

言語資源の在り方に関する主な意見

※以下は、第1回会議における委員の意見を観点別に整理・統合したものである。

(言語資源について)

- コーパスは、ある時点の言語を記録した「言語遺産」のようなものだということであり、どのような形で言語を保存していくのが望ましいのか、これまでに保存されたものをどう保持し、次の世代に受け渡していくのがよいのかなど議論しながら、よい方向を見つけていくことができればいい。
- コーパスが必要とされるのは、潜在的文化財としての価値があるからである。言語データは、時間の経過とともに消失する。逆に、文化財的な価値は時間とともに増していくということである。

(日本語コーパスの整備・発信について)

- 過去四半世紀で日本語コーパスの整備は顕著に進み、言語研究に占めるコーパスの位置は確立されたが、いまだ道半ばである。とりわけ、将来の日本語に強い影響を及ぼすと考えられるCMC関係や日本社会における多言語関係部分である。この点、整備に際しては、超領域的な熟議が必要であろう。
- 日本語コーパスの整備上の課題は、著作権に係る権利処理と個人情報への対処方策をいかに講じるかという点である。
- BCCWJは、非常に綿密に作られているが、そこに入っていないジャンルというものがたくさんある。そうした新しいジャンルについても、継続的に時間軸の中でデータを集めていくことが必要になろう。

- コーパスというのは、言語学者や自然言語処理の研究者にとってはなじみの深いものであるが、一般の方から見て、それほど身近な状態にはなっていないので、このギャップを埋めるため、コーパスの魅力を広く発信していく機会を講じたらよい。

（生成AI等これから日本語へ影響を与えていくものについて）

- 生成AIは、コーパスの中に埋め込まれた知識、規範、倫理といったものを獲得していく過程がある一方で、生成モデルの中にあるバイアスといったものが、逆に人間に影響を及ぼしていくという双方向的な関係にある。
- 日本語がこれから変化していく要因は、社会の変化にあるわけで、サイバー化、高齢化、多文化・多言語化、そしてAIのインパクトといったようなものが挙げられる。大事なことは、こうした変化への対処方策を講じる際に科学的なデータに基づいて対応できればそれにこしたことはない。

（今後における望ましい言語資源の在り方について）

- コーパスは、一義的には言語学の研究、共時的あるいは通時的な言語の研究に大変有用な研究資料であるが、同時に、より大きな社会変革、社会のイノベーションにもしっかりと貢献していける素地を持つものであるというコーパスの可能性についても見ていく必要がある。
- コーパスというのは、大きな日本語の基盤であって、その上に様々なアノテーションを載せていくことによって社会実装がされていく。それは大規模言語モデルが汎用モデルであって、その一つのモデルがいろんなタスクをこなしていくのとちょうど同じイメージで、網羅性がある、多様性がある日本語をカバーすることによって、初めていろいろな社会実証が開けていくと見られる。
- 長期的に見て日本の社会に裨益^ひするような新しい国民財産をこれから作っていく、こうした出発点をいい方向に定めていきたい。

- コーパスを作る人は「作る」ことにのみエネルギーを注ぎがちである。作ったものを広く使っていただくためには、検索のためのインターフェースの在り方なども考えていく必要がある。
- 良い言語資源を作っていくということ、作ったものは皆さんに使っていただけるようにしていくということ、さらには、作り放しにしないこと、アノテーションのこともそうだが、いろいろ人が関わって、集めたデータの質を磨いていく。

以上

言語資源の在り方に関する委員の意見（第1回分）（案）

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)」の
＜検討の観点について＞に基づく意見の分類である。

