第1回 SPARC Japan セミナー2019

「人文社会系分野におけるオープンサイエンス ～実践に向けて～」

国立国語研究所の言語資源と オープンデータ・オープンサイエンス

小木曽 智信
（国立国語研究所）

講演要旨

国立国語研究所では、コーパスや電子化辞書、言語地図や方言の調査データや音源などさまざまな言語資源を自ら構築し、保有し、公開している。オープンサイエンスの潮流の中で、これからの国語研では、これらの資源をオープンデータとして公開し、研究者のみならず広く一般に利用できるようにすることを計画している。しかし、コーパス等の言語資源のオープン化にはいくつかの課題がある。その一つは、自己収入の確保の観点から、安易にコーパスをオープン化することができないことである。コーパスは高いコストを払って内製したものであり、かつ IT 企業からの強い需要があるデータである。今日研究機関に求められる自己収入の確保や、研究機関の存在意義にもつながる内製データの保持ということと、研究資源をオープンデータとして広く公開することの間でどのようにバランスを取るべきなのか。国語研の取り組みの現状を紹介するとともに、問題提起としたい。

小木曽 智信

東京大学大学院人文科学研究科修士課程終了、博士課程単位取得満期退学。奈良先端科学技術大学情報科学研究科修了、博士（工学）。明海大学講師、独立行政法人国立国語研究所研究開発部門研究員等を経て、2017年より現職。日本語学会評議員。専門は日本語学（国語学）、コーパス言語学、自然言語処理。バックグラウンドは日本語の歴史の研究で、コーパスを活用した研究を行っている。国立国語研究所でコーパスの開発に携わり、在職中に奈良先端科学技術大学院大学で自然言語処理を専攻。現在は日本語の通時的な研究を可能にする「日本語歴史コーパス」の構築プロジェクトのリーダーを努める。

私の講演の元々のタイトルは「国立国語研究所の言語資源とオープンデータ」でよいのですが、今回の全体テーマに合わせて「オープンサイエンス」という言葉を最後に加えました。データの話だけではなく、研究方法の実践としてのオープンサイエンスの話を少し加えてお話をさせていただきたいと思います。

私は普段は国立国語研究所の言語変化研究領域で、日本語の変化・歴史を研究しています。そこで「通時コーパスの構築と日本語史研究の新展開」というプロジェクトのリーダーを務めています。コーパスというものを作る、そして、使うということは、オープンデータ、オープンサイエンスに少し関わってくるところがあります。

元々の専攻は文学部で人文社会系で、日本語学の人間なのですが、社会人になってから今度は情報系の大学院に行って、自然言語処理について学びました。大学院時代から、古い時代の日本語のコーパス、用例データベースのようなものを作る仕事を行ってきました。

国語研とオープンサイエンス・オープンデータ

国立国語研究所では将来計画委員会というのをやっ

ていて、そこで、オープンサイエンス、オープンデータを次の期の研究所の基盤として進めていこうと言っています。調査・収集してきたデータは公開を原則としオープンデータにする、所定の手続きで誰でもアクセス可能なデータにすることを考えています。ここで、本当のオープンデータであれば当然無料、無制限の利用ということになってくるのですが、そこはこの後、お話しするような事情で完全オープンではありません。

それから、方法の面でもオープンにしよう、検証可能にしようということで、主観を排するということもそうですが、研究に用いた中間的なデータやプログラムもオープンにしていこうとしています。小野さんからお話があった市民参加という意味でのオープンサイエンスもあるのですが、それ以前の話として、研究者間でのデータ共有、データ・方法をオープンにしていくということもオープンサイエンスの重要な側面だと思っています。

それから、コーパスとアーカイブを核とした研究をしていこうということも言っています。今の研究所の中で一番中心になっていることが「多様な言語資源に基づく総合的日本語研究の開拓」です。たくさんのいろいろなコーパスを作って、それを基に研究を進めていくことが中心となっていて、また、外部からも高い評価を得ていることから、そのようなことを考えています。また、今できていないこととして、危機言語、危機方言等の音声データ、録音データをアーカイブ化して、それもオープンにしていきたいと考えています。

今日はオープンサイエンスの実践を進めている方々からのお話ということでしたが、まだそんなにできているわけではないのです。だからこそ、次の期にオープンサイエンスを進めていきたいという話をしています。人文系の研究だと昔からある話ですが、データの囲い込みのような問題、昔で言えば本を見せないというような話から始まって、個人で作ったデータは出さない、カードは見せないというような状況をどんどんオープンにしていきたいということです。

