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 10 月 31 日（火）　第 4 回

「コーパスから探る英語話者の世界の見方」

東京外国語大学講師
大谷 直輝

「講座：コーパスから見えることば・社会・文化」の第 4 回目として、「コーパスから探る英語話者の世界の見方」というタイトルでコーパスの説明をしたいと思います。お付き合いくだされればと思います。

最初に、私の紹介からしていきたいと思います。私は、今日担当します大谷です。所属は東京外国語大学大学院総合国際学研究院というところで、職階は講師です。専門は認知言語学という分野とコーパス言語学になります。皆様の中に認知言語学に関して何か聞いたことがあるという方はいらっしゃいますか。（挙手なし）いらっしゃらないようですね。私が何に関心があるかというと――これが認知言語学の考え方でもあるのですが――「ことばを通して見えてくる人間の心のありようと、ことばを通して見えてくる言葉の働きの背後にある原理」に関心があります。コーパス言語学というのは――もうみなさん 3 回授業を受けてきたので大体のイメージはついていると思いますが――電子化された大量のテクストから何らかの傾向を導き出すというのがコーパス言語学の手法です。認知言語学では、言語に現れる傾向を通して人間の心のありようを知りたいという目的で、コーパスを使うことになります。

今日は、コーパスを通して、コーパスの中にある言語のありようについて考えるだけではなく、言語を使う人間の心のありよう、あるいは人間がその中で生活をし、言語の基盤となるような文化や社会のありようというものについても考えていきたいと思います。

認知言語学を知っていただくために、初めに少し皆さんと考えていきたいこと

があります。日本全国に、「上り坂」と「下り坂」は、どちらが多いと思われますか。同じですよね。「上り坂」も「下り坂」も、数は同じです。そうすると、この二つの表現というのは何が違うのかというということが重要になってくると思います。「上り坂」と「下り坂」という表現は何が違いますか。例えば、「上り坂を駆け降りる」ことはできるのか、あういは「下り坂を駆け上る」ことができるのかということを考えると、ちょっと表現として難しいように思われます。同じ「坂」というものは世界の中に一つしかないですが、それを坂の上から見れば「下り坂」、坂の下から見れば「上り坂」となります。つまり、言語表現の違いの中には、私たちが世界のどこから対象を見ているのかということが入り込んでいるわけです。

2 つ目に、“The highway goes from Tokyo to Osaka.” あるいは “The highway runs from Tokyo to Osaka.” という表現を見てもらいたいのですが、この表現では、いったい何が “go”とか “run” しているのでしょうか。例えば “The car goes from Tokyo to Osaka.” といった事態とは、“The highway goes from Tokyo to Osaka.” という事態は性質が違うように思われます。“The car goes from Tokyo to Osaka.” というのは、何かが移動して東京から大阪に行くという、モノの移動であるのに対して、“The highway goes from Tokyo to Osaka.” というのは、モノの移動ではなく、静的な事態を描写しています。静的な事態であるにも関わらず、何らかの方向性が感じられるならば、それを “go” で表したり “run” で表したりできることがあります。日本語でも、例えば「南北を走る大通り」というのは、大通りが走っているわけではありません。ひょっとしたら目で南から北に追っているかも知れない、あるいは自分が車の中にいて移動しているかも知れないのですが、「南北を走る大通り」とか “The highway goes from Tokyo to Osaka.” という表現の中には、人間の世界の認識の仕方が入っているわけです。静的な事態であっても、何らかの方向性が見つかる場合に、（移動として）追いたくなるという人間の心的な状況が入っているわけです。

3 つ目の例を見てみたいと思います、“The bird is a land bird.” と “The bird ist a ground bird.” という例です。“land” と “ground” というのはどう意味が違うのかというのを考えてみたいと思います、“land bird” というのと “ground bird” というのは何が違うのでしょうか、“land” が表すものと “ground” が表すものという

のは同じでしょうか、違うならばどの点が違うでしょうか。どなたかいかがでしょうか。（受講生「違うと思います。」）どういう点が違うでしょうか。（受講生「？？？（聞き取れず）」）“land” と “ground” というのは、背景となるものが違います。“ground” というのは、“sky” との対比で “ground” があるのに対し、“land” というのは、どちらかというと “sea” との対比から “land” というものがあるわけです。そうなると、“land bird” というのは海との対比で「泳げない鳥」、“ground bird” というのは空との対比で「飛べない鳥」ということになります。ここも、人間というのが世界をありのままに認識しているのではなくて、フレームと呼ばれるのですが、背後にある大きな知識の中でとらえているからこそ、こういう意味の違いが出てくるということになります。こういう言語表現を見ることによって、言語表現の背後に入り込んでいる人間の心のありようを探っていきたいというのが私の研究のスタンスになります。

