コーパスの普及と日本語研究の動向

著者	山崎 誠
雑誌名	文化情報学
巻	9
号	2
ページ	32-45
発行年	2014-03-31
権利	同志社大学文化情報学会
URL	http://doi.org/10.14988/pa.2017.0000014570

Journal of Culture and Information Science, 9(2), 32-45,（March 2014） March 2014

講演記録

同志社大学文化情報学研究科共通シンポジウム

日時：2013 年 11 月 13 日㈬　午後 4 時 40 分から 6 時 10 分
場所：同志社大学京田辺校地　夢告館 102 教室（ＭＫ 102）

「コーパスの普及と日本語研究の動向」

国立国語研究所准教授
山崎　　誠 氏

いまご紹介いただきました国立国語研究所の山崎と申します。どうぞよろしくお願いいたします。過分なご紹介をいただき、ありがとうございます。今日はこちらにお招きいただいてこういう話ができるのを非常に喜んでおります。

ただ、皆さんにお聞きいただくようなきちんとした内容ではなくて、どちらかというと、最近思いついたこと、あるいは、いままでやってきたことをベースにして、もしかしたらこういうことが言えるかもしれないという、研究に関する大きな、俯瞰的な状況を皆さんにお伝えして、これから若い、次世代を担う方とともに次の研究のあり方を考えていこうかなという思いでこれをつくりました。すごく大きなタイトルになっていますので、看板倒れになっていることを恐れます。

それでは始めたいと思います。

発表の流れはこのようになっています。概要をご紹介して、コーパスの普及ということについて簡単に触れます。さらに具体的に『現代日本語書き言葉均衡コーパス』がどう利用されているかということをご紹介いたします。ここまでが前半部です。

後半は、定性的研究と定量的研究という 2 つの研究スタイルについての話になります。この 2 つは、べつに言語研究には限りません。ほとんどの学術研究にあり得るスタイルかと思いますが、そのやり方、方法について考えるというのが後半になります。特に方法論の比較ですとか、それを統一、融合して、どのように持っていったらいいかという、そういうようなことが後半のテーマとなります。

こちらが概要です。大きく 4 つに分かれます。日本語研究、一応私は日本語のことしか分からないので日本語に限定しますが、日本語研究では、最近コーパスを利用するということがけっこう普及してきて、ごく当たり前のようになってきました。ただし、少なくとも書き言葉のコーパスを利用する場合は、用例を調べたりするなど、単純な量的把握のようなレベルにとどまっていると思われます。「思われます」としたのは、きちんと実証的に調べたわけではないからです。私の印象です。ただ、それほど外れてはいないような気がします。

こちらの文化情報学部文化情報学科は、むしろその例外であって、金先生をはじめとして、高度にコーパスを利用されていると思いますので、皆さんのところは例外だと思ってください。一般的にはこういう傾向にあるということです。

ということは、せっかくコーパスは普及したのだけれども、本来の研究スタイルである定量的研究の普及には結びつかなかった。少なくともいまのところ結びついていないということが言えるのではないかと思います。将来は、定量的な研究と定性的な研究をバランスよくおこない、そのために両者をつなぐ理論的な整備というものが必要ではないかと思っています。

この最後の部分ですけれども、一応日本語ではというふうに始めましたけれども、英語研究などでも若干こういうことが言われているように思いますし、この量的な研究と質的な研究というのは、日本ばかりではなくて、海外でも溝があると言われていて、昨年でしたか、国際計量言語学会

というところで会ったヨーロッパの先生方のほとんどが同じことをおっしゃっていました。つまり数理的研究と一般的な人文研究とは、どうも人のレベルで合い入れないということがあるということが世界的な傾向のようです。それをなくしていくためのヒントを一緒に考えられればと思っています。

それでは、最初にコーパスの普及ということについて申し上げます。多分皆さんも、日本語学の入門書をお買いになったり手に取ったりしたことがあると思うのですが、最近出た本を見ますと、例えば、『はじめて学ぶ日本語学』という本には、1つのチャプターとして、「コーパス日本語学」ということが載っていたりします。

またこれは、2012年に出た『私たちの日本語』という本には、「言いません」と「言わないです」をコーパスで調べるというような1つの内容が載っていたりします。2010年の『日本語教育研究への招待』という本にも、「コーパスを使った文法研究」というチャプターが現れています。

日本語学も日本語教育も、その必須の要素としてコーパスというものを入れてきているというのが最近の傾向だと思います。もちろん、本を出すには一定の時間がかかりますので、これより1年か2年前にはこういうことが計画されていたということになります。

では、もう少し実証的なレベルで、研究文献においてコーパスというのがどのくらい利用されているかということをこのグラフで見てみます。

ここに、2つのカーブがありますが、赤いのは「日本語研究・日本語教育文献データベース」によるもので、国立国語研究所の情報研究資料センターがつくっているものです。昔の方は、『国語年鑑』という本を見たことがあるかと思います。その『国語年鑑』を増補して、日本語教育の内容も含めたものがこちらのデータベースです。

もう1つ、国立情報学研究所が提供している文献データベースで、「CiNii」というものがあります。この2つで、タイトルにコーパスが含まれている文献を調べました。古くは1980年ぐらいに1つあるのですが、これはいわゆるごみでして、ここに出てくる「コーパス」は会社の名前です。ここにも小さな山がありますが、これも、法律用語のコーパス。「ヘイビアス・コーパス」かな。これは言語のコーパスのことではないのですね。言語のコーパスがちゃんと出てくるのはこのへんからです。

このカーブが両方とも1990年代に大きく上昇していて、2000年代でさらに上がっているということが分かります。2005年ぐらいでしょうか、大きなカーブがあるのは。ここで多分「日本語話し言葉コーパス」が公開されたということに関係しているかもしれません。最後にこう下がっていますのは、これはまだデータの入力化が進んでいないので最後の1、2年は無視していただくとして、平坦なのか、あるいは下がっているのかちょっと分かりません。これから、どういうふうになっていくのかがちょっと懸念される気がします。

それから、「コーパス」をタイトルに含む文献が言語研究でどのくらいの位置を占めているかということですけれども、こちらは、「日本語研究・日本語教育文献データベース」で、2000年以降の文献を対象に適宜選んだ検索語をタイトルに含む文献を調査したものです。

適宜選んだというのはどういうことかと言いますと、例えば「方言」とか「意味」とか、研究テーマとして非常に大きな領域カテゴリーをつくっているようなもの。それと、研究領域のなかでさらに「名詞」とか「格」とか、ある程度研究の数が多いと思われる、そういうものを任意に選びました。ですからこれはあまり客観的でありませんけれども、これを見ますと、一番多い「日本語教育」を筆頭にこういう順番に並んでいます。

上のほうはさすがに、研究領域そのものを表すものが多いのですが、このへんにコーパスが出てきます。数で言うと、「副詞」と同じぐらい。「音韻」よりは多いという、こういう位置になります。研究領域と比べると確かに下のほうですけれども、研究のキーワードのなかではかなり上位になります。

