第２回国語教育カフェ・講演レポート（2009.12.9　マーメイド・カフェla place）

前川喜久雄先生（国立国語研究所）

日本語を科学する　―コーパスを用いた日本語研究の可能性―

報告　井浪真吾・岡本絵里・釋　就美
校閲　小西いずみ

２００９年１２月１９日(土)、第２回国語教育カフェにおいて、前川喜久雄先生(大学共同利用機関法人人間文化研究機構　国立国語研究所　言語資源研究系)にご講演いただいた。

前川先生は１９９９年以来、国立国語研究所の職務として日本語コーパスの開発に携わられ、『日本語話し言葉コーパス』の開発を経て、現在、『現代日本語書き言葉均衡コーパス』の構築に取り組まれている。今回のご講演では、コーパスを用いた日本語研究が切り開く可能性についてお話しくださった。折しも、「国語文化学基礎演習Ⅱ」（小西いずみ先生担当）の授業において、コーパスを用いた演習に取り組んでいるところであり、先生のお話を、大変興味深く拝聴した。

前川先生は、コーパスの性質やそれが日本語研究に切り開く可能性を大変わかりやすく示してくださっただけでなく、先生のお話は私たちに、「日本語」へのとらえかた、向き合い方それ自体を問い直すきっかけを与えてくださるものであった。ご講演の要旨は、次の通りである。

講演要旨

現代語の仕組みを解明するための方法は、内省に基づく代数的アプローチから、帰納的にデータから確率的な「規則」を導く統計的アプローチへと、近年ゆるやかに移行している。現在では、コンピュータの発達と普及が進み、大量のデータを容易に保存、検索、分析することが可能になっている。これはいわゆる人工知能学全般にみられる傾向だが、言語研究の分野においても、英語学の領域などではすでにコーパスを利用した研究が盛んに行われている。日本語研究(国語学)の分野においてもコーパスを利用した研究には大きな可能性があり、現在日本語を対象とした種々のコーパスの構築が進められている。

●コーパスとは何か

コーパスとは、通常コンピュータで利用する、言語研究のための大規模データをいう。そこでは対象とする言語について、実際に用いられた話し言葉や書き言葉の用例が、その言語の実情を正確

に反映するように組織的に収集され、公開されている。品詞情報など検索用情報を付加したしたものも多く、この検索可能な電子言語資料を使用することによってデータを効率的に検索することが可能となるのである。

●コーパスの必要性

では、コーパスは何故必要なのだろうか。

現代語の仕組みを解明するために大きな役割を果たすものとして、言語行動の内省という方法があるが、果たして、言語行動を内省することは常に可能なのだろうか。例えば、「風景」と「光景」の違いは何であろうか。「問題」は「起きる」のか「起こる」のか、それとも「生じる」のか。また、「はなしあい」という語は、一体どのように表記されることが一番多いのだろうか。このような日常的に用いる言葉であったとしても、その言葉の用い方や表記に対する認識については、実はあいまいな点が多いものである。また、日本語学の専門家間においても、このような問に対する回答は統一されていないことが多い。人間の情報処理能力や、内省には限界があることは否めない。

先に挙げたような曖昧な問いについて、コーパスを用いることで実際に使用された言葉のデータから明確な回答を得ることが可能になる。類義語とされる「風景」と「光景」について、『現代日本語書き言葉均衡コーパス』(2008 年度領域内公開版)を用いた分析によれば、「光景」は、「風景」に比べて、複合語後部要素として用いられる場合が極端に少ないという結果が示される(単独:「光景」1348 例、「風景」1253 例/複合語後部要素:「光景」10 例(「読書光景」など)、「風景」337 例(「田園風景」など))。類義語である「光景」「風景」は、複合語へのなりやすさという点において、両者間に明かな相違が認められるのである。

また、「問題が/事件が」は「起きる/起こる/生じる」のいずれの言葉と共起しやすいかという問題や、「はなしあい」の表記の相違については、新聞、書籍、国会会議録また政府白書、知恵袋といった媒体の異なりによってそれぞれの数値に違いが生じることを見て取ることができる。

もうひとつ日本語の動詞分類の問題の例をあげる。金田一春彦氏は、日本語アスペクト研究の嚆矢となって有名な論文のなかで、「山が聳えている」の「聳える」は、いつも「―ている」の形で状態を表すのに用いると述べた(「国語動詞の一分類」1940 年)。しかし、コーパスを検索すると、「聳える」は、「そびえる」単独の形で用いられる例が少なくないことが明らかになる。すぐれた語感をそなえて国語学者にあっても、ときに内省が難しいこと、そのような場合でも、きちんと設計された均衡コーパスを利用すれば、現実の多様性を正確に把握できる可能性があることがわかる。

