平成 29 年度　東京外国語大学オープンアカデミー

東京外国語大学語学研究所 企画

『コーパスから見えることば・文化・社会』

2017 年 11 月 14 日（火）　第 6 回

「日本語：数億語のコーパスを作って調べてみるとみえてくる頻出語、頻出表現」

東京外国語大学准教授

望月 源

今日で最終回、6回目のオープンアカデミー「コーパスから見える言葉文化社会」と言うことで、私「もちづき」といいます。扱う言語が日本語です。

最初に私についてのバックグラウンドで話したいと思います。現在はこの大学の教員で総合国際学研究院の准教授です。本学での授業の担当は学部、大学院とあるんですが、学部は言語文化学部の担当で、大学院は総合国際学研究科というところでして、それからこの大学のコンピューターネットワークの面倒を見ている総合情報コラボレーションセンターの仕事にも携わっています。何故かというと、大学院はこの学校では数少ない理科系の出身でして、石川県にあります国立科学技術大学院大学という仰々しい名前の学校で情報科学研究科の情報処理学専攻出身でして、1番近いのは佐野先生の自然言語処理、翻訳、要約とかそういう分野の研究をするようなところの出身で学位をとりました。

専門は自然言語処理、それから計算言語学でコンピューターを使って言葉を処理するとか言葉メカニズムを解明するとかから入って、この大学にいますと周辺で皆さん言語教育を盛んにやっていられるので、自然と教育システムとコンピューターと言語教育というところでやっているとか、もともとは情報検索とかテキストの表現とかやってきていて、最近は日本語コーパスの構築というところで、後で出てきますけれども、われわれグループでやっておりますけれども、日本語の字幕、テレビの字幕放送で流れている字幕をデータとして取り合取り出して集めて、最終的にはそれを日本語教育の用例みたいな形で扱うとか、定型表現、そこに頻繁に見られる形式がどんなものがあるかということを研究しているところです。

今日の内容なんですけれども、5回連続でこのような話をなさったと思うんですけれども、それを前提として、多分前の5回とはちょっと系統が違うかと思います。というのも、どんなコーパスがあってそれを使って中でこんな表現がこれだけ見えますと言う話が続いて来たと思いますが、それだけではなくてコンピューターを使ってどうやって作ることができるかという話をします。実際作って集めたコーパスが見つかりますしデータとしてはあるので、その中でどういう見え方がするのかという品質や表現を見ていきたいと思います。

一応簡単なレジュメを皆さんにお渡ししました。細かいデータは載せていませんけれどもおおまかな流れが分かります。

コーパスとはというところで、言語学の世界では事例を見つけたりこんな発見があるとかで出てきたと思いますけれども、私のやっている計算言語学、コンピュータ処理の世界でも昔からコーパスと言うのは言語データとして広く認識され意識されてきました。多分どちらかというとコーパス

的なものが先ではないかなとデータの生まれ方としては思います。そして言語学の方でも採用されるようになってきて、コーパス言語学と言う分野が出てきた順番なのだと思います。私たちが普段喋っているテクストをデータとしてあるいはメールでもブログでもなんでもそうですが普段生み出している普通に使っている生のものを大量に集めて整理して言語データと言う形で利用できるようにしたものをコーパスといいます。ですので、いろんなものが言語データとして考えられるわけですね。具体的にはここに書きましたけれども、古くから使われている者として新聞記事、それから雑誌、小説、それから我々が前集めたものでは教科書のデータとか、ウェブページですね。こういったものはどちらかと言うと書き言葉なので、書き言葉の研究として書き言葉の特徴が出るようなデータとして集めて扱います。会話とか授業等もそうなんですが話をしているものをデータとして集めると話し言葉のデータとか話し言葉コーパスみたいな言い方もします。ここには書いていないんですけれども、話し言葉の場合は音声の状態を頭の状態をデータとして取ると音声認識ですとか、ロボットと話をするためのデータに使えるので、音声データは無声コーパスと言う時もあります。

文字にしてテキストの形で保存するのは、書き起こしと言う作業があります。テープを聞いてそれを文字にするとか、今は音声認識のプログラムが良くなっているので音を直接文字にするとか、それを文字に起こして間違っているところを修正して書き起こしコーパスと言う言い方もします。メディアが音のままなのか文字にしてからなのかと言う違いがありますが、言語の種類としては書き言葉と違うタイプの話し言葉のコーパスとして使います。