それでは、本日の内容の方に移っていきたいと思います。「1. はじめに」というところで、本発表の目的や背景を話します。コーパスというものをどのような目的で使うのかは、人によって全く違います。私の場合は、「ことばを通して心を見る」という立場からコーパスを使っています。その背景となる認知言語学の考え方について「2. 認知言語学の考え方」で少しだけ説明したいと思います。そして、「3. コーパスの種類と特徴」ということで、コーパスの種類と特徴について簡単に見ていきたいと思います。この時、ただ「言語が収集されている」というだけではなく、「実際に発話された言語が収集されている」ということがどういうことなのかということを見ていきたいと思います。「4. ウェブで使えるコーパスの紹介」として、パソコンとインターネットがあればすぐに使えるようなコーパスを 3 種類ほど紹介したいと思います。具体的には、COCA (Corpus of contemporary American English) というアメリカのコーパスと、最近できたコーパスで、TED Talk を収めた TCSE というコーパス、それから Google Books Ngram Viewer という Google Books をもとにしたウェブコーパスを紹介したいと思います。時間があれば、もう少し専門的な研究の紹介として、BNC (British National Corpus) というイギリスのコーパスを使った研究の紹介をしていきたいと思います。

1. はじめに

この講座の目的は、「人間の世界の捉え方には、人間が所属する文化や社会の影響が色濃く反映されている」というスタンスから大量の言語データが収集されている英語のコーパスをみることで、英語の機能や構造、英語話者の心のありよう、英語話者を取り囲む文化や社会のありように対する理解を深めるということです。この背景となるのが、「コーパスには、現実に使用されているありのままのことばの姿が現れる」ということです。これは、例えば、学校で習う規範文法と言われるようなものとは違います。例えば、「『ら抜きことば』は言ってはいけない」と言われたとしても、実際には「ら抜きことば」は日常生活に入り込んでいますし、1000 年単位の言語変化で見てみますと、「ら抜きことば」になっていくのが普通というような現状があります。コーパスには、規範文法ではなくて、現実に使用されるありのままのことばの姿が現れます。そのため、コーパスに収集されている言語データは、言葉の使用者である私たち人間が世界をどのようなものとしてとらえているかを知る手掛かりになります。

現在、いまだかつてないほど大量で多種多様な言語データに瞬時にアクセスできる非常に幸運な時代です。コーパスの定義については、すでに何度か聞いているかと思います。コーパスとは、電子化された言語データの総体です。現在、電子化された言語データは非常に身近にあります。例えば Kindle のような電子書籍を持っている方は、これをそのままコーパスとして使うことができます。あるいは電子新聞などもコーパスとして使われます。Facebook や LINE のような SNS も、それ自体がすべて電子化された言語データなので、コーパスとして使うこともできます。あるいはウェブというものも、それ自体をひとつの巨大なコーパスとして使うことができます。場合によっては一日あれば、100 万語程度のコーパスを作れる可能性があります。もちろん、それをすぐに研究に使ったり、商業目的で使ったりすることはできませんが、個人的に使うだけであれば、すぐに作ることができます。ウェブの言語データを使ったコーパスの例として、Google Books Ngram Viewer というものを見てみたいと思います。（受講生「Ngram というのは何ですか？」）Ngram というのは、N が 1 だったり 2 だったりするのですが、その数の語の連続のことです。2 つの語の連続だったら 2 つの gram ということで digram です。digram ならば、「語と語の連続」のことで、どの語と語のつながりが強いか、弱いかといったことが見ることができるようなものになります。

“big”、“small”、“large”、“little” という 4 つの語についてと、“flight attendant” と “stewardess” という 2 つの語について、簡単に調査をしてみたいと思います。“big” という語を検索してみると、1800 年ぐらいから使用頻度が増えていることがわかります。これだけでは面白くないので、“large” と比べてみると、“big” と比べて “large” の方が、かなり使用頻度が高いということがわかります。そこに “small” というのも加えると、“large” と “small” の使用頻度と、使用頻度の動向がかなり似ているということがわかります。最後に “little” も加えてみます。これ自体はただの事実を表しているだけなので、ここから何かを読み取るのは分析者の仕事です。

次に、“flight attendant” と “stewardess” を見てみます。昔は “stewardess” と言ったものを今は “flight attendant” と言うわけですので、そのことが結果にも表れると思いますが、こういった語を検索すると、いつ頃からジェンダー的性差をなくしていこうという動きが起きたかを知る手掛かりになります。“flight attendant” という言葉を見てみると、もともとはなかった言葉が、1967 年ごろから急激に増えていることがわかります。それに対して “stewardess” を検索すると ―― “stward” という男性の形に対して、女性を表す “stewardess” の方が、言語形式的に “-dess” という要素が多いので問題だということで是正されたわけですが ―― “stewardess” という語の使用頻度が低下すると同時に “fight attendant” という語の使用頻度が上がっているという歴史的な流れを、言葉を通しても知ることができるわけです。

