

言語教育情報研究科 20周年記念企画

講演 田窪行則先生（国立国語研究所前所長・京都大学名誉教授）

演題 「大規模言語モデルと日本語の未来 ― 方言はどうなるか」

[1. ChatGPT などの大規模言語モデルと日本の社会](#)

[2. 日本の言語状況：多言語社会としての日本](#)

[3. 言語か方言か：相互理解性を測る](#)

[4. 消滅危機言語の記録・保存・再生の試み：デジタル博物館構想](#)

[5. 記録保存と言語再活性化の諸問題](#)

[6. 市民科学者の育成と AI による危機言語再活性化](#)

ご紹介をいただきました田窪と申します。本当に久しぶりの対面の講演会ということで、すこし緊張しております。私、講演会も授業も最初の 10 分ぐらい緊張して、「あの先生、大丈夫かな？」と皆さん心配をされますが、追々普通になって、最後のほうは普通になりますので、ご心配なく。

今回は、大規模言語モデルと日本語の未来ということでお話をいたしますが、一部、3月の言語処理学会でお話しした内容と重なります。その際はオンラインで行ったんですが、最後のほう時間が足りなくなって大規模言語モデルについて話せませんでしたので、今回は、ちょっとお時間いただいてそれとの関連をお話しできるかなと思います。

私自身は工学的な背景をあまり持っておりませんで、学生時代は京大総長であった長尾先生の研究室の方々と一緒にいろんなことをやったんですけど、言語処理そのものに関しては、ほとんど知識を持っておりません。時々いろんなところで勉強させてもらっています。ChatGPT みたいな大規模言語モデルそのものに関して、どういうふうなことがちゃんとと言えるかは、分からないですが、これを機会にいま興味あること、ChatGPT の 4 に聞いてみましたので、その答えを併せながら少しお話をさせていただきます。

時間が足らなくなると困りますので、何をやるかというのを全部、最初にお話しします。まず、非常に簡単ですけど ChatGPT のような大規模言語モデルが作る世界、これからの世界へ言語多様性、社会の多様性、文化多様性というのは、どのように保証されるかというふうなことについて、ちょっとだけ述べたいと思います。

1. ChatGPT などの大規模言語モデルと日本の社会

かなり表面的には英語が支配するような社会になってしまうかもしれないです。その危険性は常にあります。後で述べますが、ChatGPT 自体に相談すると、そういうことにはな

らないんだって言うってくれるんですけど、ちゃんと他の文化も大切ですよというふうなことを言うてくれますが、本当かなと、危惧を持つことがあります。それとは別に、実は今、急速に多様性が失われている。ユネスコは多様性の保証の重要性をいろんなところで言われる。日本の政府のほうでも、そういうことは大事であるというふうに言っています。そういうふうみんなが言うということは、実は、多様性が無くなっていったということの証拠です。従って、それを食い止めるために、どういうふうなことができるか、なんらかの手当をしないといけないような状況になっているということです。

ひるがえって、大規模言語モデルのようなものが出てきて、今の ChatGPT に代表されるような大規模言語モデルは、95%ぐらい、もうちょっと少ないかもしれないですけど、英語の資料に基づいています。日本語であるとか、他の言語というのは、そこから微調整しながら翻訳していくような形でできています。

日本語自体のモデルができるかというところは簡単でない。まず最初に作る場合はスペイン語を作ると言われていて、現在は英語のデータでモデルを作って、それを調整するような形で他の言語を作るという対応をしています。

そのような状況で、例えば今問題にしているような日本国内の言語の多様性の保証をするとき、すなわち、危機言語の記録であるとか、維持・再活性化をしようとする、これに対して大規模言語モデルというのは、原理的に作れないですね。超小規模言語モデルしか作れない。何ができるかという、低資源言語処理、大規模じゃなくて小規模ですね。しかも極小規模の言語処理というのを、やらざるを得ないという問題に必ず帰着します。だから多様性の保証というふうなことを考える場合、必ずこの問題に突き当たって、今の問題は低資源言語の処理であるとか、小規模言語モデルであるとかに大規模言語モデルを調整する話になります。これは実際には困難ではあるけれども可能であるかもしれないです。

それと、そのような低資源であってもデータそのものがですね、これ多くないといけませんけど、小規模データ、低資源言語処理と言いながら、地域の言語データそのものをどうやって記録・保存するか、増やすかというのは、実はそんなに簡単じゃないです。地域それぞれの自主的な解決を地域の人自身でやらないといけません、という話になります。

低資源の言語処理というのは、原理的には可能でも、実際におこなうのは様々な困難を伴います。なぜかというふうなこともすこし後でお話しします。

2. 日本の言語状況：多言語社会としての日本

まず、日本の言語状況・文化状況というのが、どういうものであるかを述べます。最近まで、学校教育や、世間・マスコミでは、日本は単一言語、単一文化の国であるというふうに言われてきました。しかし、これからは、日本は、実際には、多言語文化、多文化国家であるということを意識しないといけません。外国の人もそういうふう意識し始めています。日本というのは、単一文化ではなくて、複数の文化圏が合わさったものであり、実際、複数の言語が話されているということが、2009年のユネスコの危機言語認定から話題になっています。

まず、先住民語としてのアイヌ語であるとか、ほかにもアイヌ語以外にも北海道にはツングース系の言語があったりします。それに、移民言語ですね、継承言語があります、朝鮮語であるとか、ブラジルポルトガル語なんかもがそれにあたります。朝鮮語に関しては多くがバイリンガルですけど、一時期は50万人ぐらいの人が話していたというふう言われます。それともちろん手話、日本手話というのも50万人ぐらいの人が話していると考え

られます。これらは日本語と系統関係が異なる言語として話されています。それとは別に琉球の諸言語ですね、それと八丈語も相互理解性の観点から、日本語とは別の言語として扱われます。これらは全て危機言語であって、これらはデータが非常に限定されていると。一番多いのはアイヌ語なんですけど。他の言語に関しては、データそのものが実はそれほど無いというふうなことです。

これがそのユネスコの Intangible Cultural Heritage の Endangered Languages の内容です。そこでは消滅危機度で言語を測るということが行われています。これは割と重要な概念なので後で述べますけれども、これは世代間継承度指標としても使えることをあとで述べます。

