特集 英語教育・リサーチのすすめ

学習者コーパスを使って生徒のアウトプットを分析したい

投野由紀夫
Tono Yukio

リサーチの目的，調査の前提

「学習者コーパス（learner corpus）」は外国語学習者の会話や作文のような産出データ（production data）を目的に応じて収集・電子化したテキストの集合体である。学習者データの種類としては，自然な言語使用データ（natural language use data）なので，特定の文法項目を意図的に導出したものではないから，回避（avoidance）などで見にくい文法項目などは別の方法で取り出す必要がある。一方で，自然な言語使用を見るために，文法性判断（grammaticality judgment）テストなどでは見られない自然な誤りや過剰・過少使用（overuse / underuse）などの現象を観察できる。

また，まとまった産出データをとるので，その学習者の発話量や語彙の豊かさ，構文の複雑さ，テキストの談話構造などを記述的に診断することができるような量的データも取ることができる。

現在，英語学習者コーパスとして一般的に利用できるものとしては，1万人の英作文データを集めた JEFLL Corpus（投野 2007）[1]，1200人余の Standard Speaking Test のインタビュー・スクリプトをコーパス化した NICT JLE Corpus（和泉他 2004）がある[2]。

もし自分の生徒のデータを分析対象にしたい場合には，学習者コーパスとしてデータの自作を試みることになる。その際には，一定の会話や作文のタスクを考案し，それによって自然な言語使用に近いアウトプットを記録して，電子化することになる。専門的にはいろいろな条件を考慮するのだが，一般の英語教師が取り組む場合には，比較したいグループ（学年や学級など）単位のフォルダ管理をすること，基本的にテキスト・ファイル形式で入力すること（Word 形式などでは汎用のコンコーダンサーで処理できないため）などに留意しておけばよい。

研究方法

(1) 抽出すべき言語特徴を決める

学習者コーパスを用いた研究では，まずどのような言語特徴を抽出するかが大事である。それによっていくつかのタイプに分かれる。以下に簡単に説明する。

① テキスト全体の特徴

テキスト全体から取り出すべき特徴としては，語彙頻度リスト，分布（dispersion），語彙密度（lexical density），総語数に占める異なり語数の割合（Type / Token Ratio）などがある。

② 特定の語彙・文法項目に特化した分析

この場合は，テキスト内の特定の語彙や文法項目に焦点を当てる。たとえば，動詞 take のコロケーション分析や発達段階ごとの受動態の定着度調査などがこれに当たる。

(2) 必要ならば言語注釈づけを行う

上記の言語特徴を取り出す際に，表面上の文字列に対して処理を行うものと，それに言語注釈づけ（linguistic annotation）を追加情報としてつけて処理を行うものとに分かれる。たとえば，take の共起語を調査したい場合に，言語注釈づ

けのない素の作文データからすべての take の用例を抽出するには，take の１つ１つの活用形（takes, took, taken, taking）を別個に検索しなければならない。しかし，言語注釈づけの１つである品詞情報付与（POS tagging）および見出し語化（lemmatization）がなされたデータならば，見出し語（lemma）の take だけを検索すれば全用例を抽出することができる。また同様に take の後に続く名詞のコロケーションを抽出する，というような課題も，品詞情報があればある程度正確に検索が行える。前述の JEFLL のオンライン版はこのような品詞・見出し語処理が施してあるデータなので共起語リストなども簡単に作成できる。また現在では TreeTagger のようなフリーの自動品詞タグ付与プログラムも公開されており，自作のデータにも品詞タグ付与を自分で行うことが可能である。

(3) 研究デザインを組む

研究デザインを組む際に最も重要なポイントは，自分が知りたいことについてどのような変数が関係していて，それに関して自分は何をどう見たいのか，をきちんと整理することである。

たとえば，take の語彙知識の習得過程を take の後ろにくるコロケーションの獲得を中心に研究する場合を考えてみよう。このテーマに関しても，実際はいろいろなデザイン上の決断をしなければならない。

a）学習者集団の定義
- (ア) 学習段階別のグループにするか？
- (イ) 指導の効果を見るために異なる処置を施したグループにするか？
- (ウ) 母語は固定するか，複数の母語話者グループを扱うか？

b）take のコロケーションの定義
- (ア) 構文の種類（take＋名詞，など）

c）正用法 vs 誤用法

これらを組み合わせることにより，いろいろな研究デザインが組めるので，それに応じた統計処理の選択をすることになる。

リサーチの準備

研究デザインによって，自分でコーパスデータをとるか，既存のコーパスを用いてデータを抽出するかを決定する。この際に注意してほしいのは，コーパス分析を本当にする必要があるのかをよく考えること。たとえば，take の名詞のコロケーションを抽出する場合，自然な英文を書かせなくとも，コロケーションを取り出すための take に特化したテストを行うことも可能だし，その方が効率がよいかもしれない。研究課題によっては無理にコーパスを作る必要はなく，省エネできる，という場合もありえる。自然な言語使用を見るということがコーパス分析の最大の利点なので，それが生きるようなテーマ設定をすることが大事だ。