Ngram というのは語の連続です。２つの語、例えば “big and” について調べると、“big” と “and” という連続自体が一つの Ngram になります。さらに “big and” の後ろにワイルドカードとしてアスタリスク記号を入れて “big and *” というのを検索すると、アスタリスクに任意の語が入ってきますので、“big and” の直後にどんな語が出てきやすいかということがわかります。（受講生「これは 3gram ですか？」）これは 3gram になります。trigram ですね。“big and small” とか “big and strong”、“big and little” などが出てきます。Google Books Ngram Viewer については後でもう一度述べたいと思います。

それでは、次にコーパス研究の背景について述べたいと思います。他の先生方からも問いかけがあったかも知れませんが、「なぜそもそもコーパスを見るのか」

という問いがあります。僕の場合は言語学者なので、言語の分析をするために見るわけですが、その際、コーパスが他の分析に対してどういう点が勝っているのかということを常に考えています。コーパスを見るうえで、考えなくてはならないことがいくつかあります。まず、コーパスにはどのような情報が載っているのかということ、それからコーパスからの情報によって、どのようなことを実証できるかということを考えなくてはなりません。コーパスにどのような情報が載っているでしょうか。（受講生「新聞なら新聞で使われる語、文学作品なら文学作品で使われる語が載っていると思います。」）そうですね、コーパスには言語レジスターといわれるような使用文脈、言語使用域が載っています。あるいは文学作品なら、どんな文学作品かというということも載っています。逆に、コーパスにはどんな情報が載っていないのかということも考える必要があります。それから、コーパスからの情報によって、何が実証できるのかということを考える必要があります。私の場合は、コーパスを通して、言葉を用いる人間のありようや、人間を取り囲み言語の基盤となる文化や社会に対する理解を深められるということが、コーパスを使う動機となっています。

2. 認知言語学の考え方

私の専門分野である認知言語学から、コーパス研究の背景を三点ほど挙げたいと思います。まず、人間は、世界をありのままに認識するのではなく、主観的にとらえているということです。先ほどの「上り坂」と「下り坂」では、同じ坂でもどこから見るかによって異なる捉え方をしていて、“The highway goes from Tokyo to Osaka.” では、もともと静的な事態であっても、動的な動きのあるものとしてとらえるといいうことです。２つ目として、言語には人間による世界の認識の仕方が反映されているということ、３つ目は、認識の仕方の違いが各言語に色濃く反映されているということです。つまり、人間は世界を、世界をありのままにとらえるわけではなく、五感を使ってとらえ、知覚を通じて概念化しており、言語には、ありのままの世界ではなく、人間がとらえた世界が反映されているという言語観が背景となっています。コーパス言語学の手法は、この言語観を逆手に取ります。もしも言語の中に、人間の世界のとらえ方が反映されているならば、言語を見ることによって、人間の心のありようや、人間が存在する社会や文化のありようを見ていくことができるということです。

まず、人間は世界を主観的にとらえるということから見ていきたいと思います。一番左の図（抽象的な紋様）を見てください。この図が動いて見えるという方はどのくらいいますか？止まって見えるという方は？同じ図であっても、人によってはそれが動いて見えるし、人によっては止まって見えます。次の図はどうでしょうか。この図は平行な線ですが、傾いて見えるというものです。次の図は、同じ大きさの 2 つの円が、左の円の方が大きく見えるというものです。次の図（ルビンの壺）は、城を背景とすれば二つの顔に見え、黒を背景とすれば盃に見えます。次の図は、三角形が見えると思いますが、三角形が存在するわけではなくて、我々が主観的な線を引いて三角形を見てしまうわけです。最後の絵（「にしこり」という文字列）は、ヤンキースにいた松井選手に見えませんか。世界は一つであっても、人間はそれを主観的にとらえている。ここで言いたいのは、人間が認識している世界というのは、ありのままの客観的な世界ではなく、人間にとっての世界であるということです。

次に、言語には人間による世界の認識の仕方が反映されているということを見ていきたいと思います。次の表現には人間による世界の認識がどのように表れているでしょう。

(1a) The glass is half empty.
(1b) The glass is half full.

(2a) This highway goes from Tokyo to Osaka.
(2b) The mountain range runs from Canada to Mexico.

(3a) The bird spends its life on the land.
(3b) The bird spends its life on the ground.

(4a) John broke the vase (with the hammer).
(4b) The hammer broke the vase.
(4c) The vase broke.

