英語教師のための コーパス入門

第1回

コーパスの種類と利用方法

明海大学　助教授
投野由紀夫

1 はじめに

この連載の目的は，中学・高校の英語教師が授業や研究に利用できる英語コーパスの紹介と，具体的な活用方法を紹介することである。「コーパス（corpus）」とはある目的をもって組織的に集められた大量のテキストを電子化し，コンピュータ処理できるように整備したものを言う。近年，コーパスは我々英語教師や英語の研究者にとってぐっと身近なものになってきた。それにはいくつかの理由がある。

1つはCOBUILD English Dictionaryに端を発するコーパスを用いた英語辞典作成（これをcorpus lexicographyと言う）がヨーロッパを中心に非常に盛んになり，現在通称 Big 4 と呼ばれる COBUILD, LDOCE, OALD, CIDE と言う4つの学習英英辞典がすべてコーパスを基礎資料として作られるほど，その活用が必須となってきたことがある。

第2の理由としては，10年前に比べてコーパスの一般利用が非常にしやすくなったことが挙げられる。Windowsの普及，PCの急速な浸透，大容量データが安価なシステムで利用しやすくなったこと，そしてコーパス自体も比較的安価なライセンス料で利用可能なものが格段に増えたこと，などがあげられる。

第3に生成文法の影響で言語統計に対する全般的な興味が薄れていた60～80年代の言語学の過程を経て，この10年ほどでコーパス言語学が再び息を吹き返し，言語使用のデータを詳しく分析する意義が再認識されてきたことがある。コーパス言語学の入門書（McEnery & Wilson 2001; Biber, et al. 1998; Kennedy 1998; 斎藤他 1998）などが相次いで出版され，一般の研究者が基礎知識を得やすくなったことも挙げられる。

こういう背景があっても，まだ巷の中学・高校の英語教師にとってコーパスというものは敷居が高いに違いない。なにかコンピューターオタクがやるようなもの，というような印象があるかもしれ

ない。この連載では，なるべくそのような先生方の抵抗感をなくし，コーパスを身近に感じてもらいたいし，自分でそれを利用して明日の授業に活かせる，という実感を得てもらいたい。実際そのようにしてコーパスを日常の授業の準備や自己研鑽に利用している教師の数はどんどん増えている。そういう意味ではタイムリーな連載だと思う。なるべくわかりやすく解説したいと思うので，じっくり読んで基礎知識を身に付けてほしい。

2 利用できる英語コーパスの種類

さて初めてコーパスについて情報を得る先生方のために，コーパスにはどのような種類があるのかをざっと紹介しておこう。表1にこの連載で紹介する主要なコーパスをリストしてみた。

表1 いろいろな種類のコーパス

(a)科学的方法で収集され英語の変種ごとに比較が可能になっているもの Brown / LOB / Frown / Flob / Kolhapur / ACE / Wellington Corpus / etc.
(b)収集方法は多少ばらつきがあるが，コーパスの規模が非常に大きいもの Bank of English / British National Corpus / etc.
(c)特殊分野に限定して収集されたコーパス London-Lund / SEC / WSC / COLT / ICE-GB / ICLE / CHILDES / JEFLL / etc.

(a)，(b)は比較的一般的な英語コーパスであるが，(a)が100万語規模に限定していろいろな種類があるのに対して，(b)は1億語以上の規模になる。検索したい項目によっては100万語では少ない，ということがままあり，いわゆるメガ・コーパスと呼ばれる(b)のタイプのコーパスがどんどん開発されてきている。これに対して(c)は目的をもっと絞ったコーパスで，たとえばESP (English for specific purposes) /EAP (English for academic purposes) 用，学習者コーパス (learner corpus)，会話コーパス (speech corpus) など，いろいろな分野のニーズに応じて作られるものである。

これらのうちのかなりのものが現在は比較的入手しやすくなってきており，英語教師が自分で10～20のコーパスを収集してラップトップに入れて持ち歩くということが可能な時代になってきている。これはもう使わない手はないだろう。

3 コーパスを活用すると何が変わるのか?

それではここまで新しい知識を仕入れてパソコンの環境を整備してまでコーパスを利用するメリットは何なのだろうか? コーパスを使うと英語教師の何がどう変化するのか?

「コーパスを使うメリット」に関して私なりの意見を書いておこう。

コーパスを使うと何が変わる?