「副詞」を研究していますというのは、日本語研究者以外でも分かる表現ですけれども、「コーパスで研究しています」、「コーパスを使って研究しています」というのは、多分日本語研究以外の人には分かりにくいようです。ちょっとこの間には格差があるのですけれども、数的にはかなりメジャーな領域になっているということが言えると思います。

実は同じことが英語研究でも起きています。これはサンプソンという人が調べたものですけれども、アメリカ言語学会の機関誌『Language』を1960年からずっと調べたものです。

なぜこの人はこういうことをやったかというと、このサンプソン自身は、コンピュータ言語学の人ですけれども、生成文法の時代を経てきていて、最近、実証的な研究が根づいたのではないかということを『Language』を使って実証しようとしたわけです。確かに、この1970年代、生成文法が隆盛だったときは実証的な研究の割合が低かったのですが、やはり1990年代から上昇してきて、いま、2011年は8割を超しています。この8割というのは、1950年度のラインを上回っています。なぜ1950年をこの人が選んだかというと、生成文法が登場する前のベースラインがここだということで選んだそうです。

ただし、この実証的研究が何を指すかというのは、かなり恣意的なところがあると本人も言っていますように、音韻研究ですとか歴史研究などはニュートラルな研究だからこれはカウントしないですとか、1つの文献のなかでも、実証と実証でないものが混ざっているのは自分の基準でどちらかに決めたと言っています。

そういう難しいところがありますけれども、アメリカを代表するような言語研究誌でも、かなりの文献が実証研究で、かつ、実証研究の多くはコーパスベースだということになっていると言われています。日本でもこれに近いような状況に、将来なるのではないかという気がします。

いま、現在の状況をお伝えしましたが、コーパスがこれぐらい普及するまでにはどういう歴史があったかということを簡単に述べたいと思います。

「コーパス」という名のつくデータは、私の知る限りでは、日本では「京大コーパス」すなわち、「京都大学テキストコーパス」が最初ではないかと思います。もし違うということをご存じの方がいたらお教えください。

「京大コーパス」は、1995年発行の毎日新聞の記事4万文に形態素情報などをつけたものです。もちろん著作権の問題がありますので、新聞データ記事は利用者自身が買って、そのタグだけを利用するということになります。これを開発したのは自然言語処理の方たちですので、いまでも自然言語処理の研究ではよく使われています。たった4万文ですけれども、係り受けの解析とかには、数的にはちょうどいいということです。

それをはるかさかのぼること50年ぐらい前、国語研究所では、設立当初から実態調査をおこなっていたという歴史があります。ただ、「コーパス」という名前は、ここでは一切使われていません。

なお、ここで書いてありますように、朝日新聞ですとか婦人雑誌、総合雑誌、雑誌90種とか、主に一般の人、多くの人が目にする媒体を中心に10万語から20万語、多くて50万語ぐらいの調査をおこなっています。この赤く書いたのは、「雑誌九十種の用字用語」です。これは、いわゆるこの当時のこういう調査・研究としては一番名の通ったもので、レベルの非常に高い、いま考えてもレベルが高い分析をおこなっていました。ただ、「コーパス」という言葉は、ここでも一切使われていません。

同じように話し言葉についても、『言語生活の実態』、これもかなり古くて、もう半世紀も前ですけれども、一般の人、農家の男性とか商家の主婦に、一日中付き添って会話を記録するということをやっていました。これはテープレコーダーではなくて、筆記です。このように24時間付き添って記録するということをやっていて、その結果が報告書になっています。これは、国語研のホームページで報告書が公開されていますので誰でも見ることができます。

こういう24時間型の調査というのは、山形県の鶴岡市ですとか、島根県松江市などでもおこなっていて、それなりの成果を上げたのですが、これ以降は、この道は途絶えてしまいました。こういうことを、いまこそおこなってもいいかと思いますけれども、個人情報のような問題があってなかなかできにくいのではないかと思います。これもかなり、コーパス研究に近いような研究だったと思います。

もう1つは、これが文系的な研究の典型的なものですけれども、『現代語の助詞・助動詞』とか、『談話語の実態』。これは話し言葉ですね。あとは明治期の新聞、文学作品など、いろいろなものを大量に収集して、それを記述するというような研究をおこなってきました。ある意味では、これなどもコーパス研究に近いものだったと言えます。

年代がみんな、1950年代あるいは1960年代であるということにご注目ください。先ほどもそうでしたけれども、1950年代から1960年代が、こう言ったら変ですが、国語研究所の研究活動の最も盛んな時期だったと思います。私が研究所に入ったのは1980年ですから、かなり停滞してい

た時期です。

この時期、1960 年代ぐらいまでについて、「第 1 期コーパス日本語学」というふうに名称をつけている人もいます。もちろんこれは後付けですから、そう呼ぶ必要はないのですけれども、データを集めて分析して、定量的な研究をおこなっていた時期は、日本では、1950 年代に既にあったということになります。

こういう一連のコーパス言語学的研究が世界に認知されていなかったのは、多分英語での発信がなかったことが一番大きいかと思います。いまご紹介した報告書はすべて日本語だけで書かれていて、英語にはなっていません。そのような観念がなかったわけですね。英語を母語とする人に発信するという考えはありませんでした。

もう 1 つ大きな点は、データを公開しなかったということです。つまり自分たちだけでデータを独占して使っていて、その結果を出している。とすると、外の人は一切使えないわけですから、出てきた結果だけを受け取る。そうすると、研究の世界、学界への普及とか研究の発展というのはかなり制限されます。この点が、コーパスと呼べない大きな理由の 1 つになっているかと思います。いまのコーパスは基本的には公開が原則ですので、そのことによって研究全体を盛り上げるという機能があります。かつてはそれが果たせなかったということになります。よって特殊な研究という立場に置かれていたというような感想を持つ方がいることになります。

時代は少し下りまして、1980 年ぐらいから、いわゆる内省をもとにおこなう現代語の理論的研究が主流となってきます。これは、いわゆるチョムスキーの生成文法の影響を受けて、日本でも、特に現代語の研究の中心は文法研究ですけれども、そこでデータによらない研究、自分の頭の中にある理想的な発話に基づく研究というのが主流になってきました。そのことによってデータではどうなっているかという、検証する必要が意識されにくくなっていました。

もう 1 つは、古典を研究する人は、自分で古典語をしゃべるわけにいきませんからデータが必要なのですけれども、大規模なデータを構築して共有するという発想はありませんでした。古典では各種の索引が若干データ共有化を役目を果たしたかもしれませんが、それ以外の研究では、自分の発話だけで研究ができてしまいますので、大きなデータをみんなで使うということは考えられなかったのです。

最後にコンピュータを扱える言語研究者が限られていたということがあります。いまも、もしかするとその可能性はあるのですけれども、当時は、パソコンは購入するとただの箱で、そこにベーシックなどの言語を入れないと動かない時代があったのですけれども、そこのハードルがかなり高かったと言われています。いまももしかすると便利なツールですけれども、プログラムを組む人は少ないかもしれません。

そういう時代を経て、外圧ではないのですけれども、英語を中心としたコーパス言語学の発展、あるいは自然言語処理におけるコーパスの利用などに刺激されたり、触発されたりして、いわゆる人文学と呼ばれてきた人たちも、1990 年代以降、コーパスに対する関心や期待も高まってきたというような状況だと、リアルタイムで経験した身としてはそのように思います。

