つ目に、国内の研究者による学術研究の中でどうなっているか、五つ目に、いわゆる言語景観、看板などでどうなっているか、最後に６番目として、実際にローマ字によって日本語を表記するといった形で作られている刊行物やウェブサイトがあるかどうかです。これは部分的にローマを使っているものではなく、全体がローマ字で書かれているものがあるかどうかということを調査できたらと考えています。

それとともに、世界の国々で、日本語をローマ字にするときにどのような書き方をしているかということも調査したいと考えています。諸外国の言語における日本語のローマ字表記、例えばガイドブックなどを調べるといいのではないか。次に、日本語の作品が海外で翻訳されるときにどのようなローマ字が使われているか。そして、特に人名などについて、各スポーツにおける日本人選手の登録名などを調べると、いろい

ろなことが分かるのではないかと考えています。海外リーグやツアー、国際大会などでの登録名、ユニフォームの表示などを調べられないかということです。

こういったものについて、実際にどのような表記が使われているかということを何らかの形で集計します。それから、長音や、「ジ・ヂ・ズ・ヅ」の表記、撥音・促音といったものの書き方についても調べたいと考えています。

最後に、どれぐらい収集できるかということはありますが、一般化した外来語や外国人名が仮名表記に合わせてローマ字表記されている場合があるかどうかということも調べたいと考えています。

こういったことを事務局で検討しているのですが、更に良いもの、あるいはこういうものも必要ないのではないかといった辺りの御意見を伺えればと思っております。よろしくお願いいたします。

○森山主査

ありがとうございました。この件に関しまして、御意見等ありましたら、是非お願いいたします。

○古田委員

１点だけです。今の資料を拝見した限りでは、例えば学術論文に関して、「日本人研究者の姓名等」となっていて、「等」の部分が気になりました。例えば、「わび」などが代表的なものですが、日本語をそのまま概念として用いているといったものを、ローマ字表記と言っていいのか分かりませんが、どのように表記しているのかというのも参考になるかと思いました。その「等」の中に、概念と言いましょうか、そういうものも調査の中に含めていただければ参考になるかと思いました。

○森山主査

ありがとうございます。ほかはいかがでしょうか。

○斎藤委員

例えば外国の日本語学者、日本文化研究者が、参考文献など日本語で書かれたものを引用したり、本文の中で、日本人の考えを書いたりしたときにどうなのかというのも参考になるかと思いました。

○森山主査

ありがとうございます。ほかにいかがでしょうか。

○成川委員

質問です。資料のⅠの３に「他言語でのつづりを除く。」と書いてありますが、この他多言語というのは、英語以外ということでしょうか。

○武田主任国語調査官

これは日本語以外ということです。例えば、日本の国籍を持っていらっしゃるスポーツ選手の中でも、外国に由来する名前をお持ちの方がいらっしゃいます。そういう

方は、例えばその国のつづりでユニフォームに名前を表示している可能性がありますので、そういうものは除くという意味でした。

○成川委員
分かりました。

○森山主査
ほかにいかがでしょうか。

（ → 挙手なし。）

では、ただ今頂きました御意見に関しては、できるだけ調査に反映する方向で検討していただきたいと思います。

次に、今期になってから話題にしてまいりました言語資源の整備という観点について、時間を取りたいと思います。今、ローマ字の調査の話がありましたが、前回の国語課題小委員会では、国立国語研究所の「現代日本語書き言葉均衡コーパス（ＢＣＣＷＪ）」について、今後、国語分科会で国語に関する課題の検討を進めていく上で、このような信頼度の高い日本語のデータベースが将来に向けて再び整備されていけば、言葉に関する調査を実施する際にも安心であるといった御意見を頂いてまいりました。

ＡＩ時代などとも言われる中で、これからの国語施策において、信頼の置ける言語資源としてどのようなものを想定し、また、必要としているのか、そして、今後新たに取り組むべきことがないのかといったことに関して、国語分科会としても一緒に考えてまいりたいと思っています。

そこで、本日は、国立国語研究所教授で研究所の研究主幹でいらっしゃいます小木曽智信さんに御参加いただき、現代日本語書き言葉均衡コーパスの内容や意義、経緯、現状などについて御説明いただくこととしました。

小木曽さんには、配布資料５「現代語の書き言葉コーパスが果たす役割 ―『現代日本語書き言葉均衡コーパス』の意義と今後の課題―（大学共同利用機関法人 人間文化研究機構 国立国語研究所 小木曽智信氏提出）」を御用意いただき、画面共有していただけることになっています。それでは、小木曽さん、御説明よろしくお願いいたします。