また、言語研究だとしばしば問題になるのが文法性の判断で、こういう言い方は言える、言えないというのが文法の記述で重要になってくるのですが、それが主観的にしか見えないと言われることがあります。これは言える、言えないと言っているのですが、その根拠はというときに、何かが見えない。それはやはり実験など、いろいろな方法でもっとエビデンスを出せるようなものにしていく必要があります。また、それらに限らず、論文の元となった中間的なデータなどが公開されないので、そう言うならそうなのだろうという話になってしまって、検証可能性という面で乏しさが残るといったことがあります。

そういうことを何とかしていきたいということが研究所としての方針でもあり、それを何とか変えていくための取り組みとして、オープンデータ、オープンサイエンスということを言っていきたいということです。

国語研のコーパスとオープンデータ

国語研究所の言語資源と最初に申しましたが、ここではコーパスのことを主にお話ししていこうと思います。言語資源と言う場合、いろいろなものがあり、公開されているものでも、言語地図、社会調査型の調査結果、音源などいろいろありますが、一番中心となるコーパスのことをお話ししていきたいと思います。

「コーパス」という言葉になじみのない方もいらっしゃると思うので、簡単にお話しします。これは、言語を分析するための基礎資料として、書き言葉や話し言葉の資料を体系的に収集し、研究用の情報を付与した大規模なデータベースです。要は、実際の言葉の用例をたくさん集めてきて、研究に必要なだけの情報を付けたものということになります。ただ何でもかんでも集めるのだったら、ウェブをクロールすればいいのですが、そうではなくてバランス良く集めたり、日本語の実態を反映できるように設計の上で持ってきたりします。

また、研究用の情報としては、誰が、どんな人がいつ発話しているというような情報から、全てのテキストに単語の情報を付けて、品詞分解のようなことをし

て、全部に付けるということをしています。今、これが日本語研究、言語研究の中で非常に重要な位置を占めていて、研究するならこういうものが必要だということが今世紀に入ってから常識化してきています。

研究所で出してきたものとしては、2004 年の「日本語話し言葉コーパス（CSJ）」、2011 年に公開した「現代日本語書き言葉均衡コーパス（BCCWJ）」、こちらは 1 億語の日本語の書き言葉を、新聞・雑誌・書籍などからバランス良く集めてきたものです。今、私が中心になっているのは「日本語歴史コーパス（CHJ）」で、これは 2013 年から公開しています。奈良時代の『万葉集』から始まり、明治・大正時代までの新聞・雑誌等を集めてきて、千数百年分の日本語の歴史を検索できる通時コーパス、縦に時代を調べられるコーパスを作っています。他にも、「国語研日本語ウェブコーパス（NWJC）」はウェブをクロールしてくるタイプの大規模な 100 億語のコーパスです。

それから、「多言語母語の日本語学習者横断コーパス（I-JAS）」という、外国人などの日本語を学習している人の日本語を集めたものがあります。さらに「日常会話コーパス（CEJC）」という、日常どんなことをしゃべっているのか、カメラとマイクを入れて食卓・居酒屋・職場などで録音してきて、それを書き起こしてコーパスにするというようなものなども作っています。また、日本各地の方言のコーパス化なども進めています。

これらのコーパスは、国語研究所ではコーパス開発センターというところがあり、ここで公開等を行っています。全てオンラインで利用可能になっています。オンラインですから、オープンといえばオープンで、いずれも無料で基本的には使えるようになっているのですが、この後お話しするような事情で、そうではない部分もあります。

こういうコーパスが日本語研究でどれぐらい使われているかということなのですが、「現代日本語書き言葉均衡コーパス」は、今、登録ユーザーが 2 万人になっています。年間のクエリ数が 50 万件ちょっとです。この辺はぴんとこないかもしれませんが、言語研究者の数は、多く見積もってもそもそも日本に数千人しかいないはずです。ですから、かなりたくさんの人が使ってくださっているということは間違いありません。また、これを利用した論文が年に約 70 本出ています。日本語の研究論文の数はそんなに多いわけではないので、かなりの割合ということです。

同じように、「日本語歴史コーパス」は今、登録ユーザー数が 1 万人になっています。年間のクエリ数、検索の数が 26 万件、利用した論文が大体年に 50 本出てくるようになっています。これも日本語の歴史という非常にニッチな部分での数なので、この分野においては研究に欠かせないものになっていると言っていいのではないか、そういう意味でインフラと言ってもよいのではないかと考えています。

