新しい波：学習者コーパス

「学習者コーパス」という用語は、この記事を読む多くの方にはまだ耳慣れない言葉であろう。英語では (computer) learner corpora (corpora は corpus の複数形)とか learner's corpus という言い方をする。簡潔に言うと、学習者コーパスとは「ある外国語を勉強する学習者が話したり書いたりしたものを大量に収集し電子化したもの」である。

学習者の output を調べてみるというのは別に目新しい発想ではない。昔から授業分析などでは、プロトコルを作って一生懸命データの分析をしてきた。しかし学習者コーパスは、何よりもそういった学習者データをコーパスとしてコンピューター内に大量整備するという点が違う。これはこの 10 年ほどで急速に進歩したコーパス言語学や自然言語処理のおかげだ。学習者データを大量に収集し、レベル別、技能別、抽出タスク別などのインデックスをつけて管理し、コーパス言語学のノウハウを応用して統語構造や語彙使用についての使用頻度や例文検索、結びつきのパタン、誤用例等々の情報を容易に取り出せるようにデータの整備をしたもの、これが学習者コーパスである。

このシリーズでは 1 年間にわたって、英語学習者コーパスがどのようにして作られるのか、英語学習者コーパスをもとにどのようなことが分かるか、またその研究を進めていくことによって英語教育がどう変わるか、といった点を深く掘り下げてみたいと思う。そして、この分野の研究が日本においてもますます発展することを期待したい。

学習者コーパスで何ができるのか？

学習者コーパスの開発はまだこの 5, 6 年で研究が始まったばかりだ。次回に詳しく世界の学習者コーパスの紹介をしたいと思うが、世界的にもまだそれほど規模や内容的に充実したものは存在していない。しかし、もしこれが本格的に整備されるならば、我々は大変な恩恵に浴することになるのである。ではその具体的なイメージを読者諸兄に持っていただくために、学習者コーパスがあると何が可能か、例を挙げて考えてみよう。

学習者コーパスの利用が期待される分野として、大きく分けて (1) 文法・語彙指導、(2) 第 2 言語習得研究、(3) 辞書・教科書・教材開発が考えられる。それぞれその可能性について見てみることにする。

(1) 学習者コーパスと英語学習・指導

はじめに、指導現場で学習者コーパスが威力を発揮するようなケースをいくつか挙げてみよう:

◎ ケース 1:

A 先生は語彙指導でいつも悩んでいる。新出の文法事項を練習させるにしても、語彙が足りない。簡単な語彙で繰り返し練習をすると生徒はすぐに飽きてしまう。生徒に身につけさせたい語彙は山ほどあるが、たとえば平均的中学 2 年生なら普通どのくらいの語彙を身につけているのか？ 果たして自分が練習で使おうとしている語彙は、この段階で導入するのに適しているのか？ といった疑問に答えてくれる的確な資料がない。

◎ ケース 2:

B 先生は生徒のよくやる英語の間違いをどうやって直してやればいいか迷っている。毎回注意してもまったく直らないような間違いもあれば、すぐ直るものもある。しかしいつもその場限りの処置で、きちんとした判断基準がない。

◎ ケース 3:

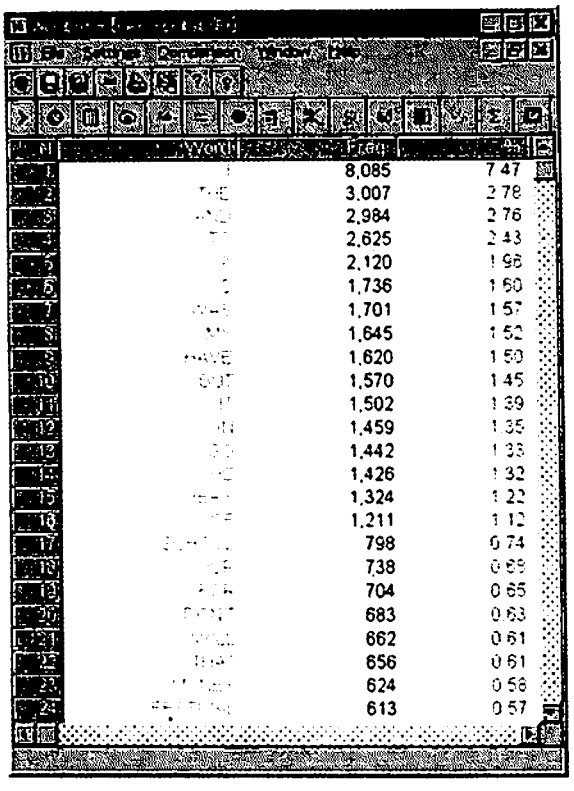
C 先生は長年スピーチの指導をしているが、いつも生徒の英語を添削するのに追われている。スピーチのテーマもいろいろ考えて指導してきたが、もっと生徒自身に英語で自己表現をしてもらいたい。しかし和英辞典を使うと変な英語を書いてしまうし、テーマ別語彙リストなども作ってみたが、トピックも語彙も広範囲でなかなか多様な生徒のニーズに答えられない。