さらにそこにパソコンが普及してきたり、新聞記事のデータベースが売られるようになったりしました。『新潮文庫の 100 冊』をコーパスとしてみなすということが盛んになってきました。この『新潮文庫の 100 冊』は、いまも日本語学会に行くと、一人ぐらいはこのデータを使う人がいます。わりとポピュラーなデータですけれども、あまり好ましくないと言いますか、コーパス的なバランスが取れていないデータだと思います。

これは、私がかつて調査したものですけれども、1990 年代の現代語の文法研究がどれくらい実証的なデータを使っていたかということです。つまり、自分の頭でつくった作例ではなくて実例を使っていた文献がどれぐらいあったかということです。

この「80 ～ 100」というのは実際のデータを使った文献がこれだけあったということです。一番多いのは、0 件です。30 何件は、すべてデータを一切見ずに、自分だけで例文をつくって、これは正しいとか、言えるとか言えないとかということをやっていたのが一番多かった。「0 ～ 20」、ここすごく多いですね。50 を境にすると、約 60 数パーセントがこういう実証によらない研究、自分の頭だけで考えた研究ということで、1990 年代はどちらかというとこれが、日本では盛んな時代だということが分かります。

いまこれを同じことをもう一回やってみようか

と思うのですが、まだ時間がなくて、できていません。

いまコーパスのことをいろいろと申しましたけれども、国語研究所関係のコーパスというのは、全部で 6 個ぐらい主要なものがあります。公開の年代順に申しますと、「日本語話し言葉コーパス」が 2004 年、「太陽コーパス」が 2005 年。そのあと「近代女性雑誌コーパス」。これは『女学雑誌』とか明治初期の雑誌を収録したものです。2011 年、ちょっと間をおきまして、「現代日本語書き言葉均衡コーパス」が公開されました。そのあと「明六雑誌コーパス」ですとか、「日本語歴史コーパス」、これは昨年の暮れに一部先行公開というかたちをとりました。

「日本語歴史コーパス」の対象年代は長く、『万葉集』から江戸時代ぐらいまでありますけれども、ここで先行公開したのは平安時代だけです。『源氏物語』を中心とする平安時代の作品です。語数はそんなに多くないのですけれども、いままでは 1 つ 1 つ、作品で見る、あるいは国文学研究資料館の「古典データベース」で見るとかいうことだったのですけれども、この「日本語歴史コーパス」の特徴は、既に品詞分解、形態素解析が済んでいますので、「思は（ない）」「思へ（ば）」などの活用の違いを無視してひとつの語彙素「思ふ」として一度に全部検索することができます。活用形を指定することもできますし、その前にどういう語がくるか、後ろにどういう語がくるかということも分かるわけです。

古典は、現代の文献よりも非常に難しいと言われています。原文には「。」や「、」がありませんから、それをどう処理したのかということを多分、皆さん疑問に思うかもしれませんけれども、そういう複雑な本文校訂を避けるために、この「日本語歴史コーパス」は、小学館の『日本古典文学全集』を定本にしています。つまり、誰かが校訂したものを電子化していますので、まずその文の認定などの問題は、一応クリアしたということになります。そうしないと、延々とそこで文献の本文批判で終わってしまうことになりかねませんのでそういうスタンスをとっています。

ここから、ごく簡単に、先ほど挙げました『現代日本語書き言葉均衡コーパス』の利用についてご説明します。

略称は「BCCWJ」と言いますけれども、全部で 3 つ公開形態をとっています。オンラインによる無償公開。「少納言」と名づけています。オンラインによる有償公開、「中納言」。有償と書きましたけれども、実は当面無償です。お金は取っていません。3 つ目が、DVD による全文公開です。

つまり上の 2 つはオンラインですから、テキストが全部手に入るわけではなく、検索した結果だけが返ってくるということになります。DVD の場合は、約 1 億語の全体のテキストを自分で検索したり、分析したりすることができます。

これは「少納言」のこのアドレスをご覧いただければどなたでも試せます。これは実際に検索した結果です。コーパスという語がこの中には 4 件入っているということです。

これは、ちょっと経緯を申しますと、2007 年に既にネット上での検索サイトを立ち上げていました。このときは「少納言」という名前ではなくて別の名前でしたけれども、白書と Yahoo! 知恵袋で 1,000 万語だけでしたけれども、毎年語数を増やしまして、2011 年 3 月 13 日に 1 億語に達しました。この微妙な日付をご覧ください。3 月 11 日に震災が起きています。そのときに実は外注業者の方に無理して作業していただいて、その 2 日後に公開したということを思い出します。

この名称ですが、なぜ「少納言」にしたかと聞かれるときがあるので事情を申しますと、研究所の中でコーパスをつくって運用、管理していたシステムを「大納言」と名付けました。それが最初にあったものですから、このコンコーダンサーを「中納言」、さらに下位のものを「少納言」と、そういう三段階のレベルに分けるということになりました。

この「少納言」の特徴は、媒体、ジャンル、期間を指定できることです。つまり、書籍とか新聞とか雑誌とか、ジャンルはその媒体によりますけれども、哲学の本であるとか、芸術の本であるとかというのを指定して、何年に発行された本かということを指定して検索することができます。

また、前後の文脈に簡易な正規表現を使うことが可能です。ただし検索結果は 500 件までしかありませんし、ダウンロードもできません。その代わり、申し込み不要なので誰でも利用できます。この利用方法は著作権者の方に了解を得ています。

この運用はかなり時間がたちますので、次のような集計が可能になりました。先ほどの公開日から今年の 7 月まで集計した結果、117 万件余りの

検索がありました。これは、誰かが 1 回検索すると「1」と、そういうふうなカウントになっています。全部で 70 の国と地域からアクセスがあったということが分かっています。

検索の月ごとの増減がこのグラフです。3 月、4 月、5 月、6 月、7 月、ここがピークになって、8 月が下がる。そして 11 月、12 月、1 月、こういうふうなでこぼこになっています。夏休みとか、休みの時期に下がり、そして、その前に上がっているということは、多分これを使って何か期末のレポートを書く人がいるということなのかなという気がします。多いときと少ないときとでは 2 倍ぐらいの開きがありますので、これだけ大きくずれるということは、例えばそのような理由があるのではないかと思います。

それから、どのへんの国や地域の人がアクセスしたかというのがこちらです。もちろん一番多いのは日本からですけれども、次は、中国、韓国、台湾、アメリカ、ロシア、フランス、イギリス、タイ、香港。このへんは日本語教育が盛んだということで何となく分かります。ウクライナとかポーランドもあります。トルコは日本語教育がさかんです。どうしてこの国がというようなところがあったりもします。70 の国と地域がありますので、下のほうはもっといろいろになっていますが、IP アドレスによる限りはこれぐらいのアクセス数があったということが分かっています。

もう 1 つ、「中納言」というものがあります。これは、先ほど有償と申しましたが、こちらは当面無償で公開されています。ただし、ここにありますように、利用申請をしていただいて、いまの段階だと、申し込み書とか契約書とか、紙を書いて郵便で送るということになっていて、ひと手間かかります。そこはちょっと改善の余地があるかもしれませんけれども、いまのところは、このシステムを採っています。

