第３回　国語分科会言語資源小委員会（Web開催）議事録

令　和　６年　９月２６日
１０時 ００分 ～ １２時 ０５分
文部科学省15階・15Ｆ特別会議室

〔出席者〕

（委員）相澤主査、石川副主査、植木、神永、武田、前川各委員（計６名）
（ゲスト）川原繁人慶應義塾大学言語文化研究所教授
（文部科学省・文化庁）村瀬国語課長、武田主任国語調査官、鈴木国語調査官、
町田国語調査官ほか関係官

〔配布資料〕

１　第２回国語分科会言語資源小委員会議事録（案）（委員限り）
２　令和５年度「国語に関する世論調査」の主な結果ほか
３　大規模言語モデルとテキストコーパス（相澤主査御提出資料）
４　生成AIおしゃべりアプリは子どもにとって薬か毒か（川原繁久氏御提出資料）

〔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　令和５年度「国語に関する世論調査」の結果の概要

〔経過概要〕

１ 事務局から定足数の確認が行われた。
２ 事務局から配布資料の確認が行われた。
３ 前回の議事録（案）が確認された。
４ 事務局から配布資料２「令和５年度「国語に関する世論調査」の主な結果ほか)」の説明があった。
５ 相澤主査から配布資料３「大規模言語モデルとテキストコーパス」に基づいて

説明があり、質疑応答が行われた。

６ 川原繁久氏から配布資料４「生成AIおしゃべりアプリは子どもにとって薬か毒か」に基づいて説明があり、質疑応答が行われた。

７ 二つの説明を踏まえ、意見交換が行われた。意見交換の後、国語分科会における言語資源小委員会の審議状況の報告については、相澤主査に一任することが了承された。

８ 次回の言語資源小委員会について、令和６年１１月２８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が確認された。その後の開催日については、令和６年１２月２６日（木）10時から12時まで、令和７年１月２８日（火）10時から12時までであること、国語分科会については日程の調整中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

ただ今から第３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、本小委員会はハイブリッド形式で開催しております。文部科学省の会議室にて御参加いただいている方々、また、オンラインで御参加の方々いらっしゃいます。また、傍聴者の方々もオンラインでこの会議に御参加されていますので、御承知おきください。

さて、本日は、お手元の議事次第にありますとおり、「言語資源の在り方について」、「有識者による事例等紹介」、「その他」という三つの議事を想定しております。本日、この小委員会のこれまでの審議内容を踏まえまして、「議事２　有識者による事例等紹介」につきましては、私、相澤からと、生成ＡＩの言語への影響について慶應義塾大学言語文化研究所の川原繁人先生からお話を頂くこととなっております。これらの話題は、言語資源としてのコーパスそのものの在り方とは少し離れる部分もございますが、収集される言語への影響ということで、今日のＡＩの急速な普及を背景に委員の皆様で共通の認識を持てるということで、よい機会となるかと存じます。意見交換では、生成ＡＩをめぐる気になっている点なども含めまして活発な御意見、御発言をお願いいたします。

本日は、初めに、先週発表された令和５年度「国語に関する世論調査」の結果と関連して、令和７年度予算の国語課関係の概算要求について報告をお願いいたします。配布資料２について、村瀬課長から御説明いただきます。

○村瀬国語課長

最近の国語施策に関するトピックといたしまして、ただ今御案内がございました令和５年度の「国語に関する世論調査」の結果について御説明したいと存じます。配布資料２を御用意いただきたいと思います。

本調査は、全国16歳以上の個人の方を対象といたしまして、本年１月から３月までの間に調査したものでございます。調査項目は、国語への関心、あるいはローマ字表記、読書、慣用句の意味などに関するものとなります。既に公表したものでございますけれども、主な点について御紹介したいと存じます。

まず、この１枚目のところになりますけれども、全体的な問いといたしまして、国語とコミュニケーションに関する意識ということに関しまして、「国語に関心がある」という回答が約８割となっております。「関心がある」と御回答いただいた方々のうち、具体的な点をお尋ねしましたところ、この上段右側でございますけれども、「日常の言葉遣いや話し方」が最も多く、次に「敬語の使い方」が続いておりまして、ここに身近な、社会活動におけるコミュニケーションのやりとりに気を付けていらっしゃる方々の意識の表れが読み取れるのではなかろうかと思っております。

続いて、その下でございますが、日本語の特徴で魅力を感じるところについてでございます。これにつきましては御覧のとおり、「漢字や平仮名、片仮名などの様々な文字」といったものが最も多く、続いて「敬語」や「季節を表す言葉」、「ことわざ」関係の回答が多くなっております。このことは昨今、インバウンドが進行しておりまして、社会が多様化する中、日本語の伝統面に対する注目の表れとも言えるかもしれないと思っております。

続いて、その下でございますが、ローマ字に関してでございます。どういったものが読み書きしやすいのかということについてお尋ねしました。御覧の結果となっておりまして、とりわけ、例えば「交番」だとか、「大江戸」といったところで御覧いただきたいと存じます。俗にオーバーバーと言いますか、これ（¯）はマクロンと

いう横棒のものでございますが、こういった、このマクロンを回答した割合が多くなっているところでございます。続いて、母音を重ねるという表記が多くなっております。ここで言うと水色の縦線みたいなものでございますが、こういったものが多くなっているところでございます。

伸ばす音につきましては、例えば人のお名前、オノさんとオオノさんが英語の影響を受けるなどして、いずれもＯＮＯという同じ表記になる―オノさんとオオノさんの表記が同じ表記となるという、正しく表記されない、正確に表記されないといったような課題があったところです。こうした伸ばす音の表記など、ローマ字のつづり方につきましては、現在、もう一つの小委員会―ローマ字小委員会におきまして、御検討いただいているところでございます。調査結果は御覧のとおり、長音符号を用いることに続きまして、いわば長音符号を代替する措置としてこの母音を重ねるつづり方に一定の支持があるのではないかと見受けられるところでございます。

続きまして、２枚目に移りたいと思います。２枚目は読書関係の問いのグループでございまして、これは５年に一度実施しております定期的な調査というものでございます。上段左側でございますが、20年度から順次調査をしております。これは途中でコロナの関係等もございましたので、面接聴取法から郵送法に変更している―これは全ての問いについて、全て同じなのですが―調査方法の違いといったようなものがあるわけでございます。「１か月にどの程度本を読んでいるか」という問いでございますけれども、６割を超える人々が「読まない」と回答しており、これまでで最も高い割合となっております。ここで言う「本」というのは、電子書籍を含んでおります。そして、紙の方も、電子書籍につきましても、雑誌と漫画を含まないということが前提となっております。

今、６割を超える人々が本を読まないと御紹介したわけでございますが、そういった方々、６２．６％の人々に対して、右側でございますが、本以外の文字情報―ＳＮＳであったり、あるいはネット記事等を読むことについてお尋ねしたところ、「ほぼ毎日ある」、あるいは「週のうち何日かある」と回答された方々の割合が８割程度あるという結果になっております。

さらに、その下でございますが、読書量の変化につきましても、「減っている」という回答をした方が多くなっておりまして、また、その理由は右側でございますが、「情報機器で時間が取られる」と回答した割合が最も多くなっております。こちら

が読書関係でございました。

続いて３枚目になりますけれども、こちらがいわゆる新語と言いますか、新しい表現、それから、慣用句の意味等に関してお尋ねしたグループになります。

まず、新語につきまして、新しい表現につきましては、一番上でございますけれども、御覧のような結果となっております。「御自身が使うことがあるか」、あるいは「他者が使うことについて気になるか」という、そのような質問でございました。今回、取り上げた言葉―擬態語と呼ばれるものでございまして、その使い勝手の良さを感じる方々にとっては、一定の使用状況が見られるといった状況になっているところでございます。

そして、中段と下段というのは、いずれも慣用句についての言い方であったり、意味であったりをお尋ねしたものでございます。そして、これらいずれも青色でマークしている部分というのが、辞書等でも本来の言い方、あるいはその意味とされてきたものでございます。御覧いただきますと、おおむね異なる言い方であったり、異なる意味を選択した者が多かったというような状況になっております。これは振り返ってみましても、このような結果となりましたことは、これらの言葉のニュアンスから生じる意味の変容であったり、あるいは結び付く単語との関係で本来の意味等と異なった使われ方がなされているのではないかと感じているところでございます。

ただ、私どもといたしましては、本来の意味を尊重することは重要であると認識した上でございますけれども、言葉というのは時代とともに変化するものでございまして、人によっては捉え方が異なるという言葉もございますことから、こうしたことを踏まえた上でコミュニケーション、これが円滑に図られるようにしていくことが重要だと認識しているところでございます。

調査結果につきましては以上でございます。これらにつきましては、本審議会における、国語政策における検討資料としての活用とともに、国民の皆様の国語に対する関心の喚起ということにつながる一助になればと考えているところです。ただ今御覧いただきましたように、とりわけ読書関係の問いにつきまして、「本を読まない」とする割合が６割超となっていることも踏まえまして、次のページに移りますけれども、文字・活字文化の振興を図るという観点から、御覧のとおり、新規事業を来年度の予算の概算要求で要求をしているところです。

また、隣のページになるわけでございますけれども、本小委員会の議論と関連するものでございます。現代日本語のコーパスの拡充に向けた予算ということで、今年度と同様に要求をしているところでございます。今後とも委員の皆様の御指導を賜りながら、国語の改善、普及に向けて、ただ今申し上げました本調査結果等も有効活用しながら、円滑なコミュニケーション実現に努めてまいりたいと考えております。

○相澤主査

ありがとうございました。

正に今の話と同じで、私もこの場で「ヨロン」か、「セロン」か一瞬迷って、失礼いたしました。また、辞書を作成している方の統計など教えていただければと思います。ただ今の説明について、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。）

よろしいでしょうか。では、本日、また後ほどまとめて議論の時間もございますので、その際にも、もし御意見等ありましたら、よろしくお願いいたします。

では、続きまして、次の議題に移ります。議事次第に従いまして「事例紹介等」に入りたいと思います。冒頭で申し上げましたとおり、本日は、私相澤と慶應義塾大学言語文化研究所の川原先生からのお話となります。まず、私から「大規模言語モデルとテキストコーパス」と題して簡単なお話をさせていただければと思います。準備をいたしますので、しばしお待ちください。