新連載 学習者コーパスと

このような場合、今まではほとんど教師が自分の経験から来るカンに頼るか、または非常に研究熱心な先生ならば、自分でこつこつ生徒の書いたデータをためて語彙表を手作りして、それをもとに判断するということがせいぜいだった。しかし学習者コーパスの開発によって、より客観的な学習者の英語使用の実態に迫ることができるようになる。学年ごとに学習者の実際に書いたり話したりした英語を大量に収集することで、どのレベルの学習者がどういう単語をどのように使っているかが客観的なデータとして見られるわけだ。「論より証拠」という発想なのである。では具体的に個々のケースにどのように対処できるのか、見てみることにしよう。

(2) 学習段階ごとの語彙使用の実態把握

学習者コーパスが整備されれば、ケース1のA先生の悩みは氷解する。全国1万人の中学生が1年から3年までの間に書いた英文や英会話のデータが1000万語のコーパスになっているとする(まだこのくらい大規模なものは実在しない)。そこにインターネットでアクセスして、自分の生徒と同じ公立中学の2年生でどのくらいの語彙を使うか、語彙リストを抽出して調べてみることができる(そのイメージが図1だ。ただしこのデータの規模は筆者が現在整備中のもので、まだ12万語ほど)。A先生はここから自分が選んだ語彙が上位何%に入る単語か、学習させてもいい妥当な語か、といったことの基礎的な判断材料を得ることができる。



	Word	Freq	%
1		8,085	7.47
2		3,007	2.78
3		2,984	2.76
4		2,625	2.43
5		2,120	1.96
6		1,736	1.60
7		1,701	1.57
8		1,645	1.52
9		1,620	1.50
10		1,570	1.45
11		1,502	1.39
12		1,459	1.35
13		1,442	1.33
14		1,426	1.32
15		1,324	1.22
16		1,211	1.12
17		798	0.74
18		738	0.69
19		704	0.65
20		683	0.63
21		662	0.61
22		656	0.61
23		624	0.58
24		613	0.57

図1 単語の頻度リスト (WordSmith による)

このように学習者コーパスは、英語学習者の異なる学習段階での語彙使用の実態把握に、今までにない客観的で一般性のあるデータを提供できる。

(3) 英語学習上の困難点を予測

学習者コーパスは単なる英文の集まりではない。そのテクストにさまざまな情報を加えることにより付加価値が増す。その1つは学習者のおかす文法的・語彙的エラーに関する情報だ。英語学習者が実際に発話したり書いたりしたテクストの中のエラーを抽出して、その頻度などを調べる。それによって自分が気になるエラーがどの程度一般的によく起こるものか、学習段階のどのへんでエラーが消失するかなどの実態を知ることができる。ケース2のB先生も、学習者コーパスを検索して、気になるエラーをチェックし、それがどの程度普通に見られるエラーか、定着するまで何年くらいエラーが続くのか、といったことを知れば、誰でもする誤りですぐに直っているから放っておこうとか、かなり深刻な誤りだから早めに手を打とうとか、指導上の判断をすることができるだろう。

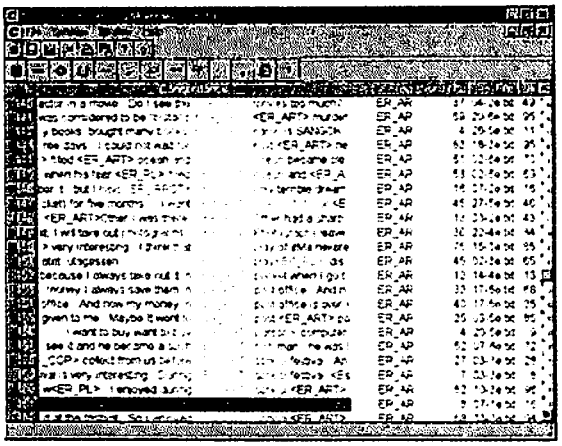


図2 冠詞のエラーの KWIC ライン (WordSmith による)

英語指導 ……(1)学習者コーパスとは何か

words 投野由紀夫

(4) コーパス・データによる表現学習

学習者コーパスは単に教師のリソースになるだけでなく、学習者自身が活用することで、和英辞典などに勝る表現辞典として活用できる可能性を秘めている。ケース3のC先生のように、スピーチやオーラル・コミュニケーションの指導で、topic vocabularyをどのように整理して与えてやればいいのか、悩んでいる先生は多い。ここで学習者コーパスが大きな意味を持ってくるのである。

