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2回 「データを集める」

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

はじめに

今回はコーパスのもととなるデータ収集のお話をしよう。コーパスの処理技術がどんなに進んでも，中身のデータの質が悪ければアウトプットの結果も歪んでしまう。逆にデータが良質であれば，研究目的によっては小規模のデータでも非常に有益な情報を提供してくれる可能性がある。現場の先生方の強みは目の前に英語を学習している生徒たちがいることだ。彼らの英語力が向上していくためにも，自分の英語指導の一環として生徒の英語使用の実態をモニタするような工夫を組み入れてみるといい。今回はとくに中学・高校の現場の先生方が試してみやすい課題の設定の具体例とそのポイントをチェックしてみよう。

研究目的をはっきりさせる

★ データ収集は大変

まずデータを集める際には目的をはっきりさせてからデータ収集の方法を考えたほうがよい。よく先生方の集まる研究会などで「作文のデータはなにかに役に立つかなと思ってとってあるんですが，ほとんど積んどくだけで…」というような話を耳にする。データとして役に立ちそうだ，と考えるだけでこの先生は研究意欲があって素晴らしいが，残念ながらなんとなくとってあるデータはほとんど処理できずに埋もれていく運命をたどることが圧倒的に多い。データ収集にはやはりそれなりの達成可能な目標の設定が大切だ。

★知りたいことにもレベルがある

そこでまずは先生方にスピーチ指導を例にして，自分はいったい「何を知りたいか」という問題について少し考えてみてほしい。次のリストを見ながら，ご自分なら何がいちばん興味のあるトピックであるか，それをどのようにして知るか，を考えてみよう。

①生徒がスピーチで取り上げたいトピックを知りたい
②生徒が話したいトピックに必要な語彙を知りたい
③生徒が事前に読んだり書いたりする作業の効果を知りたい
④自分が直すべき英語のポイントを知りたい
⑤長期的に話す力がついていくかどうかを知りたい

こうしてみると，知りたいことにもレベルがあることがわかるだろう。①，②などは比較的すぐに調査して結果が得られるのに対して，⑤などは半年とか1年とかのスパンで結果を見なければならない。つまり「単発ですぐに調べられるものか」，それともある程度準備し実行して結果を得るまでに「一定期間を要するものか」によって，データをとる体制はかなり違ってくる。長期にわたる調査を意図するならば，それなりに実施上の多くの制約（授業

時間，クラス替え，テストへの配慮など）を覚悟しなければならない。

★目的にあった道具を選ぼう

次に大事なのは，「目的に合った道具を選んでいるか?」ということである。①〜⑤のトピックのうち，コーパス・データが実際に必要なものはどれかと考えると，①などは別にスピーチを全部録音して書き起こさなくても，生徒たちにアンケートのような感じで話したい話題について聞けばすむ。②についても，むりやり英語を書かせるよりは，日本語で作文を書かせて，そこから英語にしにくい表現を抽出したほうがいいかもしれない。要するに，必ずしも非常に手間と労力のかかるコーパス作りをしなければ全部の問題がわからないということではないのである。

こう書くと「この連載の目的はコーパス作りではないか?」と怒られそうだが，**「コーパスはあくまでも教育・研究のための道具である」**という認識は非常に大切だ。コーパスがなにか万能薬のように扱われては筆者も困る。目的に応じて道具を選ぶのはどんな作業でも基本である。われわれも，知りたいことにコーパスが必要かどうかは十分に吟味する必要がある。③〜⑤についても，スピーチの質の向上などは文字化してその特徴をコーパスで調べるだけでは不十分な場合もある。音声情報が重要だし，スピーチをコーパスとして取り込む以外に別の評価尺度が必要だろう。スピーチがうまくなったかどうかを調べるのに，従来のような発音・文法・構成・表現力などの項目別評価とか全体評価のような方法は役に立つ。

★学習者コーパスの価値

しかし，この2点のこと，すなわち「知りたいことにもレベルがある」，「知るための道具は使いよう」ということを踏まえた上で，学習者コーパスの作成は非常に価値のある試みだとあえて強調しておきたい。なぜなら，きちんとした学習者コーパスがスピーチのデータをもとに作られ，それに音声テープなどの補助データと各学習者のスピーチ評価の情報が付属していたら，①〜⑤のポイントは全部コーパス・データから抽出することができるからだ。④については，エラーコーディングが威力を発揮するし，⑤に関しても定期的に同一学習者のスピーチを取り続けることで，その変化をとらえることができる。