それでは、本日は、自然言語処理、計算機を用いた言語理解の研究という観点で、大規模言語モデルとテキストコーパスについてお話をさせていただきます。本日の内容でございますが、こちらにありますとおり、大規模言語モデルにおける文脈と語の意味、大規模言語モデルの意味理解、大規模言語モデルの訓練とテキストコーパス、コーパスに由来するリスク、テキストコーパスと大規模言語モデルという五つの項目に分けて順番に御紹介させていただきます。

まず、大規模言語モデルにおける文脈と語の意味についてです。大規模言語モデルというのは、英語ではｌａｒｇｅ　ｌａｎｇｕａｇｅ　Ｍｏｄｅｌで、分かりやすくは、ＣｈａｔＧＰＴ―生成ＡＩの言語バージョンというふうにお考えください。

計算機を使って言語を解析する、あるいは言語の意味を探るという学問は非常に歴史が古く、計算機の登場当時から機械翻訳に代表されるいろいろなプロジェクトが立ち上がってまいりました。その意味を計算的に捉える、つまり、計算可能な形式表現で表せる計算可能なものであると捉えるための根幹となる仮説というのがありまして、それが分布仮説ですとか、文脈類似度などと呼ばれるものです。

分布仮説、文脈類似度、ともに1950年代、つまり、計算機が現れて人工知能という概念が広まり出す―そういった時代に提唱されてきたもので、私たちの自然言語処理の分野においては、広く知られている仮説であります。まず1954年のこの分布仮説ですけれども、英語の原文のままですが、Ｗｏｒｄｓ ｔｈａｔ ａｒｅ ｕｓｅｄ ａｎｄ ｏｃｃｕｒ ｉｎ ｔｈｅ ｓａｍｅ ｃｏｎｔｅｘｔｓ ｔｅｎｄ ｔｏ ｐｕｒｐｏｒｔ ｓｉｍｉｌａｒ ｍｅａｎｉｎｇｓなので、類似した文脈に現れるような語というのは意味が似ている―文脈類似度というのは、それに基づいて、Ａ ｗｏｒｄ ｉｓ ｃｈａｒａｃｔｅｒｉｚｅｄ ｂｙ ｔｈｅ ｃｏｍｐａｎｙ ｉｔ ｋｅｅｐｓなので、語の意味というのは文脈によって規定される―こういった考えであります。

次のスライドに参りまして、では、この文脈と意味、具体的にはどのようなことかというと、例えば「野菜」という語を考えたときに、野菜の意味は何かというと、もちろん野菜そのものの意味を語義によって、自然言語によって説明することも可能ですけれども、その野菜が現れる、出現する文の文脈によって定義することもできる―このことは今までの様々な御発表の中でも触れられてきたとおりでございます―この文脈によって意味の近さを測るというのは、こちらはたまたまＣｈａｔＧＰＴで生成した例なのですけれども、野菜を含むたくさんの文があったときに、その文の中で野菜が伴うようないろいろな語ですとか、その語との関係によって野菜の意味を定義していくということです。

したがって、例えば「果物」という語があったときに、果物についても文脈をいろいろ調べて、野菜と果物の文脈がどれくらい近いかを調べることによって、野菜と果物の意味の近さというものが測れることになります。これは単に似ているというだけではなくて、計算によって数値として似ていることを表せるということを意味しています。そういった意味を測るための数値の並びを文脈語のベクトルと呼びます。

次のページに行きまして、少し数字が出てまいりますが、例えば野菜を見たときに、「野菜を使う」、「野菜を食べる」、「野菜を作る」といった共起―野菜という言葉と助詞、動詞のペアの共起を調べていったとします。そのときに「野菜を使う」という例文が5,300回、コーパス中に出現する。「野菜を食べる」という例文が4,400回出現するというふうな頻度の数字を並べていきます。同じように果物についても「果物を使う」が2,300回、「豚肉を使う」は2,900回というふうに頻度を数字で表します。

そうすると、野菜に対して5,300、4,400、3,100、果物に対して2,300、2,200、1,100といった数字が対応付けられていくことになります。このような数字の並びを文脈ベクトルと呼び、このベクトル―単なる数字の並びですけれども、このベクトルが近いかどうかを計算することによって、意味空間と呼ばれる意味を表す空間の中での語の意味の近さが捉えられるということになります。これが計算機で言葉の意味を捉えるというやり方の基本となるものになっています。

こういった計算は簡単ですし、有効でもあるのですが、この文脈ベクトルに必要な文脈語というのは非常にたくさんありまして、大きなコーパスですと数百万語、数千万語という単位で増えていきます。しかも、非常にスパース(sparse)である―１回しか出現しないような共起関係というのがコーパスの中には大量にありますので、そういった文脈ベクトルというものを考えて、野菜に対する共起語の頻度を並べて、数字のベクトルを作る、果物に対応する頻度を並べて数字のベクトルを作るということをしてしまうと、非常に巨大な数字の行列ができてきます。ところが、これは数百万次元、数千万次元、あるいはそれ以上のオーダーのものであって、とても今の計算機でも扱い切れるものではないですし、中身はほぼ空っぽなので、効率もよくないということになります。

そこで、これを圧縮しようということを考えるわけです。つまり、ほぼゼロで、意味のあるところに数字がちょっと入っていたような、そのような文脈ベクトルと呼ばれる数字の表現をギュッと圧縮する。今の圧縮率ですと、全体を数百ですとか、たかだか10万ぐらいの次元の数字の並びに圧縮する。この圧縮したものが「埋め込み表現」と呼ばれて、現在の大規模言語モデル、あるいは生成モデルの基本になっています。ですので、言語モデルの出現当時の目的は何だったかと言うと、このようにして意味を実数値のベクトルで表現するようなやり方をギュッと圧縮する、そ

の圧縮のところにニューラルネットワークという非線形な、高度な情報処理の技術を用いたものと私は捉えております。

さて、そういった言語モデルですけれども、では、圧縮するために何を使うか、どのようにしてこの圧縮した表現を得るかというと、実に簡単な訓練の方法を用いています。素材となるのは、大量のテキストです。コンピューターが行うのは、次の単語は何かということを予測するという、次の単語の予測であります。ひたすらこれを繰り返していきます。例えば「駆け込み乗車は」という文脈が与えられたときに、次に来る語は何でしょうかということをコンピューターに予測をさせます。コンピューターは何かを答えるのですけれども、それが違えばペナルティを与えるし、合っていれば、「ああ、正解ですね」ということで報酬を与える。そういった訓練を無限回と思われるぐらい膨大な回数、繰り返して、なるべくなるべく正確に予測する。

これが言語モデルの訓練のほぼ全てでありまして、実際に例えば生成ＡＩ、ＣｈａｔＧＰＴ、あるいはＣｌｏｕｄ、最近ではたくさんのモデルが出てきていますが、そういったモデルを使う際にプロンプトと呼ばれる部分がこの周辺語入力になっていて、生成ＡＩが返してくる答えというのは、このターゲット語の部分になっていて、正にこの仕組みだけでユーザーが与えた入力に対して、コーパスを訓練して得られた最適の予測に基づいて、答えを言語モデルが返してきているというのが、プロンプトに対する言語モデルからのレスポンス生成の仕組みになっています。この予測というのは、タスク自体は簡単ですし、言語のコーパスを作る際に必要となるような高度なアノテーションという作業が全く必要ないということで、非常に学習を簡略化しています。

簡単に思えますが、実際にはこの次に来る単語を予測するために必要な言語的な知識、あるいは常識、推論といったものは、人間のこの認知的な処理の全てといってよいぐらい多くのものを含んでいます。

例えばテキストのスタイルといったものの理解も必要ですし、文脈、文脈を超えた談話構造の理解といったものも恐らく必要ですし、知識も必要、あるいは構文の理解も必要ということで、ありとあらゆる言語理解の知識がこの次に来る語の予測というタスクの中に埋め込まれています。そのために先ほど御紹介しましたような数万、あるいは10万次元の圧縮した意味の空間の中に、その知識を埋め込むこと

ができている。これが実現できていることは計算言語学の研究者にとっても非常に驚異であり、言語モデルの中で何が起きているかを調べること自体が、今、学問として重要になっています。以上が、大規模言語モデルの背景となります。

続きまして、２番目のトピックに参ります。大規模言語モデルの意味理解について、その中で特にこの場でこれまで伺ったお話に関係していると思われる語義の曖昧性と意味の最小単位というものが、この言語モデルの中でどのように扱われているかについて簡単に御紹介いたします。語義の曖昧性は、皆様、御承知のとおり、同じ言葉が異なる意味を表すことがあるという問題です。コンピューターは、文字列を読むだけ、つまり表層的な処理をしているだけです。

例えば「ｍｏｕｓｅ」という語は、全く違う意味を持っているにもかかわらず、文字のレベルでは同じになります。元々これは言語処理における一大問題でありまして、この語義をどうやって計算機に区別させるかということは、長年の研究者の課題でありました。

言語モデルが登場する前にやっていたのは、まず辞書を持ってくると、その辞書の中で複数の語義が定義されています。コンピューターにターゲットとなる語と文脈を与えたときに、その語が辞書のどの語義で使われていますかを選ぶことでした。ところが、大規模言語モデルが出てきて、その語義、曖昧性解消の考え方が一新しました。何が変わったかと言うと、先ほどの埋め込み表現でして―少し話が複雑になってまいりますが―この埋め込み表現は数字で、ベクトルなのですけれども、この埋め込み表現が関わってきます。深層学習というのは、コンピューターが言語を処理していくレイヤーが非常にたくさん積み重ねられているという意味で、「深い層」という名前になっていますが、そのレイヤーを経るに従って、いろいろな文脈の情報を次第に取り込んでいって、最後に、与えられた文脈のいろいろな側面を考慮した埋め込み表現に変換されるという処理が行われるようになりました。つまり、入力される埋め込み表現と出力される埋め込み表現は異なるものであるという、そのような文脈情報が反映された埋め込み表現をモデルが生成することができるようになったというのが大きな点です。その結果として、例えばスライドのこの例は比較的単純なモデル―生成型のモデルがこれほど一般的になる以前の話、2021年の話ですけれども―その中で「ｓｐｒｉｎｇ」という言葉が言語モデルによってどのように区別されるかというこの語義の変化を捉えた様子を論文から引用してきたもの

です。例えば「ｓｐｒｉｎｇ」はシーズンであるとか、ｗａｔｅｒ　ｓｐｒｉｎｇであるとか、デバイスであるといった、こういった語義の区別は非常にシャープにできていますし、その中でもこの「ｓｐｒｉｎｇ」の2005とか、2009といったような非常に細かな意味の違いも、実はモデルは捉えられているということが分かります。このように語義の曖昧性解消のやり方というのは、今の言語モデルで非常に大きく変わったところであります。