①用例観が変わる
②文法観が変わる
③英語の変種への認識が変わる
④ネイティブに対する劣等感が変わる
⑤役に立つ表現への姿勢が変わる
⑥学習事項に対する見方が変わる
⑦学習者の英語力観が変わる

① 用例観が変わる

コーパスに実際に触れてみると，従来の教科書や文法書に載っていた例文がいかに無味乾燥だったかが実感できる。コーパスから出てくる例文は鮮度が違う。たとえその例文をそのまま生徒に提示しなくとも，英語教師が「生きた例文」に触れ，実際の言語使用の感覚を研ぎ澄ませることは重要だ。

② 文法観が変わる

規範主義による文法書には実際の英語の使用状況からかけ離れてしまった記述が散見される。もちろん指導上どちらを教えたらいいかは熟慮が必要だが，英語教師は文法書の「～は誤り」というような記述のみを鵜呑みにせずに，実際にそれらが普通に日常会話などで使われている現実を疑似体験しないといけない。コーパスはそういう意味で凝り固まった英語教師の文法に対する感覚も変えてしまう。

③ 英語の変種への認識が変わる

日本人は教科書がアメリカ英語で統一されているので，海外に行くと逆にあまりに多様な英語の発音や表現に戸惑うことが多い。コーパスでアメリカ英語，イギリス英語をはじめ代表的な変種の英語に触れることができる。どういう新聞・雑誌が代表的な地域の英語として採取されているか，変種が異なると特徴的な語彙にはどういう変化があるか，そういったアメリカ英語以外の英語に対する神経が，コーパスに触れることで研ぎ澄まされる。

④ ネイティブに対する劣等感が変わる

コーパスで1億語規模のものを自分で操作し表現集のように活用すれば，隣にネイティブがすわっているようなものだ。自分の表現する英語が妥当かどうか，たいていの場合はコーパスをチェックすれば解決する。それに近い英語の表現が必ず見つかるからだ。今までの英語辞典にはこのような信頼は置けなかった。常に感じていた自分の英語力不足への不安感，ネイティブへの劣等感のようなものから解放される。

⑤ 役に立つ表現への姿勢が変わる

コーパスを検索することで初めて，より一般的なもの，より使用頻度の高いものに対する客観的な判断を下せるようになる。これは重要度の軽重の記述がない平面的な文法書・参考書を使っていたのとは次元が違う。

単語集や熟語集もコーパスから抽出した頻度による「有用性」でフィルタをかけることで，「この表現が役に立つのだ」というより強い確信をもてる。指導法も変わってくる。

⑥ 学習事項に対する見方が変わる

特殊コーパスの進化により目的に応じて「ESPコーパス，教科書コーパス，入試問題コーパス」などを自作することが可能になってきた。この連載でも紹介する DIY コーパス（Do It Yourself つまり自作コーパス）を最大限に活用すると，どのような学習事項を生徒に与えればよいか，といった判断をより確信をもってすることができる。

⑦ 学習者の英語力観が変わる

やはり特殊コーパスの部類に入るが，学習者の作文や発話データをコーパスにした「学習者コーパス」を用いれば，生徒の英語力の発達過程をモニターできる。ネイティブのコーパスと自分の生徒のコーパスを比べる，または，もっと上級の日本人英語学習者と比べる，といったプロセスを経て，自分の生徒の英語力に対するより深い理解，適確な判断が下せるようになる。

4 コーパスを入手する

さてそろそろ読者の皆さんもコーパスを触ってみたくなっただろうか？ これから1年間の連載をしていく中で，本格的にコーパスを利用するにはやはりコーパス本体を購入しなければならない。しかしそのためには若干の時間がかかるだろうから，今回はその手続きの仕方だけ紹介するので，ぜひ興味のある方は入手してみてほしい。

(A) ICAME Corpus Collection CD-ROM

http://www.hit.uib.no/icame.html

2 で述べたコーパスのうち Brown，LOB などの 100 万語規模のコーパスはほとんどが ICAME の Corpus Collection という CD を1枚買えば一度に手に入る。ICAME は「アイケイム」と呼ばれ，ノルウェーのベルゲン大学に置かれた英語コーパスの配布機関である。配布価格はシングルユーザー 3500 クローネで約 5 万円くら