このコーパスは、基本的には無料で、オンラインでの利用という形で公開しています。「中納言」というアプリケーションの中で、マウスでクリックしながら、調べたい言葉を入れていくと検索できます。『源氏物語』に出てくる形容詞を全部持ってくるとか、『枕草子』の中のこれを探そうとか、ある言葉の次に来る助動詞にどんなものがあるかとか、そういう非常に細かい検索ができるので、日本語の研究にとって欠かせないものになってきています。オンラインで無料だけれども、登録が必要な形で公開しています。

1 億語のコーパスやオンラインでの検索環境の提供ということで、コーパスの構築と公開には大変コストがかかります。1 億語の「現代日本語書き言葉均衡コーパス」の場合は、特定領域研究で、科研費で全体で 8 億円、プラス、国語研究所の運営費交付金が、人件費を入れると同額ぐらいになってしまうのではないかという気もしますが、かなりの額をつぎ込んで、また、5 年間丸々かけて作ったものになります。書籍等からバランスを取ったサンプリングをする、このときにはいろいろな図書館にお世話になったのですが、J-BISC からランダムサンプリングして、バランスを取るということをしています。それを電子化して単語の情報を

付けるということをしたものです。

「日本語歴史コーパス」の場合は、2013 年から私が関わっているので、大体、把握しているのですけれども、何しろ奈良時代から明治・大正時代までの日本語の全てに単語の情報を付けなければいけません。つまり、品詞分解、まさに古文の品詞分解そのものを 1,000 万語以上のものに対してやるのです。

もちろん人間ではできませんので、形態素解析という自然言語処理の技術を用いるのですが、大変な手間をかけて、それを後で修正してやるということをしています。それから、外部の画像等にリンクして使えるようにしています。最近、各地の大学図書館等がオンラインで貴重書をどんどん公開してくださっているので、歴史コーパスで検索すると、その言葉が出てきた原本を確認できるという体制が整ってきています。ですから、コーパスというのは、一面ではそれだけで独立しているようでもあるのですが、考えようによっては原本を読むためのツールとしても使うことができて、言葉を探して、それがどう出てきているかというような形で原本を読んでいくという利用の仕方も可能になっています。こちらは年間 3,000 万円弱の国語研究所の予算に加えて、科研費の基盤（A）などを加えて、やはり年間数千万円のコストをかけて作っています。

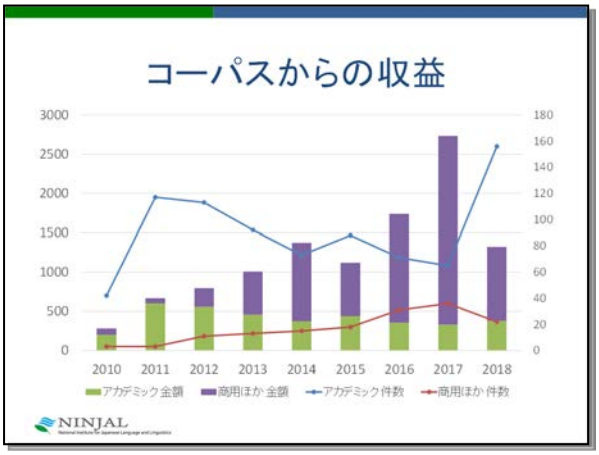
そして、サーバーで「中納言」などを公開していますが、人件費や電気代を置いておいても、サーバーのリプレース等で 1,000 万円近くかかってきます。新しいコーパスを作るというときに、新規性をもって予算獲得をするということはできるわけですが、インフラ化してしまうと、むしろそういうお金が取りづらくなってきて、必要なものであるにもかかわらず、新しく取ることは難しいということがあります。

研究所で、コーパスからの収益は上がっています（図 1）。最初に申し上げたとおり、完全にフリーではありません。「現代日本語書き言葉均衡コーパス」については、特定の社名は出しませんが、GAFA のような IT 企業数社と契約を結んだりしています。ここ 5 年以上の間、年間 1,000 万円以上、全体で収益が上がっています。こういうものがコーパス公開の原資にもなっていくことから、コーパスをオープンデータ化してほしいという話とは相反する部分が出てきます。

コーパスは、コーパスの構築を目的として研究を組むということが行われます。元々、多目的なものなので、一つの目的のためでなく、一回作ってしまえばいろいろできるため、コーパスの構築を目的とした研究がたくさんあります。ですから、副産物ではなく、研究データといっても、研究のために作られたものなのです。それ自体が目的なのです。しかも、コーパスは元々は他者の著作物であるものもあるのですが、それを集めてきて、単語レベルでたくさんの情報を付けるということをしているので、かなり高度な編集著作物ということにもなってきます。