次に、意味を担う単位の話に移ります。先ほど出てきた図の埋め込み表現は言語モデルの基本なので何度も出てくるんですけれども、例えば野菜に対して数字のベクトルが張りついたもので、８ページにございました。前のスライドにあったのは、そういった埋め込み表現が入力層に入力されるということですが、これはつまり、野菜という言葉が現れたときに、辞書を引いて野菜に対応するベクトルを深層モデルに入力しているということになります。

逆に言えば、この入力する単位というのは、ここに、この辞書に登録されているもので、埋め込み表現が事前に分かっている単位のみが入力のユニットとなるということになります。その単位を言語モデルでは「トークン」と呼びます。意味を担う最小の単位は、言語学的には形態素と呼ばれる場合が多いと思いますけれども、言語モデルにとってはトークンと呼ぶということです。

言語モデルの場合は、したがいまして、語彙というのはトークンの異なり数となっていて、辞書というのは、このトークンと埋め込み表現の対応表となっています。トークナイザーというのは、入力テキストをトークンに分割する処理となります。このトークナイザーが非常に重要な役割を実は果たしていまして、こちらはＧＰＴ－４ｏのトークナイザーですけれども、誰でも試すことができます。語というのは無限に生成されていくので、全部の語を登録するということは無理で、先ほどから出てきているとおり、数万から10万異なりトークン―語彙しか今の言語モデルでも持っていません。そのためにどうするかというと、英語の場合は、普通は単語自体を更に分割することは行わないのですけれども、その単語をトークンに分割する際に、例えばここにあるやや難しい単語、この単語はめったに出てこない言葉なので辞書には登録されていない。

これを、トークナイザーというものを使って機械的に分割をしてしまって、埋め込み表現に直すと、これは符号化と同じで、全く意味を考慮せずに統計的に決まり

ます。最小の単位は文字なので、文字単位にも分割できるようにしておけば、原理的に全ての文字列がトークナイズできるという形になります。

これは英語と、英語を日本語に翻訳したものをトークナイザーに掛けたものですが、日本語の方が、多少効率が悪い結果になっている。すなわち、訓練するコーパスに応じて英語中心のコーパスで訓練したりしたものは、日本語のトークナイズの効率が悪い、つまり、分割したトークンの数が多くなってしまうという問題があります。

ちょっと枝葉ですが、これは日本の語、日本で生成ＡＩを使う際に当初非常に問題になったことでして、このトークンの数というのは、入力の長さなので、直接コストに直結してまいります。ですので、マイナーな言語、コーパス中で出現が少ない言語というのは、実は損をしている―あるいは能力的にも入力のサイズが小さくなることによって、機能的にも制約を受けるという、その言語間の不平等性というのが大きな問題となりましたし、今でも完全には解消されていないと考えられます。これがその言語間の不平等について分析した論文からの引用となります。

というわけで、トークナイザーに関するまとめとして、このトークンという言語モデルで扱っている最小単位というのは、統計的アプローチによって機械的に決まっていることが挙げられます。我々が思うような形態素、つまり、人間にとっての意味の最小単位とは異なるものになります。例えば星座、
「ｃｏｎｓｔｅｌｌａｔｉｏｎ」という英語を、我々が
ｃｏｎｓｔｅｌｌａｔｉｏｎはどのような意味を持つかを知ろうとするとき、この「ｃｏｎ」と「ｓｔｅｌｌａ」と「ｔｉｏｎ」のそれぞれの果たす役割を知っていれば、語の意味というのは推測可能なのですけれども、言語モデルは効率のよい符号としてトークンを考えていますので、必ずしも意味のある単位に語を分割するというわけではありません。もちろん、非常に頻出するものについては分割しますが、基本的には意味の最小単位と外れているということになります。

トークナイザーというのは、言語モデルの裏方にある、ある意味マニアックな課題ではありますが、語の意味というのを考える上で、これからも重要になってくるし、実際には対象コーパスを常に意識しているところであります。対象コーパスが重要という点では、言語モデルは、言語のためだけのものではありません。例えばプログラム言語も同じモデルで処理をしていたりします。プログラム言語において

は独特な語彙が生じていますし、大きく違うところは、スペース、タブ、括弧などの扱いが違います。また、コーパスの主流となっているウェブ文書の中には、様々なウェブ文書固有の、例えば改行の扱われ方、箇条書きのやり方、指示などがありますので、この「言語モデル用のコーパス＝人間が読み書きする言語」ということではない、あるいはトークナイザーのような統計的処理だけではなく、漢字というユニットに配慮する必要があるなど、様々な問題を含んでいます。

少し長くなりましたが、次の話題に移ります。次は大規模言語モデルの訓練とテキストコーパスということで、大規模言語モデルの規模についてお話をいたします。先ほどのように単純な次のトークンの予測によって学習をしていく言語モデルですが、大きなブレークスルーというのは、その規模が非常に大きかったということにあります。その変数、つまり、この予測に有効な手掛かりを表すための重み、ここに実数値を入れて最適化するように変数の値を変更していく。その変数の値は、例えば数千億から１兆個という規模に今なっています。また、それらの変数を最適化するために必要となるテキストは、当然、膨大な量となりまして、例えば数兆トークンという形になります。

どうしてこんなに必要なのかというのは、実際にモデルを構築する立場から見ても、実感ですけれども、モデルのパラメータサイズが大きければ大きいほど、あるいは訓練に使うコーパスのサイズが大きければ大きいほどモデルの性能がよくなるという関係がいろいろな実験、あるいは評価によって確認されていて、この関係を超える技術というのは、今ないというのが現状であります。しかも、そのモデルが大きくなったときに、コーパスも大きくしないと意味がないということで、今は、更に大きなコーパスの獲得に向けてしのぎを削っている状況であります。

次に行きまして、したがいまして、コーパスというのは大きければ大きい方がいいというのが今の世界ですが、先ほど申し上げたとおり、ウェブ由来のコーパスというのが人間の使う言語そのものであるかというと、必ずしもそうでない側面もあります。

質も重要であるということは知られています。ウェブの文書を、様々な方法でノイズを取り、重複を削り、そして有害とされる情報をなるべく削る―例えば個人情報は削るとか、様々な前処理をして質を高めた上で学習に掛ける。ウェブの中で数％ぐらいしか学習に使わないという話もあるぐらい、頑張って質を高める。あるいは

学習の最後に質の良いコーパスを入れるなどの工夫が行われています。

後ほども出てくるかもしれませんが、現状でもまだトークンは足りない。今後、トークン、あるいはコーパスはどんどん足りなくなってくるだろうと言われています。今、それで何が行われているかと言うと、自分でテキストを生成したものを訓練に使うということが始まっています。自動生成したテキストを訓練に使うというのは、比較的当たり前の作業となってきていますが課題も指摘されています。ということで、コーパスの中には自動生成したテキストが入ってくることに対してどう考えるかというのは問題になってきまして、この後のスライドで少し触れさせていただければと思います。

続きまして４番目、コーパスに由来する大規模言語モデルのリスクということで、バイアスとテキストの復元について簡単に御紹介します。テキストを生成する言語モデルのリスクというのは、非常にたくさん指摘もされてきましたし、言語モデルを発表する際には、テクニカルレポートというのを出すのですけれども、そのテクニカルレポートの中でも、このモデルは、こうした問題に対してこのような問題があるという分析を添えて発表するということが行われています。例えば、差別ですとか、有害な表現、あるいは事実の誤り、それから、プライバシー侵害などなど多くの問題は認識されていて、特に今は事実の誤りが話題になることが多いですけれども、様々な改善の努力はされています。

その中でも、特に見付けにくく、言語そのものに影響を与えると考えられるのはバイアスの問題です。元々コーパスにはバイアスは含まれている、バイアスのないコーパスというのはないと思いますけれども、生成するテキストの中にバイアスがあるという影響は、リスクとしてやはり憂慮するべきものであると考えられます。バイアスはいろいろな方法で測るのですけれども、これは、職業バイアスというのは多く話題になる典型的なバイアスの例ですが、ここにあるような

Ｔｈｅ ａｄｖｉｓｏｒ ｍｅｔ ｗｉｔｈ ｔｈｅ ａｄｖｉｓｅｅ ｂｅｃａｕｓｅ ｓｈｅ ｗａｎｔｅｄ ｔｏ ｇｅｔ ａｄｖｉｃｅ ａｂｏｕｔ ｊｏｂ ａｐｐｌｉｃａｔｉｏｎｓ

という英文があったときに、この「ｓｈｅ」という言葉は、「ａｄｖｉｓｏｒ」と「ａｄｖｉｓｅｅ」のどちらを参照していますかということを質問することによって、この職業に対する性別のバイアスがあるかどうかを調べるといった調べ方もベ

ンチマークとして行われます。

それをやってみると、やはりモデルはコーパスに存在するバイアスをかなり正確に反映してしまうので、最後のセーフティーの訓練によって、それをなるべく補正するようにしているというのが今の作り方です。なので、世の中に出す前にこういったバイアスの既知のものに対しては、訓練をしてなるべく出力しない形で配慮して、初めてユーザーの下にモデルが届けられる、あるいはそういったことを義務付ける規制を掛けていこうとしているという動きになっています。

更に分かりにくいことに、例えば政治的嗜好に対してもバイアスがあるということは、報告されています。これは去年の言語処理の学会での報告からの引用ですけれども、いろいろな言語モデルがある中で、西洋の民主主義に沿ったこの基準というのでプロットしてみると、言語モデルによってやはり政治的な嗜好についてもバイアスが掛かっている。あるいはこれを、言語モデルをファインチューニングと調律、調整することによって、このバイアスは更に変わるというふうなことが報告されています。

２番目にオリジナルテキストの復元問題、これは著作権の問題で、やはり非常に気になっているところです。言語モデルがどれぐらい入力されたコーパスを暗記していて、どれくらい復元してそのまま出す可能性があるのかということは、実は研究の途中で、特に大きなパラメータのモデルについて、確固とした結論というのは、これからの緻密な分析を待つ必要があるかと思います。というのも、学習に使ったコーパスというのが公開されていない場合が大半で、その学習に使ったコーパスがくまなく分からない限り、緻密な分析が難しいからです。こちらは、その幾つかの比較的小規模なコーパスについて、何度も出てくるデータほど続きを抽出しやすいか、文脈が長いほど続きを抽出しやすいか、終盤に学習したデータほど続きを抽出しやすいかといった幾つかのリサーチクエスチョンに対して分析してみたというものです。