こっちが書き言葉、こっちが話し言葉と言うことになりますと、大体その中間的なコーパスとしてよく言われるのが、少しくだけた感じの書き方をするようなタイプのブログとかTwitterとかソーシャルネットワークとか、こういったものは書き言葉と言うより中間のもので、色々と集められています。

書いていないんですけれども、結局言語データとして自分たちが話しをしているものとか、時代を追いかけているものもあるのだけれども、その時の使われている言葉を集めることなので、ありとあらゆるものが対象になる、というところですね。あとで出て来ますが、それぞれ別々に書いてありますが、別の考え方として世の中の日本語として話されている言葉というのは何だ、と考えると、それぞれあり得るので、それをうまくブレンドしていろいろなタイプの言語が1つのコーパスとして集まっているものだと考える。バランスのコーパスと言うのですが、そうするとリアルな日本語の品質とかが分かると言う考え方で集められるコーパスがあります。でも平たく言うと普段私たちが使っている言葉を大量に集めた言語データ、ということになります。なぜこんなものを集めるかと言うと、実際使われている言葉がつまっているので、詳しく調べると多分その言葉の特徴が見えてくるだろうと。使い勝手の良いデータがそこから見えてくるという期待があって昔から集められている。大事なのは、コンピューターで扱える形式のものを今は集めていることです。ここ最近なんですが、機械化の形式で大量に集められてきたのは多分15年20年位のことです。私が学生時代は20数年前なのですが、まだインターネットというものが世の中に普及していない時代で、wwwのホームページで送られるものもなかった。インターネットは商用ではなくて研究者用のネットワークだったのです。その頃はテキストといっても電子化されている情報はほとんどなかったのです。学生の頃はテキストを処理しなくてはいけなかったので上で持ってきたものを高額文字読み取り装置OCRで読み込んでそれでテキストを作ったことがありました。その時代には大量のデータを集め

てと言う話は考えとしてはあったのですが、現実的ではないと言う事でした。そのうち、新聞とか、英語と違って日本ではなかなかデータが集まらない。新聞社は毎日新聞を電子化してワープロなどで売って打って電子データの形で存在していたので、世の中にはまだネットニュースみたいのがないので、紙媒体しかない時代ですがデータとしては電子化されたものがあったので、自然言語処理と言う分野になりますと偉い先生が当時毎日新聞とかに直接掛け合って、こういうデータを研究目的で利用させてもらえないか、と言うような交渉をして、最初は断られるのだけれども、段々認められて今では販売されて研究目的に使えるデータに提供されるようになっています。それを集めてきて、新聞というデータをまとまった形で集める。そうすると大体年間、新聞記事は20万記事位ですか、長いのも短いのもありますがテキストになるのです。10年位集めると2,000,000記事ぐらいになるのです。そういった形なのですが結構大規模で、新聞の中に出てくる言葉の特徴とかが見えてくる。それで翻訳をやったりとか、形態素解析とか、文を単語に区切るという技術が発達してきたという歴史があります。

機械で利用できると良いことがあります。最近は言わなくても当たり前のようになりましたが、検索が楽になったという事があります。コーパスの中にあるものが場所を素早くとらえることができます。機械に読ませられないと、なかなかこれだけを調べるのは難しい。それから数を数えるとか、よく一緒に使われる語とか、そういう情報を取り出して一覧したり確率を計算したりすることができるようになりました。それが非常に有効なために、コーパスが段々作られるようになったと同時に、最初新聞社のデータといった時代から、インターネットが普及するようになって、各家庭に入るようになって今のようにメールを書いたりブログを書いたりと言うことになると、それまでの時代には考えられない位の生産量で言葉が世の中に出されるようになった。しかも出したものが上と違って広く集めてコンピューターで使われるようになった。ということがあってデータが広く使われるようになった、ということです。