そして、自己収入を生み出すような経済的な価値を持っています。現代語のコーパスの場合は、現代語の自然言語処理のベースになっています。話し言葉コーパスは、今日使われている日本語の音声認識のベースになっていて、Alexa、Siri などの基礎にもなります。

というわけで、なかなかオープンデータにはそのままつながっていかない部分があります。では研究所としてはどうするかというと、所定の手続きを踏んでいただければ誰でもアクセスや利用は可能ですが、無償・無制限であることは意味しません。なぜなら、たくさんの企業から引き合いがあるからというスタンスで、コーパスについては扱っているからです。つまり、オープンデータではないのです。ただ、アカデミック



（図 1）

な料金を設けるなどして、もちろん使いやすいようにしていますし、先ほどのオンライン検索のような場合には完全に無料で利用できるようにしています。これは頑張っているところで、サーバー等はこちらで負担して、でも、学術の基盤としては公開し続けなければいけないだろうと、そういうことをしているわけです。

ですから、オープンデータとしてコーパスを語ることは非常に難しくなってしまうのですが、そうではなくて、コーパスに対して情報を付けていく、「アノテーション」は全然構わないのです。どこにどういう情報を付けるかという、付けただけの情報については本体とは切り離して捉えることができるので、これは無料です。

例えば、コーパスの中のこの単語はこうであるという情報を付けるとか、この文はこうだとか、この発話は誰がしているとか、これはどういう意図で言っているとか、そういう情報を付けていくということは言語研究ではよく行われます。それはやっていいし、付けたデータはむしろどんどん公開してほしいわけです。コーパスの利用価値の向上にもつながるからです。

さらに、もっと高度なものだと、意味情報、全部の単語にこれは食べ物だとか、こういう意味だという情報を付けるとか、統語情報、係り受けや句構造というような文法情報を付けるとか、メタ情報としてさまざまな書籍のレベルの話から人物の話など、いろいろなことを付けていくことも可能です。従って、アノテーションを基盤にしてコーパスに依拠したオープンサイエンスということが可能なのではないかと考えています。そのことをこの後お話ししていきたいと思います。

コーパスとオープンサイエンス

「日本語歴史コーパス」を例にして、コーパスを基礎に、アノテーションということを使いながらオープンサイエンスに近づけるような話をしていきたいということです。

人文学をやっている方、特に言語研究をしている方はよくお分かりだと思いますが、日本語研究をする場合、必ずデータを研究者は持っているはずです。それなしに直感だけで何かを言うことはあり得ないはずで、何かあります。特に今は、先ほどの人数からも分かるように、みんなコーパスを使っているのです。

ですから、コーパスを使って用例を分類したという場合は、研究者の手元には大体 Excel データがあって、用法分類をしたものがあるはずです。ところが、それは個人が持っていて公開されることがないのです。論文にまとめると、論文が完成品なので、基礎となったデータは捨ててしまう、少なくとも出てこないということが非常に多いです。これが非常にもったいないと昔から思っていました。

もう一つは大学院の授業です。演習で何かを読むということをやる場合は、みんなでかなり徹底的に読んで、情報付与のようなことをするはずです。そういったものが消えていってしまうということもありました。

まずは、用法分類のようなデータについて、Excel かどうか分かりませんが、手元にある Excel を共有することができれば、まずは出してもらった研究が検証可能になります。エビデンスが出てくるということになります。もちろん論文の中にも載っているのですが、さらにベースのローデータに近い部分を出してほしいのです。それが出てくると、他の人も共有することができて、新しい研究に利用できるのではないかということです。これをアイデアとして、コーパスをベースにしたオープンな研究を進めていきたいと思っています。このように、コーパスを対象にした研究のデータはほとんどがアノテーションという形でまとめることができるはずなのです。

では、そのためにどういう仕掛けをしておく必要があるかということで、これは非常にプリミティブなところから始めているのですけれども、今までの日本の人文系の研究では必ずしもできていなかったことで、まずはこういうことからやりましょうということで、1 から 5 まで挙げてみました（図 2)。「検索条件式」というのが出てくるのですが、「中納言」で検索するときの式の話、それから、用例にパーマリンクを付け