こちらは、その簡単な結果ですけれども、この調べた範囲では、正にこの三つの設問に対してはイエスであって、高頻度のテキストは再現されやすいし、長いプロンプトほど再現しやすいし、最後に学習したものほど再現しやすい。これは再現するのは著作権的には好ましいことではないかもしれないですけれども、知識を学習させるという文脈においては重要なことなので、この暗記するということと、再現

するということの間にはトレードオフ関係があって、この辺りのチューニングも言語モデルを世に出す際には、議論が必要になっています。

最後にテキストコーパスと大規模言語モデルについて、簡単にお話ししてまとめとしたいと思います。今までのお話の中で出てきたことで、例えばＫＯＴＯＮＯＨＡコーパスとの対応の中で、少し話題を拾ってみたのがこちらの３枚のスライドとなります。

まず、ＫＯＴＯＮＯＨＡコーパス中で、例えば「野菜」を検索してみると、検索文字列前後の文脈が出てきます。この野菜については、出典も明確であり、この「野菜」という文字がどのような出典の、どのような文献の中でどのように使われていることを知るための非常に重要な手段であります。

しかしながら、一方で、先ほど見たような野菜の文脈のベクトルを計算するような例文とはやはり趣旨が異なるし、使われ方も異なるという形になっています。それから、語義について、「ｓｐｒｉｎｇ」の語義をＷｏｒｄ Ｎｅｔというもので引いてみますと、「ｓｐｒｉｎｇ」というものについて異なる語義が提示されているわけですけれども、この「ｓｐｒｉｎｇ」の語義というのは、先ほど分布図の上で見たような言語モデルが見ている境界が比較的曖昧な語義の捉え方とはまた違ったものとなっています。

また、トークナイザーで御紹介したような、この統計的に決まってくるようなトークンの単位というのは、例えば書き言葉均衡コーパスの上での短単位、長単位という意味を考慮した単位とは違ってきます。ですので、かなり人間の言語学的に見ているような意味の捉え方と、言語モデルが見ている意味の捉え方は、根本の下の層の部分では異なるところがあるのですけれども、言語モデルが出力するテキストについては、非常に自然な言葉をしゃべり、知的であるという印象をユーザーの誰にも与えるようなものになっていると言えると思います。

最後に、言語モデル構築における良質なコーパスの使われ方ですけれども、今まで出てきた話のまとめとして、一つは品質のフィルタリングに使うことができます。例えば、アノテーションなしの良質なテキストは事前学習に、学習に使うコーパスの中で質のよいものをフィルタリングするということに使うことができます。２番目はモデルの能力評価で、人手によるアノテーションが付いた言語資源というのは、例えばメタ言語能力、形態素解析、あるいは用語抽出、構文解析といった、そうい

った言語モデルの能力テストや、あるいは微調整するために使うことができます。また、トークナイザーの語彙辞書を作るのに当たって、機能語などを意図的に取り込む、あるいは漢字を取り込むといった、この言語学からの知見に基づいた語彙の増強ということをすることができます。

最後に重要なのは、言語空間のモニタリングだと思います。言語空間のモニタリングについて、少しスライドを紹介してまとめとしたいと思います。このモニタリングの意味は、計算的に語の意味の変遷や変化を捉えるということで、これはモデルを見ていても駄目で、必ずコーパスが必要になります。

こちらは一時期話題になった「ｄｅｌｖｅ」という単語が増えているというＴｗｉｔｔｅｒですけれども、それにとどまらず、例えば学術論文でアーカイブというプレプリントサーバーがあるのですけれども、その中でどのような語が増えて、頻出語はどのようなトレンドがあるかというものを調べて、ＣｈａｔＧＰＴの登場以前と以後に分けたという分析もあります。例えばこういったものと言うのは、年度別に区切ったコーパスがなければ捉えることができませんし、たとえ捉えたとしても、普通に、別にＣｈａｔＧＰＴがなくても学問の変化とともに使われるようになる用語というのはあるので、それを更に一体どのように生成ＡＩが関与しているのかという分析は、コーパスを得て更に深い分析を行って初めて結論付けられるものです。

最初に前半部分で申し上げましたとおり、言語モデルでは、今、データを生成するのに言語モデルを使っている―言語モデルが言語モデルの作成に使われるということが定着化しつつあります。これに対して Ｍｏｄｅｌ　Ｃｏｌｌａｐｓｅということで、例えばＮａｔｕｒｅの記事では、生成したテキストをモデルの学習に再利用することを繰り返すことによって、テキストが本来持っていた多様性は失われていくということ、言わば、ある意味、モデルは統計的に最適なものを出してくるので、多様性が失われるというのは自然なことではありますが、それを数学的にも実験的にも試したという、そのような論文があり、それを Ｍｏｄｅｌ Ｃｏｌｌａｐｓｅというふうに命名しているという、そういったワーニングも出ています。

以上の言語空間のモニタリングについて、最後に一言、言語空間には今後ＡＩが生成したテキスト、人間とＡＩがやり取りしたテキスト、人間がＡＩの影響を受け

たテキストなどが混在していくことになりますので、その中でモデルを構築する側としては、とにかくテキストが欲しいという切実な要求の中で、どうやって品質というもの、あるいは正しい言語というものを考えていくかというのは、非常に重要な問題であると考えています。

以上で発表を終わります。

ということで、自分で発表しておいて何ですけれども、御質問等がありましたら―後で議論の時間を設けますが、この場で、もし御質問等がありましたら、よろしくお願いします。

○石川副主査

相澤主査、大変啓発的なお話をありがとうございました。幾つか教えていただきたいのですが、まず１点目として、言語モデルが新しいデータを自分で作り、それを更に学習していくという御指摘でありましたが、このことによってモデルの精度が下がっているというようなことはないのでしょうか。

○相澤主査

正に、やみくもにやるとモデルの精度が下がると言われていて、そこをいかに克服するかが、今、技術の焦点の一つにはなっております。

○石川副主査

ありがとうございます。

２点目ですけれども、差別や偏見を含むテキストも世に多く実在しているわけで、現実言語を母集団とすれば、サンプリングでそれらがコーパスデータに入り込んでくる可能性もあるかと思います。この点に関して、大規模言語モデルの研究者から御覧になった場合、伝統的なコーパスにはどのような言語データが収集されることが望ましいとお考えでしょうか。

○相澤主査

バイアスの問題は非常に奥深いものがありますが、コーパスとしては、やはりあ

るがままの姿を収集するべきであるというふうに私自身思っております。というのは、何が公平であるかという基準は、これまで人類は持ってこなかったと思いますので、その基準を決める役割をコーパスが持つのは問題があると感じるからです。何が公平であるかというのは、今、現時点で誰にも分からないわけですが、社会のノルム（norm）というのはある。しかも、そのノルムというのは社会によって、コミュニティーによって異なるので、いかにそのコミュニティーの基準に言語モデルを合わせていくかというのが価値のアラインメント（alignment）と呼ばれる操作ですが、これは、今現在は言語モデルの出力に対して○や×を付ける。アノテーターが決めているということです。それでいいのかという議論はあるかもしれませんが、それはコーパスの役割を超えた――繰り返しになりますが、というふうに思っております。

○石川副主査

ありがとうございます。コーパスの作り手の側が恣意的にフィルタリングをするのではなく、まずはあるがままのものを集めて提供する。その後、目的に応じて、フィルターを掛けていくといった方向が望ましいこと、よく理解できました。

最後に、３点目になります。ハルシネーションについて、こうしたことが起こる仕組みと、あるいはそれを抑止するためにデータ収集の際に工夫できることがもしございましたら、御教示頂ければと思います。

○相澤主査

ハルシネーションについては、ハルシネーションを全く出さないようなモデルというものは原理的に難しいのではないかということが最近は言われ始めています。ただ、ハルシネーション自体が、ハルシネーションがあるから何もできないということでは決してないので、ハルシネーションがあるということを想定して、どのように我々は言語モデルを使っていけばいいのかという議論に次第になっている面もあると思います。

○石川副主査

なるほど。

○相澤主査

一つは、やはり信頼のおける情報源を参照することを想定しつつ、言語モデルを使っていく検索ベースのやり方です。とはいえ、余りにも問題の多いテキストで学習をしてしまうのは、それはもちろんないにこしたことはないので、ある程度質は担保しつつ、かといって良質なもので学習したからといって問題が起きないとは言えないことも認識しつつ、やっていく必要があると思います。つまり、正しいものＡと正しいものＢを一緒に出力することによって正しくなくなるということはよくある話なので、しかも、言語モデルは適当に組み合わせてくるということですので、コーパスだけでの対処ではなくて、信頼のおけるコーパスは、むしろ出口の外側で検索して使われることになるかなと思います。

○石川副主査

ありがとうございました。

○相澤主査

では、ちょうど時間となりましたので、続きまして川原先生による事例紹介ということで、よろしくお願いいたします。配布資料４となります。

○川原発表者

よろしくお願いします。では、始めさせていただきます。今日は、言語学者としてお話しするべきなのだとは思うのですが、この問題を考えるに当たって、自分が父親ということに影響されていることは間違いないので、そこは素直に情報開示してからお話しした方がいいかと思います。

生成ＡＩを子供に与えるかどうかという議論なのですけれども、それを考える以前からスマホがメンタルヘルスに与える影響というものに関しては、前々から興味を持っておりました。特にここに提示した本には非常に影響を受けまして、スマホの普及とともに精神疾患の数が急増しているという可能性が指摘されています。これが具体的な値なのですけれども、Ｙ軸がアメリカの学部生の何％がそれぞれの精神疾患を持って、発症しているかということのグラフなのですけれども、2010年、

2012年辺りから急増していることが報告されています。一番上がＡｎｘｉｅｔｙ、不安感ですね。不安障害。その下が鬱で、あとはＡＤＨＤも急上昇しております。この2010年、2012年というのが、それぞれスマホが若者の手に渡った時期、インスタグラムが普及した時期というふうに指摘されております。

本題ですけれども、今、ＣｈａｔＧＰＴに代表される生成ＡＩを搭載したおしゃべりアプリというものが開発され、子供に与えられようとしています。私は、この動きに対して意識的に慎重な立場から意見を申し上げたいと思います。