といいますのは、これは、先ほども申しましたように、著作権者の方がデータを提供して下さり、著作権者の方に許可を得てデータを使っていますので、利用許諾を得た使い方以外のことはできません。つまり、国語研究所がこのようなオンラインで検索できるようなことをするということで許可を得ています。その同じ検索した結果を第三者が大量に収集して、そこでまた別のデータを公開するというようなことがあっては困るので、一応、少し敷居を高くしているということです。

「中納言」はどなたでも申し込めますので、いまこちらにいらっしゃる方が手続きをしてくだされればどなたでもお使いになれます。

もう 1 つ、DVD というのがございます。これにはテキストが全文入っています。テキストは主に 2 つの形式で記述されています。ひとつは XML というタグを多用した形式でデータを格納したもの。XML は、2004 年に公開した「日本語話し言葉コーパス」でもその方式を使っていましたけれども、まだ普及しているとは言い難いのです。特に個人レベルで XML をきちんと書いて処理する人は少ないので、そのままだと埋もれてしまう可能性があります。そこで TSV の形式、すなわちエクセルで読み込めるようなデータに落としたものも同時に入っています。

ただし、こちらのデータは行数がすごく多いのです。1000 万行以上のデータもあります。そうすると、到底エクセルでも読めませんので、それを一度データベースとして読み込んで使う。あるいはそれを自分で、ruby とか perl というプログラム言語を使って検索するということになってきます。その意味では少し、まだハードルも高いのですけれども、先ほどご紹介した「少納言」とか「中納言」ではやはり限界がありますので、自分でカスタマイズした研究、分析がしたいという方は、こちらを手に入れて、目的に合ったデータを取得して、そこから分析をするというのが理想ではないかと思います。

この場には留学生の方もいらっしゃると思うのですけれども、いまご紹介したコーパスの公開形態が日本以外からどれぐらい利用されているのかということを調べたのがこちらです。国内と国外に分けますけれども、このなかには国内にいる留学生の人もいますし、海外の日本人もいますから厳密には分けられませんけれども、内外で分けるとこのような割合になります。

「少納言」というオンラインの下位コンコーダンサーでは、約 4 分の 1 が海外からの利用です。「中納言」の場合は 5 分の 1。DVD を買った人、これが 10 分の 1 ぐらいです。この比率が高いか低いかはちょっと何とも言えないのですけれども、先ほど申しましたように、ここの研究所はあまり海外への発信というのは多くありませんでしたので、これも特に外国で宣伝したわけではないので、それでもこれぐらい申し込みがあったということは、それなりに広まってきているのではな

いかと思います。

さて、ここからが後半の話になります。

いま、コーパスが、日本語研究でもそろそろ普及してきて使う人も増えてきた、研究のテーマとしても普通になってきたということを申し上げました。それだけだとめでたしめでたしですけれども、実はコーパスを使った研究というものに、ちょっとした落とし穴があります。そこが解決すれば、もっと高度な利用ができ、研究の発展につながってくれることも可能ではないかと、このように感じましたので、それについて簡単に説明します。

定性的研究、これは質的な研究と言ってもいいですが、と、定量的、量的な研究について簡単に定義します。これは私が勝手に定義したものです。対象となる現象の属性や関係を質的に解明するのが定性的研究、一方、量的に解明するのが定量的研究。違っているのは質的か量的かというところだけです。つまりこれは、対立的な関係と捉えてもいいのですけれども、むしろ相補的な関係と見るほうが、よりポジティブなのではないかと思います。つまり、どちらか 1 つでいいということではなくて、両面を考えるほうが研究としてはよりよくなるというのがこれからお話しする内容です。

先ほどコーパスベースの研究が実証的な研究で増えてきているという例がありましたように、特に、いま日本語研究の主流は、多分まだ文法だと思うのですけれども、語彙とか文法の研究でデータをどうやって取ってくるか。自分の思いついたものだけを研究対象とする場合と、自分以外のところから、よそからデータを取ってくる場合と、その 2 つのタイプがあります。これを「作例」と「実例」と考えます。別の名前で言うと、実例がコーパスに相当し、作例が自分の頭で考えたものなので、イントゥーイッションなどの内省タイプということも言えるかと思います。研究者自身の内省によって得られるということは、よく生成文法でありますように、この文章は言えない、こういう言い方はない、という例文をつくることができるということです。

あるいは、これは不自然であるとか、自然でない、ちょっと不自然だとかという適格性の判断を、グレードをつけておこなうこともできます。場合によってはクエスチョンマークの個数が、1 つ、2 つ、3 つとか、アスタリスクが多くなったりするというように段階的に使うこともあります。

一方で、実例をコーパスから拾ってくる場合は、そういう例があったということしか言えません。つまり、あり得ない文とか、あるいはちょっと不自然な文というのを、実例だけから判断することはできません。仮に誤用があっても、それは誤用だということを判断するのは研究者自身ですので、データだけから、これは誤用だというのは、ちょっと論理矛盾を起こすかもしれません。この点については、誤用を誰が判断するかというのは、先ほど申し上げましたサンプソンという人の文献にあったことですが、2002 年にアメリカ言語学会の会長だったフレデリック・ニューマイヤーという人が、その文章が正しいかどうかを判断するのは、人間、すなわち研究者の側なので、実例があったかどうかとは関係ない。グラマーとユーセージ（用法）は別であるということを述べているという、そういう指摘がありました。

ちょっとうがった見方をすると、自分が例をつくるというのは、自分自身から得られる実例と考えることもできます。でもそれは、一般に研究者としての主観が入る可能性があるので、そういう立場はとられていません。ただし、ある研究者が、こういう文章は言えるというのを論文で書いたとして、それを別の研究者が引用した場合には、それは実例になってしまうので、そのとたんに、ある人が考えた作例が変わる、そういうトリッキーなことも起きたりします。そのへんのことは、まだ私は解明していません。

いま申し上げた 2 つの研究のタイプとデータのあり方です。これをマトリックスに書くとこういうようなものができます。定性的な研究と定量的な研究。データとしては、自分で例をつくる作例、内省を中心とするタイプ、それからコーパスなどから実例を持ってくるタイプ。これらを掛け合わせると 4 つの組み合わせができます。

（1）が一番分かりやすいと思うのですけれども、1980 年代から 1990 年代の文法研究が（1）でした。自分で例文をつくってそれを質的に研究するということです。（2）があるかどうか分かりません。どのような研究があるのか、ちょっと分からないので、これはクエスチョンマークにしておきます。（3）は、日本語の通時的研究。国語史とか日本語史とか言われているのがこれかなと思います。データがなければできない研究で、しかも、それを質的に研究する。実は近年のコーパス

を利用した文法研究の多くも、私は（3）ではないかと思っています。

といいますのは、学会発表をする院生の方々とか、学会誌に載る論文とかを見ていますと、コーパスを使って何をしているかというと、自分の研究の論旨とか目的に合った例文を拾ってきて、こういう例があったというのを紹介している、それが多いのですね。こういう量的な傾向があるとか、多い少ないというような、素朴なレベルでの考察すらないこともけっこうあります。ですから、まだ（3）のレベルなのかなと思います。

