英語教師のための コーパス入門

第3回

コーパスからの頻度リストを活用する

明海大学　助教授
投野 由紀夫

はじめに

さて，いよいよ自分の手元にあるコーパス・データを利用して何ができるのかを具体的に見ていくことにしたい。なお，今回から基本的にコーパスの処理は Oxford University Press から発売されている WordSmith Tools または Michael Barlow 氏の開発した MonoConc Pro というプログラムを用いて行う。WordSmith の入手方法およびインストール方法に関しては，作者の Mike Scott のページ（http://www.liv.ac.uk/ ~ms2928/index.htm）を参考にしていただきたい。デモ版がダウンロードできるので，まずはそれをもとにどのような処理ができるか，今回の内容をもとにチェックしてみるとよいだろう。MonoConc Pro は Barlow 氏の webpage を参考にしていただきたい。

コーパスから頻度リストを作成する

コーパスから得られる最も有益な情報の１つは，「単語リスト」だ。そのサンプル全体がどのような語彙によって成り立っているか，というデータ全体の鳥瞰図を「単語リスト」は示してくれる。早速，WordSmith を使って単語リストを作成してみる手順を紹介しよう。

WordSmith で単語リストを作成する

1) Tools Controller から［Tools］－［Wordlist］を選択。
2) Wordlist の画面で左隅の緑の丸のアイコンをクリックする。
3) Getting started の画面で［Choose texts now］を選択。
4) ここで任意のファイルを選択する。
5) Getting started の画面に戻ったら，［Make a wordlist now］のボタンをクリック。
6) 自動的にワードリストが作成される。

図１が WordSmith によって作成された British National Corpus の書き言葉部分の単語リストである。

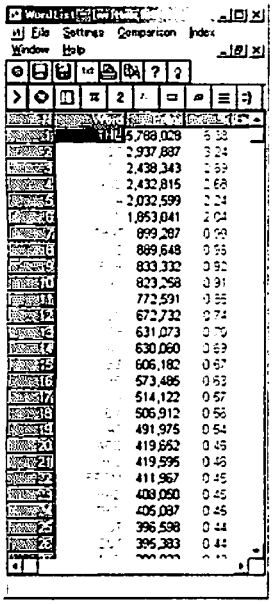


図1：BNC written corpus の単語リスト

図1の単語リストは頻度順のリストであるが，WordSmith ではこれ以外にアルファベット順のリストと，コーパス全体の総語数などの統計値をまとめたファイルを作成してくれる。図2はその統計情報をまとめたファイルである。

図2に示されているように，コーパス全体のデータサイズ（バイト数），総語数（tokens），異なり語数（types），異なり語の総語数に占める割合(type/token ratio)およびその標準化された値（Standardized type/token），平均単語長（Ave. word length），文の数（sentences），平均文長（Sent. length），段落数（paragraphs），段落の長さ（Para. length）などの詳細な情報がコーパスごとに瞬時に得られるので，大変便利だ。

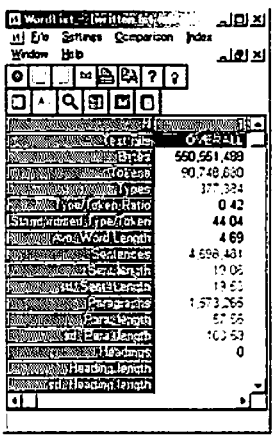


図2：コーパス統計情報のファイル

単語リストをどう利用するのか？

このような単語リストを実際にどのように中高の英語教員が利用できるのか，を考えてみることにしよう。単語リストのおもな活用分野として，今回は以下の3つの点を中心に解説をしてみたい：