消滅危機度に関しては safe は「安全」、全ての世代がその言語を話して、次世代への継承が行われている場合ですね。日本語共通語というのは、この safe に入るので、我々は今のところは安心して日本語をしゃべっている。

次に vulnerable は「ちょっと危ない」、子どもの大部分が話すのだけれど、使用領域が限定されて、例えば、家庭だけになってしまっている。学校教育は例えば全部英語で行われるとかですね。いろんな国がありますけど、高等教育は英語で行うという国が非常に多いです。日本は高等教育も今は、日本語で行われていますが、例えば、明治初期は高等教育は英語で行われたりしました。そういう状態が続けば母語の維持は危なくなります。

definitely endangered は「確実に危ない」、子どもが家庭でも母語としてその言語を習得しない、限られた家庭ではしゃべっているかもしれないけど。非常に危なくとおじいさん、おばあさんしかしゃべってなくて、父親世代は理解はできるけれど、自分たちは話さない。当然子どもがしゃべらない。こうなると「確実に危なく」になります。

瀕死というのは、一番若い世代が 65 歳とか 70 歳とかになっていて、しかも使用も部分的になる場合です。

extinct というのが話し手がいないくなる。こういうふうな指標です。

ユネスコが 2008 年頃調査を行って 2009 年にこれが発表されています。日本で話されている言語にこれを当てはめると、日本で話されている言語は 9 つある。琉球で 6 つで、八丈で 1 つ、アイヌが 1 つで、全部で 8 つ、さらに本土の日本語入れて 9 つです。このうち日本語と系統的に関係が無いと思われるのはアイヌ語だけで、あとは日本語の親戚です。ユネスコの考えるような相互理解性でいくと、これらの言語というのは、日本語の方言ではなくて言語であるというふうな認定をされます。だから日本は単一言語国では、少なくとも日本語以外の 8 つの言語が話されている国だというふうに認定されたわけです。

この発表は、これらの言語を話している人たちには非常に影響を持ちました。これまで方言であるとして、さげすまれているというふうに考えられていたこれらの言語の話者たちが、自分たちの話すことばは、「方言」ではなく独立した言語であるといった誇りをもつようになりました。我々言語学者は言語と方言という区別を実はあまりしないんですが言語であるということで、こういうふうに誇りを持つということがあって、地域言語の維持・再生に関して、潮目が少し変わったということがあります。自分の母語に対する意識

が変わったということで、実際、我々が調査をしている時もそのような印象を受けました。

アイヌ語以外は同源であることが証明されていますが、アイヌ語が日本語とどういう関係にあるのかというのは実は学者によって違います。少なくとも近い関係では絶対ないということはほぼ確かです。あとの7言語は日本語の地域変種とされており、戦前はほとんどが日本語の方言とされていました。いまでも方言とする人はいます。これらはなぜユネスコが取り上げたかと言うと、これらは消滅危機にある言語であり、このまま放っておくと先ほどの指標に合わせて確実に危ないとか、非常に危ないとか、瀕死、アイヌ語はひょっとしたら瀕死というか死んでいると考える人もいます。ユネスコのすすめる文化多様性という観点から、維持・再生の道を図るような方向を見通した、現状把握の報告です。さらにはその現状把握から、これら言語の維持再生その活動が行われることが推奨されたわけです。今のところ成功するかどうかというのは、これからの話になります。

母語だけで日常的に生活できる人というのは、実は非常に限られています。ユネスコの調査で、これらの危機言語はそれなりに有名になったというか、改めて光が当てられた訳です。実は、日本語はセーフと書いてありますけれど、自分たちの地域の言語を考えると、セーフかどうか怪しいです。私は岡山県生まれで、岡山県の南の方言をしゃべるんですけど、18の時から関西に行って、しゃべらなくなったので、もう岡山弁はしゃべれません。泥酔するか興奮した時にしか出てこない。だから、昔お酒たくさん飲んだので、お酒飲む前に「岡山弁しゃべりだしたら、もう止めて家に連れて帰ってくれ」などと言っていました。それぐらいもう心の奥底にいつかしゃべってないです。ということはを、こういう人間が増えれば消滅してしまうわけです。岡山に帰れば皆さんしゃべっているようですが、だんだんしゃべれる人間が減っています。

関西弁は比較的セーフ（安全）であると思いましたが、実は、関西弁というのが存在するかどうかというのは、微妙です。関西の方言はそれなりにテレビ、ラジオなんかでしゃべられることがあるけれど、それ以外は、実は関西弁も危ない。なぜかと言うと関西弁だけで会話を維持できるかと言うと、実はそうではないから。親子の会話というのを録音したものを聞いたことがあるんですが、二人とも完全にバイリンガルで、共通語と関西弁を行き来する、話題によって全部変えてしまうということが起きています。バイリンガルで、2つの言語間を移行しながらしゃべっていくというのは、危なくなる寸前、すなわち、一番の「セーフ」から二番の「危ないかも」というところに行く兆候なんです。

学校教育を関西弁で行うというのは、我々の時代には結構行われていて、僕の先生は京都弁でしかしゃべれなくて、京都弁でしゃべりながら「僕は共通語うまいから」と言って授業してたのですが、我々は「先生、それはみんな京都弁でっせ」と指摘してからかってたりしました。そういうことは実は最近は少なくなっています。共通語プレッシャーが非常に強くなってきだしている。だから、NHK共通語以外は危ないということが起こるかもしれない。

3. 言語か方言か：相互理解性を測る

今、方言とか言語とか、ユネスコは方言でなくて言語だよと言ったんですけど、これは用語の問題で、ユネスコは方言も言語も区別せず、みな言語として扱うと言っているのです、そんなに大きな意味はありません。言語学者も、方言は差別用語になりかけているので、地域変異みたいな言い方に書き直すことが多いです。あえて方言という言葉を使うとすると方言と言語とどう違うかと言うと、相互理解できるのが方言で、できないのが言語であるという定義がチャールズ・ホケットという人が書いた言語学の教科書には書いてあり、