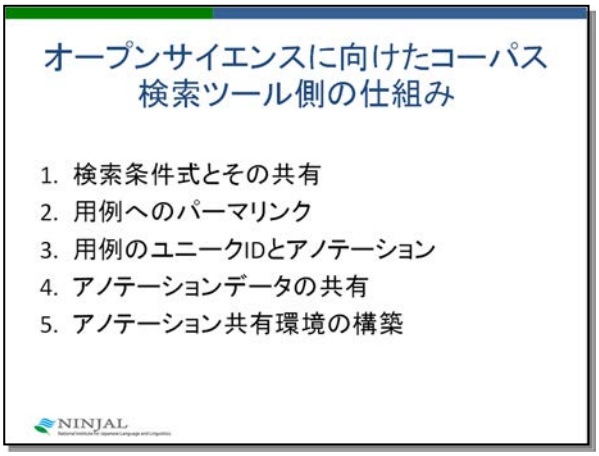
てやる、そして、ユニーク ID を付けてやって、それを使ったアノテーションができるという話になるのですが、順番に見てまいります。

まずは検索条件式です。図 3 は「中納言」というコーパスを検索するアプリケーションで、品詞が形容詞で、活用形が連体形、つまり、形容詞の連体形の次に「言葉」という語が来ている、そういうものをここで探しています。キーが形容詞になっているので、「言葉」というものの前にどんな形容詞が来るか、美しい言葉、優しい言葉、きつい言葉とか、どんな形容詞が使われるかというのを探そうという例です。こういうことが「中納言」だとできます。

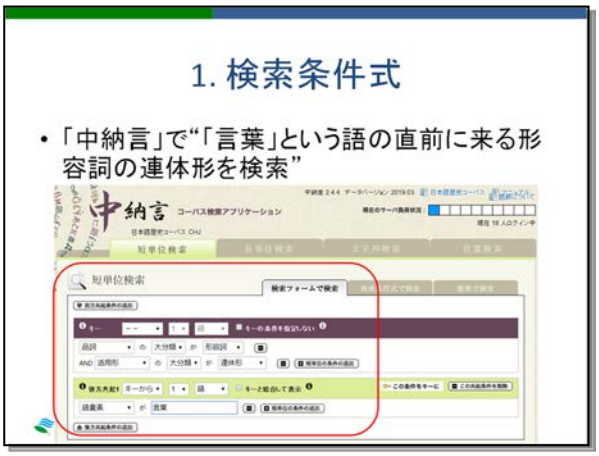
画面で説明するのは大変ですが、内部的には検索条件式というのがあって、これをすぐ表示することができるようになっています（図 4)。一見すると、これはちょっとやっかいなのですが、そんなに難しいものでもありません。コピペしてメールなどで送って、「これでやると用例が取れますよ」とか、「これだとこんな結果だけれど、あなたのと用例数が違うのですが」ということに使えます。研究の再現性、用例の共有のベースになると思います。ですから、まずこれを出すようにしました。

「中納言」の講習会では必ずこの話をして、研究の再現性のために、論文に表示するときには検索条件式を貼りましょう、そうすれば読んだ人が研究を再現できますからとよく言っています。コーパスを使っている時点で研究データの共有は一定のレベルであるわけです。それをどう使ったかという検索条件式が加わると、かなりのレベルでの検証可能性が出てきます。とんでもない間違いをしているというのはあり得る話で、それが抑止できるし、良いものにしていけるのではないかというのが一つ目の話です。これは実際に、かなり以前から「中納言」に組み込んだ機能で、使われるようになってきています。

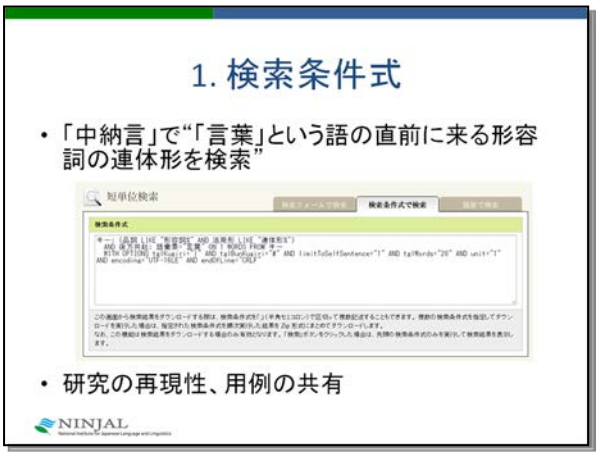
さらに、用例へのパーマリンクを作りました。個々の用例、つまり全部の作品が品詞分解されていますから、『源氏物語』の「桐壺」の最初のところに出てくる「やんごとない」という形容詞に ID が付いていて、それをクリックすれば「中納言」上で表示できるというパーマリンクを用意しました。これがあれば QR コードなどで出せます。「中納言」のアカウントがないといけないのですが、それさえあれば、リンクをクリックすれば用例が表示されます。