（4）というのは、自然言語処理とか、かつての国語研究所の語彙調査などは、間違いなく（4）に相当すると思います。分量としては、まだこの（4）は少なくて、いま（3）が一番多いのかなという印象を持っています。これはまだ実証していませんので、例えば文法研究の文献をランダムにサンプリングして、この 4 つに振り分けて分布を調べなければいけないのですが、これもまだできておりません。

先ほど申しましたように、近年日本語研究に起きた変化は、（1）から（3）の動きで、同じ定性的な研究のなかでデータの取り方が変わったと、データの取得の仕方が変わったということではないかと思います。つまり、それはコーパスの登場によって変わったわけであって、研究手法を変えたわけではない。つまり、研究手法はけっこうハードルが高くて乗り越えにくい壁である。（3）から（4）に行くのには、ちょっとまだ難しいのではないかと思っています。

またこの文化情報学科の話をしますが、多分こちらの学生さんは、（4）ができる人たちが多いと思うので、こんなことは気にしないで済むのかもしれませんけれども、これができる人は、いまの日本語研究の世界では例外だと思ってください。むしろこういう（3）の人たちが多いので、こういう人たちを納得させる必要があるわけです。どうして数量的な研究が必要で、それをしなければいけないのかと。なおかつ、ここだけをやっていたのでは理解しにくいので、この定性的な研究と橋渡しをする、つなぐようなこともやらなければいけない、そういうふうに思っています。

ここでは、いま申し上げました 2 つの研究スタイルのうちの内省による研究というものをもう少し詳しく見ていきます。

1980 年代ごろからこの研究スタイルが根付いてきたと思います。それ以前の、私もリアルタイムで経験していないので分からないのですが、多分そういう事情があったと思います。恐らくその背景には、日本語研究であれば、日本語の話者、自分自身が話者ですから、こういう言い方は言える、これは言えない、ちょっと自然でないということが直感的に分かるわけです。ほかの人に確認しなくても大丈夫という素朴な研究観があったのではないかと思います。つまりは自分が、ある意味では日本語に関してはプロフェッショナルだから、こういう言い方が可能か可能でないかは確実に分かると、そういうことだったのではないかと思います。それを後押ししたのは、生成文法の思想だと思います。

ところが、現代語、自分がいまここで話したり使ったりしている言語であっても、その幅の広がりというのは相当広いということが分かっています。

例えば、ある文献で紹介されている「あるです」とか、「するです」というような、動詞の終止形に「です」を使うという形があります。これは方言などではあるのですけれども、普通の日常会話、日常の書き言葉でも使うと、ものすごく違和感があると思うのですね。これを、もし日本語教育とか留学生の人が使ったら、「それは間違い」と言われたり、こういう言葉は言ってはいけないと言われる思うのです。

しかし、「あるです」は、インターネットで検索すると、ものすごくたくさん出てきます。そのなかにもちろん、誤用であったり、わざと間違ったり、一定の効果を狙ったりするものがあるのですが、それを除いても、なぜか、普通の文脈で使われているというのが出てきます。それは何なのかが分からないのです。もしかすると将来的にはこの言い方が普通になるのかもしれません。

でも、そういうようなことを見据えて、いまの日本語はこうなっているというのを直感で分かる人は、それほど多くはないのではないかと思います。そこで、実際にデータが必要になるということになります。

かつて、1989 年ですから随分古いのですけれども、こういうことが言われていました。この「北原 1989」、北原というのは北原保雄さんという方で、私が筑波大学のときの指導教官だった方です。北原（1989）では、「文法研究において帰納的方法が重要なものであることはいうまでもないが」

と前置きして話を展開しています。この帰納的方法というのがデータに基づいて、データドリブン方法ということになります。

北原（1989）では、演繹ではなくて帰納によって分析するということは、次のような問題点があると指摘します。1つには、客観的なものだけでは用例がそろわない。いくら資料（ここに「corpus」と注が付けられています）の範囲を広げても、具体言語の一部であることは変わりない。これはつまり構造とか体系を問題にする研究の場合は、どんなにデータを集めても、その体系を見ることができない。もうそれは演繹的に自分でつくるんだということを言っているということが言えます。

最後に、帰納的方法ではどうしても具体言語の現象を説明することに終わってしまう。③は的を射ていると思います。数を数えただけで終わる。解釈がない。説明とか意味付け、価値付けがないという指摘は、いまでもこれは通用します。これは、帰納的方法だけではなくて、演繹的な、理論先行な研究でも、実は同じことが言えるかと思います。この段階でこういう指摘がされていました。

また、この池原さんというのは、言語処理学会の会長だった方で、理系、工学系の研究者です。この方は、統計手法による翻訳ですね、自動翻訳の研究をされていましたが、そのためのデータをつくるということについてこのような指摘をしています。「いかに厳密な統計手法でも結果はアバウトである」と。この時代は統計手法による翻訳が盛んだった時期です。「これを応用した形態素解析や構文解析の研究結果は、従来の人手作成の規則に及ばないレベルにあり、解析精度向上への貢献はほとんど見られない。これは以下に示す統計の本質を考えれば、当然の結果とも言える」と。つまり統計による、統計ベースの機械翻訳の限界ということを言っています。

その理由として、「出現頻度の高い現象は、コーパスに繰り返し現れるため、統計的に有意な解析ができるが、すでに、人手による規則でカバーされている場合が多い」。つまり、人間が分かっていることを改めて統計で情報を得ることはないと言っています。

もう一方で、「コーパスに期待されるのは、出現頻度の低い現象であるが、そのような表現は十分な標本数が得られない」と、こう言っています。つまり人間が考えても分からないような出現頻度の低い現象をコーパスに求めようとするのですけれども、それは標本数が少ない、得られないということを言っています。2001年の指摘です。

いまのこととは今度、逆になりますけれども、データを自分の頭だけで取ってくるのはちょっと危険だという反対の意見が「田野村 1995」であります。

その大きな観点としては、言語知識の個人差ということが1つ挙げられます。「日本語の骨格的な部分については知識の不一致は少ないとしても、こと日本語研究において問題としなければならないような微妙な問題」、こういう言い方ができるかできないか、ボーダーラインのような言い方については個人差は無視できないと。要するに、「あの人は言えるけどこの人は言えない」と言ったって、それは水掛け論になってしまうということです。

もう1つの指摘は、内省を使って念頭に思い浮かべることのできる例文の範囲、知り得る語句の用法の範囲には限度がある。つまり実際のデータやコーパスを使ったほうが網羅性が高い、多くの現象を拾うことができる、自分の頭で考えると限界があるということです。知識を持っているということと完全に想起できるかは別であるということも田野村（1995）は言っています。

いまの赤くなったところを全部まとめますとこういうことになります。北原とか池原が言っている、データを使った場合は、用例がそろわない。あるいは十分な標本数が得られない、網羅的ではないという主張をしていますし、先ほどの田野村の主張の、内省を使った場合は限界がある、つまりデータには負けてしまうということを言っていまして、両方とも、お互いが網羅的ではないということを言っているのです。