これが引用されます。

ウィキペディアの昔の記事ですけれど、相互理解性 (mutual intelligibility) という記事の中に、相互理解できないけども方言とされている言語があつて、それは琉球語であるという記述があります。今は削除されているようです。だから、方言と言語というのは、原理的には区別はありません。ウィキペディアの昔の記述だと琉球列島の言葉全体を琉球語とよんで、その中の地域変異を琉球方言と呼んでいました。現在、研究者のなかだと、琉球のことば全体を琉球語と呼んでその中の地域変種を方言と呼ぶ人と、最初から琉球諸語と呼んで琉球には 6 つあるという人がいて、最近ユネスコは研究者へのアンケートにもとづいて、琉球諸語のなかに 12 ぐらいの言語を認定しているようです。

これは相互理解性という基準でいくともっと 12 では全然収まらないですね。宮古語の地域変種でも多良間のことばと宮古島のことばはおそらく通じない。先ほど、言語と方言の区別は政治的な定義と書きましたけれども、言語というのは、陸軍と海軍を持つ方言であるといった人がいます。要するに先祖が共通の言語の地域変種は全部方言であり、その方言が陸軍と海軍を持っていると言語になる。だからスイスは全部 3 つ本当は言語があるんですけど、1 つの陸軍、あそこは海軍ないですけど、軍隊があるので、言語というふうになるかもしれない。だから、この定義でいくと琉球語は方言になっちゃう訳ですね。琉球国が独立して軍隊、海軍と陸軍を持つと言語になるというふうな話になるんです。

これは Weinreich という言語学者の講演会が行われた時に聴衆の一人が立って言ったというふうに言われています。Weinreich 自身が言語と方言などというのは意味がない、全部連続しているという話として持ち出されているということです。

ホケットの話も言語と方言の区別はなく、方言連続体であるとか、言語連続体であるというふうな話です。だけど、もし区別をするのであれば、相互理解度による定義が使えるだろう。要するに通じれば方言で、通じなければ言語だというふうになります。

実際問題として言語と方言の区別は連続体なので、実際区別ができるかというのが問題です。ユネスコはそれは分かっているので、最初 6 つだったやつが今 12 になったりしているわけです。次の問題として相互理解度、どうやって測るんだというのがありますね。これどうやって測ったかという、言語学者にアンケートで聞くんです。実際、僕のところにも調査票来まして、調査票書いたんですけど、通じるかどうかというのを書くところがあつて。そうすると特に主観的評価、言語学者の主観的評価なので、それなりに信用できるんですけど、相互理解度を測った訳ではない。

本当に相互理解が可能かどうか測ってみる必要があつて、実際、連続体ならどういうふうに連続しているかというのでも測る必要があるというので、我々はそれやってみました。相互理解度を測るというプロジェクトを計画して、それをやってみました。方言間の相互理解度がどの程度あるかを測る尺度を作るものです。客観テストみたいなものを作りました。これは相互理解性という尺度によって地域変種間の距離をはかるものです。距離の概念というのは大切で、相互理解性の距離が離れていると全然通じない。距離が近いと通じるみたいな連続体として測ることができます。

尺度を 1 から 0 のメモリの位置のようにすると、1 が一番離れていて、0 が同じというふうな違いを測るということができます。20 で測っても 100 で測ってもいいんですけど。この場合、系統関係というのは関係ない。系統関係というのは、比較言語学という学問で元々同じであったが違うのかというのを測る。それでいくとアイヌ語と日本語は系統関係が無い。琉球語と本土の日本語は系統関係があるので親戚であることは確かですが、それとは関係ないです。借用が多ければ系統が違っても理解する部分が増えるので、全然違う言語でも語彙が取り変わってしまえば、半分ぐらい通じたりすることがあります。

この指標は系統関係がある言語が方言じゃなく別言語とみなすべきかの指標として使う

ものです。もう一つはですね、これが実は大切なんですけれど、距離は縦に測ると横に測ると両方あって。横に測ると地域間の相互理解性指標、縦に測ると自分の子どもや孫に自分の母語が通じるかという、言語継承度の指標となります。まったく通じなくて、継承されてなければ、言語が取り変わっているということです。琉球諸語ではほとんどそれが起こったということです。自分の親の言葉が子どもに通じない。おばあちゃん言葉が子どもに通じなくなるんです。

先行研究としてはですね、O'Grady 先生というハワイ大の先生が韓国の済州島の変種の相互理解テストというのをやりました。これは政治的な意味合いがあって、済州島の言葉というのは、本土の韓国語、韓半島の言語とはかなり違うので全く通じないんですけど、方言と考えられて馬鹿にされるということが起こっていた。それで済州島の人が本土の言語を勉強して、言語の取り替えが起こって、済州島のことばが無くなってしまうということが起きています。これはまずいということで、O'Grady さんは、済州島のことばは独立した言語であり、韓国語の姉妹語であるというふうにすべきだということで、調査を行った。調査のやり方は、1分程度の簡単な話を済州語で聞いて10の簡単な質問に答えるというものです。それでやると母語話者の継承度が分かるという話です。比較的若い世代の人でも90点ぐらいとっている。10問やって9問分かるという感じですよ。

全羅南道、これは光州とかがあるところですよ。慶尚南道は釜山とかがあるところですよ。6.0ということは、1分もかからない簡単な話を聞いて10の質問をするんですよ。平均すると10%以下しかわからない。つまり、相互理解性からいうと済州島のことばと全羅南道のことばは、別の言語である。ソウルの人がなぜか12、ということは、1問必ず分かった。これ謎なんですけど。ソウルのほうがドラマたくさん見ているので方言を知っているんですかねえ。

我々は同じようなことを、改良して客観性を高めたものを琉球諸語でやってみました。いろんな人に協力してもらってやりました。いろんなところでやろうということで、こういう計画立てたんですけど、ある理由であまり多くの地域ではできませんでした。

これがそれなんですけど、聞いてみますか。実際には20問やるんですよ。

(音声再生)

