連載

学習者コーパス入門

学習者コーパスのあるべき条件とその作成の具体的方法は?

英語学習者の発話や作文などの performance data を
コンピュータ処理し，教育や研究に役立てるために

第1回 「基礎と活用」

元東京学芸大学講師／ランカスター大学言語学科博士課程在籍 投野 由紀夫

1. はじめに

今回から1年間にわたって，「英語学習者コーパス」の基礎と活用について連載することになった。英語学習者コーパスとは，英語学習者の発話や作文などの performance data をコンピュータで処理できるようにデータ化し，それを利用して教育・研究に役立てようという試みである。

この連載では，とくに初心者にも十分わかりやすいように，コーパスの作成の仕方および実際の現場におけるさまざまな活用方法を具体的に解説してみることにする。なお，本連載記事では「学習者コーパス」および learner corpus と言ったときは，「英語学習者コーパス」のことを指し，わざわざ「英語」と断らないのでご了承願いたい。

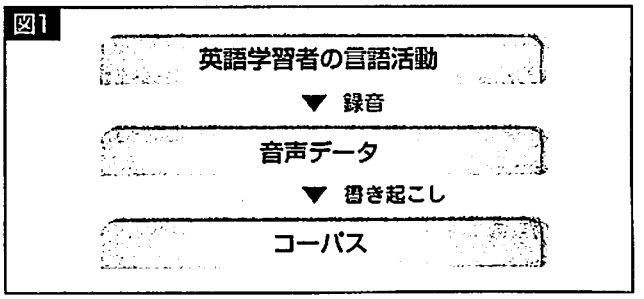
2. コーパスとは?

まずは「学習者コーパス」の具体的な問題に入る前に，一般的に「コーパス」がどのような条件を満たしているデータなのか，少々解説してみたい。コーパスはいろいろな厳密さで定義することができる。以下のようなポイントを押さえておこう。

❶データが文字化されていること

たとえば，英語教師が自分のクラスのスピーチコンテストのデータをテープに録音したとする。このままでは通例「コーパス」とは言わない。「音声データ」とは呼ぶだろうが，「コーパス」というにはこの音声データを書き起こす（transcribe）必要がある。40名のデータをせっせと書き起こしして，全部ノートに書いたとする。これはいわゆる「コーパス」であるし，英語学習者のデータだから「学習者コーパス」と言える。**データが文字化されていること**が単なる録音テープなどとの違いだ（図1参照）。コーパスという用語を使わず，プロトコルデータ（protocol）のように言う場合もある。

図1



❷コンピュータに入力されていること

いくら文字化していたとしても，ノートに書いたデータはさまざまな点で不便である。まず単語の検索とか頻度

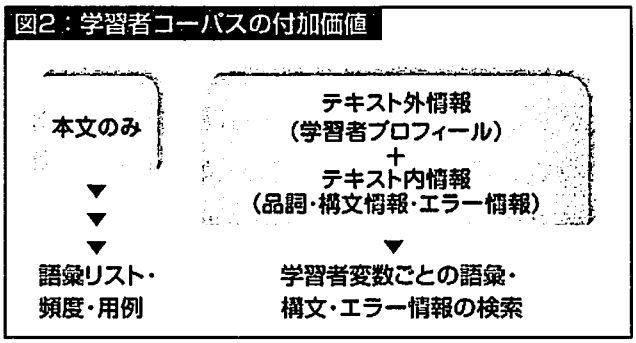
などの集計は自動的にはできない。いちいち自分で「正」の字をなん度も書いて，1個1個の単語の頻度を数えていたとしたら，その人は英語教師を廃業しなければ，一生研究が終わらないだろう。現在では「コーパス」と言えば，通例はコンピュータに入力してあるテキストデータのことを言う。コンピュータが読めること，これを自然言語処理では**「機械可読」**(machine-readable)と言う。これが条件だ。