これはなぜかというと、1つには、多分、北原の主張は1989年の指摘ですから、インターネットがなかった時代。あるいは新聞記事のデータなども、それがあったとしても手に入れられる時代ではないので、用例がそろわないということは当時の研究環境による意見ではないかと思います。いまは多分、この制限はかなりクリアされていると思います。

池原の言っている標本数が得られないというのは、これは別の考え方を取らなければいけないということは、出現頻度の低い現象を統計的なアプローチで対応するのがいいのかどうなのかということで、そこの問題点かなと思っています。つま

り、頭でデータを考えてもコーパスを使っても、どちらも網羅的ではないという主張は、そろそろ崩れ始めているかと思います。

つい最近、2013 年ですけれども、シンポジウムの原稿としてこういうものが登場しました。これは、最近の日本語研究の流れからすると少し逆を向いているというか、逆襲と言ってもいいのですけれども、そういう立場に立っています。

神戸大学の定延さんの指摘ですけれども、「従来から批判されているとはいえ、文法研究は実際のところ内省なしにはおこなえない」と明確に書いています。「文法研究の周辺領域は、心理／脳実験、コーパスを用いた計量分析、自然会話データなど、内省以外の手法を重視している」、つまり外堀が埋められてきていると、そういうことです。

さらにそういう、いま言ったような心理や脳科学や自然会話分析などの領域と、どうしても日本語の記述文法が接触すると。そうすると、そこで侵略されて、研究手法として内省を使ったものが絶滅するのではないかという恐れを抱いています。そのことは、「文法研究の本質を危うくする」とまでおっしゃっています。最後は随分過激な主張になってきますけれども、どうしてこのような主張が出てきたのかというのは、裏を返すと、それだけコーパスを使った研究が盛んになってきているということだと思います。

ただちょっと、ここで定延氏が言っている内省というものと、北原、池原、それから私が言っている内省と、ちょっと違うかなと思ったのは、ここで述べられている内省というのは、実際に例文を自分でつくって言えるとか言えないとか言っている、そういう話ではなくて、むしろ研究の枠組み、研究を進めるにあたってのアブダクションのようなもの、帰納でも演繹でもなくて、ひらめきのようなもので研究を進めなければいけないということではないかと思います。直感とか第六感とかそういう、それを重視するとという、そのことをもしかすると内省と言っているのではないかなと思います。ちょっと内省の意味が違う感じがしました。そうであれば、これはコーパスを使った研究でも言えることですし、定性でも定量でも、こういうアブダクションがなければ進みませんので、そのようなことはどちらにでも言えることであります。そもそもデータだけを集めていれば研究が進むということはあり得ませんで、その観察の理論的枠組みが必要です。そのことも含めて内省と言えるかもしれませんけれども、最初に何か理論的な整備があって分析するというような、その手順が守られないといけないであろうと思います。データを集めれば何とかなるというのは間違いだと思います。

以上、まとめますと、内省とか実例とか言っているのは二者択一で考えるべきではなくて、研究目的によって妥当な選択肢を選ばなければいけません。もちろん、自分の研究にはアンケートや意識調査が必要であるという場合もありますから、その場合はコーパスを使わなくてもいいし、内省を使う必要もありません。また、内省にもコーパスにも弱点がありますので、それを補完する必要性、どう補完したらいいかということが今後の大きな課題ではないかと思います。

これは、ジェフリー・リーチが 1992 年に主張したコーパスを利用した研究の特徴というものです。30 年ぐらい前でしょうか。ここで、いまこの生成文法に対するアンチテーゼみたいなことが書いてあったり、言語記述とか、経験主義的なというようなものが登場します。(3) にちょっと、少し私には異質と思える主張があります。「質的な言語モデルのみならず数量的な言語モデルも中心に置く」と言っています。

この (1) とか (2) とか (4) というのは、これは原文を確かめたのですが、「よりも」「よりも」「よりも」といって、A よりも B と、片方をむしろ重視する立場なのですが、(3) については、「のみならず」であって、これもこれもという、そういう、両方なければいけないと、そういう主張をしています。この書きぶりで言うと、最初にこれがあって、なおかつこれ。つまり、qualitative があって、そのうえに quantitative を付け加えると、そういう主張のようにも考えられます。

同じようなことは、日本でもちょっと前に宮島達夫さんという人が述べています。これは、論集、単行本に入っているので、あまり日の目を見ることはなく、リポジトリにも入っていないのでネットでは引っかからないのですけれども、宮島達夫氏は、かなり前にスライドで紹介した「雑誌九十種の語彙調査」を担当した人で、定性的な研究も定量的な研究も両方できる希有な研究者ですけれども、その方がこういうことを言っています。「量と質の差は絶対的なものではない。ある格とある動詞との結びつきが、あり得ないかごく稀か必須

なのかは調査抜きでは簡単に言えない。」この文献は、動詞と格の関係を数量的に分析した文献なのでこういうことを言っていますけれども、「どこどこから帰る」と、「どこどこに帰る」、どちらが多いかというのは頭で考えても分からないので、実際に調べないと分からない、そういうようなことを主張しています。この量と質の差は絶対的なものではないというのは、必須格とか任意格というのは、量で捉えるべきなのか、それとも質で捉えるべきなのか、その議論と関係しています。

もう1つの主張がこのあとですが、能力、すなわち可能性の問題としては、「行く」も「来る」も同じような格と結びつくけれども、現象的には「行く」のほうが到着点表現度が高く、主体表現度が低い。これはもちろん動詞の意味の違いに関係があると。

ちょっとこれは切り取ったので分かりにくいのですけれども、「どこどこへ行く」という表現のほうが、「誰だれが行く」よりも多い。「来る」というのは、逆に、「誰だれが来る」が多くて、「どこどこから来る」というのは少ない。そういうことだったと思います。それが、動詞の意味の違いに関係があると言っています。つまり頻度の違いは動詞の意味の違いに帰着するということを言ったあとで、このように「意味の記述は、現象における量的なちがいを説明できなければ不完全である」と述べます。つまり、文法的な意味の記述であっても、その現れとして量に差があるということを記述しなければ、記述としては不完全であるというような主張をしています。これは先ほどのリーチの言ったような質的なモデル以外にも量的なモデルを考えて、それをバランスよくおこなう必要があるということと相通ずる指摘だと思っています。もう、これが言われて随分時間がたちますけれども、このようなことを実践できる研究者はあまり多くないと思います。

以上、少し駆け足になりましたけれども、定性的研究と定量的研究を比較するとどのような状況になっているか、最近の状況をケース・スタディで観察します。

ケース・スタディですけれども、もちろん、内容のよしあしを言うわけではありませんが、あまりほかの人のものを出すと差し障りがあると思いまして、自分のものを含めて2つ挙げます。

たまたま、この私が書いた「新聞記事データに見る『つれて』『したがって』」というものとほぼ同じテーマを劉怡伶さんという方が書いています。

仮にわたしのものを定量的研究としますと、劉さんのものは定性的な研究のよい例になります。どちらも本当に典型かどうか分からないのですけれども、私の見たところ、こういうような研究が一般的な感じがしますので、この2つを研究手法として比べるということをおこなってみました。繰り返しますが、あくまでもこれはケース・スタディですので、全部こういうことになるのか、ほかもこうだとかというわけではありません。