ちょっと1問聞こえなくなっていますけど、あまり関係ないですね。聞いてみると何やっているか全然分からないでしょう。いま聞こえたのは「お母さんが食べ物をこどもにあげようとしたけれど、それをこどもが手で払った。」というものです。本当は、「子どもと親がご飯を食べています」が第1問だったんですけど、ちょっとうまくいかなかったんですよ。どっちでも全然分からないということですね。これ20問やらされるんですよ。面白いんですけど、だいたい東京の若い人でやると、0点です。東京外大でやったんですけど、100人ぐらい集まってもらって、平均点が0点ですね、20問やって。1人なんとなく自転車というのを聞いた人がいて、自転車、その人「きー」っていうのを聞いて「木」だと思った人がいて、それが0.5点ぐらいだったです。あとは全然分からなかったということですね。西原地区の人、池間方言の話者ですが、にこれを聞いてもらうと怒り出すんですよ。なんで怒り出すかというと、そんな普通にご飯を食べていますというのを聞いてですね、親と子どもが、お父さんとお母さんと子どもが夕飯を食べていますといって、登場人物は何をしていますかと聞くわけですよ。なんでこんな当たり前のことを聞くんだと怒り出す。標準語の話者にこれを聞かせると、なんで全然分からないものを聞かせると怒り出すということで、どっちも怒り出すという非常に辛い調査になるんですけど。

(音声再生)

今のが「お父さんとお母さんと子どもが晩ご飯を食べています」です。分からないですよ。これは次に。

(音声再生)

これ同じこと言っているんです、実は。「お父さんとお母さんと子どもが晩ご飯を食べています」これは沖永良部の方言です。最初のは宮古島の西原の方言です。それを聞いているわけですね。

(音声再生)

今みたいなのを聞いてもらってやるんですけど、非常に簡単なストーリーです。ピングーという粘土のアニメで、何言っているか分からない「ワーワーワー」ということばを話すものですね。あの可愛らしいアニメを使っています。今のはピングーの一部を池間方言と沖永良部方言に翻訳した物です。もう一つは言語学の実験用に作られたペア・ストーリーというものを使う。ペア・ストーリーというのは無声映画なんですけど、それを見ながら何をやっているかというのをナレートするようなやつです。これは言語学で非常によく使われます。この話の一部を当該の方言に翻訳して録音し、聞いてもらいます。なんでこのような話にするかというと、文化的な背景が中立的なものにしないといけないからです。要するに、昔話だと知っていたりするんですよ。「桃太郎」を使ったテストを作った人がいるんですけど、心理学の先生がそれを聞いて「桃太郎」と分かっちゃった。「どんぶらこどんぶらこ」というのが、なんとなく分かってしまって、想像力豊かな先生だったので、これは「桃太郎」に違いないというふうに全部問題を当ててしまう。これではテストになりませんから、何の話か絶対分からないようにしないといけない。それで借用語が多くないとか、内容が想像できないというやつを使う。簡単な質問をして、同じ形式の形でやります。

これを2地域でやってみました。西原方言というのは宮古島の池間島というところで話されている方言で、池間島から宮古島に移ってきた移住民の集落がありますので、その人が話している方言です。したがって池間方言。話し手が話者が西原方言というのを聞くと、だいたい平均年齢71ぐらいで95点ぐらい。これ95点のほうは母語の話聞いたものです。20問聞いて1問間違えるんですけど。複雑な質問すると覚えられませんから、我々が共通語でやっても1問間違えるんです。

面白いのはもうちょっと若い世代ですね。50代前半の人でやっても90点。ほとんど問題無く分かる。ちょっと単語が分からないやつもあったぐらい。ところがその人たちが先ほど2番目に聞いた沖縄以外の方言を聞くとこれぐらいです、20問中の3問。20問中の2問半ぐらい。ほとんど分からないということになります。これをですね、沖永良部のほうでやるんですけど、上平川、これ沖永良部の方言ですが、平均年齢73歳で沖永良部の方言で今の話をやってみると90点。ほとんどわかる。これは57歳から36歳だからもうちょっと下の世代ですが、やはり、ほとんどわかる。ところが池間方言を聞いてもらう。これも同じ琉球諸語である宮古語の方言だとされているやつですけど、南琉球の池間方言と北琉球の沖永良部方言でやると、4点ということで、4%ですから、20問聞いて1つ分かるか分からないかです。上平川の方言と池間の方言では全く通じないということになります。

方言学者がそれまで琉球のことばは琉球方言だと言っていた琉球語と言われているものの方言だとされているもの、南と北ではこれぐらい違いがあつて、互いのことばが全く分からない。やさしい話なので母語のほうで聞けば、当たり前のこと言っているのに、もう一つのことば全然分からない、どこの外国語なんや、と言われるくらい、全然分からないということになります。池間と沖永良部はこれは全く通じない。なんで池間の人が沖永良部の言葉をすこしだけ聞けるのかは不思議ですが、たぶん沖縄に住んだことがある人が何人かいたんじゃないかと思います。沖縄の言葉と沖永良部の言葉は多少近い。ほとんど通じないとは思いますが。

従つてですね、琉球のことばっていうのは全て、本土のことばと別のことば、琉球の中では南と北では別、全く通じない。南琉球の諸言語であるとか、北琉球の諸言語の間でも通じないということがあるかもしれない。ユネスコが言ったとおり。これを言うんだつたら青森の方はどうやねん、鹿児島の方はどうやねんということになりますけど、青森も東京の人には通じない。だけど、琉球と本土のことばの相互理解度よりはましで、15点ぐらいの感じになります。これは東京の若い人にやっただいたいそれぐらいです。津軽弁のテストもあるんですけど、時間の関係で省きます。

この結果を見ると、相互理解度で2つの言語の距離で測ると沖永良部とウチナーグチは近いかと、池間と沖永良部はというと、非常に違う、ということになります。これらは全部連続的に相互理解度のパーセントで一列に並べられるわけです。そうすると言語と方言を相互理解で決めるというのは、相対的なものなので、より通じないとより通じるぐらいの話にしか、実はならない。方言と言語の区別をすることはできますが、どこかで切るか恣意的にならざるを得ません。だいたい言語学で普通やられるのは、20%か30%。場合によっては40%ぐらいで切る。だから4割分かるぐらいのやつだと言語で、それ以上だと方言というふうな形ですね。だからポルトガル語とスペイン語ぐらいだと5割ぐらい分かると言われてますから、それだともうポルトガル語とスペイン語と先ほど言った別の国だから別の語だ、要するにネイビーとアーミーがあるから別の国だというだけの話になってしまいます。ドイツとオランダの間も全部連続で、だからそれほど意味が実はないんです。

