九州大学学術情報リポジトリ
Kyushu University Institutional Repository

言語データから見えてくることばの機微

内田, 諭
九州大学大学院言語文化研究院 : 准教授

https://hdl.handle.net/2324/1932325

出版情報：人文情報学月報. 56, 2016-03-29. 人文情報学研究所
バージョン：
権利関係：

言語データから見えてくることばの機微

（内田 諭：九州大学大学院言語文化研究院）

「big と large の意味はどう違うのですか」

これは私が大学時代に塾で講師をしていた時に中学生から受けた質問である。英語学を専攻していた身としては華麗に回答したかったが、思わず答えに窮してしまった。この 2 つの単語の意味の違いは、辞書を調べると明確になるとは限らない。例えば、LDOCE online[1]では big の第一語義として「of more than average size or amount」、また large の第一語義として「big in size, amount, or number」を挙げているが、一読してその違いはわからないだろう[2]。

「デジタル化された言語データ」があれば、この質問に対する妥当な解答を具体的な形で得ることが可能である。研究を目的として体系的に収集された言語データのまとまりは「コーパス」（corpus）と呼ばれる。分野やテーマを限定した特殊コーパスも存在するが、ランダムかつ均等なサンプリングを経て作成された汎用コーパスも多く存在する。英語の汎用コーパスでは British National Corpus (BNC)[3]や Corpus of Contemporary American English (COCA)[4][5]などがその代表例である。

コーパスを用いることで単語や構文などの頻度を数量化することが可能となる。類義語の比較をする場合、対象となる単語の前後にくる語句（コロケーション）を集計して比べれば、その違いが見えてくる。意味がほとんど同じで交換可能性が高い場合は共起語も似ていることが推測される。一方、意味が違っていれば、共起語が異なるはずで、その違いから対象の語の違いをあぶり出そうという寸法である。

では COCA を用いて big と large の比較を行ってみよう。これらは形容詞なので、「直後にくる名詞」を見出し語形（lemma）で集計する。ここでは上位 20 語を比べてみよう[6]。

・big の直後にくる名詞（上位 20 語）
deal, problem, thing, question, difference, issue, city, picture, man, part, business, bang, challenge, brother, story, mistake, change, game, hit, fan

・large の直後にくる名詞（上位 20 語）
number, part, bowl, amount, group, city, egg, scale, company, percentage, quantity, skillet, area, portion, pot, population, family, proportion, extent, sample,

一見して単語の顔ぶれが全く異なることが見て取れる。どちらの単語も日本語にしてみれば「大きい」となり、似ているように思えるが、このように比べてみると、意味の差異が浮き彫りになる。

共起語を仔細に見てみると、large は number, amount, scale, percentage, quantity, portion などの数量的な単語が多いのに対して、big は deal, problem, question, issue, challenge, story, mistake など数量

化することが難しいような単語が並んでいる。このことから、large はある基準によって数量化できる大きさを表し（客観的に測れる大きさとも言えるだろう）、big は数量化し難い、ある意味で主観的な大きさを表す傾向がある、と読み取れるだろう（この他にも語義の違いや慣用性の違いなどもあるがここでは議論を割愛する）。

大規模コーパスを用いれば、普段何気なく使っている日本語の類義語の違いを明確にすることも可能である。母語話者であれば、似ている単語の意味の違いは直感的には理解できるが、それを人に対して説明するとなると途端難易度が上がる。例えば日本語の学習者に「重要な」と「重大な」の意味の違いを聞かれたらどのように答えるのがよいだろうか。

日本語の大規模汎用コーパスとして国立国語研究所が中心に構築した『現代日本語書き言葉均衡コーパス』（BCCWJ）がある。このデータを利用して「重要な」と「重大な」の意味の違いを考えてみよう。これらの単語は、『広辞苑』第 6 版では「重要：大事なこと。大切なこと」、「重大：事柄が容易でないこと。また、非常に大切なこと」と定義されている。big と large の場合と同様、定義だけではその違いは判然としない。BCCWJ の検索システム[7]を用いて直後にくる名詞を集計すると次のような語が目立つ（「こと」、「もの」などの形式名詞を除き、頻度上位語から 10 語を任意に抽出）。

重要な：役割、課題、問題、意味、要素、ポイント、点、位置、部分、情報
重大な：問題、影響、関心、意味、過失、違反、事態、事故、犯罪、事由

まず「問題」、「意味」が共通した共起語であることがわかる。これはこの 2 つの単語の意味が似ていると感じることの証左でもあるが、例えば「重要な問題」と「重大な問題」が同じ意味かというと、そうではないだろう。共起語の違いに着目すると、「重要な」は「要素」や「ポイント」、「位置」などから「欠くことのできない、中心的な」という意味で用いられることが多いと読み取れる。一方、「重大な」は「影響」、「過失」、「事故」などの共起語から「深刻な、影響の大きい」というニュアンスで、ネガティブな意味も含んでいることがわかる。従って、「重要な問題」とは「課題解決のために避けては通れない中心的な問題」という意味であるのに対し、「重大な問題」とは「深刻な影響を引き起こす可能性のある問題」であると読み解くことができるだろう。

言語データの大規模な蓄積によって、このようにことばの微妙な違いを具体的に示すことが可能となった。これは人間の「言語直感」（内省）を裏付けることができるようになったことを示しており、言語研究において重要な意味合いを持つ。また、ことばの意味の違いを具体的な共起語のリストで示すことは、言語教育的にも大いに効果のあることだと思われる。

言語データの蓄積は研究者だけではなく企業によっても今後も継続的に行われていくだろう[8]。言語研究で扱うデータのサイズもさらに大規模化し、いわば big-data linguistics が言語学の 1 つの方向性となっていくのかもしれない。

[1] http://www.ldoceonline.com/
[2] big の項の Word Choice というコラムで big, large, great の使い分けについて記述がある。
[3] http://www.natcorp.ox.ac.uk/
[4] http://corpus.byu.edu/coca/
[5] 操作方法については次のサイトで詳しく説明している。
http://www.kenkyusha.co.jp/uploads/lingua/prt/13/UchidaSatoru1408.html
[6] COCA では類語比較の機能が提供されており、より詳細な分析が可能である。使い方については次を参照。
http://www.kenkyusha.co.jp/uploads/lingua/prt/13/UchidaSatoru1409.html
[7] http://nlb.ninjal.ac.jp/
[8] ただし、ただデータを蓄積すればよいというわけではなく、適切に分類し体系化して管理する必要がある。前述の BNC、COCA、BCCWJ 等の利用が可能になるまでには研究者の地道な努力があったことは忘れてはならない。

執筆者プロフィール

内田 諭（うちだ・さとる）九州大学大学院言語文化研究院准教授。東京外国語大学講師を経て現職。認知意味論（特にフレーム意味論）に足場を置きつつ、コーパスを使った意味分析や英語教育への応用研究も行っている。