(1a) と (1b) が表す事態は同じですか？（受講生「(1a) は『もう半分しかない

な』、(1b) は『まだ半分あるな』という感じでしょうか」）そうですね。(1a) を言った人は悲観主義者かも知れません。(1b) を言った人は楽観主義者かも知れません。つまり、同じ事態を見ていても、無くなった方に注目して「もう半分しかない」ということもできれば、残っている方に注目して「まだ半分ある」ということもできるわけです。同じ事態のどこに注目するかによって表現が異なっているわけです。(2a) と (2b) は、静的な状態でも、我々は動きのあるものとしてとらえることができるという例です。あるいは、（「>>>>」と板書）こういう記号があったとしたら、これは静的な事態ですが、我々は左から右に流れているものとして認識します。(3a) と (3b) は、同じ場所であっても、“land” は海との対比としてとらえられ、“ground” というのは空との対比でとらえられるという例です。あるいは「ケチ」と「質素」の違いも考えてみてください。「ケチ」も「質素」も、お金を使わないということですが、何が違うかというと、背景が違います。みんながお金を使うところでお金を出し惜しみするのが「ケチ」、みんなが浪費してしまうところでお金を使わないのが「質素」と、お金を使わないということは同じでも、背景の違いによって、一方は良い評価になり、もう一方は悪い評価になります。(4a)、(4b)、(4c) は、「誰かがハンマーで花瓶を割ってしまった」という事態を表しています。連続する行為のどこを言語化するのかということは、話者にゆだねられています。言語表現を見ることによって、事態のどこに注目しているのかということがわかります。(4b) は、ハンマー自体が勝手に花瓶を壊すということはあり得ないのですが、ハンマーを主語にすることで、花瓶を壊した人の責任を回避したいといった意味も出てきます。道具を主語にする表現は、道具の使用者の責任を回避しようとする文脈で使われやすいということがあったりします。

３つ目として、認識の仕方の違いは、各言語に色濃く反映されているということを見ていきたいと思います。みなさん、“How many fingers do you have?” と聞かれたら、何本と答えますか？20本という方？10本？正解は8本です。“finger” というのは、“one of the four long thin parts on your hand, not including thumb” という意味なので、右手 4 本と左手 4 本で 8 本です。“finger” とは別に “toe” という語があって、これは “one of five fingers movable parts at the end of your foot” という意味です。足の先は “toe” です。日本語ではしない区別を、英語ではしています。このように日英語では手の指の切り分け方に違いがあるのですが、共通点もありま

す。「手」と「指」を区別することは日本語も英語も共通です。英語も “finger” と “hand” を区別し、日本語も「手」と「指」を区別します。しかし、世界の言語を見てみると、「手」と「指」を区別せずに同じ語で呼ぶ言語もあります。あるいは、「手」と「腕」、“hand” と “arm” を区別せずに同じ語で呼ぶ言語となると、さらに多くあります。一見すると「手」と「腕」を区別しないなんていうのは想像できないかもしれないですが、「足」のことを考えてみると、日本語でも「足」という語は “foot” の部分と “leg” の部分を区別していないわけです。

このように、人間は世界をありのままに認識するのではなく、主観的に捉えるため、言語には人間による世界の認識の仕方が反映されていて、各言語には、言語に固有の文化や社会に根差した認識の仕方が現れているわけです。この言語観に基づき、これを逆手にとってコーパス言語学を研究している学者もいるわけです。これはどういうことかというと、コーパスに収集されている言語データには、言葉の使用者である私たち人間が世界をどのようなものとしてとらえているかが表れるので、コーパスを用いて言語の使用者である人間の心のありようを探っていくことができるということです。

3. コーパスの種類と特徴

コーパスの種類と特徴について話したいと思います。コーパスの種類や特徴については、すでに３人の先生からお話があったと思います。コーパスには様々な種類があり、さまざまな目的や用途に応じて編纂されています。代表的な区別は、汎用コーパスと特殊目的コーパスです。これは主に研究目的による区別です。総合的な目的のためか、特定の言語研究のために編纂されるのかという違いです。汎用コーパスというものは、ある言語コミュニティ全体の傾向を見るために編纂されたもの、特殊目的コーパスというのは、例えば話し言葉、講義の言語、アカデミックライティングの言語といったものの特徴を知るために、そういったテクストだけを編纂したものになります。共時コーパスと通時コーパスという区別もあります。共時コーパスというのは、同時代のデータを集めてくるものです。通時コーパスというのは、複数の時代区分からデータを集めてくるものになります。先ほどの Google Books の Ngram は、どちらかというと通時的なコーパスになります。ただ、通時コーパスとはいっても、かなり巨大なものなのであれば、探す範囲を限定すれば共時コーパスとしても使えます。また話し言葉コーパスと書