これはおまけですが、相互理解性テストの改善策と問題点を述べます。問題点は、心理言語学的な確実性を保証するために、いろいろな工夫をしたのですが、調査自体に非常に時間がかかって、調査者、被調査者、ともに疲労こんぱいすることです。1時間かかるんです。両方とも疲れ切ってしまう。疲れる原因は他にもあつて、先ほど言いましたけど。別の言語を聞かせると、分からなすぎて怒る。母語を聞かせると当たり前すぎて認知症のテストみたいに思えて、怒る。どっちにしても怒るので、疲れ切る。それを1時間やらせると怒って帰ってしまう人がでてくる。ほとんどの方は優しいので怒らずに付き合ってくださいますが、それでも調査している身としては、疲れ切ってしまうということで、この2つだけやって、あとはやっていない。

O'Gradyさんたちが、なぜやれているかと言うと、調査対象が若い人たちなんです。学生使っている。学生に単位を与えながら調査するのでできる。我々は80代とか70代とかでやっているのちょっとつらいわけです。O'Gradyさんたちはより改良したテストも作っています。これは穴埋めテスト使うやつですが、これは省略します。

だからうまくやれば、簡単なのができるかもしれないですけど、今のところはできない。やれば面白いと思うんです。例えばチェコ語とポーランド語は通じるか、ちゃんと調査した人があまりいない。それと大阪弁と東京弁はちょっと聞きすぎているからダメだけど、鹿児島弁と東京弁だとどれぐらい通じるのかというのをテストしようと思えばできるんで

すけど。ちょっと10人ぐらい疲労困憊してもらえればできるかもしれない。だから面白いですが、その調査をできるといいかもしれないですが、実際にやろうとするとつらいです。

それともう一つ、実はこの相互理解性テストというのは、先ほど言いましたけど、言語継承度に使えるということを述べました。言語継承度というのは、おじいちゃん言葉が孫に通じるか。継承が無くなるということは、切れるということですね、おじいちゃんとか、おばあちゃん言葉が、そのまま子どもに通じないということです。だからおじいちゃん、おばあちゃんは、下手くそな共通語でしゃべって、孫たちと遊ぶということになります。しゃべれない場合は、もう遊べないことになるわけです、切れてしまう。そういうことが日本中に実は起こっていたり。特に琉球では起こっています。

それで我々がやろうとしているのは、消滅危機言語の危機度を世代間継承度として使うことです。言語の方言の区別とかというのは連続体なんだから、そんなものやっても仕方ないことはないし、面白いかもしれないけれど、手間に対してそれほどメリットが少ないかもしれないということです。

それで、相互理解性テストを世代間継承度を測るよう使うということをしました。継承度がそれほど低くなければ、実は上げることが可能だということが分かっています。それは、例えばこんな感じで測ることができます。世代別に並べると全ての上の世代の話者は全員80点以上取れる。下のほうの世代だと65点とかで、分からない部分がいぶ増えている。何してるのというふうな簡単なお話を聞いてもらって内容を質問するだけのテストですから、話せなくてもある程度理解できればだいたい取れる訳です。これが、全然点数がとれない場合は、祖父母の世代と孫の世代とが別言語を話しているというような形になってしまう。そういうふうな形で世代間継承がうまく言っているかどうかの指標としてみることができます。

我々が使いたいのは65点、50点より上ぐらいの場合です。英語でも、日本語でもリスニングコンプレヘンションで合格点を取れるぐらいの点数です。だいたいやさしい話だったら筋が分かるぐらいなので、大量に聞けば使えるようになるかもしれない。だから消滅危機が迫っている言語を再活性化できる可能性があるということです。理解度テストをやってもらって、ある程度高得点を取れたら、どの部分をもっと聞いてもらえば分かるようになるかを考慮して、その部分に対するテストみたいな作れる。何が苦手なのかというのが分かるかもしれない。そんな形で利用すればいいんじゃないか。

つまり危機言語再活性化への道が拓けるといことです。我々がテストを行ってびっくりしたのは、言語学の教科書には琉球の言語というのは消滅危機言語で、流ちょうな話し手は65歳以上であると書いてある。若い人はいないとされているんですが、先ほどのテストでやると35歳ぐらいまでだいたい分かる人いるんです。これ以上下になるとむずかしいですが。このテストをやったのは5年ぐらい前だから、年齢は上がっているんですけど、比較的下の年齢でもかなり分かるわけです。点数にすると85点ぐらい。ほとんど分かっているいくつかの単語が聞き取れないぐらいです。そうするともっと聞けば分かるようになる。35歳、40歳なら小さい子の親の世代です。ただ、分かるんだけど、しゃべろうとするとしゃべれないとか、しゃべらない。なんでしゃべらないかと言うとしゃべろうとすると上の世代の人が注意するんですね、間違っているところを。一番語学教育で良くないことです。間違いをすぐ注意してしまう。もっとしゃべってもらおうと、実はこの35歳の85点ぐらいの人は、あっという間にしゃべれるようになる可能性があるわけです。特に敬語がもんだいで、一番大切なのは、敬語を間違えると強く非難されて萎縮してしゃべる気がなくなります。