ここまでの話をまとめよう。言語行動は、話し言葉であっても書き言葉でも、個人の内省によって正確に認識することが難しい。内省だけでは、現実の多様性を認識することができないのである。よって、言語行動に向き合う場合、我々をとりまく外界のデータが必要になるのだが、一個人が外界をすべて見渡すことは不可能といってよいだろう。そこで、外界を正確に代表したデータ、殊に代表性をもつ均衡コーパスが求められる。また、コーパスは言語学の領域だけでなく、言語情報処理や、語学教育の領域への応用が期待される。非母語話者のための、日本語教育の現場においてはもちろん、母語話者に対する国語教育の局面においても、貢献が期待されるのである。

●言語意識と行動の乖離

内省の難しさは、日本語を用いる話者による、言葉に対する自らの規範意識や認識と、実際使用する言葉との乖離からも示すことができる。

例えば、「日本」の発音について、「ニホン」と「ニッポン」どちらを発音することが多いだろうか。「NHK」はどのように発音するだろうか。これらは、実は人々の認識の上での発音や辞書における発音の表記と、実際行われている発音の割合に、大きく差が見られる例なのである。

「日本」を、「ニホン」でなく「ニッポン」と発音するのは、『日本語話し言葉コーパス』を用いた分析によると実は全体として3%以下にすぎない。しかし、2004年のNHK放送文化研究所のアンケート調査の結果によれば、37%の人が「ニホン」ではなく「ニッポン」と発音すると回答している。また、「NHK」は、国語辞典の見出しではおしなべてエヌエイチケー」と示されており、アクセント辞典では、「エヌエッチケー」とも併記しているものが見られる。しかし、『日本語話し言葉コーパス』を用いて実際なされた発音の用例を窺うと、一番頻度の多い発音は、「エヌエチケー」(132 回)、続いて「エネーチケー」(24 回)であり、辞書類に示されていた「エヌエイチケー」と「エヌエッチケー」は、それぞれわずか9回と7回にすぎなかった。

さらに、いわゆる「ら」抜き言葉については、2001年の文化庁国語課による世論調査によれば、「コラレル」と「コレル」の割合が逆転したのが、1971〜1980年生まれの人とされている。しかし、『日本語話し言葉コーパス』における行動においては、すでに1940〜1949年生まれの話者において、「コレル」の割合が「コラレル」を大きく上回っていた。

発音は無意識に行われているものであるし、そこには社会的規範の影響も働いている。つまり、自らが用いている言葉であっても、自らの言葉への意識や認識と、実際の行動とは、必ずしも結びついているとは言えないのである。

●文の正しさの判断—「グレーゾーン」の存在

話し言葉だけでなく、書き言葉においても、同じことが言える。

例えば、次のような文を、一見してどのように判断するだろうか。

(1)昨晩、あるいは昨夜おそく、このあたりは雨が降ったです

(グロルラー著、阿部主計訳「奇妙な跡」)

文法的に間違っていると判断する人が少なくないだろうが、これは実際に用いられた、しかもロングセラーとなった著書に見える用例である。このような「動詞+タ+デス」、あるいは「だったです」との用例は、検索してみると他にも見つかる。次にあげる三例は、雑誌の座談会において記録された話し言葉のデータである。

(2)まさに正岡子規だったですよ

(3)それだもんで参っちゃったですよ

(4)ああ、これは本腰を入れなきゃいかんと思ったですね

他にも、次のような例が見られる。

(5)政府は一体具体的に何をやったですか　(国会会議録)

(6)初めて海外に行ったですよ　(『日本語話し言葉コーパス』)

このように実際に話し言葉として存在する用例を順に見ていくと、このような用例もだんだんと自然に思えてこないだろうか。また、話し言葉ならこのように用いる場合もあると思えば、(1)への許容度に変化が生じくるだろう。他にも、「～に信頼する」という用法も一見文法的に正しくないものと思われるが、このような日本語は、明治生まれの文筆家の用例に多く見つかる。

(7)僕たちは警察に信頼して好いと思う　(今東光「赤線消ゆ・東光辻説法」1948)
(8)生活を維持するに足る詩的天才に信頼したために胃袋の一語を忘れた
(芥川龍之介『河童』1927)
(9)安心して、僕に信頼したらよかろう　(夏目漱石『二百十日』1906)
(10)あまりに現在の脆弱な文明的設備に信頼し過ぎているような気がする
(寺田寅彦「石油ランプ」1924)