データ解析はパソコンのソフトウェアに依存する。現在では，コーパス・データを処理する検索プログラムは，大規模コーパスほど web 検索ソフトが多く用いられているのでその使用方法に十分慣れておくことが望ましい。また AntConc のような無料の汎用コンコーダンサーでも様々な処理が可能なので使用方法に習熟しておくとよい。コーパスからの文字列の取り出し方が不正確だとせっかくの頻度情報も信頼できなくなる。専門的にやりたい人はぜひ「正規表現」（文字列のパターンを表現する表記法）による検索技法を学ぼう。

統計処理は当然のことながらソフトで行う。エクセルにも多変量解析のアドオンが発売されている。SPSS，SAS などの商用ソフト，R などの無料ソフトなどでも上記の多変量解析のパッケージはすべて利用可能だ。使用方法に関しては，当然のことながら勉強しなければならない。

実際の研究手順

実際の手順は以下のようになる：

a）実験手続きが必要であれば被験者の集団のグループ分け，統制群，実験群を設ける。

b）必要な実験群への処置を施す。

c）コーパス・データを各集団から得る。

[実験をしない場合にはｄ)から始まる]

d）対象となる言語特徴（例「take＋名詞」のコロケーション）の頻度を抽出する。

e）頻度をグループごとに比較する。
f）必要な統計処理を行う。

リサーチ結果の分析

学習者コーパス研究で最も大事なところは，デザインに応じた適切な分析方法を選ぶことである。大きな分析のステップとして：

(1) 変数を１つずつ分析する
(2) ２変数の関係の分析をする
(3) 多変量解析

がある。

(1) １変数としての分析

まず，第１に１つ１つの言語特徴ごとのきちんとした統計指標をとることが大切。take の名詞のコロケーション頻度であれば，take の後の名詞（たとえば，place, time, part, care, notice など）のそれぞれの頻度をカウントし，もし異なる集団から頻度データを得た場合には，コーパス・サイズが異なるので，生頻度（raw frequency）では比べることができないから，相対頻度（relative frequency）または正規化頻度（normalized frequency）に換算して比較可能にする。この場合の換算は当該コーパスが100万語前後ならば100万語，10万語前後ならば10万語，といったように，決まったサイズはない。また研究者によっては，グループのテキスト全体から頻度を抽出すると質的データとしてしか処理できないので，サブコーパスの平均値（たとえば１学年100人ならば100サンプルの平均値）を取ったりする場合もある。

(2) ２変数の関係の分析

たくさんの変数を扱う場合でも，いきなり多変量解析を行うのではなく，２変数の関係を見極めておくのが賢明である。たとえば，学習段階の異なる学習者グループから動詞 take のコロケーション頻度を抽出した場合，２グループずつ比較をしてみる。もしテストの点数のような間隔尺度のデータを取っているならば，散布図を描き２つの変数のばらつきが楕円形（２次元正規分布）をしているかを確認し，必要に応じて相関係数を取っておく，といった作業をしておくとよい。

(3) 多変量解析

調査の対象は通例複数個あるのが普通である。take のコロケーションにしても，複数の学習者集団，母語話者との比較，take のコロケーションとなる名詞候補，コロケーションがエラーか否か，take 以外の動詞との比較，など，さまざまな特性を関連付けて比べたいという場合がある。このように，たくさんの変数をデザインに応じて適切な分析方法を用いて処理するのだが，その中心となるのが「多変量解析」という手法である。

通例はこれらの変数が独立変数（原因）→従属変数（結果）という形で「因果関係」を表すモデルを作ってそのモデルの当てはまり度や予測力を調べたり，または単純に多変量のデータを類似度，潜在変数などを基準に縮約したり，といった手法に分かれる。

いずれにせよ，多変量解析は複数の変数をまとめて取り扱うので非常に複雑になりやすい。事前に散布図行列や相関行列を作って，データ全般の傾向を観察しておくとよい。表１に独立・従属変数のデータの性質と潜在変数の有無，取り扱う変数による手法の分類を示す。