この話題なのですけれども、父親としていろいろな方とお話しする機会がありまして、結局、どっちなんでしょうか、役に立つんでしょうか、危ないんでしょうかということを聞かれます。正直な答えとしましては、今のところ分からないというのが一番誠実な答えだと思っております。日本や世界を救うような特効薬になる可能性もありますし、毒にも薬にもならないかもしれません。ただ、臨床試験を経ていない、リスクがはっきりしていないという点は自覚して進めていく話であろうと思っております。

何でこのようなことを申し上げるかと言いますと、やはり我々の世代で、子供にスマホを見続けさせるのか、どれだけ自由に与えるのかというのは、個人差があり、私の周りの印象ですけれども。やっぱりこの話になりますと、特に公共の場、レストランなどで子供が騒ぐと良くないので、仕方がないから静かにさせるためにスマホを見せるのはしようがないじゃないというような意見も聞かれるんです。その気持ちは同じ親として痛いほど分かるのですけれども、同時に臨床試験を経ていないけれども、ここに子供が静かになる薬がありますというものがあったときに、それを本当に飲ませる親がいるのかという話になると思うんですね。私、この二つに関して原理的な違いを感じられないんです。

ですから、このような考え方を土台として、やはり生成ＡＩを子供に与えるときに、どのようなリスクが考えられるのか、しっかり考えていかなければならないと思っております。もちろん、前提として母語獲得というのは人間の成長、それから、全ての学びの土台になるという前提でお話ししていきます。

先ほど相澤主査からじっくり説明いただいたので、繰り返すのも何なので少し簡略化しますけれども、まず、言語学をずっと研究している身からすると、生成ＡＩが出力するもの＝人間言語だと考えるのは非常に危険だと思っております。ここで

違いを数点、明確化したいのですけれども、一つは、生成ＡＩは書き言葉が主な訓練データとなっております。人間は音声から学ぶ場合がほとんどです。生成ＡＩには身体がございません。人間は、ほかの身体感覚とともに言語を学んでいきます。生成ＡＩ、先ほどじっくりトークンの話が出てきたので、私は安心してトークンと単語は違うよという前提は置いておいて、数億、数兆のデータが必要というのが今の生成ＡＩの現状でして、これ、人間で置き換えますと、非常に雑な計算なんですけれども、１秒に２単語読み続けたとして６０００年程度掛かる。これがＧＰＴ３ですね。ＧＰＴ４は規模が公表されていないので分からないですけれども、恐らく１万年以上掛かるであろうと。

こちらも相澤主査のスライドに出てきましたけれども、生成ＡＩは学習データが増えれば増えるほど性能が上がるということが示されています。人間は、そこそこの量でも基本的な言語知識が誰しも身に付くというのが近年の言語学の知見でございます。もちろん、読み聞かせとか、語り掛けが大事ではないとは言っていません。大事なのですけれども、６，０００億と７，０００億で違いが出るというようなことは恐らく人間には当てはまらないであろうと。そう考えると、やっぱり人間言語と生成ＡＩの出力を同じものとみなすのは危険であろうと考えます。これが懸念点の第１ですね。非常に確かに自然な日本語というふうに見えるのですけれども、その日本語というのは、果たして本当に日本語なのかというのは一度立ち止まって考える必要があろうと思っております。

２番目の点が、簡単に言うと、うまくいくんだけれども、何でうまくいくか、そしてそれぞれのパラメータが実際何をしているのかということが分かっていないというのが、私個人としては少し心配に感じられるところであります。これは個人、個人の考え方だと思うのですけれども、その正体不明のものを子供に与えていいのかという不安は、親としてはございます。例えば新しい車が開発されました。何だか分からないけれども、ちゃんと動いて、ちゃんと止まりますという車が開発されたときに、それに乗れるかというと、私は正直乗りたくないし、子供を乗せたくないなと思います。ただ、こう言ってしまうと、自然言語の正体、何で自然言語はうまくいっているんですかと問われるとちょっと難しいんですけれども、自然言語は少なくとも今までの実績があるので、親から子に継がれていったという実績があるので、その点においては信頼できるのではないかなと思っております。

３番目の点です。先ほど簡単に触れましたけれども、生成ＡＩの学習データは主に書き言葉です。人間は養育者の音声的な語り掛けから言語を学びます。簡単に言えば、新聞や小説を読みながら母語を学ぶ赤ちゃんはいないわけです。それぞれのポイントは、クスッとなってしまうポイントかもしれないですけれども、重要だと思うんですね。音声と書き言葉という違いだけではなく、養育者の語り掛けというのは独特な言語学的な特徴を持つことが分かっていて、例えば声が高くなるであるとか、イントネーションが強調されるであるとか、文が短めで分かりやすくなっている、繰り返しが多いなどの特徴を持っていて、これらの語り掛けに対して赤ちゃんがより強い反応を示すということも分かっているので、この語り掛けの重要性というのも見逃せないのではないかと思います。

少しずれてしまうかもしれないのですが、今、この書き言葉と音声表現という比較について考えてみますと、文字が生まれたのはすごく最近のことなんですよね。人間言語がいつ誕生したかというのは諸説あるのですが、大体５万年前から200万年前と言われています。文字の誕生が5,000年前ですね。ですから、人間言語の歴史を短く見積もって、文字の歴史の長さを多く見積もっても10分の１、もしかしたら400万分の１で、しかも、この文字が誕生した頃というのは、文字というのは非常に限られた人たちが使っていたもので、多くの人は読めなかったんですね。人間言語のコミュニケーションにおいて、書き言葉が占める割合というのは非常に少なかった。50年前はほとんどのコミュニケーションが対面で行われていました、もちろん手紙などもありましたけれども。

1980年ぐらいですか、電話が普及して、そこから声だけで、表情とか身振り手振りを含まないコミュニケーションというものが生まれ、2000年に入ってインターネットが普及したことで音声すら捨象されたコミュニケーションが占める割合というのが急増したんです。これ、たかだか25年ですから非常に最近なことでありまして、そこも人間はそんなに早く変われるのかという懸念点はございます。人間は本当に書き言葉だけで分かり合えるのかということですね。

書き言葉で捨象されるものは何かというのを考えてみますと、例えばイントネーションですね。あとは抑揚、声の強さの変化、声色や、あとはブレスの使い方ですね。息づかいなんかも、音声で話していれば伝わることですけれども、書き言葉ではこれらが表現できません。特に感情というのは、これらによって表現されること

が多いと思います。ですから、生成ＡＩは人間言語のこれらの特徴を本当に再現できるのだろうか。まずはできるのかどうかをしっかり精査すべきだと思いますし、もしできないのであれば、生成ＡＩの出力を子供に与えると悪影響につながる可能性もあると思います。これに関してですが、生成ＡＩに限らず、イントネーションに関しては様々な研究がなされています。ＹｏｕＴｕｂｅで今はやっている人工読み上げソフト―非常にイントネーションが不自然なものがありますけれども、あれが当たり前になっていいのだろうかという不安感は正直ございます。

懸念点の四つ目としまして、生成ＡＩの背後に感情がないということですね。恐らくと言うか、これ、試しましたけれども、生成ＡＩに「ただいま」と言えば、「おかえりなさい」と返してくれます。これは「ただいま」の後に「おかえりなさい」と来る確率が高いから、そう返すのであるんですね。ただ、例えば私の娘が小学校から帰ってきたときに「ただいま」と言って、「おかえりなさい」と私が言ったときに、それは確率じゃないんですよね。まあ、確率もあるのかもしれないですけれども、事故に遭わずに無事帰ってきてくれてよかったという安心感もあるし、学校、お疲れさまという気持ちもあるし、今日はこれから何しようという期待もあるし、このような感情を全部含んだ「おかえりなさい」なんですが、生成ＡＩはそのような感情はない。ないならないでいいのかもしれないのですけれども、子供は、その生成ＡＩの背後に感情がないことを理解できるのだろうか。そこで不利益は生じないのであろうかというのも懸念事項ではございます。

５番目としては、共同注意の機会の損失という、これは心理学でよく言われていることなのですが、共同注意というのは、右側の図の３項関係の方でして、２項関係というのは、子供がいて何かを見る。自分と何かという対象、自分と対象という二つの関係なのですけれども、この共同注意というのは３項関係を基に成り立っていまして、子供がいて、養育者がいて、子供と養育者が同じものに対して経験して注意を払って何かをするという機会が共同注意です。なぜこの共同注意が大事なのかというと、心理学で有名な心の理論というものの発達に重要であるということが知られています。

この心の理論というのは、どのようなことかと言うと、他人がどのような知識を持っているか。他人がどのような気持ちでいるのか―簡単に言えば、空気を読む力ですね―他人の立場になってみるという人間として非常に重要な力なのですけれど

も、これが前のスライドのこの３項関係を持った共同注意によって育まれるということは、心理学の分野でかなり明確に示されているのですが、生成ＡＩ、スマホに搭載された生成ＡＩとは恐らくこの生の体験を共有できない。よって、生成ＡＩとの対話では心の理論は発達しないだろうと私は考えております。

懸念点、こちらは拍車を掛けるという、生成ＡＩだけの問題ではないのかもしれないのですけれども、大前提として人間の成長には五感全てへの刺激が重要ということが指摘されています。ここでは小泉英明先生―神経学者でありながら教育について非常にしっかりと考えている方の本からの引用になりますけれども、五感への刺激が非常に重要だということが神経科学の立場からも指摘されております。生成ＡＩは視覚情報及び音声情報のみですね。視覚情報もどれだけ提供できるのかは定かではありません。しかも、どちらも二次元でデジタル化されているという点において、実世界の刺激とは大きく異なるということを肝に銘じておくべきだろうと思っております。

これがなぜ大事なのかというと、これも小泉先生の受け売りなのですが、人間、動物一般そうなのですけれども、発達期に刺激を受けないと、それを感じる器官というものが発達しなくなってしまう。脳の刈り込みの過程で、このような刺激が入ってこなければ、ここは要らないんだよというふうに思ってしまう。そのように成長してしまうんですね。日本人がＬとＲの違いが聞こえなくなるというのは有名な話ですけれども、あれはＬとＲの違いは必要なのだというふうに脳がもう覚え込むわけですね。これは我々痛感していると思いますけれども、取り返しがつかないことが多いです。

有名なのがネコの実験で、縦ジマだけを見せて育てると、横シマが見えなくなってしまうということも報告されていますし、これはイギリスでの実例らしいのですけれども、結膜炎がはやったときに、子供に眼帯をして育てた時期があるらしいのですが、その子たちが弱視になってしまったということで、現在は眼帯を付ける期間というのは、できるだけ短くしているらしいんですね。というのは、視覚からの情報を小さいときにシャットアウトしてしまうと、脳が、視覚からの情報は重要ではないというふうに判断してしまう。ですから、五感全てへの刺激を与えるというのは、本当に重要なことで取り返しがつかないことである。言語に関して言えば、その表情を読み取る力であるとか、ジェスチャーであるとかという力を生成ＡＩの