また、この「～に信頼する」は、日本国憲法の前文にも用いられている。このように、「～に信頼する」の用例をある程度の数体験すると、これを適格な文として許容しても良いような気がしてくる。また、はじめに感じた「～に信頼する」の違和感を、ただ単なる通時的変化とのみみなしてしまうことはできない。現在生きる人々のうちにも、先に挙げた明治生まれの作家と同世代の大人達と接触しながら育ち、また読書体験として、先にあげたような資料に触れている人はいくらでもいるはずである。

従来の言語研究や文法理論では、文の正しさを判断するにあたって、その「正」と「誤」の境界を、明確に定めてきた。しかし、先の例に見たように、その境界は言語との接触経験によって変化し、またその変化は比較的短時間でおこる可能性もある。文法的な規則や、規範意識が絶対のものであるとは言えず、言語の変化は、このような境界のグレゾーン(連続性)を見据え、連続的な変化としてとらえる必要がある。これからの言語研究は、このような境界のグレーゾーンを見据え、その研究対象として積極的にとりあげていく必要がある。そこで、外界をうまく代表した言語コーパスが大きな役割を果たすのである。

●日本語コーパスの構築と応用

国立国語研究所では、日本語データベースの長期整備計画(KOTONOHA計画)が立案され、様々なコーパスの構築が進んでいる。すでに、明治・大正期の総合雑誌『太陽』の語を収録した『太陽コーパス』、また現代日本語の現実に行われた自発音声を大量に格納した『日本語話し言葉コーパス』などが公開されている。

『日本語話し言葉コーパス』(Corpus of Spontaneous Japanese＝CSJ)は、1999年～2003年度に、国立国語研究所、通信総合研究所(現在の情報通信研究機構)、東京工業大学とで共同開発された。工学上の目標は、自然な話し言葉(自発音声)をコンピュータに自動認識させるための学習用コーパスとして活用できるものとすること、また言語学上の目標は、自発音声の言語学的、音声学的研究のための基礎研究用コーパスとすることであった。この『話し言葉コーパス』を用いることで、アンケートで知ることが難しい、話し言葉の語形(発音)のゆれの実態やその背景を分析することが可能となる。

日本語の書き言葉コーパスについては、すでに新聞記事テキストデータベースや、「青空文庫」、国会図書館の国会会議録、インターネット上の文書を仮想的コーパスとして利用する研究などがあ

るが、日本語全体を上手く均衡的に代表しているコーパスは存在しない。そこで現在、KOTONOHA 計画において、代表性を持つ『現代日本語書き言葉均衡コーパス』(Balanced Corpus of Contemporary Written Japanese＝BCCWJ)が、開発途中にある(開発期間：2006～2010 年の５年間)。「代表」性のあるコーパスとは、サンプルを母集団から無作為抽出(ランダム・サンプリング)し、統計的推計を行うものである。この BCCWJ は、出版(生産実態)サブコーパス*1、図書館(流通実態)サブコーパス*2、特定目的(非母集団)サブコーパス*3の三種のサブコーパスから構成され、著作権処理の施された、全体で一億語以上の言語が資料となる予定である。公開に先駆けて、検索デモンストレーションサイト(http://www.kotonoha.gr.jp/demo/)において、BCCWJ の試験公開が行われており、実際に検索を体験することができる。

●日本語コーパスの問題点と、可能性

日本語には分かち書きの習慣がなく、日本語の表記には「語」に類する単位が表示されない。そのため、語によっては、コーパスの検索の際に問題が生じてしまう。例えば、「リズム」を検索するとき、音楽にかかわる「リズム」だけを(「マンネリズム」「アルゴリズム」「フォルマリズム」などを排除して)取り出したい場合には、データをあらかじめ「語」に分割しておく必要があるが、そもそも日本語の「語」をどのように規定するかという問題につきあたる。例えば、「国立国会図書館」は、全体で一語と言うこともできれば、「国立／国会図書館」、「国立／国会／図書館」、「国立／国会／図書／館」とも考えられる。このように単位認定の問題をめぐっては、コーパスの全体において首尾一貫した方法で語を抽出することが求められる。また一方では、著作権処理のありかたも、コーパスの規模が大きければ大きいほど、コーパス構築にあたって容易ではない問題である。

このような構築過程における問題点はあるものの、BCCWJ のような大規模均衡コーパスが完成すれば、日本語研究に新たな局面が開かれることが期待できる。先に挙げたようにコーパス言語学によって、言葉の「グレーゾーン」が研究対象となり得、「美しい日本語」を求めるのではなく、「リアルな日本語」に迫っていくことが可能となってきた。書き手・話し手や社会的文脈との関係を重視した研究が促進されることになろう。また、日本語教育や国語教育の場においても、従来のように「正」や「誤」、創られた規範意識の枠組みでは捉えられない「リアルな日本語」とどのように関係をきり結んでいくのか、どのように向き合い、どのように教材化していくのかなどを、顧慮していく必要があるだろう。