○小木曽氏
御紹介ありがとうございます。国立国語研究所の小木曽です。本日はこのような機会を設けてくださり、ありがとうございます。国語研究所で作っている言語資源について知っていただき、また、今後について御検討いただければということで、御説明申し上げます。

配布資料を画面共有いたします。「現代語の書き言葉コーパスが果たす役割」ということで、ＢＣＣＷＪ（現代日本語書き言葉均衡コーパス）の意義と今後の課題などについてお話ししたいと思います。

御存じの方もいらっしゃると思いますが、現代日本語書き言葉均衡コーパスは、国立国語研究所が中心になって構築した初めての大規模なコーパスで、１億語規模のコーパスです。2006 年に構築を開始し、５年間掛けて一応の完成を見ました。１億語の

多様な現代語書き言葉を収録したもので、出てくる全ての語について、短単位と長単位という二つの単語情報を付与しています。現代の書き言葉の代表性を有する均衡コーパスとして設計されたもので、ＢＣＣＷＪと略称で呼んでいます。

ＢＣＣＷＪの特徴ですが、均衡コーパスとは、これを調べれば対象とする言語、この場合は現代語の書き言葉の全体像が分かるという、代表性（Representativeness）を持つ、バランス良く収録されたコーパスというものです。全体像は巨大で茫漠としていますが、これを見れば大体分かるというように設計して資料を収集してきました。

具体的には御覧のような設計になっています。一つは、どのような書き言葉が産出されているのか、生産されているのかという実態を反映するものとして、2001 年から 2005 年に刊行、出版された書籍、雑誌、新聞の全体を母集団として考え、そこからランダムサンプリングによって3,500万語分を作った出版サブコーパスというものです。もう一つは、流通の実態、どれぐらい読まれているのかについてのものです。本当は需要の実態が知りたいのですが、そこを調べるのはとても困難なので、どのようなものが広く流通しているのか、すなわち読まれているのであろうかという点について、図書館に協力を仰ぎ、都内の 52 自治体の図書館から、どのような本が図書館に入っているのかを調べさせていただき、それを母集団として収集したものです。こちらは 3,000 万語です。2005 年当時に読まれていたものということで、出版されたのはもっと古く、20 年ぐらい前のものも含まれています。

この二つですと、まだどうしても媒体が偏るということで、さらに様々な媒体から 3,500 万語分集めてきています。これは対象期間が偏ってしまっている部分もあるのですが、現代の書き言葉を調べる上で入れるべきと考えた様々なレジスター（使用域）のものを、母集団設定はできませんが、できるだけバランスを取りながら集めてきたもの 3,500 万語です。このような設計で、全体を調べれば日本語の現代書き言葉の全体像が分かるようなものとして設定しています。

予算についてです。非常にコストの掛かるものでして、今、研究所の所長を務めている前川が、科研費（科学研究費助成事業）の特定領域研究「日本語コーパス」というもので外部資金を得て、そして、国語研究所の運営費交付金を合わせて、５年間掛けて構築してきたものです。

構築当初から予定していた用途についてです。御覧のように、日本語学が中心です。現代日本語の書き言葉を中心とした研究に利用するということです。そこから、日本語教育や国語教育のような教育面、辞書の編纂といった部分、それから、周辺の分野として心理学・認知科学、また、自然言語処理、音声合成といった工学的な分野まで応用されるだろうと期待されていました。さらに、国語分科会に関することで言えば、国語政策にも資するものであろうということで、常用漢字表の見直しや正書法というものを考えるときに、現代書き言葉の実態調査というのがとても重要だろうと期待されていました。そして最後のところですが、文化資源としての価値のあるものとして作ろうということです。未来から見たときの文化資源として、今の書き言葉を調べられる形で残しておこうという考えがありました。

これらは「中納言」という、オンライン上で人文系の人でも使いやすい形で公開されていて、単語の情報、それを組み合わせた検索などができる形で公開をしています。

これが公開されてから十数年たちます。予想を上回ると言ってもよいかと思いますが、非常に多くの利用をしていただいているところです。一つ目は研究の分野です。日

本語学を中心とする分野、そして教育の分野です。登録している利用ユーザーが４万5,000人です。今月までの１年間で調べたところ、大体100万件近い回数の検索がされています。そして、ＢＣＣＷＪを使った研究論文が、2022年の刊行分で、少なくとも180本公開されています。大学等での授業で利用するアカウントの発給数が126授業、4,200名分ほどです。このように日本語研究のインフラとして非常に重要な位置を占めるようになってきたと考えています。この数字はまだまだ増えているところで、毎年、検索クエリの数などは前の年を上回って増えている状況にあります。