データが大きくなってもコンピューターが使えるようになって比較的いろいろな処理ができる。うちの学生さんにも最初の時にこういう話をするのですけれども、例えば今日の新聞記事の中で、車と言う文字が何回出てくるのか、新聞ぽいと渡されて調べてと言われても、まぁ引きますよね。手作業で調べると言う人もいたんですけれども、それがデータの形で今日の新聞の全部のデータがあれば、コンピューターの力を借りれば割と楽にできます。では1年分の記事の中で車が何回出てくるかと言うと、コンピューターの力を借りれば10分で可能な作業です。前にとったデータをちょっと見てみたいのですけれども、手元には古い新聞記事のデータしか持っていないのですけれども、1995年には7,085回出てきて、2000年には7,050回車と言う単語が出てくる。「車」だけではなくて「なんとか車」というのまで入れるとすごい数になる。と言う話ができます。

そろそろ次の話に行くのですけれども、そういったコーパスをもし使おうと思ったら、研究者などがどうやってコーパスを手に入れるかと言うことになります。研究目的の利用で無償提供されたコーパスと言うのは、国立国語研究所が提供しているコーパスがあります。どういう括りとかどういう単位で申請するか調べてこなかったんですけれども、公開はされているので、研究目的であれば使えることになっています。

この環境でやるのは初めてなのでできるかどうか分かりませんが、これが国立国語研究所のコー

パスです。（コーパス提示）こうやってコーパスが並んでいまして、さっき言いましたがいろいろな言葉を集めるというのはここにあると思います。現代日本語のコーパス、バランスコーパスです。バランスよくいろいろな言葉を集めようにしているコーパスです。集める話の前に、結局この中に出てくる実際に使われている言葉を調べようということなので、自分がどういう目的で研究したいかによってそれに合うコーパスを探すと言うことになります。話し言葉の研究をしようとしているのに書き言葉のコーパスは合わないですよね。既存のコーパスというのもいろいろな人が自分の研究目的に合うようにいろいろな人が工夫をして一生懸命集めたものです。

それ以外に、楽天とかライブドアとかリクルートとか、こういったIT系の企業では、自社のサイトにいろいろな商品購入サイトとかレビューとか、利用者が商品の評価を書いているレビューのデータを研究用に提供していたりします。評価する言葉は、どういう言葉があると良い評価になるのか、どういう言葉が入っていると悪い評価になるのかとか、などを研究することに使われるのに提供されています。後は翻訳とかいろいろなものに使うウィキペディアがありますが、ウィキペディアの記事を一括してダウンロードできるサイトが用意されています。そこに行くとウィキペディアの全部のデータが、日本語なら日本語のデータが、一括して自分のコンピューターにダウンロードして使うということができるようになっています。実際これもよく使われていて、同じ項目の日本語とか英語の記事の説明が全く対訳になっているわけでは無いのですけれど、同じことを説明しているので文章が似ていることがあるので、２つを自動的にアライメントと言ってどことどこが似ているところにつき合わせて研究するということもできます。その他、百科事典なので物事の説明の情報源として使われるとか、言葉を調べるだけではなくていろいろな使われ方をしています。こういうのも広い意味ではコーパスです。後は市販されているコーパスを購入するということもあります。さっき言いましたが新聞記事のデータと言うのはもう20年分ぐらいは売られています。また研究目的で買うと高いです。200,000円位します。なかなか個人で買えるものではないですが、新聞社と使用権の契約をして、データにアクセスすることもできます。

それから後はコーパスとして出されているわけではないのですが、著作権が切れていてパブリックドメインになって、青空文庫とかというのもあります。それを使って小説のデータとしていろいろなことを調べたりするということもできます。こんな感じで1つは既存のコーパスを、データを入手して利用するというのがあります。既存のコーパスを利用すると良いこと悪いことがいろいろありますが、良いことはまとまったデータが比較的容易に利用できるということです。著作権処理がされていて、コーパスとして使うには契約をしてルールに基づいて使えば良いのです。ただ有料なものもあって比較的それが高いです。最後にこれが問題なのですが、誰かが用意したので自分の目的に合うとは限らないという問題があります。研究によっては合わないということもありますし、それよりもっとここにあるものを自分の研究に利用した方が良いと言う場合もあります。

今日は自分でコーパスを作成すると言うことで用意してきているのですが、1つは1番大きいものとして、wwwのウェブのホームページをコーパスにすると言うことができます。利点は大量に存在しているのでたくさん集めることができることです。次に、これは他人が書いたページなので著作権がそれぞれあるはずです。初期は日本の法律では研究目的であれ、ダメなところがあったのですが、今はクリアになっていて、研究目的であれば収集して中身を解析して使うということができます。昔はそれが駄目だった時代は、どうしてもそれを使いたいときは日本国内で収集するとまずい