(1) 学習語彙の選択・重み付け
(2) プロダクション用語彙の選定
(3) 英作文指導と評価の資料

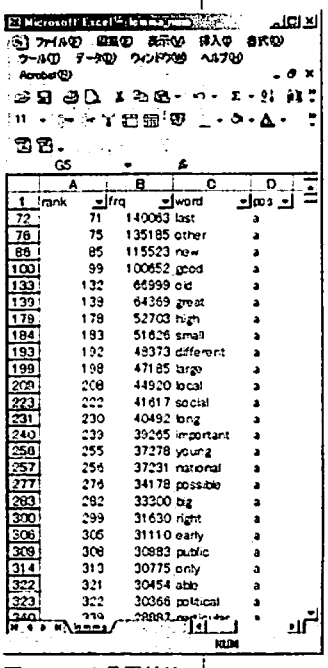
では，早速実際の応用例を見てみることにしよう。

(1)学習語彙の選択・重み付け

コーパスから単語リストを作ることのメリットの1つは，どのような単語がより頻繁に用いられるかをデータによって直接確かめることによって，学習語彙の選定に役に立てられる，ということである。「学習語彙の選定などは，文科省とか教科書の著者がすることだ」と思うかもしれない。しかし，コーパスを用いることで，中高の教員が自らの手で選択したコーパス・データをもとに語彙表を作成し，重要度に関する明確なイメージを持ったり，教材の配列や選択を考えたりということが可能な時代になってきている。実際に試してみれば，それほどむずかしいしいことではなく，誰もが利用できるのだ。

たとえば，Brown Corpus と LOB Corpus のデータが手元にあれば，100万語のアメリカ英語とイギリス英語の語彙リストが瞬時に作成できる。これを合成して両者に共通な基礎語彙リストを作れば，英米の違いを加味したリストを自分なりにアレンジできる（注:合成するためには2つのコーパスを同時にWordSmithなどの検索ソフトにかければよい）。これにBritish National Corpus 1億語のデータを加えて，3つの標準コーパスに載っている上位2000語くらいを抜き出せば，それだけでかなり一般的な基本語リストができ上がる。

さらに品詞タグ（POS tag）を使って語彙リストを作成できれば活用度は飛躍的にアップする。品詞タグ付コーパスは単語1つ1つを品詞解析プログラム(POS tagger)によって品詞タグ付与を施したもので，BNCはその代表的なものだ。BNCから単語リストを作成することで，たとえば品詞ごとの頻度表を作成できる。こうすることで，単純な頻度リスト1000語を持っているよりも，そこから基本動詞リストだけを抜き出したり，最もよく用いる形容詞100を抽出して早めに指導する，といったようなことが可能になる。図3はBNCからの頻度リストをエクセルに移し変えて，オートフィルタという機能を用いて品詞ごとのリストを自動的に作成できるようにしたものである。このような工夫を施すことで，単語リストそのものが非常に利用価値の高いものになる。



1	rank	frq	word	pos
72	71	140063	last	a
76	75	135185	other	a
86	85	115523	new	a
100	99	100652	good	a
133	132	66999	old	a
139	139	64369	great	a
179	178	52703	high	a
184	183	51626	small	a
193	192	48373	different	a
199	198	47185	large	a
209	208	44920	local	a
223	222	41617	social	a
231	230	40492	long	a
240	239	39265	important	a
256	255	37278	young	a
257	256	37231	national	a
277	276	34178	possible	a
283	282	33300	big	a
300	299	31630	right	a
306	305	31110	early	a
309	308	30883	public	a
314	313	30775	only	a
322	321	30454	able	a
323	322	30366	political	a

図3：BNC品詞付リストで高頻度形容詞を見る

(2) プロダクション用語彙の選定

語彙リストが役に立つ第2の例として，英作文とか英語スピーチのようなプロダクション活動の場面を考えてみよう。最近はライティングの授業もコミュニケーションの手段や目的を重視したライティングや，1文ずつの和文英訳ではなく一定のまとまった英文を書かせるような発展的なライティングが志向されている。スピーチでも，ある特定のテーマやトピックに関して，まずライティングで基本的なアイディアをまとめ，発表活動を行ったりすることが多い。この際に問題になるのは，その課題となるテーマに関するトピック語彙が不足して表現できない，という問題だ。多くの場合，教師は英作文のテキストのモデル・パラグラフの中のボキャブラリーで何とかまかなおうとするのだが，実際の作文ではより広範囲の語彙を与えておかなければ，とても対応できない。