（図 2）



（図 3）

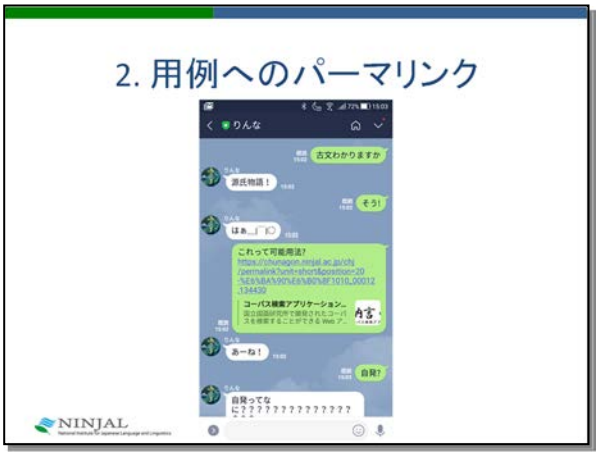


（図 4）

用例が表示されると何がうれしいかというと、まず単語の情報が付いています。品詞など、そういうものが付いていて、文脈が分かるだけではなく、ジャパンナレッジや各大学の原本データへのリンクが付いています。これを使うことで、「この用例なのだけど」という話ができるわけです。SNS でもできるので、みんなやろうよと日本語学会ではよく言っています。

図 5 は LINE の画面ですが、私は友達がいないので、マイクロソフトの AI 女子高生とやりとりをしたところを表示しているものです。要するにこうやると、「この用例はどうなの」という話がお互いにできるのです。この基盤が今までなかったわけです。もしやろうとすると、「『源氏物語』の新編全集の何ページの 3 行目なのだけれども」という話をしなければいけなかったかもしれませんが、これがあれば、そこを出して、「これはこうじゃないの？」という話ができます。基礎になるようなものがまずなければいけないだろうということで、作った機能です。このパーマリンクは何とか維持していかなければいけません。維持するのは大変で、勝手にずらしたりしないようにしなければいけないのですが、そういうものを頑張ってやりはじめました。

それから、今の ID にそのまま使われていることなのですが、パーマリンクに組み込まれているのはユニークな ID です。われわれは「サンプル ID と開始位置」と呼んでいるのですが、これは先頭から何文字目かという情報です。ですから、本が変わらない以上、



（図 5）

ずれないのです。「これは用例のマイナンバーだから、みんな使ってね」ということを日本語学会へ行ってよく説明しています。先ほどの「中納言」の検索結果で表示される部分です。ID さえあれば他の人は表示できるから、文脈や品詞などがなくてもいいからこれは出してほしいと言っています。これがあると何がいいかというと、必要な用法を番号だけ並べてやれば、特殊な用例集のようなものができることです。

ちなみに、先ほどの AI 女子高生とのやりとりは、『源氏物語』の一部分の「海見やらるる廊に」で、「るる」「られる」なのですが、受け身、尊敬、自発、可能のうち、「海を見やることができる」と訳せるから、可能に見えるのです。しかし、中古に可能の肯定用法はないはずだといわれているので、どう考えるかというと、「自然と見渡される」という自発ではないかといわれています。議論の対象になるものをリストアップするだけで、日本語研究者にとって貴重なデータになるのです。

ちょっとおかしな自発用法のリストとか、さらにその横にこれは自発だと一言付けてやれば、それはアノテーションとして使えるようになって、「られる」にこういうものを並べて、受け身、尊敬、自発、可能のどれかを自分で付けましたというデータができれば、結構使えるデータになるのではないかと思います。

コーパス上では助動詞の「る」「らる」という情報しか付けていないのです。それが受け身、尊敬、自発、可能かというのは研究者によってもかなり議論が分かれる部分があるのですが、これを付けてやると、例えば小木曽（2019）のデータで可能用法は何例あるとか、小木曽（2019）のこの ID の可能は間違っているとか、そういう議論につなげていけます。このデータを共有することで、他の人はその次に来る可能用法の動詞だけを持ってくるなど、そういうことができるので、みんなでどんどん共有しようよと言い歩いています。

要は、Excel のデータなのです。日本語の研究者はみんな多分、これを作っていたのです。そうでないと、可能用法が何例あっても出せないはずなので、これを

集計しているに違いないというか、私も実はやっていました。ですから、これを表に出しましょうということを一生懸命みんなに勧めているところです。

できたらそれを公開しましょうということも言っていて、自分もやらないわけにいかないので、隗より始めよということで始めました。researchmap の「資料公開」のところで実際にこのデータを公開してみました。そんなにどんどん利用が進むようなものでもないのですが、まずはそういうところから始めることで、研究用のデータの流通、共有、再利用が進むのではないかと考えています。