のでオーストラリアで収集したデータを日本に送ると言うそういうことを考えなければいけないこともありました。コーパスは生のデータなので人が書いたものには著作権があります。そこを注意して使わないといけない。使っている分にはいいのですが、それは外で発表するときには気を使わないとまずいのです。

では、ウェブのホームページを使ってコーパスを作るとどんな感じなのかという話をします。順番だと3段階ぐらいあるのですが、最初はもとになるデータをウェブから収集するという作業になります。これを最近流行の言い方だとストレイティング、フローリングと言うものです。収集したデータがテキストというものなんですが中は見たままのデータではなくていろいろなものがついているのですね。なので、そこから純粋に自分が使いたい部分を取り出すという作業です。これがデータの加工とか抽出という作業です。データを整理して単語の数を数えるとかとかのための作業をします。

あとウェブページの構造という話をします。例えばYahoo!ニュースのページでタイトルと内容に自分の目的で取り出すということもできます。1枚1枚のページであればコピーして外にペーストすると内容は取れますが、こんなことをしていたら、何十万、何百万という量は集められないわけです。そういうときにどうするかということですが、例えばこういうところに各ニュースへのリンクがありますね。リンクというのがこういうようになっているのですが、ご覧になったことがありますか？これは設計図といったようなもので、htmlと言うもので、一応プログラミング言語と言うコンピューターが読む言語なのです。これはGoogle Chromeなのですが、この文字がどういう意味なのかということを翻訳できる、解読できるソフトウェア、プログラム言語なのです。解読しながらどんどん書いていくとこういうページができるのです。こういう仕組みなのです。この中に画像ファイルがあって、Yahoo!ニュースなどたくさん入っていますね、一個一個画像をここに貼りこみなさいとか、書いてある。これはトピックのページなのでそれぞれのニュースにリンクと言うもので飛べるようになっています。こういったすべてのページを世界中のリンクから全部収集して、収集したものをインデックスを作って集めたら、データとなります。

画面がこういうものですが実際ファイルとして中身をとってくるとこういうデータです。実はこの中でずっと見ていくと、ここに画像が入ってイメージファイルがあって、見出しの上野のパンダと言うテキストが来て…と言うように、こういうルールになっているのですが、htmlでそれを見極めた上でプログラムを書いていくわけです。特にタグと言うものがありますがhtmlタグで、学生に教えるのですがなかなか理解してもらえないんです。データを抽出するという作業は時間がかかるので今日はやらないのですが、1つ、さっきのパンダのところの同じファイルをデータとして取り出されたとします。それをプログラムで書いてあるのですが、そのプログラミング言語でこう書けばYahoo!のニュースがみんな同じプログラムで書いてあるので一気に取り出せます。ブログについてもいろいろな人が様々な記事を書いているのですが、ブログ自体は構造が同じなので一気にプログラムでデータとして取り出せます。こんな単純なことなのですが理科系の世界です。

次は、もっとちゃんとやるべきなのですが、この中の単語を取り出してみたいと思います。データができたので、ちょっとこれを形態素解析という処理が入って取り出してみます。先程のテキス

トが、左側の部分がこんな文になっているのですが、見えますか？これは日本語の文章を入れると、その中を単語に区切って各単語の品詞の情報とか読みの情報とかを解析して行き、こういう研究を形態素解析といいます。そういう研究成果として一般に公開されたプログラムを使ってそれを処理しています。それで何をしたかと言うと、もともとこの状態です。そこから必要なテキストだけを取り出しました。上でデータを作りました。できたデータを今度は形態素解析という形で単語に区切るということをしました。それで1番簡単なのだと、「上野」という文字が出ていますが、このテキストには「上野」と言う単語が検索できます。これが単語にするという状態なのですけれど、もう1個は、これで単語が分かるので、「上野」が何回出てきたか、「パンダ」が何回出てきたかというのを数えるのもプログラムを書けばできます。