◆「日本語のデジタル言語資源の整備に関する国語分科会の見解」（文化審議会
国語分科会 令和5年9月29日）に対して

- 「4」の最後のところで、各国語のコーパスのサイズがいろいろ紹介されていて、BCCWJが相対的に見て小さい、その事実自体はそのとおりであるが、例えばドイツのDeReKoを取り上げて、550億語としている。これ、実は大部分がウィキペディアのデータで、ある時期に突然数百億語増えたというような感じである。チェコ語で10億語となっているが、実は現在600億語ぐらいになっていて、これも中身はほとんどウィキペディアを中心としたウェブのデータである。ちょっと比べる内容と時期がばらばらになっている。
- 「4」の頭のところで、BCCWJの開発が止まった原因の一つとして、大学共同利用機関になった時に、よりアカデミックな方向を目指したからということが書いてある。もちろんそういう要因もあるが、一番大きな原因は、もともとBCCWJというのは外部資金で作っているため、その後、大きな予算が取れなかったことが一つの大きな原因であろう。
- 前年度に作成した「見解」のペーパーを、土台にしてブラッシュアップを図っていく。

◆言語の果たす機能やその本質をどのように考えるのか、その上で、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。

言語、言葉というのは、人間が思考し、他者と意思疎通を図る上で不可欠なものであり、アイデンティティーに関わり、生きる力の土台となるものである。そして、それは歴史と伝統の中で培われてきたと言える。したがって、これまでの歴史の中で用いられてきた言語の在り方を振り返ることで、言語の変遷であったり、あるいは往時の諸相であったりといったものを捕捉することが可能となる。こうしたことから、言語資源を保存し、通覧・分析可能な状態にしておくことは、言語研究や言語政策、言語教育の分野はもちろん、社会全体としても意義深いものである。

このような意義を踏まえると、言語資源を「未来へ伝える文化遺産」（仮称）と位置付けた上で、言語資源の果たす役割と今後に向けた在り方を検討していくということがあるのではないか。

多様な言語コーパスを言語資源として構築して活用していく土壌を整えていく。

【定義】

- ウィキペディアに「言語資源」というページがあるが、一つの問題は、ここでは言語資源というものが自然言語処理の技術のためのリソースだと位置付けられているということ。元々そうではない。この小委員会、分科会としてもそれだけを考えるわけではない。
- ウィキペディアに「言語資源」というページがあるが、作成に非常に手間が掛かるために、その数がいまだに少なくとても高価であるということが書かれている。高価かどうかよりも、現状における問題は、一部の特権的な組織とか、会社とか、そういうところしか特にウェブ上のデータを大量に備蓄することができないこと、つまりアクセシビリティに極端な差があるということ。そういう意味でも、公的な組織の関与というものが何かの形で期待される。
- ウィキペディアに「言語資源」というページがあるが、ウェブから取ってきたデータに関して、一般的にノイズが多くて価値のある情報が…ということが書いてある。これもそのとおりであるが、現実には非常に多くのものがウェブのデータを基に開発されている。

それに対して、BCCWJは、大規模言語モデルが作り出すものに対するフィルターBCCWJを一つのかがみとして、ある程度ちゃんとしたものができているかどうかというものを評価するという形で使われている。

- コーパスというのは何かと言うと、実際の言語活動を体系的に記録して、電子的に検索可能な状態にして公開した言語資料である。
- 言語コーパスとは具体的にどういう要件があるかということをまとめると、七つポイントが挙げられる。真正性(Authenticity－実際に生じた言語活動の記録であること)、均衡性(バランス－一つの言語にいろんな側面があり、様々なレジスターをできるだけ広くカバーしているということ)、代表性(－母集団が決められるのであれば、その母集団をできるだけ正確な縮図として取り出してきているかどうかということ)、規模(－どういう目的に使うかが決まったときに、その目的のために十分な量が確保されているということ)、電子化(－単なる文字列検索しかできないのではなく、もうちょっと高度な検索ができるようにいろいろな付加価値を付けること)、公開(－誰でも利用したい人が利用できるということ)。これを全部満たすコーパスは実はそう多くはない。もうちょっと条件を緩めると、真正性、規模、電子化、公開ぐらいが満たされていれば、コーパスという最低限の条件ではないか。
- アーカイブズというのは、典型的に言うところ国会会議録のようなもので、はっきり限定された対象を全て保管していくというものである。余り付加情報は付けないで、メタデータ、誰がどこで言ったぐらいのことは付けるが、それ以上の検索用の情報は付けないことが多い。
- 言語データベースというのは、必ずしも真正性を前提としないコーパス的なものや、存在するコーパスを加工して2次的に派生されたデータといったものである。