❸目的にあった一定規模のデータであること

40名のデータを書き起こした全体量が1万語に満たないとした場合，このような小さな規模ではコーパスと呼べないのではないか，という人もいる。これはある意味では正しい。というのは，現在の現代英語コーパスの最低基準は100万語（たとえば，Brown Corpus とか LOB Corpus, ICE-GB などがこれに当たる），そして Mega corpora と呼ばれる大規模コーパス（British National Corpus や Bank of English など）の水準は1億語に及ぶ。1万語はこれらに比較すればあまりにも少ない。しかし，「学習者コーパス」のように収集がむずかしく，またその研究用途によっては意味のあるデータを抽出できるような特殊コーパスの場合には数万語のレベルから実用に達するものもある。ましてや，英語教師が自分の生徒の英語力の特徴や発達を把握したり，作文やスピーチなどの評価にコーパスを活用しようという場合には，一般性が低くても十分価値のあるデータになりうる。そういう意味で，数が多ければいいというわけでもない，という点も頭に入れておいてほしい。ただし，本格的に研究に学習者コーパスを使おうと思っている方は，コーパスから得られるデータの信頼性という点からも，データの規模やサンプリング方法，得られたデータの統計処理などに格別の配慮が必要になることを明記しておきたい。コーパスは道具である。要は目的に合った道具を選ぶことである。

❹データに有益な情報が組み込まれていること

図2は，2種類の学習者コーパスの情報量の違いを説明している。たとえば，スピーチコンテストのデータを書き起こしてコンピュータに入れて，それをそのままコンピュータでプロセスすれば，非常に基本的な語彙リストや単語の出現頻度などの情報はすぐに得ることができる。これが仮に図2の左側の場合だとしよう。しかし，それらのデータを個人の前学期の定期考査の成績別に上位・中位・下位に分けて，その3グループでのスピーチの違いを調べてみたいと思ったら，どうしたらいいだろうか？ 1つの方法は，単純にコンピュータ上で成績ごとにファイルを分類して，異なるフォルダにファイルを整理しておけばいい。ところが，同じデータをスピーチコンテストのための早朝練習に出席した生徒とそうでない生徒でどう違うかもみたいと思ったらどうすればいいか？ また別個のフォルダを「早朝練習組」と「出なかった組」で作ればいいかもしれないが，こういうことがある度にフォルダを作り直していてはあまりに煩雑になってくる。



実はこのような情報は，スピーチを取った際の学習者情報のデータとして一元管理したほうがよい。これが図2の

右側のデータの場合である。このような学習者情報をうまく活用することによって，コーパスにさまざまな分析の観点や特徴を付加価値として与えることができる。

同様の点は，スピーチの中身に関しても言える。スピーチになんの情報も与えられなければ，コンピュータは speak も spoke も speaking も別々の単語として処理してしまう。しかし，もしこれらの単語が speak の変化形だという情報を与えられたとすれば，コンピュータはこれらを1つの単語 speak の異なる形として認識できる。このような情報は，実は人間はすぐに判断できるけれども，機械は文字列を見ただけではそれを認識できないから，人間がテキスト内にコンピュータに理解可能なようにそのような情報を与えてやらねばならない。

今後の連載で紹介するように，実は現在では英単語の品詞などを解析して，品詞情報を付加するようなことが自動でできる時代である。こういった**テキストの本文外情報（学習者プロフィールなど）と本文内情報（品詞情報，構文情報など）**が豊富になればなるほど，コーパスとしての付加価値が高くなってくるのである。

❺一定の談話構造や文脈をもったデータであること

最後に，上記4つの条件をすべて兼ね備えていたとしても，コーパスと呼びにくいものに入試問題の文法穴埋め問題とか部分英作文のデータを集めたものがある。これらは学習者の文法問題の誤答データベースとは呼べるだろうが，普通「コーパス」とは言わない。コーパスはできるだけ一定の発話なり文章なりのまとまりをそっくり記録したものであることが多い。であるから，コーパスから採取した辞書の例文集というのはありえるが，辞書の例文だけを何万も集めても普通はそれを「コーパス」とは言わないのだ。

3．コーパス利用の環境を整える

では，これから学習者コーパスの初めの1歩を踏み出す皆さんに，コーパス利用の環境のことを少々説明しておきたい。この連載は，今後非常に具体的に学習者コーパスの作成のノウハウや実際の検索や活用の事例を紹介していく。

皆さんがこの記事を読んで，できれば2か月のインターバルの間に，そこで得た情報をもとにデータをいじってみることができたほうがいい。もし，私の記事を読んで「おもしろそうだ」と思ったら，ぜひ以下に書くようなマシンの環境を整えて，この記事の続きを読み，自分でプログラムをいじってみていただきたい。

ポイント1：Windows か Mac かはたまた Linux か？

まずは OS の話をしておこう。この連載ではもっとも普及しているマイクロソフトの Windows での環境で，基本操作やソフトの紹介をしていく。

Mac でも Linux（UNIX）でも，awk や perl のようなプログラム言語を使えば，柔軟な文字列操作が可能であるが，英語教師にこれらのプログラム言語を習わせるのは少々気が引ける。そこで，Windows で比較的ストレスを感じずに扱える範囲をとりあえず目標とし，適宜，必要に応じてそれ以外の処理方法などに触れたいと思う。