表１：多変量解析の手法とデータの性質

独立変数	従属変数	潜在変数	解析手法
量・質	量的	なし	回帰分析
量・質	質的	なし	数量化Ｉ類
量的	質的	なし	判別分析
質的	質的	なし	数量化II類
量・質	量・質	なし	回帰木
量的	量的	なし	正準相関分析
量・質	量・質	なし	一般化線形モデル
量的	なし	なし	主成分分析
量・質	なし	なし	クラスター分析
量・質	なし	あり	因子分析
質的	なし	なし	対応分析 数量化III類

表１で特に注目してほしいのは，最後の４つの手法には従属変数がないという点である。すなわち，表の最初の７つの手法は従属変数が「あり」ということは前述の「因果関係モデル」のヴァリエーションだということがわかる。一方，下の４つの手法は，複雑なデータの縮約（data reduction）が目的である。一見，複雑に見えるデータ

から何かそこに規則性や背後にあるパターンや潜在因子を取り出そうとする手法なのである（詳細は中村 2009参照）。最近，応用言語学でもよく用いられるようになった共分散構造分析は，ある意味でこの「因果関係モデル」と「潜在因子などの特定」を組み合わせた手法だと言える。

統計結果の分析，評価の方法

結果は研究デザインに依存する，と言っても過言ではない。よいデザインを組めれば，結果の統計処理もそのデザインの変数関係で判断ができる。因果関係モデルであれば，独立変数の影響で従属変数に差が生じたか，を検定する仮説検定の統計（カイ２乗検定，*t* 検定，分散分析など）を行う。また因果関係モデルを線形モデルとして，単回帰・重回帰分析を行う。さらに，因果関係ではなくデータ縮約が目的であれば，多変量データから合成因子や潜在因子を得るために主成分分析，対応分析，因子分析などを行う。

注意しなければいけないのは，特にデータ縮約の統計を使った場合に，新たに合成因子や潜在因子が数値処理されて出てくるのであるが，それらの解釈は研究者に任されている，ということである。たとえば，take とその他の共起語の頻度から take と get が１つのグループだとする結果が出るとする。その場合，このグループにどのような名前をつけるかは，take と get の似たところは何だったかを共起語の頻度の分布とこの分類に与えた影響度などの統計情報をにらんで，研究者自身が解釈しなければならない。有名な Biber (1988) の話し言葉 vs 書き言葉のコーパス研究で用いられている因子分析などもまさしくこのようなことを行って結果を出しているのである。

研究の留意点

学習者コーパスを用いた研究では，いくつか注意が必要である。１つは学習者コーパスの被験者の属性をよく理解すること。JEFLL Corpus では中１から高３までのデータが各1,000件以上あるが，実際はかなり進学校や中高一貫などのハイレベル校のデータが中心になっている。中くらいから下のデータはサンプル数もきわめて少ない。

統計処理の場合，異なるサイズのコーパスから頻度を抽出した際の比較には注意を要する。たとえば，カイ２乗検定の場合にはコーパスの総語数を分割表のセル内に明示するので，実頻度でもかまわない。ところが，主成分分析，因子分析のような場合にはデータの単位が異なる変数が混在しているため，相関行列を分析に用いる際に標準化得点（standardized score）に換算してから処理することが多い。このような単位の問題は多変量解析にはつきものなので，研究する際には基本的な知識を得ておき，適切な処理を事前に行っておく必要がある。

また，どの程度の大きさのコーパスを用意したらいいか，という質問をよく受けるが，これは調べたい研究目的にもよる。take の名詞コロケーションを複数集団のテキストから取り出して差を検定するだけならば，カイ２乗検定で期待度数が５を下回らない程度の頻度が得られていれば検定可能である。また，多変量解析の場合には行列の複数項目の頻度がゼロになることは嫌われるので，最低限でも各項目頻度が１以上になるようにコーパスの最適規模を決める，または頻度の低い変数をまとめる，などの処理をすることになる。

テキスト・マイニング的な思考法とテクニックが英語学習者の産出データと合体したもの，それが学習者コーパス研究と言える。これらの手法を駆使した研究が学校レベルでもますます盛んになることを期待している。

なお，具体的な研究事例については，本誌連載「進化する学習者コーパス(1)〜(12)」（2008年４月〜2009年３月号）を参照されたい。

◆参考文献

Biber, D. (1988) *Variation across speech and writing*. Cambridge University Press.

投野由紀夫（編著）(2007)『日本人中高生１万人の英語コーパス：JEFLL Corpus』小学館.

中村永友 (2009)『多次元データ解析法』(R で学ぶデータサイエンス２) 共立出版.

1）http://scn02.corpora.jp/~jefll04dev/

2）付録 CD にデータと検索プログラムが同梱

（東京外国語大学教授）