先ほど、自然言語処理などの工学面のことがありましたが、近年、情報学が情報産業として応用されていることもあり、産業界においてもこのＢＣＣＷＪが非常によく使われるようになってきています。有償版、これはＤＶＤなどにＢＣＣＷＪの全部のデータを入れて商業利用していただくという契約数ですが、今のところ70件ほどあります。70件というと小さいように思われるかもしれませんが、これは１件当たり40万円から800万円という有償契約をしていただいているもので、中には国際的な大手ＩＴ企業、いわゆるＧＡＦＡと言われるような企業の中の数社とも契約を結んでいるという状況です。

実際にどのような利用をするかというのは、公表できない部分もありますが、言語の解析器や検索システム、機械翻訳のモデルを作るといったような目的で多く使われていて、日本語の情報処理のために広く利用していただいているという状況です。

もう一つの面として、先ほどの構築時の目的の一つにあった文化資源というようなことがあります。言語文化の記録としてコーパスは位置付けられるのではないかということです。そのときの言語の実態をいつでも確認できる形で記録・保存するというような役割も今果たしていると言えるかと思います。ＢＣＣＷＪは2005年当時の日本語の書き言葉の全体像を記録したものとも言えます。文化庁の「国語に関する世論調査」のような意識調査も重要なものだと思いますが、実態調査のための基礎資料として確実なエビデンスを示せるという点で、コーパスはとても価値のあるものかと思います。

しばしば言語について、言語意識と実態が乖離（かい）するということがあります。よく知られている例だと、ら抜き言葉などは、使わないと言っているが実際にはよく使っているというようなことがあるわけです。意識と実態の調査は両輪としてあるべきかと思いますが、そういった点でも、このコーパスは役立つものではないかと思います。

これだけではないと思いますが、まとめますと、学術的な価値、言語研究に欠かせない基礎データとしての役割を果たしているということ、産業面では、情報処理の学習用のデータとしての役割があって、今どんどん大きくなりつつあるというところがあるかと思います。また、文化的な価値として、言語文化を継承・保存、発展させる基盤としてのデータという役割も果たせるのではないかと考えています。

このようなＢＣＣＷＪについて、ここまではとてもうまくいっているという話をしてきましたが、限界もあります。実は先ほどまで授業をしていて、学生にＢＣＣＷＪを紹介していました。2005年当時の「現代語」と言っていますが、2005年というと、もう少しで20年たってしまうわけです。検索すると、「スマホ」という言葉が出てきません。ですから、今の学生たちにとっては現代語と言えなくなっているような部分もあります。彼らが自分たちの言葉で、新しい言葉だと思って検索すると言葉が出てこないという状況があります。

また、ＢＣＣＷＪは、当時の日本語を共時的に切り取ったコーパスなので、変化という形で見ることができないものです。当時の日本語をスタティック（静的）に見る分にはいいのですが、時代別にどのように変化しているのかということには向いていないコーパスです。その点は少し残念な部分と言えるかもしれません。

それからもう一つは、規模が必ずしも十分ではありません。１億語というのは、当時としては大変大きいものでしたが、研究対象によっては、まれな現象はこれでは十分に探せないこともありますし、特に自然言語処理の分野ではもっと大きなコーパスが求められていて、そういう部分では全然足りないということも言えます。

また、ほかの言語のコーパスを見てみると、もっと大規模なものが多いということがあります。幾つかの代表的なナショナル・コーパスと言われるようなものを挙げてみました。元祖と言っていいイギリスの British National Corpus が１億語のコーパスとして作られたもので、ＢＣＣＷＪもこれを範として作ってきたところがあります。その後できてきている他国語のコーパスなどを見ますと、アメリカの COCA と呼ばれる Corpus of Contemporary American English は 10 億語、ドイツ語の DeReKo と言われるものについては 550 億語という規模です。スペイン語では 20 億語、ロシア語でも４億 5,000 万語、ポーランド語で 10 億語というように、もっと大きな規模のコーパスがいろいろな言語で作られています。そして、そのうちの多くのものは、定期的に追加やデータ更新がされているという状態です。それもあって、ここに挙げた語数は概数であって、少し異なっているものもあるかもしれませんが、少なくともこういった規模で作られています。