誰もが便利だろうとは思っていてもコーパス作成は大変時間と労力を取られるので，簡単に言えば敬遠されてきたのだ。しかし，パソコンの処理能力がどんどん上がり，ネットワーク環境が整ってくるにしたがって，リソースとして整備された学習者コーパスの価値は計りしれない重みを持ってきている。われわれはきちんと取り組まなければならない時期に来ているのだ。

具体例に見るデータ収集

それではここから具体的なケースを見ながら，データ収集のポイントを押さえていくことにしよう。今回は学習者コーパスは研究目的からいろいろな作り方が可能であることを知ることが主眼だ。

ケース・スタディ 1

教師Aの発言:

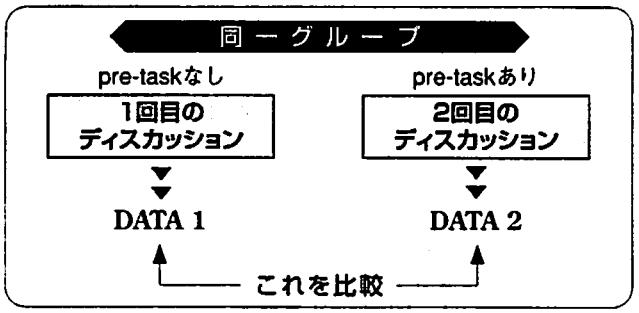
英語によるディスカッションはいきなりやろうとしても生徒にはむずかしすぎます。そこで，ディスカッションの pre-task としてどのような作業をさせると話す量とか内容が豊かになるかを調べてみたいんです。pre-task としては，関連する英語の記事を読ませるとか，英語でメモを取らせるなどを考えています。

【アドバイス】

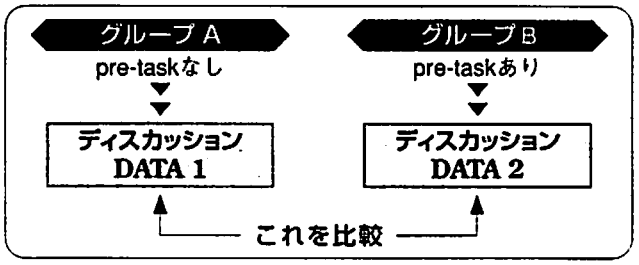
これはディスカッション・データを非常に限定された規模ではあるがコーパスととらえて，指導法の効果をそのコーパスをもとに検証してみようという試みである。この際には，pre-task の効果を見るわけなので，**タスクを与えた場合と与えない場合の比較**を行わなければいけない。

基本的には次のような2通りのデザインを組むことができる:

(1) Within-subjects design



(2) Between-subjects design



(1) は1クラスしか使えないが, 1か月に1回はディスカッションをさせるような時間を取れるというような場合に当たる。最初, 事前のタスクがない形でいきなりディスカッションをさせる。しばらくたってからの授業では pre-task を与えてやってみる。それぞれのディスカッションをグループごとに録音しておいて比較する。トピックは可能であれば同じ, もしくは似たものであるほうが望ましい。このデザインは, 別にディスカッションの指導に限らず自分の教えているクラスでちょっとした指導法の工夫を試してみるような場合に有効だ。必ず新しい工夫をしたグループとしないグループで対比させ, 学習者の performance data をコーパスまたはその他の方法で採取することが肝心である。このデザインの欠点は同一の生徒を使うので, あまり頻繁に行うと, practice effect といって繰り返しやることそのものの影響が出てきてしまう点である。

(2) の場合は, 異なるグループに異なる pre-task を与えられる場合に有効なデザインだ。複数のクラスを教えていたり, 同じ教室内でもグループによって pre-taskの内容を変えられる場合にはこちらのデザインがよい。長所は同じグループが2度やる場合の, トピックの選定のむずかしさや繰り返しによる慣れの影響を排除できる点。短所は, グループ間の能力差が出やすい点だろう。