例えば先行研究の把握ですが、論文ですから、どちらも最初にこういうことが書いてあるのですが、この部分の違いは起きません。ここで違うということはまずあり得ませんが、そのあとにどういうデータを使ったかというようなチャプターがくるのですけれども、私のほうはデータの属性とか検索方法などを挙げます。

劉さんのほうは、どういうデータを使ったかということの記述は一切ありません。突然、用例のなかにこういうのが出てきますということと、自分のつくったこの文章は言えないというのをたくさん使うことになります。これがデータの取り扱いについての大きな違いとなります。

もう1つ重要な点は、データです。ここでは文法研究ですから用例になりますけれども、これをどう評価するか。質的に評価するか、量的に評価するか、その2つの違いがかなり大きくなってきます。質的な評価というのは、正誤判断や的確性のことですけれども、私の場合は全部実例ですので、正しい、間違っているとか、的確かどうかというのは言えないわけですけれども、劉さんの場合は、この例は自然だ、不自然だ、あるいは間違っているということを、積極的に判断して、それをもとにして論を展開しています。

一方、量的な評価、用例が多い少ない、あるいは頻度に関する記述とかその意義については、私のほうは、目的は全部それですので、これはまさにそれに当たりますけれども、劉さんの場合は、こういう例は多いとか少ないという記述はほとんど見られませんでした。注記のなかで1件だけ、この1例しかこういう表現はないという指摘があっただけです。つまり、用例が多いか少ないかについてはまったく関心がないと思われました。言える言えないとか、的確でないからこういうまとめができるのだということが主体であって、多いか少ないかについてはまったく無関心と、そう

いう極めて対照的な違いが見えてきます。

このことを、いまはちょっとはしょってしまいますが、全体的にまとめたのがこの図です。先行研究のところ。データをどういうふうに扱っているかということ。それから非文を使うかどうか。仮説の提示、これはいま申し上げませんでしたけれども、多分劉さんの研究は認知言語学のフレームでやっています。認知言語学とは書いていませんが、それにのっとったかたちが提示されているように思います。そして最後にはその仮説を検証したということになっています。ということは、一応理論を使ったことになります。私のは特に理論はないので、「なし」「なし」「なし」となっています。用例の評価としては、質的なものか、量的なものか、この2つの対立が随分大きなものとなります。

つまり、この定性的研究と定量的研究、必ずしも典型とは言えませんけれども、あるところで随分違ってくると。もちろんこれを両方全部やる人もいるかもしれませんが、ここで挙げた2つのタイプの研究がそれぞれでおこなわれる場合がまだ多いのではないかという気がします。

ここで重視したいのは、この用例を質的に評価するか量的に評価するか。この2つは、必ずどちらかでなければならないというわけではなくて、両方可能であると思います。あるいは、こちらで言ったことをこちらで、別のかたちで表現するということもあり得ますので、この関係をもう少し整理するということが今後重要なことになるかと思います。

また定量的研究にははっきりした理論がないように思います。理論がないというのは、探し出してこなかったのでなくて、定量的研究に関する理論というのは、まだあまり発達していないのだと思います。言語モデルと言ってもいいのですけれども。古くは、言語モデルもなくはないのですけれども、いまそれが適用できるかどうか分かりませんし、語彙調査を中心とする、語に関する定量モデルは存在しますけれども、こういう構文とか文法とかという感じの量的な言語モデルは、まだあまり開発されていないという気がします。このへんが大きく違う点です。

これが最後のチャプターですが、その2つを併合するような立場というのがあるのかどうなのかということを最近考えています。そこで大きな意味を持ってくるのは、定量的研究のほうです。定性的研究はもう古くからおこなわれていて、理論とかモデルもしっかりしているのですけれども、定量的研究というのは、ちょっと後れを取っていたので、まだあまり理論的整備が進んでいないという気がします。特に使用頻度の持つ意味をどう捉えるか。このへんがまだあいまいさが残っています。

卑近な例で言うと、名詞などの内容語というのは、そこで述べられている話題に影響を受けますから、何が語られているか、あるいは、どう語られているかということで関係する。これは誰にも明らかだと思います。

この例ですね、何を話題とするか。これはBCCWJで、長単位で曜日を計算したものです。月曜日から日曜日まで。英語のコーパスなどでは、英語のみで言うと、日曜日が一番多くて、土曜日がこのようになっています。ウィークデーの、特に真ん中が少ないということが言われたりしますけれども、日本語でも同じ傾向があることが分かります。

ところが、このBCCWJを構成するレジスターの1つである広報紙を見てみると、なぜか「金曜日」がものすごく多いのです。2倍まではいきませんけれども、多分、統計的にもここは有意差が出るのではないかと思いますが、なぜ広報紙で「金曜日」が一番多いのか、お分かりになる方はいらっしゃいますか。これは、コンコーダンサーで文脈を見るとすぐ理由が分かります。

この広報紙というのは、イベントとか、役所とか窓口とか、そういうことの情報を伝えるのが一番の目的なのですが、そこで多く書かれているのは、月曜から金曜、月から金、「月金」という表現がものすごく多いのです。そのために「金曜日」がトップになってしまうと。そのあおりを受けて「月曜日」も高くなっています。つまり月曜から金曜におこなわれる何かを知らせるというような、そういう内容が多かった、そういう話題が多かったということになります。これはかなり特殊な例ですけれども、そこを見てみないと分からない。つまり、そういうレジスターによる差の評価は、ここの指標で確認しないと分からないことがあります。「日曜日」はむしろ低いほうに入っています。

もう1つ。これは答えが出ているのではないのですけれども、「書き言葉コーパス」のそれぞれの、1つ1つのレジスターの頻度表です。これは

短単位での集計です。一番多い格助詞の「の」から、ここで言うと書籍の場合は、これは形容詞の「ない」まで。出版･雑誌は「の」から数字の「二」までということになっていて、いろいろばらけていて、それぞれに特徴が見られたりもします。法律などは随分特徴的な語がきていますけれども、いま注目したいのは、一番多く使用されているところです。

日本語では､ほとんどのデータが､格助詞の「の」が何を調査しても 1 位なのです。そういうことをいままでも自分でも思っていましたけれども、ところがここで逆転が起きていて、Yahoo! 知恵袋と国会会議録は、接続助詞の「て」が 1 位です。「の」は 2 位になってしまっています。なぜ「の」よりも「て」のほうが使われたのかということは、先ほど申し上げた定量的研究における解釈が必要な事項ではないかと思います。

仮の考えとしては、国会会議録などで話す言葉を文字化したもの、Yahoo! 知恵袋というのはかなりラフな、書き言葉のなかでも相当くだけたものですから、そういう話し言葉的なものは「の」ではなくて「て」が多いのかなという感じもします。ただし、Yahoo! ブログは、話し言葉的とも思われますが「の」が 1 位です。