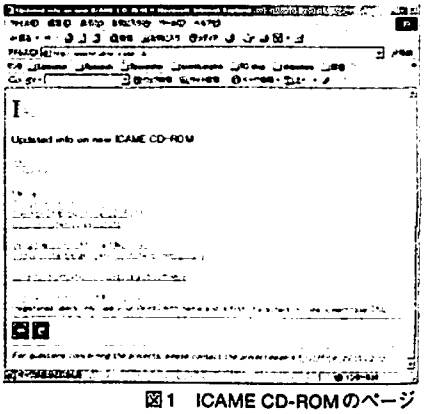

図1 ICAME CD-ROM のページ

い。これには後述する Windows ベースのコーパス検索ソフト WordSmith のユーザーライセンス料も含まれている。この1枚で自作したら何十年とかかるだろう主要コーパス21種類が手に入るのだから安い買い物である。現在はクレジットカードで直接オーダーできるので便利だ。

(B) British National Corpus (BNC) World Edition
http://www.hcu.ox.ac.uk/BNC/

BNC は1億語のイギリス英語のコーパスだ。90年代の前半に開発され，つい最近まで EU 圏内でのみの利用制限があったが，ついに全世界に公開された。BNC を利用した辞書には OALD, LDOCE, CIDE などがあり，その信頼性と権威は折り紙つき。1000万語の話し言葉データはその半分近くが日常の自然な会話を特別に採取したもので，それに非常に詳しい話者の年齢・性別・社会階級などの属性が細かく記録されている。

BNC World Edition は研究用にライセンス料は10ポンド（約2000円），CD の配布手数料に50ポンド（約1万円）という信じられない安い値段で入手可能だ。

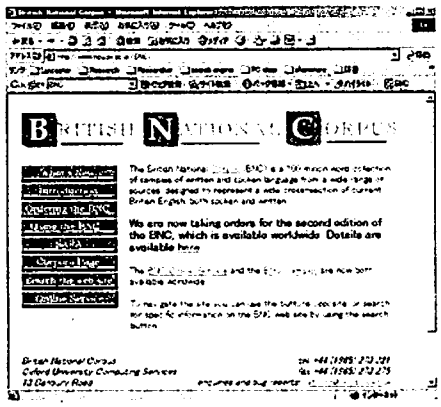

図2 British National Corpus のページ

(C) WordbanksOnline (COBUILD Direct)
http://titania.cobuild.collins.co.uk/wbanks.html

かの COBUILD シリーズの元となった Bank of English は現在4億語以上の規模を誇るが，一般へのサービスは COBUILD Direct として5,600万語に限定されている。最近，ハーパーコリンズは英語，フランス語，スペイン語の3種類のコーパスを WordbanksOnline という名称で新たにサービスを始めた。英語部分は基本的には従来の COBUILD Direct とまったく同じ物だが，インターフェースに従来の telnet 以外に JAVA を用いた web 経由のアクセスを提供する。ただしライセンス料は1年間500ポンド（約10万円）とかなり高価。

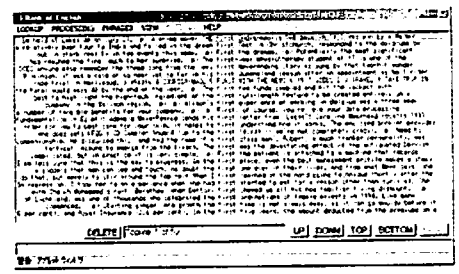
図3 Bank of English の JAVA による検索画面

(D) Web での検索サービス

ICAME も BNC もライセンスを購入したユーザー向けに簡単な web 検索サービスを提供している。もし複雑なソフトウェアのインストールなどが苦手な場合は，このようなインターネット経由でのサービスを利用するのも手だ。しかし，大規模コーパスの検索はやはりインターネット経由では相当時間がかかる。現在のように常時接続の環境が整って，自宅でブロードバンドのインターネット利用ができる場合はぜひ試してみるとよい。

これから1，2年のうちに国内外でもコーパス検索のための web サイトが利用可能になっていくだろう。小学館では筆者が監修になって現在大容量コーパス検索の専用システムを開発しており，BNC のサービスなどを開始する予定。一般ユーザー向けの環境はこのように着々と整いつつあるのだ。