ここでのポイントは, このように比較する目的や対象が非常に明確になっている場合には, **小規模のコーパスであってもそれなりにおもしろい結果が出てくる**という点だ。グループ間のディスカッション・コーパスを集計して, (a) 全体の発話量, (b) 語彙の豊富さ, (c) 文の長さや複雑さを測ってみれば, 事前タスクの効果による量的・質的な違いがディスカッションに現れてくるかもしれない。先生方の中には, 4技能の活動をさせる中で, このような pre-task を盛り込んで効果を上げようといろいろ工夫されているはずである。それらを少しこのような科学的に自分のやっている授業実践にメスを入れてみる, という発想でデータを取ってみると, 案外おもしろいことがわかるのではないだろうか。

そして, もし可能であればこのようなことに関心のある先生仲間で共通のディスカッション・トピックやタスクを設定してデータを取ってみて, 10校くらいが集まって結果を比較してみたらどうだろう。きっと, そのときにはタスクの効果についてより一般的な結果が得られそうだし, そのデータそのものが非常に貴重な学習者コーパスとなりえるだろう。

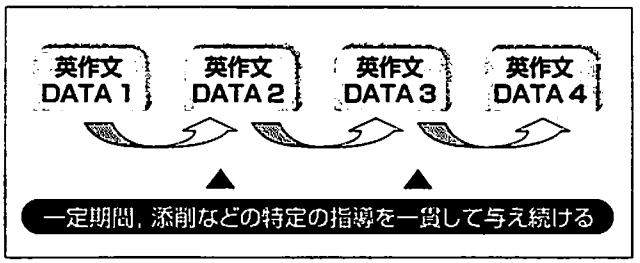
ケース・スタディ2

教師Bの発言:

私は英作文をする際に, 教師がどの程度添削などをして英語を修正してあげたらいいかをいつも悩んでいるんです。あまり細かく直しすぎると生徒は全然書けないという風に落ち込んでしまうし, 直さなければ逆に適当になってちゃんとやらない気がします。ある程度, 長期的に英語を書いていると自然に英語の誤りも直ってくるという意見もありますが, ほんとうなのでしょうか?

【アドバイス】

英作文の添削の効果については従来からさまざまな研究がされている。この種の研究は, デザイン的には一定期間繰り返しデータを取っていって, そのデータの変容を見るというタイプになる。



これは当然ながら，スタディ1に比べると長期間であるし，採取するデータも量が多くなるので，あまりたくさんの人数ではできないだろう。せいぜい1クラス40人のデータを半年とか1年とか取り続けるということになる。興味のある人は新年度の最初にこのように研究テーマを設定して，それに必要なエネルギーをかけられるように調整をしておくとよい。単発のリサーチでない場合にはなおさら新年度最初から準備してコツコツとデータの積み重ねをしていかねばならず，中高の現場の先生方の忙しさを考えると，この手のリサーチは大学の研究機関などとタイアップしてやらないと実際はむずかしいかもしれない。しかし，それほど大規模でなくとも，自分のクラスの生徒の作文を添削し，繰り返し書かせて作文の質を見たりするのは，実行可能なプランであり，後述するような入力の手間暇をうまく調整できれば試してみる価値がある。

定期的に取っていったデータでは baseline のデータになるいちばん最初の原稿を基礎に，添削した箇所が直っているか，全体的な文法の誤りが減少しているか，などの観点から作文データに分析を加えることができる。この誤りの情報を作文データに組み入れて，毎回の作文ごとに比較するようなコーパスを作るにはタグ付けの知識を多少必要とするのだが，これについてはまた機会を改めて解説しよう。

ケース・スタディ3

教師Cの発言:

私は中高一貫の私立校に勤めていますので，中学から高校にかけての生徒の単語力の伸びのようなものを見てみたいです。かなり個人差があると思いますが，認識語彙と発信語彙ではどの程度違うかなどに興味があります。

【アドバイス】

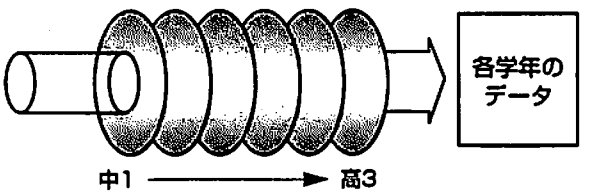
このような問題意識はかなりの先生方が持っていると思うが，なかなか実際に適当なデータを取ることがむずかしい。中学から高校にかけての比較可能な学習者コーパスの構築は，筆者の目標でもある。とくにいろいろな目的に使えるようにコーパス自体を汎用的な性格にするのには，タスクやトピックの選定，データを提供してもらう学校の選定，学習者情報などをいかに詳細に得るか，などむずかしい点が多い。