また、もっと小さな、話者の少ない言語でもコーパス構築は行われていて、チェコ語では 10 億語のコーパス、スロベニア語も 10 億語です。ウェールズ語はサイズが小さいのですが、こういった小さな言語でもコーパス構築が行われています。ここには言語文化の保存や振興という観点があるのだと思いますが、たくさんのコーパスが作られている状況にあります。

これは先ほどのドイツ語のコーパスの例ですが、毎年このようにどんどんコーパスが定期的に追加されて大きくなっています。そして、これだけの時代幅がありますので、経年変化も追うことができるものになっています。ただ、これは均衡コーパスという形で作られているものではなく、協力している新聞データなどがどんどん入っていくというような形であるようです。しかし、550 億語という、ＢＣＣＷＪと比べると規模が非常に大きなものになっています。

こうしたことを見ると、先ほどの問題点としたことは、実は発展や拡張の必要性としても見えてきます。古くなってしまっているからには、最新のデータを収録して、是非更新したいとも思われるわけです。また、経年的な調査ができないというのも、経年調査ができるように、継続的に拡張していくことができるとよいのではないか、規模が足りないというのも、追加していくような形でどんどん大きくなっていけば、規模も大きくなっていくだろうというようなことが考えられます。これができれば、極めて高い文化的な価値と応用の可能性を持つコーパスとしていくことができるのではないかと考えております。

ただ、そういったものを作るには、コストの問題が大変大きくのし掛かってまいります。先ほど紹介したとおり、ＢＣＣＷＪでは大きな科研費が取れて、運営費交付金と合わせて国語研究所で何とかこれを作ることができました。しかし、今、このＢＣＣＷ

Ｊのようなものをもう一回作りたいというようなことを考えても、一つは、研究費として取ろうとしたときに、学術的新規性が乏しいということがあります。それまで全くなかったものを作ったＢＣＣＷＪは大変評価していただけましたが、同じものをもう一回作りますという意味でこれだけのものを要求することはとても難しいということがあります。ほかの国の言語などでも見られるように、コーパスの価値に着目して、行政的なサポートがあって継続的な拡張ができると、先ほどのようなものが実現して、大変望ましいのではないか私などは考えるところです。アカデミックな面だけではなかなか難しくなってきています。

そして、もう一つ、最近話題の、ChatＧＰＴに代表されるような大規模言語モデルによるＡＩの話とコーパスについて少し触れておきたいと思います。ChatＧＰＴなどで知られる大規模言語モデルですが、これを作るのにもコーパスが必要です。国語研究所が作っているようなコーパスももちろん役立つということは言えますが、注意しなければいけないのは、これらで使われるコーパスのサイズが、文字どおり、桁違いに大きいということです。桁が４桁から６桁ぐらいまで違う規模で、大量のデータを集めてきています。ＢＣＣＷＪは約１億語で、テキストのタグ付きファイルが０．５ギガ程度ですが、ChatＧＰＴを作っている OpenＡＩのＧＰＴ－３という二つぐらい前のモデルで、570 ギガバイトのテキスト、コーパスを使っているという話が出ています。テキストの質と言いますか、タグがどう付いているかにもよりますが、恐らく 1,000 億語以上というような規模になります。それから、最新で使われているＧＰＴ－４については、非公開なので正確なところは分かりませんが、更にずっと桁が違うぐらいに多いと言われています。ただし、これらの大規模言語モデルで使われるコーパスは、ウェブ上で集められてきたデータです。また、ChatＧＰＴは日本語も使えるわけですが、日本語のデータはせいぜい５％程度しか含まれていないというようなことが言われています。

最近になって、国内で日本語に特化した大規模言語モデルも公開されるようになりました。マイクロソフト系の企業で、rinna や CyberAgent、LINE といった会社、それからＮＩＣＴという公的な機関も作り始めています。これらを見てみますと、学習のコーパスはウェブ上から集めてきた大量の日本語データで、やはり 1,000 億語規模のものを使うということのようです。物によってはもっとたくさんのものから作っているということが分かります。