き言葉コーパスという区別があります。これは伝達手段に応じた区別です。話し言葉コーパスには、書き言葉コーパスにはない特徴を持ったコーパスもあります。例えばイントネーションが入っているコーパスや、息を吸ったり吐いたりといった情報が入っているコーパスもあります。Santa Barbara Spoken Corpus という話し言葉コーパスがあるのですが、このコーパスを使って “well” という表現を調べてみたことがあります。その時、“well” という表現の前に 100%起こっていた非言語的特徴が見つかりました。何かというと、“well” という前には、必ず息を吸っていました。それが何に役立つかというと、ここから “well” というのは少し場が硬直したところで言うものなので、いったん息を吸って間を取り、注目を集めたうえで使われるのだということがわかります。その他のコーパスの種類として、ウェブコーパスというウェブ上に存在するものと、CD-ROM や（聞き取れない）を使うようなものの区別があります。それから、学習者コーパスというものがあります。これは何が面白いかというと、普通のコーパスは、ネイティブが実際に使った表現を集めているので、正しい表現しかないわけです。ネイティブが使った表現なので、間違いだとは言われないわけです。学習者の作文などは間違いもたくさんあるので、学習者の母語に引っ張られた間違いであるとか、学習の段階における間違いというものがわかったりします。こういうものがわかると、それが教育の場において役立ったりすることもあります。最後にタグ付きコーパスというコーパスです。これは、ある文がどういう言語使用域で話されているのか、どういうタイプの話者が言ったのか、あるいはある語がどういう品詞であるのか、どういう文法関係にあるのかというタグがついているコーパスです。こういうものがあると、一気に調べられるものが多くなります。特に英語は、名詞と動詞の形が基本的に同じなので、品詞がついているようなコーパスでないと結果にゴミのようなものがたくさん入ってしまって困ることがあります。

コーパスの、データとしての特徴について話したいと思います。これを意識しないで研究すると大変なことになります。コーパスのデータとしての特徴は、大きく 4 つ挙げることができます。1 つは、電子化された言語データが収集されているということです。電子化されていることによって、検索ができるようになっています。2 つ目が、大量の言語データが収集されているということです。人間が生きている間に触れることのできる言語の数が限られているとしても、場合によってはそれ以上の言語データに触れることができるという点で優れています。

3 つめは､産出された言語データが収集されているということです｡このことは、4 点目に挙げる、談話を構成する言語データが収録されているということと関係しています。コーパスのデータは、文脈から切り取られた単文の集合ではなく、それ自体がより大きな単位の中の一部、例えばパラグラフならパラグラフの一部をなしています。このような特徴を利用して、言語に関するどのようなことがわかるのかということを見ていきたいと思います。

まず、電子化された言語データが収集されているという点に関してみていきます。電子化され、数値として扱えるため、コーパス内の語句の分布状況や使用頻度を定量的に表すことができます。また、電子化されているので用例の検索や収集、コロケーションの調査等が容易に行えます。

２つ目の、大量の言語データが収取されているという点ですが、コーパスに収録されている大量の言語データにおける言語分布を調べることで、人間が持っている言語知識のありようを調べられます。これは、現代のコーパス言語学において最も注目されている特徴です。コーパスには大量の言語データがあるわけですが、この言語データが、ひょっとしたら、我々の頭の中にある言語の構造に近いのではないか、コーパス内での語と語の結びつきの強さが、頭の中での語と語の結びつきの強さと関係しているのではないかといった仮説を論証することができます。

３つ目の特徴が、産出された言語データが収集されているということです。収集されている用例が作例ではなく実例であることから、コーパスを用いることで、言語の使用実態に基づいた一般化を行うことが可能となります。このことは４つ目とセットで考えてください。

４つ目は、コーパスには談話を構成する言語データが収録されているということです。つまり、収集された言語データには、独立した単文の集合ではなく、より大きな談話の一部を構成するという特徴があります。このためコーパス内の各文は、文の内部で働く意味論的な機能と同時に、単文を超えて、談話内部で働く語用論的な機能を持ちます。a.–d.の 例を見てください。この 4 つの表現は、何が違って、何が同じでしょうか。

a. John lost his wallet.
b. As for John, he lost his wallet.
c. What John did was lose his wallet
d. It was John who lost his wallet.

これらの文は、事態としては同じことを表していますが、先行する文脈でどのようなことが話されていたのかということが違っています。c. の文が言われた文脈では、前提として、John が何かをなくしたということがすでに述べられているのに対し、d. では、誰かが財布をなくしたということが、前提として述べられていると考えられます。すなわち、a.–d. の文の違いは、これらの文の中だけをみてわかることではなく、先行文脈との関連の中で見ていかなければわかりません。単文の集合では、このように文を超えた特徴については見ることができないのに対し、コーパスは、すべての文が談話を構成する部分なので、単文を超えた語用論的機能・談話的機能についても論じることができます。