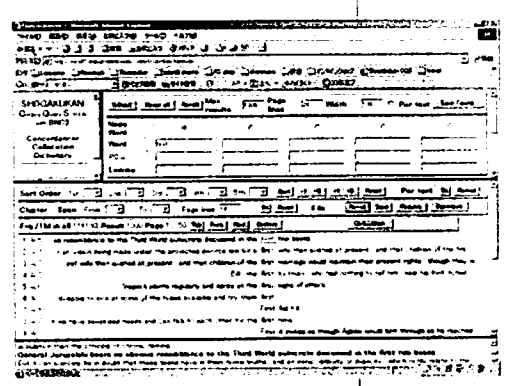
図4 小学館コーパス検索システムで BNC を検索しているところ

連載ではこれらのサービスのうち主要なものを取り上げて，より詳しく使いこなしを解説をしていきたい。それまでに興味のあるコーパスの使用ができる状態にしておくといいだろう。これらの正規に提供されるコーパス・データ以外に，自作コーパスや特殊コーパス（学習者コーパスなど）の可能性も連載では触れていくので乞うご期待。

5 検索ツールを準備する

最後にこの連載には不可欠のコーパス検索ツールに関して解説しておこう。コーパスは通例普通のテキスト・ファイルにいろいろな書誌情報や品詞などの単語情報，名詞句などの構文情報，そして文・段落などの情報がタグの形でテキスト内に付与されている。検索ツールは一般ユーザーがこのような複雑なデータを見やすく表示したり，高速に検索したりするために用いる。極端な話，専用の検索ツールがなくても，汎用の UNIX コマンド（grep, sort, uniq, wc, sed, awk）などを使用すればかなりの処理が可能だ。しかし一般の英語教師に UNIX コマンドを習いなさい，とはちょっと言えない。そこでここでは現在入手できる検索ツールの主要なものを紹介しよう。なおお断りしておくが，ここでは紙面の都合上マイクロソフトのウィンドウズで動作するものに限定する。

(1) 商用コンコーダンサー

■ WordSmith（ワードスミス）
http://www.liv.ac.uk/~ms2928/index.htm

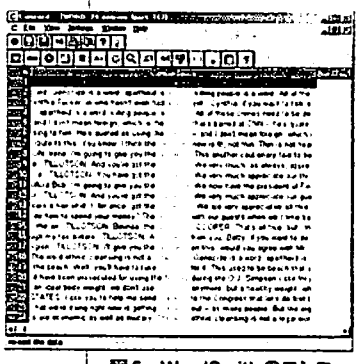
図5 WordSmithのコンコーダンス画面

Windows 用のコンコーダンサーとしてはモノコンクとともに最もよく知られている。リバプール大学の Mike Scott が開発。現在バージョン3が OUP から正式にリリースされている。単語リスト作成・比較，コンコーダンス，キーワード分析の3種類のモジュールを組み合わせて分析できる。BNC, ICAME などに標準添付され，事実上スタンダードと言ってよいだろう。現在最新ヴァージョンを鋭意開発中である。

■ MonoConc Pro（モノコンク・プロ）
http://www.ruf.rice.edu/~barlow/mono.html

ライス大学の Michael Barlow の作。彼のコーパス言語学のページは非常に有名。モノコンク・プロはワードスミスと人気を二分するコンコーダンサーで，タグ付きテキストの処理や正規表現検索（テキスト検索の際に複雑な記号を用いて柔軟な検索式を書ける）が得意。アメリカでは彼のプログラムのほうが普及しているようだ。

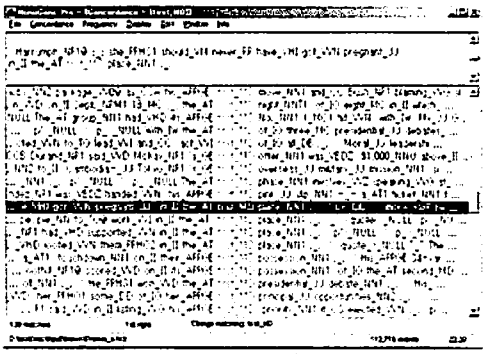
図6 MonoConc Proのコンコーダンス画面

■ TXTANA Standard Edition（テクスターナ）
http://www.biwa.ne.jp/~aka-san/

国内では最も人気のある国産コンコーダンサー。翻訳家の赤瀬川史朗氏の作。クエリー機能を使って検索条件を絞り込んでいく探索的な処理が直感的に使いやすく，またワードスミスよりも複雑な正規表現検索が可能。コンセプト辞

