

ことばだけでは、 ことばは理解できない

持橋大地 国立国語研究所 次世代言語科学研究センター 教授
統計数理研究所 統計基盤数理研究系 教授



——2024年8月に発足した次世代言語科学研究センターに、統計数理研究所（統数研）とのクロスアポイントメント制度で教授として着任されました。専門や経歴を教えてください。

統数研という理系の研究機関に所属していますが、大学受験時は文系で、東京大学文科三類に入りました。言語学を学びたかったからです。しかし、言語学の講義は「〇〇語にはどういう規則がある」といった具体的なことばかりでした。私が興味を持っていたのは言語の抽象的な問題でしたので、講義中はよくノートに数式を落書きしたりしていました。数学も好きだったからです。

文系では言語の抽象的な問題の研究は難しいと感じ、理系に進むことにしました。しかし東京大学では当時、理転と呼ばれる文系から理系への枠は1,500人中わずか2人。猛勉強を重ねて、理転を果たしました。そして、ことばをコンピュータで扱う自然言語処理という自分に合った分野を見つけたのです。

大学院に進み自然言語処理の研究をしていくと、自然言語処理は機械学習の一部だと分かりました。そこで、NTTの研究所で機械学習の研究を進めました。すると、機械学習の本質は統計学だと分かったのです。こうして統数研にたどり着き、自然言語処理の数理的な基礎研究を行っています。

——自然言語処理の数理的な基礎研究とはどのようなものですか。

私の代表的な研究の一つが、教師なし形態素解析です。形態素解析は、文を単語に区切る自然言語処理の基礎技術で、検索エンジンや文章生成にも使われています。それらの場合、教師あり機械学習といって人間が正解の区切りを教えておき、コンピュータはその通りに処理しています。私は、いかに正解に近づくかではなく、コンピュータで人間のことばを理解することに興味があります。そこで、正解を教えない、教師なし形態素解析に挑戦しました。20文字の文では区切りの組み合わせは 2^{20} 通りにもなります。非常に難しい問題ですが、統計的な方法を用いると人間の直感に近い解にたどり着けることを示しました。自然言語処理は工学であり、実用的なことに使われています。私はそれを、科学に使うことを意識しています。

——この研究の面白さは、どういう点ですか。

ことばは複雑ですが、数理を使うときれいに表現できて、理解の糸口が見えてきます。それが面白い点です。ことばだけでは、こ

とばは理解できないと思うのです。道具を使わないといけない。その道具が数理や統計です。それを使うことで、データの欠損を埋めたり、効率的なデータ収集も可能になります。

——なぜ次世代言語科学研究センターに？

統数研と国語研は隣接していて（写真は国語研の屋上で撮影。奥が統数研）、知り合いがたくさんいますし共同研究もしてきました。その中で、国語研には長年かけて収集した膨大な言語データがあるけれども、従来の言語学の手法だけでできることには限界があり、もったいないと感じていました。だから、統計的・数理的な方法を使って言語学のフロンティアを広げることを目指す研究センターの構想を聞いたときは、すごい、ぜひやるべきだと思いました。そして責任ある立場で深く関わりたいと考えたのです。

——どのような研究に取り組もうとお考えですか。

方言の研究では、どのような語形や発音がどこで使われているかを言語地図にプロットし、ことばの変化とその要因を読み取ります。私は最近、方言地図に空間統計学を拡張して適用すると、従来は読み取れなかった変化の潜在因子を推定できることを示しました。この手法を使って新しい方言研究をしたいというのが一つ。もう一つが敬語の数理です。国語研や統数研などの研究者が、敬語と敬語意識に関する調査を愛知県岡崎市で1953年、1972年、2008年に行い、数理的な研究もされています。さらに最新の統計的・数理的な手法を加えることで、55年間の変化の要因などこれまで分からなかったことが分かったと考えています。

——言語学に興味を持つようになったきっかけは？

特別なきっかけがあったわけではなく、子供のころからことばに興味がありました。例えば、小学1年生で習う数を表す漢字は「千」までですが、自分で調べて「万億兆京垓杼穰…不可思議、無量大数」までをすでに憶えていました。高校生のときは、羅和辞典を持っていました。合唱部に入っていて、ラテン語の歌詞の意味を知りたいと思ったのです。吹奏楽部ではなく合唱部に入ったのも、合唱にはことばがあるからではないかと思います。

ずっとことばに興味があったわけですから、言語研究の本場である国語研で研究できることに大きな喜びを感じています。研究者同士の日常的な交流、議論も楽しみです。次世代言語科学研究センターを新しい言語学研究の中心地にしたいと考えています。