4. ウェブで使えるコーパスの紹介

以上が、なぜコーパスを使うのか、コーパスを使って何を知りたいと思うのか、そしてコーパスに何が載っているのかになります。これから、実際にウェブで使えるコーパスを紹介していきたいと思います。具体的には 3 つ紹介したいと思います。1 つ目が、COCA (Corpus of Contemporary American English) というコーパスです。これは、おそらく現在研究者の間で最も使われているような汎用コーパスになります。アメリカ英語約 5 億語からなる均衡コーパスです。これは言語使用域や類義語の比較に便利です。2 つ目が TCSE というコーパスです。これは、TED Talk という、有名な学者や俳優が自分の活動について聴衆の前で話すコンテンツがあり、インターネット上で流行っていて、研究や教育の現場でも使われているのですが、そこからデータを持ってきたコーパスになります。3 つ目が、先ほども見た Google Books Ngram Viewer です。

まずは COCA から見ていきます。このコーパスでは、まず語や句の検索ができます。それから、言語使用域別の語や句の検索、コロケーションの検索、語と語の振る舞いの比較といったことができます。それぞれ実際に見ていきましょう。

まず、語や句の検索です。アスタリスクを使って任意の一語を検索するワイルドカード検索を使って、“more * than” を検索し、“more than” と組み合わされる語句を見てみましょう。

（COCA のウェブサイトを表示）

“more * than”を検索すると、“more important than”、“more often than”、“more likely than”、“more so than” などが上位に来ます。ここから何か面白いことを導くのは、分析者の仕事です。例えば、短い語は比較級に “-er” という語尾がつくと言われますが、“harm” のように短い語でも “more” と結びつく語があるということがわかります。あるいは、英文を書くときに、実際にその表現が使われるのだということを確認するといった使い方もできます。他には、Chart という検索機能を使うと、検索項目が文体ごとにどの程度使われるかということを調べることができます。例えば、“you know” のような、ディスコースマーカーと呼ばれるものを検索すると、spoken で最も使われていて、magazins や news paper など書き言葉では少なくなることがわかります。また、時代を見てみると、最近多少増えているということがわかります。それに対し、“she knows” のようなものを調べると、“you know” が間を埋めるためのディスコースマーカーとして spoken に多いのとは異なり、こちらは字義的な表現なので、違う傾向が見られます。fiction に多いようです。それから Compare という検索機能があります。この検索機能では、類義語の比較ができます。例えば、“boy” と “girl” という――類義語というよりも反義語ですが――2 語を検索すると、どういう言語表現が “boy” と結びつき、どういう言語表現が “girl” と結びつくかということを見ることができます。今表示されているのは、「“boy” とは結びついて “girl” とはあまり結びつかない表現」と「“girl” とは結びついて “boy” とはあまり結びつかない表現」です。ここから何を読み取るのかは分析者の仕事です。反義語や類義語がどのように使われているかを知りたいという場合には、この検索機能を使って単語を 2 つ入力することで調べることができます。

2 つ目のコーパスが、TCSE です。このコーパスでは、検索から、TED の該当場面を起動することができます。また、日本語の翻訳を用いた検索ができます。例えば “Tokyo” と検索すると、“Tokyo” が関係する文が検索できるだけでなく、TED のその場面に飛ぶことができます。英語教育にも利用することができます。Translations という検索機能を使うと、翻訳からの検索もできます。例えば、「も

しかして」と検索すると、日本語への翻訳で「もしかして」に対応する表現を検索することができます。

最後に、3 つ目のコーパスとして、Google Books Ngram Viewer を見てみます。先ほども見ましたが、ある言葉がどのように使われてきたのかということを通じて、歴史や社会についても知ることができます。これを使って、差別用語を調べてみたいと思います。「障害を持った人」を、もともと “handicapped” や “disabled” と言っていたのが、今では“challenged” と言うようになった歴史があるわけですが、このことが確かめられます。もともと “disabled” と言っていたのが、“handicapped”という表現が後から出て来て、しかしそれもおかしいということで、“challenged” というようになったということが読みとれます。あるいはジェンダーに関する語を見てみましょう。“gay” のような語を調べると、2000 年代から使われるようになっています。“lesbian” という語を調べると、こちらは新しい語であるということがわかります。それから “transgender” を調べると、頻度は低いですが、1990 年代から用例が増えているということがわかります。

あるいは、イギリス英語とアメリカ英語の影響力の違いというものも調べることができます。例えば “favor/favour” というのを調べると、この語はイギリス式とアメリカ式でつづりが違いますが、もともとイギリス式が多かったのが、アメリカ式のつづりが多くなっているということがわかります。検索範囲を British English にしてみると、イギリス英語ではやはりイギリス式の favour のつづりが優勢であるということがわかります。American English にすると、もともと favour というつづりだったのが 1840 年頃から逆転して、favor というつづりが多くなっているということがわかります。