出現が多い順に並べてみますけれども、ワードというのが全部で255ワードあります。Shangというのも数えられます。なんとなくこれだけ見ていてもどういったタイプの文章なのかなぁということが想像つくのが面白いです。画像情報には品詞の情報を用いていますので、似たような名詞だけを数えましょうということになると、似たような名詞だけが並ぶということになります。一般的にはトピックなので、それだけ見ると文章の内容がなんとなくわかる。そういうことを見ていくと似たような名詞を集めると似たような文章が集まるということになります。私が最初に大学院で言語処理という分野に行ったときには、言葉の興味があって行ったのですが、やればやるほど言葉が文だったものが単語にバラバラにされて、それが順番もなくなって数に置き換えられていって、言葉なのになぁ思って解せない部分もあったのですが、実際やってみるとなんかこういうものに結構まとまってくるのですね。だから数を数えるだけでも関心が出てきました。今1単語を見てきましたが、2語連続で出てくる場合があって２語に限らず連続して数えると、言葉の塊が見えてきて、よく出てくる塊が見えてきて、それがよくある表現ということにだんだん結びついています。それでわかるようになってきたという話を延々とやっているわけです。

イメージとしては一個一個はこういう感じなんですけれども、ウェブページからデータがいろいろな形で集まるのが、データが沢山集まって1つのまとまりとしてデータを扱うようになります。ウェブは億の単位ですので大きいのでなかなか個人レベルでは難しいですが、こんな感じで作るのです。

次にテレビ字幕というのに入りたいのですが、知っている人は知らない人は全然気がつかないと思うのですが、字幕放送というのがあります。番組表を見ると、字と書いてあるのは、リモコンで字幕をオンにすると、画面に文字が出るようになりますね。ちょっと古いですがこういう文字が出てきます。このデータを集めているのです。これがどのくらいあるかと言うと、日本の地上波デジタル放送だけでやっているのですが、東京7局キー局がありますが、総務省によると全体の番組の大体60%位が字幕放送になっているようです。この文字なんですが、見ますと、要約とかはあまりされてなくてしゃべっている音の文字がそのまま書かれている。それで、この間までやっていた朝の連続テレビドラマの『ひよっこ』などでは、そのまま訛りの文字が書かれています。テレビなので話し言葉の情報が多い字幕の文字データを収集しています。これもあまり細かい話はないんですけれども、どういうことをやるかと言うと地味な作業が続くんですけれど、番組表を自動的に解析して、字幕の文字が書いてある番組だけをピックアップして、コンピューターを利用して普通のパソ

コンなんですけれども、パソコンの裏にテレビチューナーが刺さっていて、テレビの画像が取り込めるんです。パソコンの上でもプログラムを動かして、自動予約プログラムが動いていて、字幕の文字が書いてある番組のみを片っ端から録画するのです。録画されて記録されたデータの中から、字幕のデータを自動抽出するプログラムを動かしていくのです。7局あるので1つのコンピューターにチューナーが4つあるので、2台のコンピューターを並べてずっと24時間動かしっぱなしで記録しています。コンピューターが大したコンピューターでなくともいろいろプログラムを駆使してやっています。その後ですが、このデータからテレビプログラムのテレビの説明、番組の説明とか放送時間とか、あとジャンルがあるのでドラマとかバラエティーとか、それが書かれているのと、これが画像自体圧縮して映っているんですけれども、JPEGファイルで、、、データがどういう状態であるかという話なんですが1つのファイルの中にテレビ画面の字幕を表示するためのテレビの解像度はどうでとか、字幕がどういうフォントを使うとか、色を何色にするとか、字幕を画面のどの位置に何秒後から何秒表示するとか、そういう情報が全部のっていて、それと実際の字幕のデータが入っているというそういう形のデータなんですね。（画面の説明）実際はもっと細切れになっているのですが、連続している場合はこんな感じです。

7局の字幕放送は月に約4,200番組あります。それで私たちのグループは2012年の12月から収集を開始しまして、いろいろあって安定しなくてコンピューターが落ちたとか、停電に見舞われたりとか取れない時もありましたが、大体平均して月に4,200番組収録しました結果、今もそれを続けているのですが集計がなかなか大変で、2017年3月の古い状態なのですが、その時点で232,000番組分の蓄積があります。それで文に直したら8900万文ぐらいのデータになります。全部テレビからとったデータです。では、さっきの形態素解析で単語で区切る場合、何語ぐらいになるでしょうか？この段階で約9億1300万単語になります。2012年末から2017年まで4年延々とテレビのデータを録りためていって、やっと9億語の規模になりました。今年で10億のオーダーに入るかなということになります。