ご講演を聴いて

●井浪真吾（広島大学大学院博士課程前期２年在学）

前川先生は、コーパスとは何か、コーパスの研究上の意義とは何か、コーパス構築上の課題とは

---

*1 2001～2005 年の間に出版された書籍、雑誌、新聞を母集団とし、そこから無作為に抽出された約 3500 万語を対象とする。

*2東京都の 13 自治体以上の図書館に所蔵されている書籍で、1986～2005 年に発行出版されているものの集合を母集団とし、そこから無作為抽出された約 3000 万語を対象とする。

*3前述の二種類のサブコーパスの母集団に含まれないか、頻度が低く無作為抽出が不可能であるが言語研究上の必要性が高い言語資料のコーパス。ウェブ上の文書、白書、教科書、国会会議録、ベストセラー等である。対象期間はさまざまで、最長 30 年となっている。全体で約 3500 万語となる。

何か、を多くの具体例を交えながら私達に丁寧に示して下さった。膨大な量の用例を持つコーパスは、前川先生が示して下さった「聳える／聳えている」の例から窺えるように、過去の日本語学研究を見直し、更新することになるであろう。また、日本語学研究上の見直しや更新のみならず、人間の発話行為やことばを通じての人間理解の見直しや更新にも繋がるのではないだろうか。

現在、どの研究領域においても情報機器の利用を避けては通れなくなっている。文学研究の領域においても、コーパス程の大規模なものではないが、種々のデータベースが構築され、過去の文学研究の見直しや更新が既に為されている。ただし、それらのデータベースは高額な金額を払った人等、一部の人しか利用出来ない場合も多く知の独占のような状況も生まれている。コーパスの良い点は膨大な量の用例を持つとともに、全ての人に公開されるようになるという点である。そのようなコーパスの利点を考えると、現代語のみならず古典語に関わるコーパス構築も待ち望まれるところである。

●岡本絵里（広島大学大学院博士課程前期２年在学）

先生のお話において、最も興味深く感じたのは、コーパスデータからうかがえる実際の発話の実態と、アンケートなどから見られる言語への認識が必ずしも一致しないということである。コーパスデータが照らし出していた「グレーゾーン」の存在は、言葉への規範意識や認識が、創られたものであり、現実の言葉は、その枠組みからはみ出す部分があること、また自己の言葉への意識ですら、接触した言語の刺激によって、また連続的に変化することを同時に示すものであった。言葉について考えるということは、言葉それ自体を自明のものとして捉えることではなく、自らの言葉を支えていたり、また造り出していたりする背景や、それを用いる人間について考えていくことにつながるものであるということを、前川先生のご講演を通して強く考えるようになった。

前川先生はたくさんのコーパスを用いた用例を提示してくださり、先生のご講演中、私は、私自身が持っていた言葉への規範意識や、認識が塗りかえられる局面に幾度も立ちあうことになった。「リアルな日本語」がもたらす発見は、とてもおもしろいものだと身をもって体験させていただいた。先生のご講演を、自らの言語生活や研究に活かし、また先生のご講演中に体験したような、自分の言葉に対する認識をぬりかえられる局面に生徒を立たせることのできるような言葉の授業をどのように構築するかということを自らの課題としていきたい。

●釋　就美（広島大学大学院博士課程前期２年在学）

前川先生のお話は大変面白く、引き込まれるように拝聴させていただいた。ご講演を拝聴する前に、小西先生の授業の中で、前川先生が執筆された論文を読ませていただき、実際にコーパスを使用させていただく機会を得ていた。そのため、コーパスを作成する際に大変なご苦労があったことを聞き、私たちは大変恵まれた研究環境に身を置くことができていることに感謝しなければならないと感じた。それと同時に、今後の日本語学研究のためにも、コーパスのさらなる発展を願わずにはいられない気持ちになった。

言語学という学問に携わった身として、また、今後は教壇に立つものとして、生徒が自己の言葉に対して関心を持つことができるような授業を行っていきたいと考えている。また、これまで当たり前とされてきたことに対して疑問を向けてみる、といった考え方も身につけさせることができればと考えている。このような授業や教育を行う際の一助として、コーパスを使用したい。また、使用することでコーパスという素晴らしい研究データを少しでも支えることができればと思う。

（校閲者：小西いずみ：広島大学）