最後に、少しだけコーパスを使った研究の紹介をしたいと思います。使用するのは British National Corpus という、一億語（書き言葉 9 割、話し言葉 1 割）のイギリス英語からなるコーパスです。特徴としては、品詞のタグ付けがあるという点が挙げられます。3 点見ていきたいと思います。一つは、“arrive” という動詞についてです。“arrive at”、“arrive in”、 “arrive on” といった表現がありますが、これらはどのように使い分けがなされているのでしょうか。二つ目が、いわゆる第二文型 SVC についてです。いわゆるコピュラと言われる be 動詞の代わりに使

われる “go”、“come”、“fall”、“turn”、“grow” といった動詞がどのように使い分けられているかということを見ていきたいと思います。最後に、類義的な句動詞である “burn up” と “burn down” の違いについて、コーパスを通じて実証したいと思います。

まず、“arrive at”、“arrive in”、“arrive on” に意味の違いがあるかどうかについてです。実際に BNC で見てみると、“arrive at” が約 3200 例、“arrive in” が約 1800 例、“arrive on” が約 390 例あります。“arrive on” は少し傾向が異なるので、“arrive at” と “arrive in” を比べてみると、どんな違いがあるでしょうか。“arrive in” の方は、“Paris”、“country”、“city”、“France”、“Narita”、“York” といったある程度大きな地域を表す語と用いられています。それに対し、“arrive at” の方は “house”、“airport”、“station” など、どちらかと言うと建物です。では、“arrive on” はどうかというと、“arrive on scene” や “arrive on earth” や “arrive on island” など、もしかすると日本語の「上」というのに近いかも知れません。つまり、「上陸」であるとか、舞台への「登場」のように、もしかすると on ということで、表面が強調されるのかも知れません。一般に、対象に対して at は点的に、in は容器的に、on は表面を強調してとらえるということが言えますが、こうしたことも、コーパスのデータを根拠として推測したり、あるいはあらかじめそういった直観を持っている場合には、コーパスのデータからこれを裏付けたりすることができます。

2 つ目が SVC のパターン、いわゆる第二文型についてです。SVC で使われる動詞は少なく、be 動詞や “become” 以外では “go”、“come”、“fall”、“turn”、“grow” など数えるほどしかありません。英語学習者にとって、これらの動詞の使い分けは難しいのですが、コーパスを使って、これらの動詞の補語 (C) として何が現れるのかを見ていきたいと思います。はじめに “go” と “come” の違いです。BNC に、“go” や “went” は約 20 万例、“come” は約 14 万例あるのですが、その中で、直後に形容詞が現れるものの形容詞を比べてみました。こうしてみると、少し傾向が見えてきます。“go” の方は、“go wrong”、“go mad”、“go bust”、“go bankrupt”、“go crazy” など、どちらかと言えば悪い形容詞が多いようです。それに対し、“come” の方は、“come true”、“come alive”、“come complete”、“come clean” など、どちらかと言えば良い意味の形容詞が来ています。なぜ “go” が悪い意味の変化で、“come” が良い意味の変化なのかということを考えると、“go” と

“come” のもともとの意味が関係しているのかも知れません。つまり、人間は、自分の周りの領域については、よく知っていて慣れ親しんでいるのに対して、外側の領域についてはよく知らなかったり、悪いイメージがあったりすると思います。その点で、その領域に入ってくるものをよいものとして、その領域から出ていくものを悪いものとしてとらえているのだと考えられます。こうしたことも、コーパスのデータからも言えるわけです。つづいて、“fall”、“grow”、“turn” を見てみます。それほどはっきりとした傾向は見えてきませんが、“fall” は “asleep”、“short” とか “silent” といった、どちらかと言うと「静かで力のない状態」と結びつくようです。これはもしかすると、人間が横になって休むといったことと結びついているのかも知れません。“grow” はどのような語と結びついているかというと、これはあまりはっきりとしたことは言えないのですが、“old”、“strong”、“tired” など体に起こる変化や、“big”、“large”など目に見える体積の増加を表す形容詞と結びつく傾向がみられると言えるかも知れません。また、これらはどれも比較級にできる形容詞です。つまり、段階的に増えていくような性質です。“turn” については、これもはっきりとしたことを言うのは難しいのですが、“white”、“red”、“blue”、“green” など色の変化を表すことが多いようです。

最後に “burn up” と “burn down” についてみます。“burn up” と “burn down” の`ように類義的な句動詞というものは、他にも “come up” と “come down” などたくさんあります。これらがどのように違うのかということを、コーパスを使って調べることができます。句動詞なので、“He burned up his house.” という VPO 型の言い方と “He burned his house up.” という VOP 形の言い方があります。あるいは “The fire burnt up.” や “The house burned down.” のような自動詞としての使い方もあります。自動詞になったときに、他動詞の主語が主語として残っている（主体主語）のか、他動詞の目的語だったものが主語になっている（対象主語）のかという違いもあります。あるいは受動態もあります。動詞 burn は BNC 全体で約 5400 例あり、その中に burn up と burn down が 460 例ありました。“burn up” は VPO 型が非常に多いのに対し、“burn down” の方は受動態が多いという傾向が見られます。また、自動詞の場合、“burn down” では対象主語であることが非常に多いのに対して、“burn up” の方は、主体が主語になっているという違いがあります。“burn up” と “burn down” では、“up” と “down” が修飾するものが違う