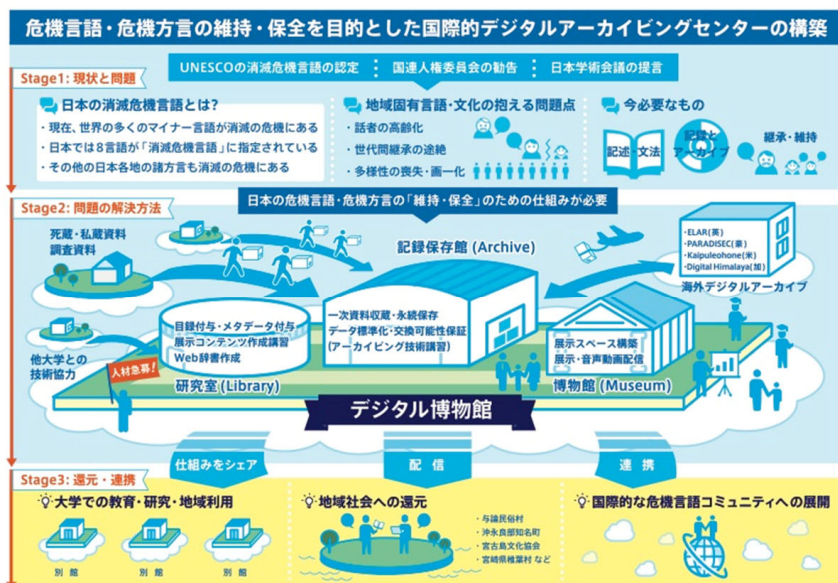
敬語というのはそんなに難しくないですが、なかなか聞いただけでマスターできませ

ん。だから、敬語の使い方をちょっと練習するだけで、しゃべれるようになる。同時に敬語を間違えてもまわりの人がすぐに注意しないというのをやればいいのかも。理解度テストで、100点満点の65点から95点とれる人なら、実はそんなに難しくなく話せるようになります。若い人はネットにアクセスできますから、たくさんコンテンツを作って聞いてもらえば、言語知識の再活性化の余地があるということです。パッシブな知識がある人をアクティブな人にするためには、たくさん聞けばいい。それこそクラッシュェン (Stephen D. Krashen) のインプット仮説 (input hypothesis) みたいなものです。

4. 消滅危機言語の記録・保存・再生の試み：デジタル博物館構想

それで我々はですね、それに初めて気がついてた訳じゃないんですけど、最初の頃からデジタルな博物館みたいなものを作って、そこでオンラインで音声とか動画とかをたくさん見てもらうことを考えてきました。隣のおじいちゃんとおばあちゃんが面白い話しているのを聞いてもらおうというものでして、そのプロトタイプみたいなものを作りました。

これがちょっと聞いてもらおうといいんですけど、最後まで行って、時間が余れば聞いてもらおうということにします。こんな感じで計画を作って文科省のほうに出したんですけど、認められませんが、残念ながら頓挫しているんですけど、世界中の消滅危機言語をこういうふうな形で再活性化するようなプログラムを作ろうと思ひまして。プロトタイプとしては、たくさんアーカイブ、音声とか動画データというのをアーカイブして、それに字幕とか付けたり説明をしたり、いろんなモノ作りの解説とかしたりしている。あとは、お祭りとかたくさんありますんで、お祭りの解説をしたりするような解説部分をこれまで作って、プライバシーを保障しないといけないので、全部は公開できないんですけど、一部でも公開すればいい。外国にはそういうふうなアーカイブが一部ありますから、そこらと協力して、こういうふうな一大プロジェクトを起こそうというので、10億円くださいと言ったんですけど、くれませんでした。10億円ダメなら、じゃあ500万円って言ったらくれたので、それで細々とやっているような感じです。



どういふにやるかというんですけれど、モノとか技術というのは、割と博物館で展示されています。民博とか歴博とかの博物館に行くとかたくさん展示されているんです。昔のやつも今のやつも。あまり言語がついていないんで、言語もつけて全体として見られるような形にしよう。モノと技術と言語をセットで保存するようなことをやればいい。だからモノ作りとか踊りとか歌とかをそれを説明している人たちのことばと一緒に見てもらう。ものや技術だけでなくことばも、全部解説みたいなやつをつけないといけないです。なんでかと言うとですね、細かい技術なんか説明書があれば、場合によっては再生できる。モノ自体は実物があればリバースエンジニアリングで再生することができるかもしれない。実際そういうことなされています。しかし言語は消えてしまうと再生できない。言語はそれを話す人がいなくなると消えて無くなるから。今どんどん言語が消えて無くなっているのがまずい。皆さん安心してますけれど、岡山弁は、例えば消えて無くなるかもしれない。みんな標準語しゃべるようになって。大阪弁だって危ないです。だからそれともっと地方の言語だと危ないですから、そういうの残しておかないと、消えて無くなります。

言語だけだと宙に浮いてしまうので、モノと技術と合わせて全部保存しましょう、記録しましょう。そいつを見せるというか、一部展示することによって、面白いコンテンツができますから。たくさん作りましたけど、我ながら毎日何時間見ても飽きないようなコンテンツができています。それを聞いてもらうとさっき言ったようにもちろん素地があってある程度理解でき状態の人なら再活性化できるわけです。私は、ここに17年通ってますけど、だいたい聞いて全部分かります。しゃべるのはそんなに上手じゃないですけど。朝から晩まで聞いてるんで。土地の人は毎日聞こうと思ったら聞ける訳ですから、それができるようになる。アメリカの人類学者ボアズ (Franz Boas) の3点セットというのがあります。この人がフィールドに行ってるのは、文法と辞書とテキストを作ります。これが3点セットと言われるもので。我々は言語学科を出ると、フィールドに出る時はこの3点セット作りましようねと言われる。今は動画とか音声ができるので、こいつを加えるということができます。これをやればいい。

5. 記録保存と言語再活性化の諸問題

実は、先ほどの言語再生とか、言語再活性化という観点からは、この3点セットでは足りなくて、教科書とか電子辞書とか動画コンテンツも作らなきゃいけない。これらが今足りないところです。だから語学教育としてやらないといけないわけです。語学教育として地域言語を再活性化みたいなのをしないといけない。そのためにはこれまでの方言調査の方法ではダメで、これまでの方言調査は行って調査して自分で抱えているだけなので、再活性化の資料には使えない。論文書くためだけに使っている。再活性化のためのドキュメンテーションの仕方というのは、ちょっと違うことが必要になります。

もっと問題はですね、我々はだいたい1,000時間ちょっとぐらいの動画音声データを集めました。これ全部開示しないとイケなくて、その中の一部だけを公開をしているんですけど、プライバシーをコントロールしないとイケないので、あまりプライバシーに触れないような公開してもいいという承諾書を書面で取らないとイケません。だから、先ほどの編集無しの原データと記録保管のためだけに最小限加工したデータとコンテンツのための編集加工データみたいなの全部です。動画なんかは、フォーマットが変わってモノによってはすぐ再生できなくなりますから、変換しないと見られない。結構技術がいることになります。