でもそれの割にはまだ成果が出せていないので、これからなんですけれども、大体内訳をみると公式に番組表に載っているジャンルがあるのですけれど、13ジャンルです。アニメーション、カルチャー、スポーツなどが入っていて、９億のオーダーになります。字数が多いのはニュースとかバラエティーとかが多いのですが、やはりニュースは書き言葉に近いので字数が多いのですが、バラエティーとかドラマは会話をするので話し言葉が多いです。

単語数は数えるのは大変なんですが、数えました。出てきた単語を何回あるかという延語数集計をするやり方は同じです。この流れの中からどういう単語が出てきたかということを見せないと詐欺みたいなので、何回出てきたかと言う頻度の回数が多い順に並べた表がこれです。これはトップ20から飛び飛びになっています。当たり前なんですけれど、「てにをは」がやっぱり多いというのが日本語の特徴であることがよくわかります。内容語ではなくて機能語が多いのは文の区切りを「。」として表示していることが多いので、区切りがついていない「。」が入っていない番組もありますけれど、大抵「。」が入っているので「。」がやっぱり多いのです。文数とは一致しないのは「。」が入っていないものもあるからです。ですからここで見る限りは、今のところこういうものが多くて、あとここにカバー範囲と言うものが書いてありますが、これは結局何かと言うと、全9億の出現のう

ちのどのくらいの割合で埋まっていくかというのを見ているわけです。そうすると、「も」まで入れて20位までの33%、20個の単語で9億のうちの33%となります。例えば出ている単語20個をマーカーでマークしていくと33%がこれにあたります。仮にこれが表自体で全部で64万697と出ているわけですけれど、これで出現数全体で9億と言う数になります。そのうちのトップの10,000種類を集めてくると全体の81%になります。すごく荒く言うとですね、頻出1万語でテレビのスクリプトの81%は理解できると言う話になります。もっと言うと覚えるべき語の順番をという話をするときに、英語の勉強をするときにこの語から覚えると言うようなことがありますが、それと同じようにリアルに実数で実際の4年とか5年分のテキスト、テレビの字幕と言うとこのくらいの一万種類だと言えるわけです。

話としてはここまで申し上げて実際のものを見てみると、40位までで見てもこういう感じで、全然内容は当然出てこないので面白くないかもしれない。唯一ここで「人」というのが出てきますが、これが人なのか何とか人なのかわかりません。40位までで40%埋まっていると言うそのくらいの内容です。

せっかくなので、データを持ってきているのでさっきの内容を見ましょう。さっきこれに基づいて表を作っていたので同じデータです。順位が左側に出て出現頻度が右側に出ています。「日本」が105位で結構日本と言う言葉が多かった。こんな感じですが数の数え方は、単語の数の数え方というのはちょっと難しくて、さっき形態素解析で切った状態で数で数えていましたが、表層の文字のままで出ている状態で活用なら活用したままを数える数え方と、活用しているものを基本の形にまとめた形で数える数え方と大きく2つあるのです。これは表層のままで活用しているものを数えているので、「わかる」が基本の形ですが、「わかって」など出ているままの状態で数えているので10,000といっても、辞書的にまとめるともっとまとまってしまうのです。これはそういう数え方で全部で600,000語位になります。ただこれだけデータを集めてどう使うのかということは悩ましかったりしますが。結構高頻度語はあまり意味がないということもいえます。低頻度の語は滅多に出てこないし、あっても大きな意味はないかもしれないのです。中頻度の語でちょうど使い勝手の良い言葉が並んでいると言えるでしょう。そんな感じでやっているのですけれど、これはさっきのYahoo!のデータで、1ファイルで数を数えるというのをやったように、これの場合どうなるかというと、ファイルの数というとこれは番組の数ですが、232,000の別々のファイルの中にそれぞれ各単語が何回出たかと言うデータを出します。その中の同じ単語をまとめていって、各単語の集計をしてというのを何回かに分けてやりますとこのようなデータになります。

次に1つの単語ではなくて複数の単語で出現数を数えるということをみます。例えば文として「幾重にも芳醇な香りとその味わい」という例では、一個二個と数えてブロックにしていって、ここはどんどんずれていくのですが「広がる芳醇」とかに分けてまとめるのですね。こういう数え方があってエムグラムといいます。エムグラムの上には数を表しています。これはIDナンバーで一個なので1グラム2グラムとあります。12345678というふうにグラムが付いて数を数えるというやり方をすると今度はまとまりでよく出てくる表現とかがわかります。それをやった結果の集計表がこんな感じですが、細かい話で、実際にこの中で4567あたりのこの辺を集計したものがこれです。こんな表