ポイント2：パソコンのスペックは？

これはこれから新しくパソコンを購入する方はほとんどスペック的には問題ないと思うので，あまり触れる必要がないかもしれない。旧型マシンを持っている方でも，Pentium

レベルの CPU ならば十分これからの連載のいろいろな作業が可能だ。そのかわり RAM と ハードディスクは多ければ多いほどよいだろう。最低でも RAM 32MB，ハードディスクは500MBくらいは余裕があるといい。現在のマシンはほとんどこのレベルは優に超えているので安心できる。1つ重要なのはモニタ画面の大きさだ。コーパスをいじるようになるとかなり細かい画面を見るので17インチ以上のモニタをぜひ勧めたい。リフレッシュレートも85Hz くらいのチラツキの少ないものを使用するようにしたい。コンコーダンスなどを眺めていると案外目に疲労を覚えるのでこれは気をつけよう。

ポイント3：インターネットは？

今回の連載でネットワーク上のリソースを紹介することはあるが，直接ネットワーク経由で使うようなサービスは紹介する予定はない。ただし，インターネット上にあるフリーのプログラムなどをダウンロードすることや，私のホームページにある資料の紹介などはありえるので，インターネットが使える環境を整えておいてほしい。

ポイント4：タイピング！

最後に余計なことかもしれないが，パソコンが苦手な人はまずタイピングがのろい人が多い。タイピングが遅いといろいろな作業に時間がかかってストレスがたまる。コーパスを作る場合にはとくに入力作業がその手間の大部分をしめる。

また一度作っても頻繁なエディタでの修正作業などがある。できるだけブラインドタッチでタイピング練習をしておくべきだ。その手のフリーソフトがいろいろ出回っているので，パソコンを購入したらまず1週間くらいはタイピング・ソフトで基礎的な練習をしてイライラを解消しよう。

4．学習者コーパスでできること

今回は第1回目であるので，学習者コーパスでどのようなことが可能になるかの数例をざっと紹介してみよう。今後も機会のあるごとにこんな研究活用例があるということをご紹介していきたい。

❶さまざまな語彙表（wordlist）の作成

英語学習にとって語彙表は重要な資料である。読者諸兄も，現場で教えていていろいろな機会に語彙表の必要性を感じた方は多いだろう。市販の単語集などはみな似たり寄ったりで十分授業などのニーズに合っていないことが多い。たとえば，中学だったら，新出事項の構文を用いていろいろな表現を広げてあげたいと思うが，たとえば旅行についてなにかを表現させたいときにすぐに必要な単語リストを20くらいピックアップしてプリントしてあげられたら便利だろう。あるいは，高校で導入が勧められている自由英作文なども，ただフリーに書かせるよりは，具体的なテーマとそれに関連するトピック・ボキャブラリーを与えて書かせるような指導が望ましい。そうなってくると，いろいろな機会に自分のパソコンから語彙表を抽出できるようなデータがあると便利だ。学習者コーパスは，教育用に整備していけばこのような中高の学習者のニーズに合った題材で作文や会話をしたデータをもとに，実際に使用される語彙などのリストを抽出して，将来の授業に活かすことが可能になる。

これら学習者データの語彙表と，たとえばすぐに電子

データとして入手できる教科書や英検の過去問などの語彙表を対照させたりして，より適切な語彙の選択と提供をできれば限られた時間内でより効果的な語彙の導入が可能になってくる。

❷学習者特有の語彙使用の特徴を知る

学習者コーパスがあると，たとえばネイティブ・スピーカーとの単語の使い方の違いを知ることができる。

<表1>

make の目的語	中学生	高校生
decision	−	−
money	+	+
difference	−	−
mistake	−	+
sense	−	−
food	+	+
comment	−	−
progress	−	−
change	−	−
effort	−	+
sound	−	−
film / movie	+	+

たとえば，表1は動詞の make の次に来る名詞のコロケーション・パターンをネイティブ・スピーカーのコーパスで調べたものである。make sense, make a difference, make a decision, make a change などのような「make + 抽象名詞」である動作を表す表現が頻繁に出てくる。それを学習者コーパスでチェックしてみると，この種の make の用法は非常に使用頻度が低く，ネイティブと共有しているフレーズは make money, make foods, make films / movies といった比較的具体的な名詞が多いのである。高校生になれば，mistake, effort などを使い始めるが，それでもネイティブのような使い回しをできていない。このような資料が簡単に得られれば，中学から高校への語彙指導への重要な指針になるはずである。