中高一貫校の場合には，汎用性を考慮せず自分たちの教えている生徒の英語力の伸びを見るだけならば比較的データを取りやすいので，このケース・スタディのようなデザインも可能かもしれない。大別するとこの種のデータは長期的にデータを取る方法と一度に縦割りでデータを取る場合とがある。

(1) Longitudinal study

中1から高3まで特定の個人／集団を追いかけてデータを取る

(2) Cross-sectional study



串刺しにして1度に大量にデータを取ってしまう

この2つの方法のどちらにも長所と短所がある。長期的にデータをとるのは学習者の習得過程を見るにはより望ましいが，実際問題としてデータが集まるまでに数年間を要する長大な計画になってしまう。その間に，クラス替えや担当教師の交代などで，現実には最初から最後までデータを取り続けられる生徒の数はごく少数になってしまうのが

通常だ。幼児の第1言語の習得のデータなどはほとんどこのようにして取られているのである。

逆に，串刺しで一度に異なる学年のデータを取るのは実施上は効率的で実行できる可能性も高いだろう。ただ，この場合の欠点はあくまでも各学年のサンプルは別人なわけで，個人の習得の推移を見るという確実さは失われてしまう。むしろ，各学年のサンプルが十分その学年のレベルや状態を代表しているという仮定の上に立って議論することになる。であるから，たまたま各学年1クラスしかデータを採取できなかったような場合には，そのクラスの学年における出来不出来があまりに違うと，正しい結果を反映できなくなる可能性もあるわけだ。

ケース・スタディ3の先生の発言に認識・発信語彙のことが出てきたが，コーパス・データはあくまでも発話や作文データになるので，認識語彙の理解度に関してはまた別のテストを作成しなければならない。たとえば，Paul Nation などが開発している Vocabulary Levels Test のようなもので，もっと基礎的なレベルの語彙を細かく扱ったテストを自作するなどして，補完しなければならないだろう。

データ入力

このような具体的なデータ収集の実例を見ながら，おそらく大半の先生方は「でも作文のデータをパソコンに入れるのが大変だ」とお考えになるだろう。これは現実問題としてほんとうに大変である。とくに会話コーパスの構築は世界的に見ても，この書き起こし部分の労力が膨大なのでなかなか進んでいないのが現状だ。しかし，コーパスがある程度小規模であれば工夫しだいで入力の手間を省力化することができる。最後にデータ入力のいくつかのヒントを挙げておこう。

★ 直接パソコンに入力させられないか？

まず最短距離として考えたいのは，自分の授業の中でなんとか工夫をして，直接生徒にパソコン端末から作文を入力させられるような方法を考案することである。CALL ラボがあればほとんど問題ないだろうが，もし限られた台数しかパソコンがなければ，クラスがある課題に従事している間に5人くらいずつ毎回コンピュータ室に行かせて打たせるとか，昼休みに打たせるとか，いろいろ場面設定をしてやれるといい。これがむずかしければ，データは紙で取ってしまってパソコンに興味がある生徒にタイピングの練習だといって打たせてもいいだろう。

IT の時間があれば，担当の先生と相談してタイピングの練習にそのデータを入れさせてもらうように交渉してもいい。実際，最近は家でもパソコンに触れている生徒は増えてきている。ひょっとすると，授業で書く作文を定期的にコンピュータ処理して結果をフィードバックしてやるという条件でならば，クラスでも興味を示して手伝ってくれるボランティアがいるかもしれない。宿題のような時間制限のないものならば，自宅でパソコンで打ってきてフロッピー提出というのもあるだろう。

高校・大学レベルになると生徒はかなり電子メールなどを活用できるようになる。全員が電子メールのアドレスを与えられるような環境の学校ならば，ぜひ電子メールを活用するとよい。香港科学技術大学の英語科などは専攻の学生のエッセイはすべて電子メールで提出を義務づけており，John Milton という人の集めている中国人の大学生のコーパスはもう2000万語を超えている。それほど大規模なプロジェクトではないが，学生のパソコン環境さえよければ，どんどん電子データを活用できる時代だ。