アノテーションのデータをオープンデータとして公開しようとしています。アノテーションを引用しようというのも、先ほどから小木曽（2019）と言っているのですが、これを作るのは結構大変なのです。何しろ、平安時代後の全部の「れる」「られる」について、受け身、尊敬、自発、可能を付けるという話になるので、これだけで一つのプロジェクトになってもおかしくないような話なのですが、そうやって作ったデータはちゃんと引用しましょうということも言っているわけです。そうしないと、オープンにしてデータを出してくれる人がモチベーションを保てませんし、作ったデータを再検証していく、再利用していくときに他の人が見られないということです。こうやっていくことで、コーパスをみんなで育てることができるのではないかと考えています。

コーパス本体が必ずしもオープンでなくても、アノテーションをベースとして公開していくことで、それが可能なのではないかということを言いたくて、こんなことを春の日本語学会で一生懸命主張してきて、まだそれほどでもないかもしれませんが、それなりの反応は得ているところです。コーパスを育てていくことをしたい、アノテーションを共有して、学会の共有財産になるようにしていきたい。この作成・公開は研究の業績として正当な評価を得られるようになってほしいということです。

そういう話をしていると、これはもっと簡単にできるようにしないと駄目なのではないかという気がします。ダウンロードしたデータで Excel 上に横に入れてやる分にはいいのですが、それを共有したとしても、それを再利用するのは大変なのです。先ほどの Excel データのような一工夫が要ります。VLOOKUP ぐらいは使えないと駄目、Access が使えればいいのだけれどとなってくると、いきなり人文系の研究者は「それは困る」という話になってしまいかねないのです。だから、それはアプリケーション上で実現したいし、もっとサポートできないかということで新しく始めたいと思いました。

そして昨年の今頃、一生懸命、科研費の書類を書いて、この 7 月に幸い採択されたので、そのお話を最後にしようと思います。

「挑戦的研究（開拓）」というものです。3 年間かけて、「日本語コーパスに対する情報付与を核としたオープンサイエンス推進環境の構築」と、すごく大きく出たタイトルなのですが、つまり、先ほど私がお話ししたようなことで、研究者間のアノテーションとしてのデータ流通を行いたいのです。先ほどから話が出ている「みんなで翻刻」の橋本さん、人文情報学研究所の永崎さん、歴史民俗博物館の後藤さんなど、いろいろな分野の主立った方に入ってもらって始めました。

全く新しいものを作っても難しいので、既に 1 万人、2 万人ユーザーがいる「中納言」に新しい機能を追加するという形で、アノテーション機能を追加するということを始めたいと思っています。本当はもう少し画面など、できたものがあるといいのですが、現時点では、お金が来ただけで、頑張ってこれからやるところなので、まだ何もありません。例えば「中納言」上で間違っている部分を指摘するということでもいいと思うのです。それをきちんと使えるようになれば、「みんなで翻刻」ではなくて、「みんなで品詞分解」という話になります。

それから、先ほどのような情報をさらに追加することもできるといいなと思っています。付けたら、それを他の人が画面表示できるようにして共有できるよう

に、引用できるようにしていく、そういう仕組みを作りたいと考えています。こういうものが実践できるようになると、もっとコーパスを使ったエビデンスベーストな、オープンな研究環境ができていくのではないかと期待しているところです。これから頑張ります。

まとめ

最初、オープンデータの話と言ってしまったのですけれども、実は日本語コーパスはあまりオープンではありません。CC ライセンスで言うと、一番緩いものでも SA が付いているぐらいです。ND が付いているものも多いですし、そもそも CC ライセンスを付けていないものが多いです。それは先ほど申しましたような理由です。

だから、コーパス本体はなかなかオープンとして出せないのですけれど、アノテーションという切り離した関係にはなるのですが、それでオープンサイエンスを推進する基盤となし得るのではないか、それを拡張して、環境の整備を今後進めたいと考えているところです。