こんなときに威力を発揮するのが，自分のクラスで書かせた作文データをコーパス化しておくことである。たとえば，1年目にライティングで書かせたいトピックをシラバスとして組み，1年がかりで自由英作文を書かせるとしよう。それらのトピックごとの作文を夏休みなどの余裕のあるときを見て，コンピュータに入力する。(もちろん，コンピュータ上で書かせる作業をできればこの手間は省ける。) こうしていくつかトピック別の作文データがミニコーパスとして集まったら，個別に前述のような方法で単語リストを作る。ただ，これだけだと単純なリストなので何がトピック語彙として重要なのかわからない。そこで登場するのがキーワード分析 (Keyword Analysis) というWordSmithの機能である。これは簡単に言うと，自分の作ったミニコーパスと基準になる大きなコーパス (たとえばBrownとかBNCなど) の2つを比較して，基準になるコーパスに比べて極端に高頻度に出てくる単語をそのミニコーパスの特徴を示すキーワードとして抽出する方法である。

WordSmith でキーワードを抽出する

1) Tools Controllerから [Tools]−[Wordlist] を選択し，それぞれ2つのコーパスに関して単語リストをまず作成する。
2) 続いて，[Tools]-[Keyword]を開く。
3) 2種類の単語リストを選ぶウィンドウが現れるので，左側に生徒の英語コーパス，右側に基準となるBNCなどのコーパスの頻度リストのファイルを選択する。
4) [OK]を押すと，双方のリストを比較した結果が出力される。

さて，実際のアウトプットは図4のようになる。

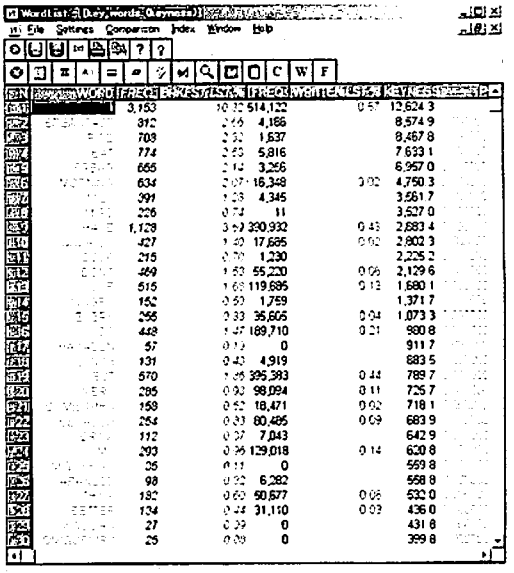


図4：Keyword 分析の結果画面

Keynessという部分のスコアが高ければ高いほど，生徒の書いた作文中で一般の英文に比べて特に高頻度に出てきている単語と言える。図4の例では「朝食」に関する作文なので，リストを見ると朝食に関して表現したい語彙 (breakfast, rice, eat, bread) が上位に来ていることがわかるだろう。

より精密にするためには，このリストからいわゆる機能語と基本語のリストを除いてしまう (これはstop wordといって普通の単語リストから上位の機能語・一般語をピックアップしてファイルにしておき，フィルタに使うわけだ)。そうすると，トピック語彙が鮮明に見えてくる，というわけだ。

この方法を用いて，1年に5〜6本程度の作文の課題をもとにトピック語彙を抽出できれば，翌年の授業に使える。今度は，トピック語彙のリストを与えて書かせてもいいし，それらのトピック語彙のリストとコーパスを両方用意しておき，実際にコーパスで実例を検索させながら指導してもよい。いずれにしても，もっと具体的に「書きたい表現」が何かを科学的に知ることができる方法である。

私が中心に収集しているJEFLL Corpusではこの情報からさらに進んで，「どういう単語が英語にできないか?」という情報をコーパス内に組み込んでいる。この情報があれば，語彙リストとして「生徒が英語にできなかった単語リスト」を作ることが可能だ。それらの詳細な分析はまた別の機会に譲りたいが，そのようなリストの中からもトピック語彙として生徒が