【言語コーパスと言語資源】

- グーグルのNgram Viewerを使って、Language resources、Language corpusとcorporaの時代の変化を見ると、90年代の半ばぐらいから急激に伸びてくる。

つまりここ数十年の間に急激に普及した概念であり、世の中でそれだけ重要性が増えているものである。この線が相関していることから、言語コーパスを論じることによって、ある程度言語資源全般を論じることにも通じることを示している。

【言語コーパスの必要性】

- 何のためにコーパスが必要とされるのか。最初の理屈として言語研究がある。言語というのは非常に多様なものであって、その全体を我々が考えるだけで一頭の中で内省するだけで捉えることはほとんど不可能だということが出発点としてある。全部、本当に間違いなんですかという、そういう話である。それを研究する必要がある。
- 2番目は、「レジスター」という言葉を何遍か使っている。言語というのは、それはどういう状況で使われるかで初めて使われる形が決まってくる。状況とともに言語を把握するというような意味でもコーパスは必要になる。
- 3番目は、言語というのは変化していった、なおかつ人間が意図的に新しく創造する、クリエイティビティーを発揮することもできるということである。変化を捉えるためには、やはりコーパス的なものが必要になってくるということである。
- 最後に「Phraseology」－フレーズのレベルがあって、そこが非常に重要である。チャンクとかフレーズとか単語のある程度の固まりで文末満のようなものを重視しようという見方である。
- 何のためにコーパスが必要とされるのか。次に、情報科学の世界である。昔から人工知能の研究は行われてきているわけで、1950年ぐらいから始まって、80年代の半ばぐらいまでは大体、いわゆる記号ベースの研究が中心であった。80年代の終わりぐらいから、いわゆる統計ベース(学習ベース)の研究が主流になってきて、その学習用のデータとしてコーパスというものが非常に注目を浴びるようになってきた。
- 何のためにコーパスが必要とされるのか。潜在的文化財としての価値ということである。言語データは放っておくとどんどん消えていく。時間とともに文化財的な価値がどんどん増していく。日本語の研究に限らず、文化財的な価値は時間とともにどんどん増していくということである。

- 何のためにコーパスが必要とされるのか。最後に、科学的な言語政策の礎としてのデータとしてのコーパスも考えられる。
- 辞書と言っても、規範性ばかりを追求するものではないので、現状を認識しながら現状をきちんと捉えていく一助として、コーパスの必要性は非常にある。
- 規範と外れるようなデータを示すようなとき、辞書に書かれていること、文法書に書かれていることが正しいんだと思っているような方々にとってはちょっと拒否反応みたいなこともあるのかと思うので、その辺りもしっかり説明していくような必要もある。
- コーパスの整備は、日本の言語政策、文化政策にとって基盤的な部分に関わることだと思う。
- 昔はできなかったことが今、情報処理技術の発達によって、許容されるコストで実行できるようになってきた。

- ◆日本語のコーパスに関する整備状況を確認し、その活用の在り方について、我が国の社会経済情勢や教育研究活動等を踏まえ、どのように考えるか。
- ・ 言語研究や言語政策、言語教育、そして大規模言語モデルとの関係での整備・活用の在り方

【日本語のコーパスの現状】

- 国立国語研究所では、細い線であるが、奈良から現代まで言葉の歴史的な変化をたどることもできるようになり、近代における進行中の変化を見ることもできるようになり、書き言葉と話し言葉の両方でいろいろなレジスターを把握するというのが、十分ではないが、できるようになってきた。これが二十数年間の成果かと考えている。
- 国立国語研究所以外にも非常にたくさんのコーパスがあり、国外で開発されたものもある。たくさんのコーパスは、大きな配布機関として四つあり、ここで入手できる。