●**フロア 1**　勉強になるお話をありがとうございました。東京外国語大学附属図書館の職員です。

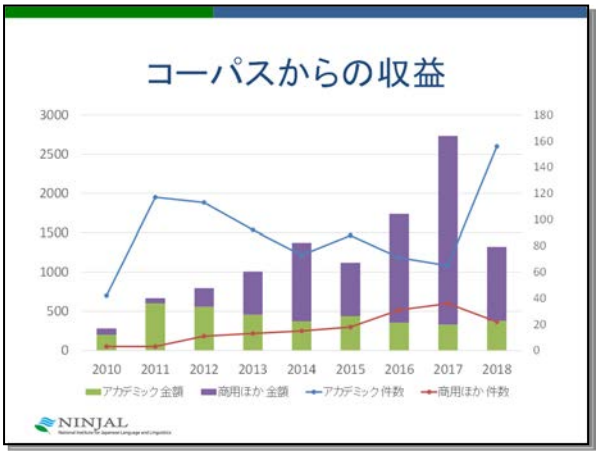
当館でもこの間、大学院生と「現代日本語書き言葉均衡コーパス」の使い方のガイダンスを作ったばかりで、本学の研究分野では本当に欠かすことができないインフラだと思ってお話を聞いておりました。

どうしても商用利用というか、有料化は外すことはできないというお話だったのですが、「コーパスからの収益」のスライド（図 1）を拝見すると、アカデミックユースが一定数であって、最初はアカデミックが多いのですが、2017 年から商用ユースがものすごく増えています。先ほどの登録者のお話でも、2 万人だけれど研究者は数千人しかいなくてというお話をされていました。アノテーションという参加をするのは多分、研究者がメインだと思うのですが、コーパスを育てている中のコミュニティの人からも現行ではお金を取っていて、商用ユースの人からも取っている、多分、価格の差はあると思うのですけれども、そういう状態かと思うのです。

それをせめてアカデミックユースは無料化するということで、アノテーションを付けるコミュニティとしては、一種、国語研究所を超えた、コンソーシアムではないですが、共同体として研究者の世界ではオープン化する、そういう発想はあるでしょうか。

●**小木曽**　実は、先ほどのアカデミックなのに有料という部分は、「現代日本語書き言葉均衡コーパス」と「日本語話し言葉コーパス」だけのケースです。これはどういうものかというと、「中納言」のようなオンラインではなくて、ディスクに入れて生データをそのままお渡しするというタイプのものなのです。これについては配布の費用もかかるということもあるのと、そもそもこれのアカデミックライセンスは相当安いです。しかも、研究室単位、大きな単位で使えるものなので、企業などと比べると相当廉価にはなっていると思います。

そうはいってもオープンにできないかというときに、結局のところ、オンラインに置けないのです。パッケージで配るということにはなってしまっていて、アカデミックにオープンということは完全にはしづらいかと思っているのですけれども、お答えになったでしょ



（図 1）

うか。

●**フロア 1**　ありがとうございました。パッケージのものをオンラインに置けないというのは、容量的に大き過ぎてということですか。

●**小木曽**　流通してしまうのではないかということが基本だと思います。二つのコーパスについてはそのようなことなのですが、近代語のデータなど、一部のものは CC ライセンスで、オープンとは言えないかもしれませんが、公開しているものもございます。

歴史コーパスの方は、また別の理由で完全公開ができなくて、中世以前のデータの大部分が小学館の『新編　日本古典文学全集』のライセンスといいますか、契約の下で使わせていただいているということがあって、われわれにもできないということがあります。同じようなことで言うと、「現代日本語書き言葉均衡コーパス」については、これも原著作者がいるわけですけれども、その人たちに許諾を得たときに有料での配布ということも書いてしまっていて、いろいろと権利関係、原著作権者との関係があるということです。

●**フロア 1**　大変よく分かりました。ありがとうございました。

●**フロア 2**　東京大学の教員です。コーパスで人文系では最大級のデータセットということで、非常に素晴らしい活動だと思っております。毎年 50〜70 本ぐらいの論文が出ているというお話だったのですけれども、これは一体どのようにして実際に使われているかというのを捉えられているか、そのプロセスが分かりましたら教えてください。

●**小木曽**　本当はコーパス利用の契約書の中に、論文を書いたら送ってくれ、少なくとも情報を送ってくれということを書いているのですが、送ってくれません。来ることはありません。また、コーパスのようなものは、特に人文系の研究では参考文献とか、引用するという習慣があまりないのです。本文中に書いてあるとか、脚注に付いているとか、言及があればいい方です。中には、「国語研の現代語のコーパスで検索したところ」などと、固有名詞なしの書き方しかないこともあります。われわれは仕方がないので、基本的には Google Scholar のようなところはもちろん、国立情報学研究所のデータベースの他に、論文集など、そういう主立ったものは見て、それでリストアップするという作業を毎年、年に 2 回やっています。その数字ということになります。