知りたい単語を抽出する貴重なリソースとなるだろう。

このような作文のトピック語彙を先生同士で共有すればよい。できればもとになったコーパス・データもシェアできるともっとよい。そうすることにより，英語教師が準備不足で悩んでいるライティングやスピーチ，ディベートなどの活動にボキャブラリーの面で大いに貢献することができるだろう。

(3) 英作文指導と評価の資料

最後に私がこの２年間行ってきたライティングの授業の話をしよう。昨年明海大学に赴任して初めて担当した Writing Ⅰの授業では「一定のまとまった英文を継続的に書かせる」ということを目標にした。コンピュータ・ルームを利用して，テキストはなし。２週間で１本のトピックに関して自由英作文をさせる。その媒体としては，いわゆるBBS（掲示板）のフリーソフトを自分のサーバーに設置して，そこに書き込みさせるようにした。

これだけではあまりおもしろくない。そこで，全員でBBSに投稿後，１人が最低５名の友人の作文を読んで，それに対して「もっと知りたいこと」をコメントすることを義務付けた。こうすると最低でも５人くらいから自分の作文に対するpeer reviewを得られるので，revisionの際の参考になる。

これだけでもまだおもしろくない。そこで，自分の書いている作文の文法の誤りに関する意識を高めるために，私が各作文に対して問題箇所をアステリスク（*）で印をつけてBBSにコメントとしてアップした。添削はしない。考えさせるためである。

学生はこの私の文法の誤り箇所のマークと，５名の友人からの参考意見をもとに再度自分の作文を書き直す。その結果を最終バージョンとしてBBSに再度アップする。

今年度はさらにこの方法から工夫をして，最終的にアップした作文をプリントアウトし，１日10回音読させ，１週間後に抜き打ちで皆の前で作文を見ないで口頭で発表するという宿題を課している。これで，書いたものを口で言えるようになるサポートを何とかしようというわけだ。

この方法はBBSというコンピュータ環境を使っているために，教師の負荷が非常に少ないので大変効果的だ。同じことを紙と鉛筆でやろうとすると，作文のコピーを作ったり，添削の記録をつけたりという教員の単純作業の時間が大幅に増えてしまう。熱心な先生はそれをやっていると思うが。

そして最後に登場するのがコーパスである。学期末になると，BBSのデータを個人データとして個別に整理して一人一人の作文コーパスを作る。そして，その個人の作文の評価に単語リストを活用するのだ。学生全員のデータを単語リストにすることで，WordSmithによる個人の作文の統計情報が出る。これで，学期中に書いた作文の総語数はどのくらいかがわかる。１つ１つの作文で書く量が増えたか減ったかもわかる。異なり語数を見ればどのくらい語彙が豊富かもわかる。キーワードを抽出すれば，一般の学生の作文に対して，どういう表現を特に使いすぎているか，どういう表現が出てこないか，といった大まかな語彙使用の状況もわかる。これを全体のコーパスを処理した平均データと併記して，一人一人にレコードとして返却する。自分がクラス全体でどの程度作文をがんばったかが実に鮮明に数値化されて，学生は非常に感心するのである。図５はその学生に返却するデータの一例である。

評価項目	平均	学生A
総語数(半年の作文総数)	1,702	2,436
異なり語数	526	687
語彙の豊富さ（全体）	32	28.2
語彙の豊富さ（作文毎）	36	40.3
平均の文の長さ	7.3	8.9
平均の単語の長さ	4	4.27

（このあとに他の学生との語彙比較リストを添付。）

【コメント】クラス平均以上にたくさん書けた。単語もいろいろな表現を使う努力が見られる。

図５：作文評価の返却例

まとめ

コーパスは研究の道具だけでなく，実際の英語指導と結びつけたときにも威力を発揮する。コーパス言語学といっても，それはデータを解析したり，指導法に活かしていくための「方法論」なのであるから，自分の自由な発想で活用すればよい。今回は単語リストにしぼって話をした。次回はもう１つのコーパス検索プログラムの大きな機能であるコンコーダンスをとりあげて，具体的な用例を見たり調べたりすることで英語教員にとってどのような活用が可能か考えてみたい。