【活用】

- 言語資源がどれぐらい使われているか。国立国語研究所が公開している「中納言」というシステムのユーザー数と利用件数を見ると、時間とともに非常に単調にどんどん伸びている状態である。ユーザー数にして、現在だと5万人を超えていて、年間検索数で250万件を超えて、もう300万に迫るような勢いで伸びている。
- 日本語話し言葉コーパスであるが、論文が英語で書かれているか日本語で書かれているかの違いを見ると、1期は日本語の方が多いが、2期になると英語が圧倒的に多くなる。それだけ世界的に今は利用されているということを表している。
- 実際に辞書の編集において、コーパスを利用する機会は非常にある。頻度、傾向、コロケーション、言葉の結び付き、そういったところを確認しながら編纂している。
- 歴史コーパスを使っているが、古典の場合は書写されて残ってきたという点で、一定の基準をクリアしていると言うか、そこまで言葉の乱れというものや庶民的な俗語がない、

ある程度洗練されて淘汰されたものが残っているので、取る範囲というところに余り迷うことがない。それが現代のものと大分違う点である。

【課題】

- 外字の問題は、検索のしやすさと、言語資料としての価値との間にトレードの関係がどうしても生じてしまう。検索のためには代表的なもので置き換えておいて、そこに元々の字はこれであったというようなことを何かのメタデータ、埋め込みデータとして残しておくというような方策が現実的ではないか。
- 国立国語研究所には分類語彙表というとてもすばらしいものがある。アノテーションに関わるようなことであるが、辞書編集者としては、分類語彙表の情報が付いていると、それを活用しながら検索ができるので非常に有り難い。
- 基本の分類語彙表といろいろな形態素解析したものとの対応が取れるようにはなりつつある。
- 分類語彙表自体はまだまだ語彙数が足りない。これからも増やしていただきたい。
- どこまでコーパスに含めるのかという問いに対する答えは一通りではなく、いろいろあり得る。BCCWJの場合、「サンプルに引っ掛かったものはみんな入れる。それは、日本語全体のゆがみのないデータを作ることが目的だから、恣意的に省いてしまったら、そこが台無しになってしまう」という判断をした。一方で、リファレンスコーパスを作るのであれば、外的な基準を導入する必要があるということは間違いない。
- 書籍とか雑誌とか新聞、有償で世の中に出てきたものとそれ以外のものというのは、レジスターの性格がすごく違うかもしれない。目的によるが、これまでのコーパスというのは、基本的には有償のものを中心に多分作られてきていた。それ以外の日常会話は、余り規範性にとらわれない日常に使われている言葉で、本当は言語学の対象であるというのが結構昔からの言語研究のテーゼになっているが、実際は必ずしもそういうものを分析していなかった。

- 絵文字は、日本語コミュニケーションのある特徴、一部でもあるので、コーパスで扱う価値がある。
- 様々なコーパスが開発されているが、大規模なものは余りなくて、小規模なものが多い。継続的な開発が行われていない。
- 言葉自体がどんどん変わっていく。辞書の編纂も、改訂のペースは思ったより早く、新語や新語義、新しいそういった情報に対応していくということがあるので、コーパスも常時改訂をするというような体制を取っていただければ辞書編纂にも役立つ。
- 書き言葉の場合、ウェブを使ったものか、あるいはBCCWJを加工したものが多くて、それ以外のものが意外と少ない。デジタル・ヒューマニティーズでよく言われるFAIR原則(Findableで、Accessibleで、Interoperableで、Reusableであること)のうちのインターオペラビリティ(相互利用性)ということが余り確立されていない。(仕様が必ずしも標準化されてないことが多い。)
- 歴史コーパスの場合、現在はまだ古いところは文学作品が中心になっている。
- 足りないレジスターというものが一杯ある。
- 過去四半世紀で日本語コーパスの整備は顕著に進み、言語研究に占めるコーパスの位置は確立された。しかし、いまだ道半ばであり、今後データ整備が望まれる多くのレジスターが残されている。とりわけ、将来の日本語に強い影響を及ぼすと考えられるCMC関係や日本社会の多言語に関する部分が弱いと考えられる。一方で、情報処理領域での需要も著しく増大していて、国費を組織的に言語データの整備に投入するのであれば、整備の順序について超領域的な熟議が必要であろう。
- 権利処理がいまだに大きな問題であり、まだまだ整備されてない。
- 現代語のコーパスは著作権から逃れられない。30条の4というので大分楽になったが、それを具体的にどう解釈するかということではまだ見解が分かれるところがある。事例の積み上げでもよいので、ここまではオーケーだというようなガイドラインである程度は