おしゃべりアプリが提供できるのだろうかというところに懸念を感じております。

こちら、まとめになります。生成ＡＩの出力は、確かにかなり自然な日本語に見えるのですけれども、それを日本語とみなすのには、言語学的には抵抗があるという点。仕組みが分かっていないものを子供に与えて大丈夫なのかという不安。書き言葉がベースになっているという点。背後に身体もなければ感情もないという点。共同注意の機会損失及び聴覚・視覚情報への偏向という点が私個人的には心配に思うところであります。

繰り返しになりますけれども、これらが子供の発達に悪影響を与えるという証拠はございません。これは本当にやってみないと分からない。ただ、与えないという証拠もございません。これも繰り返しになりますけれども、今までの議論を踏まえて臨床試験を経ていない薬とどう違うのかということを改めて自分に問うても、納得いく答えが私には返ってきません。加えて、やはり依存性なども心配です。あとは、生成ＡＩとの会話が原因で自殺につながったケースもありますので、このような個別のケースをやり玉に挙げてどうのというのは、私は余り好きではないのですけれども、このようなこともありましたよというのは考えておくべきではないかなと思っております。

私だけの意見だと不安ですので、独りよがりになっている可能性もありますので、最近、早稲田大学の折田先生と明治大学の桃生先生と共同で言語学者相手にアンケートを取ってみました。いろいろな項目を聞いたんですけれども、今日、取り上げるのは、簡単に言って生成ＡＩを発達段階の子供の話し相手として用いることに言語学者として賛成しますかというのを５択で答えていただきました。４しか入っていないですけれども、これ、何で４回答しかないかというと、「賛成」といった方はいらっしゃいませんでした。「反対」が７人、「どちらかといえば反対」が１０人で、これで過半数ですね。「どちらかといえば賛成」という意見もございました。「どちらでもない」という意見もありました。「どちらでもない」という意見の方は、どちらかというと毒にも薬にもならないだろうという意見が多かったです。悪くないんだからいいんじゃないという印象でした。

具体的な声が挙げられて、私、非常に興味深いと思ったコメントを実際に取り上げてみました。一つの、なるほどと思わされたのは、やはり全ての家庭で子供に十分に話し掛けられるとは限らないので、補助的にならばよいのではないか。

あとは、もう少し極端な例で言えば、親に虐待されるぐらいであれば、スマホに救ってもらった方がマシなんじゃないのという意見もありまして、それは一理あるなと思わされました。ただ、正直、これは妻の意見なんですけれども、そのような状況の子供こそ、人間関係でうまくやれるように育ってほしいよねという理想論と言えば理想論なんですけれども、そのような懸念もあるということがあります。

次の意見としては、生成ＡＩの方が個々の人間よりも知識量が多いので、その点で学びにつながるかもしれないという意見があったのですが、こちら、先ほど相澤主査の御発表で、じっくり説明がありましたけれども、生成ＡＩが提供する知識が正しいとは限らないという、これをプロの言語学者が知らなかったというのがちょっとショックではあるのですけれども、そこは使用者はしっかり知っていなければならないと。先ほど相澤主査がＱ＆Ａのときにインディペンデントなファクトチェックができていなければいけないというふうにおっしゃったんですけれども、子供は恐らくそれができないであろうと。だとすると、子供に与えるというのはどうなんだろうというふうに感じます。

現状、生成ＡＩの質問には工夫が要る。聞き方をうまくやらなければならないということですね。ということで、子供の質問力の向上が期待されるかもしれないと言った方もいらっしゃるんですけれども、私としては、生成ＡＩとの付き合い方を学ぶ前に生の人間との付き合い方を学んでほしいなと思います。人間にどうやって質問するかも難しいですから、それこそ表情を読みながら、どのようなイントネーションで質問するかは非常に重要になってくるので、まずそっちではないかなと個人的に思います。

批判的な意見ですね。私が思い付かなかった批判的な意見と申しましょうか、

一番上は言語能力の発達の観察は子育ての楽しみの一つである。これは少し言語学者に偏った意見かもしれませんけれども、自分の子供がどうやって言葉を身に付けていくのか観察するというのは楽しいことでありまして、それを機械に任せるというのは子育ての楽しみを減らしているのではないかという意見もありました。

２番目の点は、相澤主査の均一化、画一化ということに関わると思うのですけれども、データが画一化されていくわけですね。そうすると、方言の消失に拍車を掛ける可能性もあると思います。これはこの小委員会の趣旨とも非常に密接に関わってくるのではないかなと思います。

最後に、言葉は人を傷付ける可能性がある。これは試してみましたが、生成ＡＩは傷つかないですね。どんな暴言を吐いても傷つかない。人間関係で、こう言ったら友達を傷つけてしまったとか、友達が怒ったとかという経験というのはすごく大事だと思うのですけれども、生成ＡＩは怒らないです。何を言っても許されるというような原則を学んでしまったら困るなと思います。

このＶＲのヘッドマウントディスプレイに関して御指摘くださった方がいらっしゃいまして、これは年齢制限が推奨されているらしいです。これは医学的な見地から１２歳だったかな、ごめんなさい、具体的な数字が出てこないんですけれども、科学的な知見に基づいて、ここまではやめておきましょうというような提言が既になされています。ＡＩ搭載の自動運転車も、今すごく長い間じっくり安全性をチェックしているので、生成ＡＩのおしゃべりアプリに関しても同じような態度を持ってくれるとうれしいなというふうに言語学者として、親として思います。

批判だけしているのは簡単ですので、少し建設的なことを最後に申し上げますと、やはりガイドラインを検討すべきだと思います。これに関しては、本当にたたき台なんですけれども、やはり使用は母語獲得がある程度確立してからの方が安全であると考えます。母語獲得がいつ確立するかというのも諸説ありますし、個人差もあるのですけれども、やはり五、六歳というのが一応、その言語の知識が定着するという基準になっていますので、例えば未就学児は使用を控えるというような考え方、捉え方が安全ではないかなと考えます。そして、頻度も大事で、人間との会話の大事さを忘れてしまっては困るので、少なくとも人間との会話の頻度が多くなるように、飽くまで補助的に使ってくださいということですね。もし生成ＡＩのおしゃべりアプリを子供に与えるという開発チームがあるのであれば、正しい情報公開とリスクアセスメントをお願いしたいと思います。これは継続的に行っていただきたい。

そして、使う側です。これはもう養育者へのお願いになると思うのですけれども、子供に何を与えようとしているのかをしっかり、最低限理解した上で与えるということが肝要になると思います。ハルシネーションも一つ大きな問題ですし、バイアスですね。相澤主査が解説してくださいました。あとは、これは飽くまで機械であって、アドレナリンもセロトニンも出ないから、怒るとか幸せを感じるということができないものなんだよというのを分かった上で、子供に与えるというような姿勢が最低限求められると思います。

最後、取って付けたようななんですけれども、事前の打ち合わせでどんなコーパスが欲しいですかと聞かれたので、これはつながっているんですけれども、私の中でちゃんと。どんなコーパスが欲しいですかと聞かれましたら、まず、養育者から子供へどんな語り掛けがなされているのかというコーパスが欲しく思います。実は理研（理化学研究所）で、もうあるのですけれども、権利の関係上、理研でしか使えないということで、親としては、ほかの親がどんな語り掛けをしているのかというのを知るというのは非常に心強いことで、このようなのがあったら助かるなというのは、言語学者としても親としても思います。

あとは、先ほど生成ＡＩが感情をちゃんと表現できるのかというような問題提起をしましたけれども、私、幸いなことに声優さんであるとか、俳優さんとの付き合いもございまして、彼ら、彼女らの演じ分けを音声学的に分析したりしているんですけれども、プロがどのように感情を表現しているのかというコーパスがあれば、彼らは感情表現のプロですから、それに学べることは多いのではないか―そのようなところを子育てに生かせていける可能性というのは大いにあると感じております。

私の発表は以上になります。ありがとうございます。

○相澤主査

大変興味深いお話をどうもありがとうございました。

それでは、ただ今の川原先生の御発表につきまして、御質問等ありましたら、よろしくお願いいたします。

では、私から一つ。先生が対象としておられる「子供」という言葉の中で想定されているのは主に未就学児でしょうか。

○川原発表者

そうですね。はい。

○相澤主査

いや、つまり、小学校以降は教育のカリキュラムという中に入るので、その前の子供たちについてガイドラインというのは本当になるほどなと思ってお伺いしていたんですけれども、具体的に例えばガイドラインを作るというと、どのような形で

出ていくのが望ましいですか。

○川原発表者

いや、それは皆様方の方がお詳しいのではないかと。私が個人的に叫んでもいいんですけれども、何かしら、このような知見に基づいて正式にガイドラインを出せるのであれば、どの媒体でどの機関を代表して言うのかというのは、非常に重要な問題だとは思うのですけれども。

○相澤主査

本当に大事な問題だということを痛感いたしました。ありがとうございます。

ほかは、いかがでしょうか。

○川原発表者

当面は、個人的に叫び続けるということで。

○相澤主査

ありがとうございました。

それでは、本日の発表等も踏まえまして、自由な意見交換の時間とさせていただきます。本日の御発表についてもですけれども、お手元の参考資料に三つの観点、ございますので、そのうちの一つ目と三つ目、言語の果たす役割、その本質、あるいは今後における望ましい言語資源の在り方なども踏まえまして、皆様の御専門のお立場から御自由に御発言を頂ければと思います。川原先生からも是非よろしくお願いいたします。また、皆さん、１回は御発言いただけますようによろしくお願いいたします。

○神永委員

では、私からでよろしいですか。

○相澤主査

はい。お願いいたします。

○神永委員

今日は、とても私の専門外で、私は辞書編集者ですので、全くこのような世界というのは分からなかったので、どの程度理解しているのかなんですけれども、大変興味深いお話を二つお聞きしました。どうもありがとうございました。

やはり辞書編集者でありますので、辞書のことが気になりまして、それで伺いたいのですけれども、まず、相澤主査なんですけれども、最初、語義の曖昧性を判断するときに、辞書を与えていたというようなことをおっしゃっていましたね。その後、文脈情報が反映されることによって、そのようなものではなくなっていったというようなことなのですけれども、現時点では、その辞書を何かするとか、そのようなことはもう一切なくなってきているのでしょうか。