現が上位にきます。あまり面白くないですか。でも重要なものとしては、日本語は文末が重要なので何か意思を表したりとか主張を表したりとか、この文末の形というのがかなり高頻度で固まってくるということがデータからは見えます。単純なランキングで4グラムで「います」というのが、5位のが「しています」というのがあります。もう一個増やして「出ています」とか、もうちょっと広げると6とか7になります。これは特殊なんですが、NHKの「ナイナイバー」と言う教育テレビの番組があるのですが、「ナイナイバー」が回数が多く出ていますがこれは、これは一般的にこういうような沢山あることは無いので番組のバイアスがかかっているといえるでしょう。それ以外は、普通の会話の中で文末に出てくる普通の文末のパターンであると思います。

なんでこれをとっているかと言うと、日本語教育の教材をここから取り出して使おうということを考えているのです。よく使う表現というのが単語とか言うのではなくて実際に出てくる塊で、前に出てくる表現はどういうものかと言う事がリアルにわかります。実際に出てくるものを分類してデータとしようということです。たとえば6グラムのところで「されています」というのが、ではその前に何が来るのかというのを調べてみると、「期待されています」「展示されています」「掲載されています」「販売されています」「目撃されています」「公開されています」などと、これが何百万もある中でこれだけしか出てこないのですよね。意外とバリエーションが多いのだからそんなに単純ではないんだなと言うぐらい価値があるのですけれど、これは定型表現とかという言われ方をする決まり文句とかを探しているので、「されています」とくれば、前に期待なのか展示なのか掲載なのか、が多いといった具合ですが、ちょっとあんまりピンとこないですかね。

それから「こんにちは」を含む表現と言うのを取り出して調べてみると、「どうもこんにちは」とか「皆さんこんにちは」と言うのが多いです。「皆さん」の漢字とひらがなの違いはありますが。「すみません」は、「すみません失礼します」「すみませんでした」とかですが、あまり回数は多くないのですがそういう表現が出てきます。それを数字があるのですがこのデータを用意するのに結構時間がかかってしまったのですが、このようになります。「よろしくお願いします」とか「ありがとうございます」とかこういう表現が多いです。

ある部分に注目すると、とても良い感じで頻出の単語でそれさえ覚えれば使えると言う単語が出てきてくれて、これからここから覚えていこうみたいなリストが簡単に作れたら最高だったのですが、実際にこれだけのデータをとって見てみると、バラバラでどう見たらいいんだろうということで悩ましいものがいっぱいあるということで、それを含めて研究というのはなかなか簡単にはいかないということですね。それをどうやって捉えたらいいのかなという試行錯誤です。単純に数を数えるだけでも非常に大変な作業ですが、プログラムでこんな感じでやっています。

これだけだとピンとこないと思うので、最後のほうにちょっと関係付けると、自然言語処理のスクールで非常に有名なものがありまして、今のようにデータをたくさん単語区切ったものを使うと、単語と単語の間の関係性を計算するというのがあるのです。そうすると類語の類が出てくるので、今日はちょっとそれをお見せしようかと思います。最初は、字幕の単語単位で4年分ぐらい、ここに例えば全ジャンルをまとめたデータですけれども、「アメリカ」を入れると、アメリカと意味や用法が近いものが出てきます。それを取り出すプログラムで元のデータがコーパスです。試したことがないのでどんな感じになるでしょうか。これは距離計算をしていてたまたま近い順に出てます。韓