つきりするようになってくればやりやすくなる。

- 著作権は著者の権利を守っているものなので、著作物に価値がある場合に特に問題になるのかと思うが、一方でウェブ情報を扱ったりするときに、最近は個人情報も非常に問題になってきて、法律という意味では、著作権に加えて、個人情報の方も大きな障害として考えていかねばならなくなっている。
- 権利関係のことを中心に、かなり生臭い話もどうしても出てくる。生臭いものの代表は権利関係と、規範性の話の二つだろうと思う。そこら辺をこの場で話し合っ、いいアイデアをまとめられればと思う。
- 近年の生成AIというものが出来た後の自然言語処理の研究分野における変化は非常な勢いであるが、研究分野の変化は、大きな社会全体の流れの中で言えばごくごく一部にとどまっ、より大きく社会の人間の生活の至るところで生成AIというものが人間の言語活動に影響を及ぼしていると感じている。
- 生成AIは、コーパスの中に埋め込まれた知識、規範、倫理といったものを獲得していく過程がある一方で、生成モデルの中にあるバイアスといったものが、逆に人間に影響を及ぼしていくという双方向的な関係にある。
- いかに言語が変化していくかということを観察、捉えていくことが今後ますます重要になってくるであろう。
- 最近、生成AIの中で、「文書AI」という言葉があっ、文書のレイアウトだとか、例えばどの部分が太字になっていて、どの部分が箇条書になっていてということで、人間は目で文書のレイアウトそのものを見て理解するようなことをAIもできるようになっていくという流れがある。
- BCCWJは、段落の構成だとか。レイアウトを完全に再現できるような情報は提供できないが、部分的に再現するための情報はかなり苦勞して付けた。しかし、実際の利用例を見ていると、それを活用した研究はほとんどない。ある程度自動化できるのであれば、是非残しておきたいデータではある。
- これから日本語が大きな変化期に入っていく。その際に、どういう言語問題が生じて、有効な言語政策を考えるためにどんなコーパスが必要になるか。

日本語がこれから変化していく要因として、大きなものとしてインターネットの影響、AIの影響、そして日本社会の多言語化といったものが考えられる。例えば、「どの国籍の女の子を育ちたい？」とネットの広告で出てくる。この日本語は何だろう、おかしいわけだが、どうしてこういったものがパブリックな面で出てくるのかという、その原因が面白い。こういったものが世の中に結構たくさんこれから出てくると、単純接触効果で、触れていると自然に思えてくる。それで言語が変わっていく。こういう影響は決して無視できない。

- これから日本語は非常に大きく変化していくんじゃないかと思う。その原因は社会の変化にあるわけで、社会の変化をもたらす幾つかの要因というものは、サイバー化、高齢化、多文化・多言語化、そしてAIのインパクトといったようなものが挙げられる。大事なのは、そういった変化が起きたときに、それに対する何かの施策を行うときに、科学的なデータに基づいてできればそれに越したことはない。
- 日本語は今、非常に大きな変化期に入る時期にある。その過程は、言語を研究している人間としては、是非緻密に研究したいことであるし、一方、言語の変化は必ず社会の変化と同期しているわけで、そこに生じるであろう多くの言語問題に対して、我々が何かの形で貢献できればそれに越したことはない。それを科学的に研究するために必要なデータとしてやっぱり足りないから、いろんなデータを作る必要がある。