わけです。“burn down” の方は、“down” が表しているのは目的語の結果状態です。“burn” の結果、目的語が “down” したということです。それに対し “burn up” は、“burn” の結果として目的語が “up” したということではなく、“burn up” というのが一つのまとまりとして機能しています。“burn down” の方は目的語と強く結びついているので、受動態で使われ、自動詞になった場合には他動詞の目的語を主語として使われるということになります。

5. まとめ

最後にまとめです。人間は事態をありのままに客観的にみるわけではなく、自分なりの視点で見るので、言語表現には人間の認識の仕方が現れます。コーパスから取り出したデータは、認知言語学的な視点を用いることで、言語を話す人間の心のありようや人間が属する文化や社会を映し出す鏡となります。そのため、大量の言語データが収集されているコーパスは、人間による世界の見方を考察する上で、多くのヒントを与えてくれるということになります。ウェブで簡単に使えるコーパスがたくさんありますので、今日紹介したようなものを、実際に使ってみると面白いと思います。また、それを使って言語を観察するだけではなく、その背後にある文化や社会についても考えてみると面白いのではないかと思います。(完)

コラム †　　最近の英語のウェブコーパス

…・＊ ＊・…

現在、言語コーパスは、科学の発展によって、以前に比べてずっと身近な存在になっています。以前は、一部の研究者が有料で購入するなどして手に入れるものであり、なかなかアクセスが容易にはできないものでしたが、電子化された言語データがウェブ上に蓄積された現在、容易にアクセスできる様々なウェブコーパスが見られます。同時に、コーパスは規模の点でも、多様性の点でも、多種多様になっています。英語のウェブコーパスについて見ていきましょう。

最初に、コーパスの規模について見ていきます。1960 年代に作成された初期のコーパスである Brown Corpus は 100 万語からなっていたのに対して、現在の大規模コーパスである iWeb コーパスは 140 億語からなり、規模で言うと、1 万倍以上になっています。また、コーパスの種類に注目すると、従来は書き言葉が多かったのに対して、現在は、話し言葉のコーパスや学習者コーパス、歴史的な英語を収集したコーパス、個人語を収集したコーパス、さらには、言語データと同時に、映像などを含んだコーパスも存在します。また、従来に比べて、コーパスの使用用途も多用になり、言語研究だけでなく、言語教育にも大いに活用されています。

現在、無料で使用できるコーパスはたくさんありますが、有益な情報源として、Mark Davis のサイトを紹介したいと思います。このサイトには、British National Corpus や Corpus of Contemporary American English のような、言語研究でよく使用されるコーパス以外に、前述の iWeb コーパス、報道の英語を収集した NOW コーパス、世界の 20 の英語の方言を収集した GloWbe コーパスなどが収集されています。これらのコーパスはウェブ上の資料からできていることもあり、規模が非常に大きいです。また、このサイトには、Time Magazine の記事や Soap Opera だけを集めたような、特定の言語使用域に特化したコーパスも見られます。

コーパスを使って英語の姿を見てみたいものの、どのように始めていいか分からない方は、最初に、Mark Davis のサイト（https://corpus.byu.edu/）を訪れてみてはいかがでしょうか。

コーパス言語学を知るための３冊

・*・…‡…・*・*・…‡…・*・…‡…・*・

石川慎一郎「ベーシックコーパス言語学」 ひつじ書房 2012 年

1 冊目は石川（2012）です。石川（2012）では、コーパスを使った言語研究や言語教育をするうえで必要となる情報がコンパクトにまとまっています。大学の授業での教科書として使われることを想定していますので、分かりやすい文体で書かれており、参考文献や問題なども豊富に挙がっています。

…・*――――――――――――――――――――*・…

斉藤俊雄・赤野一郎・中村 純作（編）「英語コーパス言語学―基礎と実践」 研究社出版 1998 年

斉藤・赤野・中村（1998）は日本で初めての英語コーパス言語学の概論書です。編者の 3 名はそれぞれ英語コーパス学会の会長を務め、執筆者の多くも英語コーパス学会の関係者です。分量も多く、本格的にコーパスを使って英語を勉強したい人向きです。

…・*――――――――――――――――――――*・…

高橋英光「言葉のしくみ認知言語学のはなし」 北海道大学出版 2010 年

最後に高橋（2010）ですが、認知言語学の入門書です。人間の認識が言語の構造に影響を与えるという認知言語学の考え方を採用すると、コーパスが単なる言語データの総体ではなく、人間の心を映し出す鏡に見えてくるでしょう。

…・*――――――――――――――――――――*・…