と言いますのは、いわゆる誤用的なものが結構いろいろと存在していて、この間の「国語に関する世論調査」なんかでも出てくるようなのがありますよね。そうすると、私自身もどっちにすべきか、辞書の中でも判断を迷っていまして、完璧に誤用としているものもありますし、それから、これは新しい意味だとして載せているものもあるんですけれども、そういったものを実際にはどのように考えてやっていらっしゃるのか、お聞きできたらなと思ったのですが。

○相澤主査

ありがとうございます。言語モデルを作るという立場からは、正直申し上げて、一切反映されていないのが現状になります。

○神永委員

やっぱり、そのようなことになってきているわけなんですね。

○相澤主査

何かのルールを入れて、それを守るように振る舞うということが非常に難しい世界なので、例えば辞書でこう決めているから、あるいはこの言葉はこのように使うべきだから、それを守ってくださいということ自体が言語モデルに組み込むことが難しい。なので、もしあえてやるとすると、出力のときに間違ったことを言おうと

すると修正するというふうな仕組みを入れていくことになるんですけれども、流暢(ちょう)性も損ないたくないということで、誤用は誤用のままで普通の人が見ると気付かないレベルの日本語というのが、今、それがまかり通っている理由になるかと思います。

○神永委員

そうですか。

○相澤主査

もし気になるようなことが多々目に付くという御指摘が頂ければ、多分、言語モデルというか、ＡＩを構築する立場からも、そのような間違ったことはしないように最後に訓練をするということですね。

○神永委員

なるほど。いや、私たちも、その辞書を作っている出版社、私もそこにいましたし、それから、武田委員もいらっしゃいますけれども、やっぱり辞書を作っていく、そういったものを何らかの形でお手伝いというのか、一緒にできないのかななんていうことを、その一つの会社だけではなくて、辞書、出版社全体の課題として何かできないのかななんていう思いがチラッとありまして。

○相澤主査

それは大変興味深いと思います。

○神永委員

でも、現時点では、そのようなことなわけですよね。

○相澤主査

作り方の時点では。例えば語義を挙げてくださいという質問をすると、実にきれいに挙げてくるんですけれども、それはもしかしたら、辞書そのものを学習しているからできるのかもしれないんですね。

○神永委員

ああ、そのようなこともありますか。

○相澤主査

その辺りの解明もできていないので、辞書をきちっと体系付けて学ばせるというのは大変面白いと今お伺いしていて思いました。

○神永委員

辞書もいろいろと変わるものですから。

○相澤主査

そうですね。

○神永委員

ええ。だから、例えば「国語に関する世論調査」なんかで取り上げた語の新しい意味を誤用としているものもありますし、一方では、新しい意味もどんどん載せている辞書も当然あるわけで、だから、その辺の揺れというのでしょうか、辞書によって違いがあるので。

○相澤主査

そうですね。大変興味深いと思います。

○神永委員

いや、余計なことで失礼しました。

○川原発表者

辞書関連で相澤主査に質問があるんですけれども、よろしいでしょうか。

○相澤主査

はい。

○川原発表者

実はＣｈａｔＧＰＴの有料版の方を試していないのですが、先日、アクセントをどの程度再現できるのかというのを実験して、少なくとも無料版では聞き取りはできているようなんですね。「橋」と「箸」ってどう違うのとか、「雨」と「飴」ってどう違うのというと、意味は分かってそうなのですが、それを読み上げる段階になると、ハシ、ハシ、アメ、アメというふうに言い分けができていないようなんですね。それが有料版でどうなるのか、どうなっているのかということと、今後、改良の余地あるのかということと、そのアクセント辞書なしで、それが可能になるのかということを個人的には知りたいです。

○相澤主査

では、私から少し簡単な、作る方から言うと、本当に素材となる音声のコーパス次第というところはあるかと思います。ただ、使い分けというのがまた、この場面では大阪の方言でしゃべるべきとか、そのような使い分けまで必要になってくると、まあ、でも、できますかね。私たちが思うより能力が高いので、入力に合わせて出力というふうな訓練をすることによって、素材があればある程度は対応できるかもしれません。ただ、先ほどのお話と関係して、アクセント辞典のような正しい使い方ができるかというと、モデル側からは完璧なコントロールができるかどうかは、やってみないと分からないところがあります。辞書の御専門の方からはいかがでしょうか。

○神永委員

例えば最近、ＮＨＫなんかでもニュースで、自動音声で読み取っていますけれども、あれも微妙なんですよね。確かにＮＨＫですから、ちゃんとＮＨＫのアクセント辞典にのっとってやっているはずなんでしょうけれども、何かやっぱりちょっと違うなって。あれは何だろうといつも思いながら聞いているのですが、だから、そのような違和感というのがやがてなくなるものなのかどうかということなんですよね。やっぱりずっとそれは続いて、あのような感じで行ってしまうものなのでしょ

うか。もちろん、アクセント辞典を基にしたからちゃんとアクセントがしっかりと表現できる、正しく表現できるかというと、それはまた違うのかもしれませんけれども。

○相澤主査

最後には多分、身体性のような話になって、例えば濁音はこのような感覚と結び付いているとか、そのようなサウンドシンボリズムと呼ばれるような、トゲトゲした音はこのような、例えばトゲトゲという言葉は、このような音に結び付くという辺りまで人間は踏まえて、音声を認識している可能性はあります。そんなところまで踏み込んでいくと自然さはなかなか、特に長い対話の中では難しいのかなという気はしますね。

○神永委員

そうかもしれませんね。長い対話、はい。

○相澤主査

ただ、本当に予想しているよりも、いつも上を行った性能が出てくるので、この場でそんなにできませんというのはリスクが高いかもしれません。

前川委員、いかがでしょうか。音声の話でございますが。

○前川委員

今の件ですね。まあ、そのうち慣れてくるんじゃないかなという気が。

○川原発表者

慣れですか。

○神永委員

こっち側が。

○川原発表者

人間側が。

○前川委員

いや、人間の言語って、そのようなものだと私は思っております。もう既に大分慣れているのではないかなという気はしますね。それから、まあまあ、不自然さの原因が何かというのは、確かに音声的な問題として、とても興味深い問題だとは思いますけれども、与えられると慣れていくのが人間だと私は思っています。

それで、質問、よろしいですか。

○相澤主査

はい。是非お願いいたします。

○前川委員

相澤主査の御発表で二つ教えていただきたいことがあるんですが、一つは、言語空間のモニタリングという話が最後の方に出てまいりましたね。あれ、非常に重要な話だと思うんですけれども、同時に生成ＡＩが学習目的でどんどん生成したものを使うという、それが一つ御指摘のとおりあるわけですけれども、同時にいわゆる日本語を母語としない人たちの日本語というものも、あるいは量的には、多分、生成ＡＩほど多くはないと思いますけれども、あるわけですね。そういったものもやっぱりモニタリングの対象として、いい悪いということを言うのではなくて、それが今増えつつあって、そして、今の、正に先ほどの話になりますけれども、それに慣れてしまうんですね、我々は。それは結局、日本語の変化につながっていくわけで、私は変化していけばいいという立場なんですけれども、それのモニタリングも必要、できたらやれるといいなと私は個人的に思っております。

それで、ここから質問なのですけれども、生成ＡＩが作り出したもののモニタリングの手段、具体的な、実効的な手法としては、どのようなことになるんでしょう。つまり、何か刻印を押しておくとか、透かしを付けるとか、そのようなことになるんですか。

○相澤主査

本当に難しい御指摘だと感じています。むしろ、人間が作ったというふうに確信を持って言えるものはこれだという方が正しいかもしれないです。何かやっぱり生成ＡＩ、至るところに入り込んでくるので、しかも、自動的に分類するということが、恐らくまずできないので、やはりどっちを取るかですけれども、人間を作る方をチェックした方が早いかなという気もします。

○前川委員

一種の何か法的な規制で、その透かしを入れる。

○相澤主査

ああ、透かし。

○前川委員

考えられないことはないかなという気がするんですけどね。

○相澤主査

はい。

○前川委員

この小委員会なんかでは、そのようなことも考えてもいいのかもしれないなという気はしました。

○相澤主査

はい。ありがとうございます。そうですね。

○前川委員

それが１点。もう１点、別のことでお尋ねしたいのですけれども、御発表の中に著作権侵害の話が出てまいりましたね。どれぐらい記憶しているのかという観点からのお話だったんですけれども、あれがなぜ、どれだけ記憶しているかがなぜ著作権侵害に結び付くのかが、いまいち理解できなかったんですけれども。

○相澤主査

分かりました。

○前川委員

あれは、要するに全部記憶しちゃっているから、そのことが問題だという、そのようなことなんですか。

○相澤主査

いえ、記憶したことを出力してきて、そのときに丸ごと再現しているにもかかわらず、リファレンスは付かない可能性がゼロではないということに対して、恐らく危機感を持たれる著作権者の方がいらっしゃるということだと。

○前川委員

つまり、ある作品、適切なプロンプトを与えたら、ある本か何かならすぐ出ちゃうし、それから、短い小説ならほとんどそのまま出てしまう。

○相澤主査

丸ごとはないとは思いますが、例えば出たとして、はい。

○前川委員

それが問題だということなんですね。

○相澤主査

はい。そうです。それにリファレンスが付けばいいけれども、大量のコーパスから学習して、それが言語モデル自身も分かっていないと無断で出してきて、それは誰かが作ったものだというものが受け取った側にも分からないという状況になってしまうという。

○前川委員

それはよく分かりました。ありがとうございました。

○相澤主査

ありがとうございます。

では、ほかは、いかがでございましょうか。

○石川副主査

では、よろしいでしょうか。

○相澤主査

はい。

○石川副主査

コメントですが、コーパスを作る側として、生成ＡＩの産出する言語データに対して、複雑な気持ちを持っているということをお伝えしたいと思います。コーパス構築に関して言うと、やはり自然言語の総体からランダムサンプリングして集めるという基本方針がございます。その際、書き手や話し手の素性で価値付けを行わず、言語として実在しているものは全て等し並みに扱うことになります。これがコーパス言語学者の言語観の基本になってきました。

この方針に従うと、生成ＡＩが作った言葉だから駄目だというのは、かなり言いにくいのではないかと感じます。一方で、リアルなデータは集めなくてよい、生成ＡＩの産出したデータで十分だと言ってしまうと、自分たちがやってきたことの自己否定にもつながりかねないわけで、ここに私自身、葛藤を覚えます。川原先生のアンケートに答えた際にも、非常に深く考え込んでしまいました。