こうして見ると、では、ＢＣＣＷＪやその拡張で作られるような、億とか何十億とかいう程度のものだと役立たないのではないかというと、必ずしもそうではないと思っています。ウェブから得られるデータというのは、サイズはとても大きいのですが、内容は明らかに偏っています。ウィキペディアなどが丸ごと入っているから、知識面ではいいかもしれませんが、書かれている内容、文体などは大きく偏りがあると思われます。また、場合によって、著作権や個人情報の問題も付随して残ります。最近は、ＡＩ自体が生成したテキストがウェブ上にあふれてきて、そのテキストをＡＩが学習に使ってしまうというような、自己データによる汚染や、学習データに悪意ある改竄（ざん）を行って攻撃を行うデータ・ポイズニング（data poisoning）と言われるような危険も指摘されています。したがって、ウェブから離れたところ―ウェブも含めてということになるかと思いますが―そういったところから集めてきて丁寧に整備したＢＣＣＷＪのようなコーパスというのも、大規模言語モデルの時代にも役立つものだろうと思っ

ています。

ＢＣＣＷＪのように高品質で安心して利用できるコーパスは、比較的小さい―小さいといっても億単位なのですが―小さくても、今後の大規模言語モデルにも貢献できる可能性があるだろうと思います。これは著作権処理なども行っていますし、内容についても精査しています。例えば大規模言語モデルでは、ファインチューニング（fine-tuning）と言われるような、再学習によって調整するということが行われるわけです。そこで使われる高精度なコーパスには、ＢＣＣＷＪのようなものが価値あるものとして使えると思われます。また、評価用のデータとしても、しっかりしたデータが必要であるという点で、ＢＣＣＷＪのようなものが少なくとも今までのように利用価値があるものとして引き続き使われるだろうとも思います。

最近では、大規模言語モデルの学習にもいろいろな方法が提案されています。これはつい先月出たものですが、教科書的な高品質のテキストを用いることで、コーパスサイズなどがずっと小さくても、小さなモデルであっても高い性能を引き出せるというような研究もあります。そうした高品質なテキストがあれば、ウェブ上の雑多なテキストの何百分の１といった規模でも、それに匹敵する性能が出るというような研究もありました。今後、様々な研究手法が提案されてくると思われます。そういうときに、高品質なコーパスが役立つ場所は間違いなくあるだろうと思います。

このように、現代日本語書き言葉均衡コーパスを例に、コーパスが果たしてきている役割について見てきたわけですが、その上で、今後もこうした役割を発展させていくためには、継続的に拡張していくことが望ましいだろうと考えます。特に、国語政策のことも考えますと、意識調査と実態調査というのが調査の両輪として必要になります。今の日本語をコーパスとして継続的に記録・保存して実態調査を行うというようなことを、今後、文化行政の中で進めていただけるのであれば、国語研究所は、今までのノウハウを生かして協力しながら、良い言語資源を作っていくことができるのではないかというように思っています。今回、このような機会を頂きましたので、ＢＣＣＷＪについて御紹介すると同時に、希望、期待のようなことまでお話しさせていただきました。

こちらからの御説明は以上です。

○森山主査

小木曽さん、ありがとうございました。非常にお忙しい中、御準備くださったと伺っております。お礼を申し上げます。

残された時間も限られておりますが、ただ今のお話に関して、御質問、御感想、御意見など、自由に御発言いただきたいと思います。いかがでしょうか。

○山本（真）委員

小木曽さん、今日はどうもありがとうございました。感想と言いますか、意見なのですが、今お伝えいただいたこのコーパスというものを文化資源、言語資源として拡張・更新していくことの必要性というものに強く賛同するものです。ネット情報は、国民生活に身近な存在になっていて、手軽で便利である一方、漠とした危険性がよぎるということがあります。そのような中で国語の重要な政策の方向性を見定めるのに、あやふやなデータや情報基盤に依拠することはできないと思われます。

今御紹介のあったように、国語研究所のこのコーパスの利点を見ますと、今求め得る最も信頼の高いものであって、継続性が強く望まれると思われます。2005年でストップしているということになりますと、お話にもありましたように、そのとき生まれた人たちが18歳で、学生になってしまっていることになります。もう現代の言葉ではないというようなことになってしまっては大変もったいない。幾つかの利点もおっしゃっていましたが、著作権処理や個人情報の保護にまで周到に留意されているという点も強調しておいて良いことではないかと思います。そういうことも踏まえますと、是非文化庁に頑張っていただいて、この継続ということを強く訴えていただければと願います。