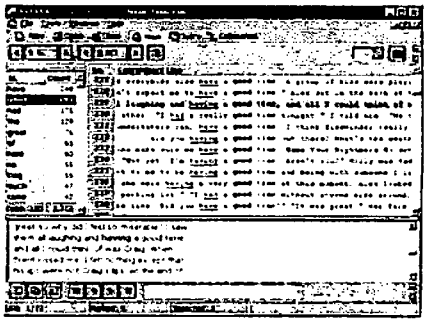
図7 TXTANA Standard Editionのコンコーダンス画面

書という検索語の変化形・活用形の辞書をカスタマイズすることで検索対象にさまざまな条件を付加できる。日本語版ウィンドウズだとワードスミスは多少不安定でテクスターナのほうが堅牢である。

■ WordPilot (ワードパイロット)
http://home.ust.hk/~autolang/

香港科学技術大学のJohn Miltonの手になる商用コンコーダンサー。ウィンドウズ 版での安定感はワードスミスを凌ぐ。純粋に研究用というよりは現場密着型の作りで，直感的な操作でコンコーダンスラインのパタンを調べたりするのに向いている。ミルトン本人は，大学内での英作文の支援ツールとして活用している。学生に検索させたい単語一覧をインタフェースの一部として自作することが可能で，単語学習とコンコーダンサー機能をリンクした学習教材を気軽に作成できる。また，その単語のコンコーダンスラインから自動的に穴埋めテストを作る機能等もある。

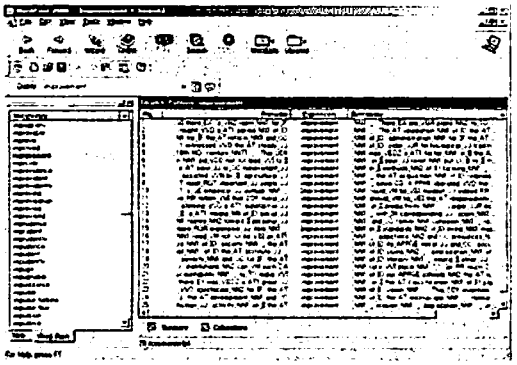

図8 WordPilotのコンコーダンス画面

(2) フリーウェア

■ ConcApp for Windows
http://vlc.polyu.edu.hk/scripts/concordance/WWWConcapp.htm

香港工科大学の Chris Greaves 率いるグループの Virtual Language Centre というサイトから web concordancer のページに行き，その一番下からダウンロードできる。ウィンドウズ版で機能はいたってシンプルだが，一応ソートやコロケーションのパタンも出せる。日本語のフォントも扱えるので，日本語コーパスも検索可能。

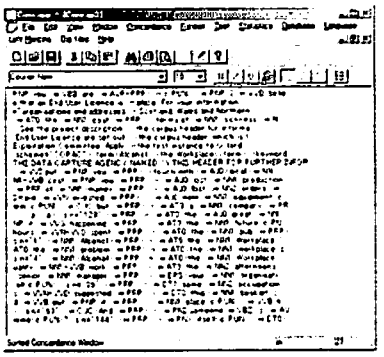

図9 ConcApp for Windowsのコンコーダンス画面

■ Text Finder
http://www.biwa.ne.jp/~aka-san/textfinder.htm

TXTANA の作者赤瀬川史朗氏の手による簡易コンコーダンサー。TXTANA の旧バージョンとほぼ同じインタフェースで日本語もこなせる。正規表現も使えるので，かなり複雑な検索式も自分で書け，基本的なコンコーダンスの作成はこれで十分まかなえる。

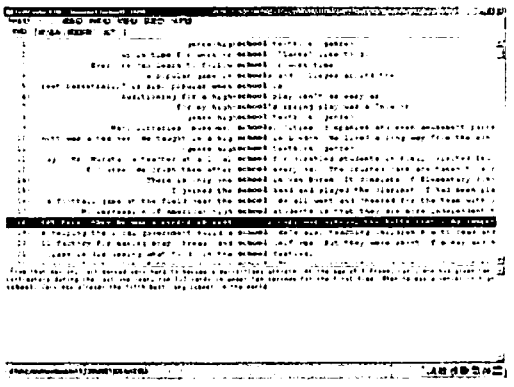

図10 Text Finderのコンコーダンス画面

参考文献

Biber, D., S.Conrad, & R.Reppen (1998). *Corpus Linguistics.* Cambridge University Press.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics.* Longman.

McEnery, T. & A. Wilson. (2001). *Corpus Linguistics.* Edinburgh University Press.

斎藤俊雄 他(編)(1998)『英語コーパス言語学:基礎と実践』. 研究社出版.