❸エラー情報の活用

学習者コーパスのもう1つの大きな特徴は学習者の語彙・構文のエラー・パターンがわかることである。最近は学習英和辞典にも文法的誤りの情報などが盛り込まれるようになってきているし，「日本人のよくやる誤り」的な本も以前からよく売れているが，中高生の実際の作文や会話データから抽出した組織的なデータというのはまとまったものがない。個別の語彙的な誤り以外に3単現の -s が導入後3年間でどの程度の定着率なのかとか，関係詞はむずかしいというけれどもどのくらい実際に高校1年生ではできているのか，などの客観的なデータが学習者コーパスを整備していけばわかるようになる。

自分の生徒のデータをある程度集められれば，それと全国規模の学習者のデータとの相違なども比較でき，自分のクラスの今後の指導に指針となるデータが得られる。

エラー情報は実際はコンピュータに判別できるように人間が工夫してテキスト内に盛り込んでいかねばならない。このエラー情報のデータ入力や検索方法についても，今後の連載で具体的に例を示しながら解説していく予定である。

5．コーパス関連の参考図書

最後に，この連載は2か月に1回なので基礎的な知識を身に付けておきたい方のために参考図書をご紹介して第1回目を締めくくろう。

『実践コーパス言語学』
（鷹家秀史・須賀廣 著，桐原ユニ，1998）

高校の英語の先生が書いたとは思えないほど詳しい内容の入門書である。学習者コーパスに関しての本ではないが，一般のコーパス利用について手っ取り早く知りたい人には最適だ。忙しい授業の合間にこれだけのデータを蓄積加工して自分たちの授業や研究に活用しているのはほんとうにすごい。具体例が豊富なので，とっつきやすい。インターネットのリソースに関する解説もなかなか参考になる。

『英語コーパス言語学:基礎と実践』
（斎藤俊雄・中村純作・赤野一郎 編，研究社出版，1998）

日本で最初に出版された英語コーパス言語学の概説書だ。一般の英語教師向きというよりは，コーパス言語学の分野で今後研究したい人向きの本である。内容的にも，概論書としては世界的レベルで見てもよく書けていると思う。とくに第3章の赤野氏の「コーパスを編纂する」は，学習者コーパスの収集にも非常に参考になるので必読。第3部は関連分野として「英語教育とコーパス」という章を東海大学の朝尾幸次郎氏が書いている。この部分が日本では初めての学習者コーパスについてのまとまった解説といってよいだろう。コーパスについて知識が深まる度になん回か繰り返し熟読するといい。

Learner English on Computer.
Ed. by S. Granger. London and New York: Longman. 1998.

世界で初めての学習者コーパスのための専門書である。もし英語でどんどん第一線の研究動向を知りたいという方にはこれがいいだろう。編者の Granger 女史は International Corpus of Learner English (ICLE) というプロジェクトのチーフで，非常に精力的に学習者コーパスのプロジェクトを進めており，この論文集でもいろいろな応用例を紹介している。言語教育への応用例はまだ少なく，学習者英語の特徴解明というのが中心的テーマになっている。なお，同様の論文集がまもなく Cambridge University Press からも出版される予定だ。

「学習者コーパスと英語教育（1）〜（12）」（投野由紀夫）
『現代英語教育』vol.35.　研究社出版.

1年連載したら雑誌が休刊になってしまったので，責任を感じているのだが，日本で具体的な学習者コーパスの研究活用事例を紹介した最初の連載特集記事だ。今回の連載と違うところは，コーパスの作成部分にはほとんど紙面をさかずに，もっぱら最新動向と研究事例を多数紹介することに的を絞って書いたつもりである。少なくとも，今読んでもランカスターでの最先端の研究のようすがわかってもらえるのではないだろうか。興味のある方は図書館などで…

＊　＊　＊　＊

今回は導入ということで，学習者コーパスの前提となる「コーパス」を作るということはどういうことなのかを，現場の英語教師の環境で考えてみた。次回から，いよいよ実際のデータ収集に関して解説していきたい。コーパスを収集するには，いくつか考えておくべきデータ収集のポイントがある。そのへんをじっくりとお話しよう。なにか私の書いた内容で質問があれば，電子メールで問い合わせていただいて結構である。大いに議論して，みなでよいデータを集めていきましょう。

問い合わせ：y.tono@lancaster.ac.uk