以上です。

○森山主査

ありがとうございました。ほか、いかがでしょうか。

○滝浦副主査

小木曽さん、ありがとうございました。言うべきことを全てコンパクトに言ってくださったという感じがして、大変心強く感じた次第です。

山本（真）委員の御感想とかなり重なりますが、今回、日本語の今を知りたいということで、大規模調査に先立って、小さな調査をしてみようと思ったが、それができないということが分かったというようなことがきっかけとしてあるのだと思います。私たちの意識としても、現代の日本語が入っているＢＣＣＷＪというように思っていても、もう20年近くたっているので、必ずしも日本語の今をそこに見ることはできないかもしれないということに気が付いたというのは、一つ、貴重なこと、意味のあることだったかと思います。改めて考えてみれば、これまでは国語研で作ってくださったものを研究者も国もみんなただで使わせていただいていたということであるわけです。それでは本当の今というのを常に追い掛けていくことはできないということで、そこに対して国・文化庁もお金を出して支えていくということができるのであれば、良いことでもあり、あるべきことでもあるかと思う次第です。お礼とともに、是非これがうまく進んでいくように願っております。

○森山主査

ありがとうございました。ほかにいかがでしょうか。

○村上委員

ありがとうございました。時間が余りないので、感想を手短に申し上げたいと思います。

このコーパスの特徴として、先ほどおっしゃったように、ウェブ以外の分野の調査が広くなされているというのは、とても心強いと思います。今言われている生成ＡＩは、大体ウェブがその調査資源の軸になっているので、それ以外のところの分野を幅広く調査されているというのは、非常に参考になると思います。

もう古くなっているという点に関して、私は小説を書くことをなりわいにしていますが、その日本語の現在を捉えられるのかということには、少し懐疑的なところがあ

ります。言葉というのは常に変化しているもので、捉えたと思った時点で既に古くなっているというようなことがあるわけです。古事記や万葉集などで当時の日本人が何を考え、何を感じていたかというようなことを我々が知ることができるように、1,000年たったときにこのコーパスは更に重要な価値が出てくるのではないかと思いました。なかなか予算を取り難いという話でしたが、これは是非我々も応援して、潤沢とはいかないまでも、経年変化が追えるような、そういう継続研究できるための資金を獲得できるように努力していきたいと思いました。

○小木曽氏

ありがとうございます。

○森山主査

ありがとうございます。ほかにいかがでしょうか。

（ → 挙手なし。）

生成ＡＩなど新たな技術について、政府内でも各方面で取扱いに関する当面の方針を検討しているなど、対応を迫られていると伺っています。それとは別に、国語施策においては、信頼の置ける日本語の言語資源を確保していくということで、ただ今多くの委員の皆様からお話があったように、ＢＣＣＷＪの意義と、その継続的な拡張・更新の必要性ということが焦点化されていくのではないかと思います。

御意見の中にもありましたが、例えば国、文化庁で、国立国語研究所のこのＢＣＣＷＪの整備・拡張に乗り出すといったことも、是非検討していただくように強くお願いを申し上げたいと思います。

それでは、本日の協議については、以上で終わりたいと思います。ヒアリングのためにお話しくださった斎藤純男委員に、また、国立国語研究所の小木曽智信さんに改めて御礼を申し上げたいと思います。ありがとうございました。

そのほか、全体を通して何か言い残していらっしゃることはございませんでしょうか。

（ → 挙手なし。）

最後になりますが、今後、ローマ字についてのこの国語課題小委員会で検討する内容を整理するため、主査打合せ会を設けたいと考えております。滝浦副主査を含む五、六名ぐらいの委員で、小委員会と小委員会の間に１回といった頻度で、御相談できる場ができると有り難いと思っております。最近では、「分かり合うための言語コミュニケーション」や「公用文作成の考え方」の検討の際にも、この主査打合せ会が設けられていました。

この主査打合せ会を設ける件については、いかがでしょうか。設けるということでよろしいでしょうか。

（ → 国語課題小委員会、了承。）

特に御異議がないようですので、主査打合せ会を設ける方向で進めたいと思います。この主査打合せ会に是非参加したいという委員がいらっしゃいましたら、今月中をめどに、事務局に御連絡くださいますようお願いいたします。メンバーが決まりましたら、メール等で御報告し、もし日程の都合が付くようであれば、次回の国語課題小委員会までに一度顔合わせをしておきたいと思います。

では、本日の国語課題小委員会はこれで閉会とさせていただきます。本日は、内容が盛りだくさんでしたが、無事に終えることができました。心からお礼を申し上げます。ありがとうございました。