書き言葉的なもの、それから略語、白書とかいうのは､かなり堅い書き言葉ですが､それでも「て」は 3 番目に来ています。同じようにいま堅い書き言葉と思われる新聞では､「て」は 8 番目ですから、書き言葉性が強くても「て」が多いか少ないかというのは関係なさそうだということになって、どういう性質が「て」の使用頻度に影響を与えるのかというのはまだ分かりません。

いま申し上げてきましたように、例えば文体ですとか、述べ方の主観性、客観性、あるいは用いられやすい文型など、こういうようなこと、もっとたくさん要素はありますが、いろいろな特徴が絡み合って使用頻度に影響を与えているという可能性が高いと思います。特に現象文であるとか、「は」と「が」を使った文の文型などについてはそういうことが指摘されたりもしています。こういう特徴がどう関係しているかを具体的に明らかにしていって 1 つのモデルとして確立する。そこで定量的な使用実態のモデルができるのではないかと思います。

これは完全に私見ですけれども、いま定量的な観察から得られた事実というのは、個々に独立していて、言語研究の体系のなかで位置付けが不明確です。私が先ほど自分の例を挙げましたけれども、そこで分析して出した結果も、ただこういう結果が出ましたということで、それを研究の体系の中に位置付けていないというのが問題なのです。それを、同じような研究スタイルでまとめていって、1 つのモデルとして確立できるように情報を集約すれば本当はいいかもしれません。こういったものが今後もとめられるのではないかと思っています。

ここでもう 1 つ実例についてご紹介します。書き言葉コーパスで「たまねぎ」という表記を見ると、このようにばらけています。交ぜ書き、カタカナ、ひらがな、漢字。これだけですと、交ぜ書きが一番多くてということになるのですけれども、例えば、それぞれ媒体別に見るともっと具体的な事情がかかわってきます。

例えば新聞の用例数ですが、これだけではまだ多い少ないは言えませんが、新聞では、カタカナしかなかった。教科書では、赤ですからひらがなが多かったということが分かります。このデータについても、ある程度、各レジスターと言われているものとの相関が分かるのですが、このなかで一番下の雑誌についてさらに詳しく見たのが次のスライドです。

雑誌のなかは、サンプリングによってもう少し細かく分かれているのですけれども、ジャンル、学習雑誌とか園芸雑誌とか女性雑誌とか、そういう細かいジャンルで見ていくと、この赤で囲ったところ、例えばこれは家庭医学・健康雑誌というのは、ひらがなの「たまねぎ」しか使っていないということが分かります。育児・家庭教育とかもこれはひらがなだけ。一方で、婦人誌はほとんどが、これは交ぜ書き。ラジオ、芸能はカタカナになっています。

全部が全部ではないのですけれども、ある特定の 1 つに実は集中していて、それが全体でまとめるとバリエーションのように見えるということなのです。個々の､もっとレベルを低くしていくと、どこかに集中していたと分かる。それが、まとめてしまったためにバリエーションに見えたということになります。必ずしもこの分け方がいいのかどうか分かりませんけれども、こういう事情が存在しているということが分かります。これを、レベルを上げていくともっと曖昧になりますので、なるべく細かい要素に分けて分析することが必要

だと言えます。

この例は外的な基準として情報がつけられていますのですべて分かるのですが、実際にもう1つ、文脈まで下りていくと、誰が書いたかとか、どういう文脈で書いたかということによって違うファクターが出てくるかもしれません。

そろそろ終わりに近いのですけれども、内省とコーパスをつなぐということで、最後まとめたいと思います。

自分で用例をつくって、これは言える、言えない、自然、不自然だというような質的な評価をするということは、見方を変えると、それは量的な評価に変換できる可能性があると思います。安易に単純化するのはいけないのですけれども、自然な言い方であれば用例数が多く、不自然な言い方なら用例数が少ないというのがまずぱっと思い浮かびます。あるいは出現しないとか。単純に一次元のスケールに置き換えるというだけでも、内省ベースで考えているタイプの研究者と、コーパスを使った研究者との橋渡しできる可能性が出てくるということが言えます。もちろんこれは一次元のスケールではない可能性もありますので、違う尺度を考える必要があるかもしれません。

それは、どちらかのタイプの研究でしかおこなわれていなかったことは、もう1つのタイプの研究に変換できる、マッピングできるという可能性が今後考えられることではないかと思います。

いま言ったことは日本語研究のなかの、特に現代語研究、もう少し細かく言うと、文法研究などを中心にする話でしたけれども、似たようなことが、この前参加した日本語学会でも話題になっていましたので、別の動きを紹介します。これは日本語史研究の動きです。言語の歴史を研究する立場の話です。これは予稿集に、1973年の山口佳紀氏の文献が引用されていました。

「『言語の実態を明らかにする』と主張する研究には、『実態』や『事実』は一定の前提にすることなく把握され得ると考えているのではないかと疑われるものが」と、ちょっと分かりにくいかもしれませんが、「実態」とか「事実」を把握するためには何らかの立場が必要であると。立場を前提にしないで「実態」や「事実」は摘出できるはずはないということです。つまりこの場合は、ある程度、理論的枠組みがないと、「実態」とか「事実」を抜き出すことはできないということを言っています。

さらに明確にそういうことを述べたのは小松英雄氏の、同じ1973年ですけれども、40年前ですが、漠然とデータを集めている、そういう漠然とした目的における実態解明というものは、地図を縮尺されている限り「実態」ではないと考えて、一分の一の地図を目指して精力を傾けるようなものであると。つまりこれは歴史研究を例に取っていますけれども、もちろん現代語研究などの研究もそうです。100パーセント、データを網羅しないといけないというような態度は誤りであると。それは目的が漠然としているからそうなるのであって、一分の一の地図をつくる必要はない、必要なサンプリングをおこなって結果が有意であるものだけのデータを集めればいいのだ、統計的に解釈すればそういうことを言っています。

ここの赤いところが重要なところですけれども、文献のなかに盛り込まれたあらゆる「事実」を明らかにしておけば研究に役立つということを考えているのは、それも一理あるけれども、本来は、「秩序のなかにおいて個々の事実の価値をあきらかにする」。これが何度か述べようと思った理論化とか、あるいは解釈とか、データをどう捉えてその理論のなかに位置づけるかということ、体系立てるかということの本質を捉えていると思います。データだけを集めて、ただ結果が出ましたというのでは、現代語研究でもこの歴史研究でも同じような失敗に陥る。こういうことが、先月、今月でしたか、日本語学会のシンポジウムでこういうことが述べられていますので、コーパス研究の側と軌を一にして似たような危機感というか、学問に対する反省がなされているということがおこなわれていました。これはあながち偶然ではないのではないかと思います。

最後に、これもまだ何も形になっていないので、お題目にすぎないのですけれども、定性的研究と定量的研究を融合して、統合したような立場の新たな研究スタイル、もしかすると金先生がされるかもしれません。

いまコーパスが普及していますから、リーチの言う数量的な言語モデルの構築というのが早急に必要ではないか思っています。このようなことを、最近ちょっと考え始めています。これは参考文献でございます。

とりとめのない、長い話で申し訳ありませんでした。以上でございます。

（終了）