再活性化ということになると、公開しないとイケないんですけど、公開するのが実はそんなに簡単ではなくて、先ほど言ったように西原地区の動画というのは、私のYouTubeにだいたい20時間ぐらい公開されています。全部字幕、日本語の字幕と書き起こし字幕、場合によっては英語の字幕もついています。けど他の地区はほとんど無いんですね。それはなぜかと言うと、動画に映っている人が許可をくれないということと、研究者が死蔵してしまって、公開しないとすることがあります。どちらも相まって、結構データあるはずなんですけど、公開されてないということです。

プライバシー保護するのに技術的なことがたくさんありますから、この間、録りに行ったやつ、一緒に行ったこの研究科長が映り込んでしまって、「これどうする？」って言ったらずっぱりまずいでしょうというので、消そうとしたんですけど、YouTube上ではなかなか消えなくて、まだ彼女が映っている。彼女が映り込むぐらいだったらいいんですけど、映り込むべき人でない人が映り込んだら消さないといけないです。それと面白い話すると、だいたいプライバシーに突っ込んでしまって悪口になったりしますから、それも消さないといけないので結構時間がかかる。面白ければ面白いほど、大変なんです。なかなかできない。私の場合地域と親戚がいるわけでもなく、もう本当に偶然に入って17年間やっている訳ですけど、それだとやはり限界があります。おじいちゃん、おばあちゃんたちは普通、機器の操作とかできないですから、地元の若い人たちがやってくれないといけないです。本当にその言語を習得したいとか、残したいとか思っている地域の人やらないといけない。そのためには我々の技術をそこに移さないといけないというふうなことがあります。技術的なこともそうですし、書き起こすためには言語学の知識が要りますから言語学者にもなってもらわないといけない。

6. 市民科学者の育成と AI による危機言語再活性化

そういう活動を実際に行っている人たちがいます。国語研の「ことばのミュージアム」ですけど、これですね、市民科学者の育成というのがあるんですけど、これは国語研の山田准教授が仲間と立ち上げているようなサロンです。これは地域の人が勉強会をして、自分たちで記録できるような形でやっている訳ですよ。そういうふうなことをやっている。集落ごとに違うので、集落の人が集まってやらない限りはできないわけですね。そういうふうな市民科学者の育成をやっています。

先ほどのやつは国語研のページなのでそこから行けます。これは、しまむにサロンという今のやつを説明したものですけれど、横山晶子さんという、国語研も属している人間文化研究機構の助教をしている人ですけど、沖永良部にずっと入って地域の人と一緒に、地域の人たちに言語学を教えて、地域の人が自分で書き起こしとかができるような実践をしている人です。

まま場合によっては今の言語再生とかに使おうと思えば使える。そういう計画もあります。

あと、ChatGPT であるとか、書き起こしソフトの Vrew です。Whisper というのは音声認識のソフトなんですけど、そういうやつも使おうと思えば使えます。地域の若い人が市民科学としてやってくれればいいんですけど、かならずしも若い人は自分の方言なんか残したいと思っていないので。お年を召した方がやることになります。さっきお話しした活動を見るとみんな 60 代、70 代です。それでもちゃんと言語学の入門講座を受けてできるようになってたりするので、それはそれで素晴らしいんですけど。いろんなところの敷居が高すぎますよね。動画作成するにしても。年をとると AI 関係の技術に追いつけない。使いやすいテンプレートとかマニュアルとかを作るのがいいんじゃないかというのでそれもやっています。辞書編集は語彙と定義、用例などを集めたやつを Excel で作って、それをテフ(TeX)で組版して辞書を作っているんです。だからそういうふうな技術は使えるんですけど、AI 使えばもっと簡単に実はできるはずなんで、ちゃんとやれば。そういうふうな形でやろうとしています。

しかし、最初にお話ししたように大規模言語モデルを危機言語に応用するにはいろいろハードルがあります。これは現在の音声認識の実力というのを示すために、ファイルを出そうとしたものです。方言をそのまま日本語の書き起こしソフトにかけたやつなんですけど、先ほど見れば分かるように、日本語共通語用に微調整されていますから、池間方言でやると、5%ぐらいしか当たらないんですね。それを調整するのは、非常に技術が要りますんで、恐らくは簡単にはできないということになります。

それともう 1 つおまけですけど、ChatGPT4 にですね、危機言語再生がどうやったら年寄りでもできるかと相談してみた。これはグローサリーの付け方ですけど、見えますかね。The Leipzig Glossing Rules というのは言語学者が対象の言語に対してグロス、要するに簡単な説明ですね、を付けるようなシステムなんですけど。ChatGPT は賢くて The Leipzig Glossing Rules 自体は知っていました。それを使ってやってやると、当然、ちゃんととはできない訳ですけど、なんとかやってくれるんですね。ChatGPT はときどき嘘をつきますから、嘘つきながらやるわけです。これは「ばが なゆーぎゃー たくぼ ゆきのりていどうあい」で、「私の名前は田窪行則と言います。」にあたる池間方言です。自己紹介の時に使うやつなんですけど。「私はチョコレートを食べています。」というふうに解釈したようです。それはまずいだらうということで、でてくる単語を教えたんですね。「ば=が」は「私=の」という意味で、「な=ゆ=ぎゃー」は「名前=対格=主題」です。訂正できますか、」と言うと、「私の回答に誤りがありました。」とって訂正してくれるんですけど、なぜか「チョコレート」が直らない、当然教えてないから。「僕の名前はチョコレート」になっちゃって。「チョコレート」を直してもらおう。「たくぼていどう」は「田窪=引用=焦点助詞」、「あい」は「言う」です。訂正できますか。」という「おっしゃるとおりです。間違っていました。」と直してくれるんですけど、こんな感じで直します。結局どうしても私の名前を「チョコレート」にしたい。いくら教えても「チョコレート」なんです。これが ChatGPT3.5。さすがに、4 はまだましなんです。同じことを聞きました。「知識データベースに入っていないから、正確な分析ができない。」という答えです。3.5 はファンタジーでやりましたが、今度は知らないから教えてもらわないとできませんよと、ちゃんと言ってくれます。