国が近いとか、この数字が距離で、これはどういう計算かと言うと、「アメリカ」という語が出て文の周りにどんな語があるかということを調べています。アメリカという語の周りに出てくる語と、韓国という語の周りに出てくるものの数値が近い、出方が近いということです。もう少し正確に言うと、アメリカと言う語の周りに出てくる語を数字に置き換えてあるのですが、ベクトル表現で、それと同じような出現傾向にある別の語が近くになるような計算をするのですが、ざっくり言って国の名前とかが出てくるのが、やはりその言葉の周りに出てくるのが自然だと言う使われ方をするのです。何か入れてみましょうか。大学とか高校とか、これは仕組みとしては知能とかが入っていないのですが、たださっきのようなデータを大量に集めて文章としてデータを集めると、そうすると大学と言う単語を含む文と周りの文にどういうものがあるか、そういうもの同士で同じような周りにある語の出現傾向を調べて、その場合だと「高校」というのが大学と近いことがわかります。ここでは大学とは何かということを教えているものではなくて、たくさん集めたテレビの字幕データの中に入っている文字の出現の仕方だけを使って計算したらこれだけのものが近いものとして出てきます、ということです。実はこれが今衝撃を与えてすごくこの情報というのが盛んに使われるようになっています。これだとわからないのですけれど、計算する場合に大学なら大学というものがどういうものであるかということを、図がなくていうのは申し訳ないんですけれど、ある言葉を表現するために100個の何かというのがあるとします。その100の何かというのが数字に全部なっています。それぞれが何かを表す数字が100並んでいます。それで高校を表すものが100個並んでいます。その100個の同じ場所同士がどのくらい近いかということを、お互いの距離の近さを計算すると、他の単語もそれぞれ100個の軸を持っているので、そのたくさんあるものに対して大学に1番近い傾向を持っているのが高校だ、ということがわかります。今これが、文字が100個の数字に置き換わっているというイメージですが、何かわからないけれども大学を表す100個の数字なのだということで、それぞれ何かの側面を表している数字が並んでいるという感じです。それが今流行の人工知能のdeepラーニングの研究になりますが、それが今までは100個の数字が並んでいるというよりは、「高校」と近い言葉が並んでいたのです。近い言葉が多いとその２つの言葉が近いと言うことになっていました。これは記号処理と言うのですが文字と文字同士の近さを使っています。それに対して今は、何かわからないけれども大学を表す何かの数字が並んでいる。高校なら高校という言葉を表す何かの数字が並んでいる。数字と数字の近さを計算しようという話なので、記号の処理から数字の処理に変わったのですが、それによって何が起きたかと言うと、あのGoogle翻訳のように、文字としては違うものなのにそれに近いものがすごく出やすいようになっています。属性が数字になっていて、その数字と数字を組み合わせて何度も何度も繰り返して、良い組み合わせを作るというようになっていて、その出来上がったものもすごく近いものになるようになっています。それにも結構このプログラムが使われています。

こういう話をしていると、どんどん文系の世界から離れて行きますね。私もなかなかついていけないみたいで、もともとこの分野は文理融合と言われていていますが、最近は融合しにくくなってしまっています。

もう一つ今のもので、単語単位ではなくてもっと長い表現になっている物を扱います。これはドラマのジャンプだけになっていますが、「こんにちは」「お疲れ様です」とかと言うのがドラマの中でどういう風になっているのか。それから、「どうも」というのにつながるものが沢山出てきて、

どういうつながりで使われるのか、意味ではなくてつながりを見ます。「どうも」を説明するのにどのくらいのバリエーションを考えれば良いのか？「どうぞ」もたくさんのつながりのバリエーションがありますね。これもジャンルが変わると見え方が違うのです。同じ「どうぞ」でも、ドラマでなくて情報番組系のものに変えますが、上がドラマでセリフのようです。下は情報番組なので「ご参考になさってください」とか「次回もご覧下さい」とか、こんな感じです。場面とかジャンルが変わると近い言葉も変わっていく、ということが沢山データ集めると使い分けのようなものもわかります。

それで何をやっているかというと、ジャンルとか場面ごとに言葉が違うということを取り出して、最終的にこれ自体が狙っているのは、Can-Doのように、これを覚えるとこういう場面でこの言葉が使えますよ、とか、レストランで食事を頼むときに使うスクリプトとか、ある場面で何を目的にするとこはどのように喋れば良いかということを語学教育で行われることが多くてそれは標準になっているのですけれど、リアルにいろんな場面でいろんな言葉の会話がこのデータの中に入っているので、その場面ごとにどういう表現か、ということを調べて、Can-Doというのを使うときの事例集に結びつけようと思ってやっている最中です。

このデータは他ではないので、私たちのグループでやっているのですけど、これだけ字幕データを扱っているところは日本語では見たことがないので、他ではやっていないことをやっているのですが、まだ成果を出していないので、これからもうちょっとちゃんとデータを出さなければいけないのです。（完）