この話を突き詰めていくと、もうコーパスなど集めなくてよいという話になってしまうかもしれません。例えば、今あるBCCWJを機械学習させ、「これと同じようなもので2020年代らしいデータを作れ」と言えば、BCCWJの2020年版が即時に完成するのだとすれば、コーパス開発の意味は根底から失われます。この点についても、自分の中で考えていかないといけない問題であると、本日、お二人の先生のお話から感じたということでございます。どちらも大変良いお話で、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。サンプリングという意味では、均衡コーパスの均衡なんですけれども、ＡＩが生成したものを例えば入れていくときの均衡のバランスはどうなるかというのは、とても難しいなと思いながらお伺いしました。

○石川副主査

そうですね。「この比率で」と言うと作ってくれそうではありますけれども。

○相澤主査

そこが非常に増幅されている。１対１でパーソナライズされた空間でＡＩと人間が話しているということが世界中で起きるとすると、コーパス、そうですね。その辺りの量的な問題も関係してくるような気がします。

○石川副主査

そうですね。ありがとうございます。

○相澤主査

ありがとうございます。

ほかは、いかがでございましょうか。

○前川委員

もう１点、よろしいですか。

○相澤主査

前川委員、はい。

○前川委員

川原先生の御発表に一つ。これはあえて意地の悪い観点から質問させてもらいた

いんですけれども、今日の御発表の中で、最初の出発点になる問題意識として、臨床試験を経ていないのに使っていいのかというような表現を取っておられたと思うんだけれども、それを聞きながら、私、思ったのは、例えばテレビというものがあるじゃないですか。テレビというのは、子供の言語発達に非常に影響することが知られているし、それから、例えば日本語のアクセント変化なんかにもものすごく直接的な影響を及ぼしたものなんですよね。それは実証されているわけなんですけれども、あれも別に臨床試験を経ていないので、それで、テレビに子育てにさせるという話は昔からよくあって、正に子供、ずっと、テレビを見ていてくれるから、その間、親が家事に集中できるみたいな話は昔からあるわけだけれども、その意味では、もちろん与えられる言語の性質そのものは違う。テレビで流れているのは人間が生成しているものだから、そこが違うわけだけれども、品質保証のないものをずっと使っているという点では、それは同じではないかなという気がして、あえて少し質問するんですけれども、そこら辺は何かお考え、ありますか。

○川原発表者

ります。まとめられるか分からないですけれども、一つは、テレビは許したんだから、スマホも生成ＡＩも大丈夫だろうというのは、あいつも万引きしたから俺も万引きするみたいな、あれも駄目だったから、こっちも、この駄目なものも許そう的な発想なのかなと私は、私と前川先生の関係性をもって、前提にお答えしたいのと、もう一つは、中毒性ですね。怖いのは。テレビ中毒もいるんでしょうけれども、やはりずっとスマホに付きっ切りになっている子、離れられない子、今の大学生で、授業中にスマホをカバンの中にしまえない子などなどを見ていると、テレビにはないポータビリティーというんですか、身体、自分の一部になってしまっているような感覚というのは、テレビとスマホの大きな違いではないかなと個人的には感じております。

○前川委員

最初のあいつが万引きしたんだからという話で言うと、テレビで何となく今までうまくやってきたから、こっちも大丈夫だろうという論理も同じ論理という気がしますよね。

○川原発表者

その点に関しては、今度、ゆっくりお話を。ありがとうございます。

○前川委員

まあまあ、どっちの論理も成り立つものですからね。

○相澤主査

どうもありがとうございます。

では、植木委員、武田委員から、あいうえお順となっていますが、順番に一言ずつ頂ければと思います。

植木委員、よろしくお願いします。

○植木委員

大変貴重なお話、ありがとうございました。とても勉強になりました。質問とコメントを数点お伝えします。相澤主査のお話について、本当に瑣末なことなんですが、語義の曖昧性を排除していくというお話で、例えば「マウス」と言ったときに動物のネズミなのか、パソコン周辺の機器のマウスなのかというところで、曖昧性を排除していくという方向が重要なのはもちろんなのですが、一方で、私は古典を勉強しているので、掛詞のような多義的な言葉に親しみがあります。例えば小町の有名な歌の「花の色は移りにけりないたづらにわが身よにふるながめせしまに」の「ふる」は、時間が経つという意味のフルと雨が降るのフル、「ながめ」も物思いという意味と長雨という意味があって、もちろん中心の意味はあるのですけれども、そのような二重の意味を持っているという場合にどのように扱っていけばいいのか、どのように扱っておられるのかというのが質問です。

続けて申し上げてしまいますと、川原先生のお話についてはコメントなのですけれども、非常に感銘を受け、私も共感するところが大きかったです。特に懸念点の４点目ですか、背後に感情が存在しないというところで、子供が生成ＡＩの背後に人格も感情も存在しないということを理解できるかという懸念は、これは子供だけ

ではなくて大人にもあって、実は既に非常に危険な状態ではないのかなと感じております。2014年に公開された「Ｈｅｒ／世界でひとつの彼女」という映画を御存じでしょうか。生成ＡＩに恋をしてしまう男性が主人公なのですけれども、その生成ＡＩの声をスカーレット・ヨハンソンがしたというので話題になりました。映画では妻をなくした男性が寂しさを紛らわすためにチャットをするのですけれども、その相手がサマンサと名付けたＯＳで、サマンサに恋してのめり込んでいってしまうという話なんです。10年前の映画ですけれども、現在を先取りしたような作品で、最近、2023年でしたか、ベルギーでイライザというＡＩとチャットをし続けた男性がＡＩにのめり込んで、自殺してしまったというニュースがありました。依存性が高くて、冷静な判断ができなくなっていくような、危険性は大人にもあることなので、ましてや発達段階にある子供たちについて、もちろん心配し過ぎるのもよくないとは思いますし、適切な使い方というのはあるとは思うのですけれども、やはり非常に注意しなければいけないことなのではないかと思いました。したがって、そのガイドライン―川原先生がおっしゃったようなガイドライン、そして啓発活動というのは非常に重要だろうと考えます。

それからもう一つ、川原先生がどのようなコーパスが欲しいかというところで、養育者から子供への語り掛けについてお話ししてくださっているのですけれども、そこで前回の、石川副主査のレクチャーを思い出し、興味深く思いました。レクチャーの中で、コーパスの社会実装、社会にどのように貢献できるかという観点から、医療従事者の方々の話し言葉についても例を挙げてくださって、私は大変感銘を受けました。医師とか看護師が患者にどのような言葉掛けをしているか、当事者に経験的、実感的に理解されている治療に効果的な話し方とか、言葉の選択というのがコーパスの構築によって客観的に多くの人に示すことが可能になるというお話でした。川原先生の養育者から子供への語り掛けというのも、そのような教育の実践というか、社会の貢献というところでは非常にいいものになるのではないかなという感じがいたしました。それは一般の子供もそうですし、障害者差別解消法が改正されて、いろいろな合理的配慮が必要になる中で、我々も学ぶことが非常に大きいのではないかと考え、前回の石川副主査のお話と関わるところもあると思いながらお伺いしました。

まとまらないコメントになってしまいましたが、お話ありがとうございました。

○相澤主査

大変ありがとうございました。

では、最初の御質問について。例えば駄じゃれを作れるかといったこともあると思うのですけれども、今までそんなに研究としてはやられてこなかったというのは、評価がすごく難しいので、評価することができれば、もう少し進むかなと思いながらお伺いしていました。

コメントはございますでしょうか。

○川原発表者

少し補足になってしまうのですけれども、医療従事者の話を聞いて、あっと思ったのは、子育て、語り掛けコーパスで便利だなと思ったのは、お父さんで苦手な方、少なくないんですよね。私は抵抗感なくできるんですけれども、「○○ちゃーん」というあれですね。あれは、お母さんは自然にやっているんですけれども、お父さんは恥ずかしがって、できないという方が多いので、そのモデルになるためにも何かしらあると、本当に今の令和の世の中には非常に役に立つものではないかなと思いました。

○相澤主査

ありがとうございます。

では、武田委員、よろしくお願いいたします。

○武田委員

相澤主査の大規模言語モデルに関するお話、規模も非常に大きなお話になりまして、コーパスの１億ですとか、そのような数字から桁も相当違いまして、また違う大きな世界があるなということで、大変参考になりました。

また、生成ＡＩに関する川原先生の御説明も非常に興味深い観点を示していただいたかと考えております。

人間、人類における言語の歴史を踏まえたときに、このような形で生成ＡＩなどに代表されるような言語環境の変化が起きたときに、どんどんと言語そのものが変わっていく状況に今いるわけですので、それに対してどう対応していくかというこ

とは非常に示唆に富んだお話であったかと受け止めております。

ガイドラインのお話も、別に子供に限らず、全ての世代が影響を受けているわけですので、アクセントの問題ですとか、気になるところは非常にございます。少し前の20年、30年前の映画では、例えばバグというふうに、プログラム上の問題のことをバグというふうに発音していますが、今は、当然、バグという平板な発音になってしまっている、アクセントになってしまっているというようなこともあったりとか、どんどんと変わっていくところが見えておりますので、その意味でもこのＡＩ、どのように受け止めていくのかということが重要になるのかなと思っております。

簡単ですけれども、以上です。

○相澤主査

どうもありがとうございました。

では、お時間となりましたので、議論の時間、こちらで終了いたしまして、本日の言語資源小委員会の議題２まで終了しました。議題３は、その他ということですけれども、次回の言語小委員会よりも前に国語分科会を開催するということで、日程の御調整をいただいているということです。国語分科会におきましては、この言語資源小委員会の審議について報告をすることとなります。つきましては、国語分科会での報告について、御意見を内容ごとに分類、整理いたしまして、資料を基に報告することで主査である私に一任していただくことでよろしいでしょうか。

（→「異議なし」の声あり、了承）

ありがとうございます。では、そのように進めさせていただきますので、どうぞよろしくお願いいたします。

本小委員会につきましては、引き続き議論の基盤を共有するために事務局とも相談しまして、日本語の言語資源を考える上で有益と思われる事柄について、委員も含めた有識者からお話を伺うということで検討しております。

もし何か言い残したこと等ございましたら、この場で御発言いただければと思いますが、よろしいでしょうか。（→　挙手なし。）

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。