現在，私のプロジェクトでは web ページから直接データを入力して，それを皆で共有するような方式を考案中である。そうなれば，授業中にインターネットでつないだサイトに直接作文を入力すれば，自分たちのデータがその場で見られるだけでなく，同じトピックで書いた他の学校の生徒の作文なども閲覧できるようになる。学校内のインフラがもっと整わなくてはいけないが，今後の中学生のコンピュータ・リテラシーの伸びを考えると，このような方法を活用できる時代もそう遠くはないと思う。

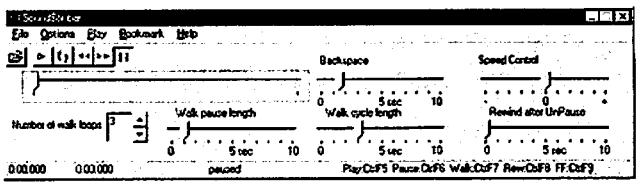
★ 手書き作文の場合のちょっとした工夫

どうしても紙に書かせたものを入力しなければならないという場合にも，いくつか工夫しておいたほうがいいことがある。1つは punctuation である。とくに中学生の作文では句読点がはっきりしないものが非常に多い。文の区切りの情報をコンピュータに認識させる際に，書き起こす人の主観に頼らねばならないものがあまりに多いのは考えものだ。そこで，中学生くらいの作文では，文の終わりには改行するように指示したほうがよい。段落の構成など，高度なことは中学生ではむりなので，1行1文でも気にしないで，むしろ途中で切れたような文が，途中でほんとうに切れたのか，自分では文だと思っていたのかなどがわかったほうがよい。

★ 音声データの場合はインターフェースも重要

スピーチなどの場合は実際に録音したテープをカセットで聞きながら書き取りをすることは非常に大変な作業である。そこでなるべくならば MD や ICレコーダーのようなデジタル媒体での録音が望ましい。最近は Smartmedia のような記憶媒体に音声を2時間くらい記録できるものも出てきている。これらを使って，WAVE ファイル，または MP3 という方式の音声データに加工できれば，パソコン上で入力すると同時に録音テープの操作などもできるようになる。

1つ参考になるページを紹介しておこう。The Michigan Corpus of Academic Spoken English（MICASE）というプロジェクト（http://www.lsa.umich.edu/eli/micase/micase.htm）では，音声データをパソコン上で入力処理できるための補助ツールを無償で公開している。Sound Scriber というツールであるが，これを用いることでほとんどのサウンドファイルの再生が可能で，かつ普通の Media Player と異なり，スピードのコントロールや再生箇所を指定して繰り返して聞くなどの細かい設定ができるようになっている。音声データの加工を考えていて，デジタル録音が可能な場合には威力を発揮するはずである。



▲Sound Scriber の起動画面

おわりに

今回は，データ収集の基礎的な考え方と具体例を示しながら，自分はどのような目的でどんなデータを取ろうか，ということをよく考えてもらう機会を設けた。データは取ってもパソコンに入れる部分がなかなか骨が折れる。私の学習者コーパスのプロジェクトでは常に協力者を求めている。皆さんの中で，データ収集には興味があるが，タスクや入力のことが不安という方は，データ収集のプロジェクトに参加してもらうだけでもよい。基本となるタスクを相談の上お送りし，入力作業は私たちのほうでさせていただく。その代わりにデータは共有という方式だ。また後述するが，データの品詞タグ付けなどのサービスも現在検討中である。今後はデータは個人レベルで集めているばかりでなく，それを多くの他の先生と共有していく必要がある。そのような趣旨に共鳴できる方はぜひこの記事を読みながら，自分の研究目的や興味について考え，共同プロジェクトの盛り上がりを作っていってほしい。詳細は私に電子メールをくださるか，以下のページ（http://www.lancs.ac.uk/postgrad/tono/）をご覧になっていただければと思う。

次回は，いよいよデータを一応取れたという仮定で，エディタなどによるファイルの最低限の体裁を整えたり，学習者情報を管理したりする方法を解説しよう。いよいよ実データに触る部分になる。できればそれまでの2か月間で先生方のほうでもデータをなにかしらとられて準備されておくとよいだろう。

問い合わせ：y.tono@lancaster.ac.uk