でもまあ、結果はですね、教えても似たようなもんです。「チョコレート」は直るのかな。最終的に教えるとそれなりにちゃんと出てきます。これもっとちゃんと教えると実は使えたりします。セッションをちゃんとして、ちゃんと指示を与えて、場合によっては辞書も与えるとできてきたりする。こういうふうな調整をすれば。だから 3.5 はアホたれで

したけど、ChatGPT4はこれは最初から65点ぐらいいける。直らないところはあるんですけど。

それですね、「この文は宮古語池間方言です。辞書があればこの言語の分析ができますか。」っていうと、突然英語に変わるんですけど、「やろうと思えばできます。」と返ってきます。どういうふうにやればいいのかと聞くと、アドバイスをくれます。

池間方言には辞書があります。私とUCLAの岩崎教授が元宮古高校の仲間博之先生とた6,000語ぐらいの言語学的に作った辞書です。そいつを使えば今の文の解析ができるかってChatGPT4に聞くと、こういうふうにやればできますというのを教えてくれる。だから、Excel形式の辞書を整理して検索しやすい形にして別々の行・列に分ける。このとおりなってるんですけど、データベースに変換する。これはSQLというQuery言語、要するにデータベースを扱うための言語を使って辞書を引くシステムを作り、それで単語に分析してデータベースで検索して、それで構造を理解するというふうにやれば出てきますというふうに、ちゃんと教えてくれます。

これはこのプロセスは特定のプログラミングスキルや、データベース管理の知識を必要とします。また、宮古池間方言に関する専門的な知識も必要となるでしょう。これは難しいけれど、今はChatGPT4自体にはできないのですが、こうすればできるというふうなことを教えてくれる。実際どうやればいいのかというのを聞いていると、こういうことに関しては、一応信用できるんですね。ほとんど信頼に足ることを言ってくれました。

最後ですけど、もっと一般的なことを聞きました。「大規模言語モデルと言語の多様性についてどう考えますか。」と言うと、要するに微調整みたいなのですればできますというふうなことを書いてあります。

ChatGPTは95%ぐらいが英語のデータに基づいているので、日本語自体が危なくなるとか、小規模のデータしか持たないような言語は危ないんじゃないかと言いました。ChatGPTには、モラルの問題や言語処理に関する倫理の問題がちゃんとデータ自体に組み込まれているので、ChatGPT3.5にこの問題を聞くと「多様性を保証しなければならない」、「低リソース言語へのアクセスというのを保証しないといけない」、「言語多様性と文化理解みたいなのが非常に大切なので、今このChatGPTみたいなものは、それに使えますよ。」というようなことを書いてきます。非常にまともです。

ただ、ここからどうやったらいいでしょうという質問をしたら、3.5ではすみません、分かりませんというのが返ってきます。4に聞くと「文化多様性みたいなことやらないといけないし、バイアス公平性、みたいなのもやらないといけないし、accessibilityも大切だし、継続的な学習と進化みたいなものもあるし、多言語化の翻訳と理解が必要である。」というふうな非常に一般的な答えが返ってきます。

ここで、質問者のアノニマスというのは、僕のことですが、「マイナーな言語使用は限定されていますが、どうやって大規模言語モデルを構築するのでしょうか。」と聞いてみました。先ほど言ったような一般的な形でトランスファー学習というのがあるんですけど、今の英語のデータを日本語や韓国語に直すというトランスファー学習です。

言語間転移技術の利用で、微調整みたいなのを言語間転移の際にやるというふうなこともできます。だから類似の言語、日本語から非常に近い言語へのトランスファー学習ならそれは類似度を最大限に使うこともできる、という答えでした。

具体的にどうやるんですかと言うと、事前学習とか微調整であるとかと技術的な答えが帰ってきます。実際に危機言語に適用した例がありますかと言うと、例として次の3つをあげてきました。

Facebook AI の Wav2Vec 2.0: Facebook AI は、Wav2Vec 2.0 という自動音声認識システムを開発し、トランスファー学習を使って多様な言語、特にリソースが限られている言語に対応させました。このシステムは、大規模な多言語データセットで事前学習され、その後、特定の少数話者言語のデータで微調整されています。

Google の Project Euphonia: Google は、Project Euphonia を通じて、話し言葉の障害を持つ人々のための音声認識技術を改善しようとしています。これは直接的に危機言語と関連しているわけではありませんが、少ないデータで効果的なモデルを作るという点で、似たようなアプローチが取られています。トランスファー学習は、一般的な音声認識モデルを特定の発話スタイルに適応させるのに用いられました。

Endangered Languages Project: このプロジェクトは、世界中の危機に瀕している言語を記録し、保存することを目的としています。AI とトランスファー学習の技術は、これらの言語の記録と解析に利用される可能性があります。例えば、既存の言語モデルを事前学習してから、特定の危機言語のデータで微調整することで、その言語の音声認識やテキスト解析が可能になります。

もっと具体的に聞くと「分かりません。」と言います。非常に一般的な話で聞くとむちゃくちゃ賢いんですけど、具体的にそんなペーパーありますかと言うと、すいませんわかりませんと言ってくる。要するに、一般的に書かれたことをデータから引っ張ってきて生成していく。こういうこと言いそうだということで先読み予想するので、ほんとに具体的なことに答えるのはなかなか難しいということがわかりました。

それにしても、これはほとんど学部のレポートぐらいだったら、そのまま使えそうな答えです。一般的ではありますが、ほとんど間違っていないです。それに実際に何がしたいかを具体的にちゃんと聞けばPythonとかのプログラムを組んでくれます。実際ちょっとやってみたんですけど。そうするとかなりプログラミング言語のこと知ってて実装できる人が使えば、先ほどの大規模データも小規模のほうに直すことができます。それを工学部の人やってくれるかというやってくれません。なぜかと言うと、これはやればできると言うことがほぼ分かっているから論文にならない。やればできると分かっていることは、工学部の人、やってくれません。それでは誰がやるかという、自分でやるしかない。我々はもう歳取ってPythonのコード書くの辛いですから、若い人がやらないといけない。国語研の人たちに聞くとですね、対象に対する愛がなければ何事もできませんというふうに言われます。だから対象に対する愛がある人がこういうことをやれば、技術的には可能であるということです。

以上。