◆日本語に関する既存の言語資源の整備状況を踏まえ、今後における望ましい言語資源の在り方について、どのように考えるか。

- ・言語資源の果たす役割を踏まえ、社会課題の解決等に資するような専門的・分野特化型コーパス等の在り方（例：教育的ニーズ、文化的ニーズ、大規模言語モデル開発からのニーズ等に即した在り方）

【方向性】

- コーパスというのは、言語学者や自然言語処理の研究者にとってはなじみの深いものであるが、一般の方から見て、それほど身近な状態にはなっていないので、このギャップを埋めるため、コーパスの魅力を広く発信していく機会にできたらよろしいのではないか。
- 国の予算を使って何か事業を行うというとき、今回ではコーパスを作るとか整備する、発展させていくというようなときに、研究者のためになるというだけではなく、使いやすさとか、一般の方に魅力が伝わるような取組をしていくのがとても大事である。
- 良い言語資源を作っていくということ、作ったものは皆さんに使っていただけるようにしていくということ、さらには、作り放しにしないこと、アノテーションのこともそうだが、いろいろ人が関わって、集めたデータの質を磨いていく。
- コーパスは、一義的には言語学の研究、共時的あるいは通時的な言語の研究に大変有用な研究資料であるが、同時に、より大きな社会変革、社会のイノベーションにもしっかり貢献していける素地を持つものであるというコーパスの可能性についても見ていく必要がある。
- コーパスというのは、大きな日本語の基盤であって、その上に様々なアノテーションを載せていくことによって社会実装がされていく。それは大規模言語モデルが汎用モデルであって、その一つのモデルがいろんなタスクをこなしていくのと同じイメージで、網羅性がある日本語をカバーすることによって、初めていろいろな社会実証が開けていくと見られる。

- コーパスは、ある時点の言語を記録した「言語遺産」のようなものだということであり、どのような形で言語を保存していくのが望ましいのか、これまでに保存されたものをどう保持し、次の世代に受け渡していくのがよいのかなど議論しながら、よい方向を見付けていくことができればいい。
- 長期的に見て日本の社会に裨益^ひするようなそういう知財、新しい文化財、社会にすぐに役立つことも含めた文化財みたいな、そういった新しい国民財産をこれから作っていく、その出発点をいい方向に定められればいいなと思う。
- 定期的にデータを取り続けていくことの重要性。BCCWJのVersion 2というものが計画されているということで、これは非常に有り難いことではあるが、逆にVersion 2が出て、そこでまた打ち止めになってしまっては困る、
- BCCWJは、ある一時期のデータで止まっていたが、新たにまた構築される。こうした継続性ということが非常に重要である。
- これからずっと続いていく時間軸の中で、将来にわたって定期的にデータを取り続けていく、また、場合によっては、前だけでなく後ろにも遡り、同じサンプリングフレームで過去のデータも集めていく取組が非常に望ましい。
- BCCWJは、非常に綿密に作られているが、そこに入っていないジャンルというものがたくさんある。そうした新しいジャンルについても、継続的に時間軸の中でデータを集めていくことが必要になろう。
- 継続性という点で、10年というのは十分長期ではあるけれども、それを超えて、20年、30年とモニタリングしていくようなファクターも重要である。
- コーパスを作る人は「作る」ことにのみエネルギーを注ぎがちである。作ったものを広く使っていただくためには、検索のためのインターフェースの在り方なども考えていく必要がある。
- 日本語のコーパスの研究成果が、日本語教育であったり国語教育であったり、あるいは、社会の様々な層にどのぐらい具体的に影響を及ぼしてきたのか。本当はもっとあっ

ていいはずのものがまだそこまで至っていない。社会実装を考えていくことも重要な課題である。

- いろいろなことを可能にしていくためには、継続的な財政支援と、それを継続的にサポートしていく国の仕組みが必要である。その都度その都度の競争的資金だけでやっていくということでは、長いスパンでものを考えられない。そうした体制が整備されるように声を上げていくことが必要ではないか。
- どんなコーパスが必要ですか、今後必要になりますかという問いに対する意見はみんなほとんど一致している。社会の多言語化によって、日本語以外の方が話している日本語のコーパスはどうしても必要であろう。それから、AIの影響なんかを評価できるようなデータもやっぱり必要だろう。
- アニメなんかでネット上で公開されているものに違法なものもあるが、ファンの人が字幕を付けて、それで、ネット上に公開してしまうみたいなものもある。その中で日本語を勉強したい、元のアニメのこと、作品をよく知りたいという動機が強いんだと思う。
- 今後構築が望まれるコーパスに挙げた、漫画、アニメ、ゲーム—そういうもので日本語に興味を持ったという外国人が圧倒的に多く、人によるが、日本語教師や日本語の学校なしで、字幕の流れる中で日本語を聞いて、それで覚えてしまったという人もいる。翻訳されたものに関してまで入れてしまえば、もう圧倒的な日本の今、最も影響力のあるメディアであろう。外国に一番流れている日本語はここであろう。
- 多文化、多言語の日本の社会の中において、誰もが同じ言語を知っているという社会じゃなくなってくる。そういう多様性をはらんだ社会における、より許容性に富んだ、柔軟性に富んだ新しい言語観が日本の社会の中で生まれるような、そういう方向に長期的なインパクトとしては持っていきたい。
- 在日の外国人の方の言語の問題がある。今日本でどれだけの外国語がしゃべられているか、実はデータが分からない。かなり多くの国で行われている、国勢調査には「あなたはどの言語をしゃべりますか」、「家庭では何をしゃべりますか」というような問い

を、日本でも加える方法が一番いい。

- 外国の方が日本語で日本の情報サービスを利用するような場合というのは、当然今後考えられる。音声認識が典型だろうと思うが、音声のデータベースがあって、それと辞書との掛け算で確率を計算して、確率が一番高いものを認識するというやり方なので、外国の方が話された日本語の音声データがある程度ないと利用できなくなってしまう。
- 音声認識にしても、機械翻訳にしても、ある程度データがカバーしてないと対応しない。どこが抜けているかを調べることはできるはずで、どういう部分に対して、例えば生成AIの反応が悪いのか、音声認識であれば、ある種の単語が欠けているか、音声の面でのバラエティーが欠けているか、あるいは発話速度に対する対応が欠けているか調べれば見付かる。
- プレーンランゲージ(「やさしい日本語」的なもの)は、日本語は必ずしも十分でない方に向けて、どういう日本語－行政なんかがどのように言葉を使えばいいかというようなことで、それに関しては世の中にいろいろな試みがある。そういったものを集積していて、望ましくは、その中の本当に分かりやすいものと必ずしも分かりやすくないようなものに対する評価値を付けるようなことをしたコーパスみたいなものを構築できれば、それは自動化機械みたいなことも含めて、かなり役に立つのではないかな。
- 評価値については、日本語が母語でない人のためにやさしく書き直しましたという例が、一杯あるので、それを日本語が母語でない人たちが読んで、「分かりました／分かりません」という評定を付けていれば、目的を達成しているかどうか分かる。評定者を雇って、その評定者にいろんな言語(日本に滞在している数の多い順に言語を選べばいい)の人が入っていて、評定を付けた結果をまとめたデータベースは有益であろう。
- 日本語教育では、どの範囲の漢字を教えるか、どういう順番で教えるか、どういう単語を先に教えるか、全部言語資源コーパスがないとできないことだろう。
- 日本人自体が今、どれぐらい読み書き能力があるのかということの調査が必要である。