コーパス・データが中学・高校・大学というレベル別、また「自己紹介」「物語の要約」「エッセイ」などのジャンル別に例文や単語の使用例を検索できるようになっていれば、学習者自身がコーパスにアクセスし、それらを比較しながら自分の作文の表現を発見していくことが可能だ。自分と同じ中学生や高校生の書いた文章を読むことで親近感も沸き、また自分が言いたいことがどのような英語になっているか、その英語が正しいのか正しくないのか、などについて学習者コーパスは情報を提供してくれる。このような方向性はdata-driven learningと呼ばれ、ヨーロッパではCALLの1つの可能性として注目を浴びてきている。

また後の連載で詳しく解説するが、日英パラレル学習者コーパスを開発することにより、日本語の表現がどのように英語に実現されるかの例を数多く見ることができるようになる。もしこのようなパラレル・コーパスが学習者用に使えるようになれば、学習者自らが自然と日本語を英語で置き換える際の構造の違いなどを発見していく可能性も秘めている。和英辞典を引いてもその単語の使い方がよくわからないという次元をはるかに越えて、自分の言いたいことを実際に学習者の書いた英語でああも言える、こうも言えるという例が何十も出てきて、一番ぴったりくる表現を選んだりできれば、これはもう和英辞典の比ではない。パラレル・コーパスがうまく整備できれば、これを学習用に利用することで大変便利なツールになることは請け合いである。

このように学習者コーパスを教育目的に使おうという試みも現在盛んに研究されてきており、大きな可能性を秘めているのである。

新しい第2言語習得研究の可能性

このように現場の英語教師が悩んでいたいろいろな語彙指導や文法指導の問題に対してヒントやデータが提供できるだけではない。学習者コーパスは、第2言語習得研究全般にも大きな貢献ができる。例を挙げてみよう。

(1) 英語学習者の中間言語の記述

学習者コーパスは、学習段階ごとにコーパスデータを整備することにより、中間言語の記述に役立つ基礎データを提供できる。たとえば、表1は中学2, 3年および高校2年の主要な文法的形態素の定着度の推移をコーパスで分析した結果を示している。かつてのDulay & Burtらの形態素の習得研究などもこのような大量のデータによって再検証が可能だ。コーパスを用いることにより、さまざまな語彙・文法の学習段階ごとの使用状況や誤りの様子などが、大量のデータによりかなり客観的に分かるようになるわけである。

表1 主要な形態素の定着率の推移（数字：%）

表 1

	中学2年	中学3年	高校2年
Copula be	94.17	96.26	94.74
Aux be	89	96.1	92.45
Possessive -s	76.67	76.19	95.24
Plural -s	80	81.04	88.51
3rd person -s	70.83	69.57	89.36
Irregular past	82.28	79.62	83.69
Article	63.02	70.24	79.62

(2) 学習者データの蓄積と共有

従来の第2言語習得の研究では、一度に大量のデータを取ると言ってもせいぜい100名単位、経年的データを取る場合には10名以下が限度であった。ドイツのZISA Projectなどはそれをいろいろ組み合わせてかなり大量のデータを取っているが、これもコーパス化はされていない。以前の研究者が取ったデータの質や中身については案外正体不明なのである。

学習者コーパスは研究者のリソースの共有も目指している。研究目的と方法がある程度一致すれば、データを共有し公開したほうがより建設的な

議論ができる、発見したことの再検証もしやすくなるというわけだ。

(3) 母語の影響とエラー分類

外国語習得では必ず母語の言語知識の影響が研究される。現在、ベルギーのルバン大学では International Corpus of Learner English というプロジェクトが動いている。これは 11 の異なる母語の英語学習者の作文データを組織的に収集し、コーパスを用いたエラー分析をすることで母語が異なるにもかかわらず共通に起こる「普遍的エラー」(universal errors) と、母語の干渉による「母語関連エラー」(L1 related errors) の仕分けをしようという試みだ。

このような試みによって今まであちらこちらでばらばらに行なわれていた個別の習得研究を大量のデータベースをもとに再検証する試みが行なわれつつある。この連載でも、今後そのような習得研究における発見も紹介していきたい。

辞書・教科書・教材開発

最後に、学習者コーパスが貢献できる分野として、辞書編纂、教科書・教材の開発を見てみよう。学習者コーパスが整備されてくることによって、これら英語教材の開発はより科学的な根拠をもったものになるだろう。

(1) 辞書編纂

ご存知の方も多いだろうが、現在出版されている学習英英辞典のほとんどは現代英語のコーパスをベースに作られている。*OALD, LDOCE* などは British National Corpus を参照し、*COBUILD* は独自の the Bank of English という超大型コーパスを用いている。このような状況で、どの学習辞典も次回の改訂での利用が期待されているのが学習者コーパスなのである。学習者のおかす誤りのデータから、学習困難点を予測し、その語彙項目に特別なコメントを付することで特徴を出そうというわけだ。その先鞭を付けているのがロングマンで、*LDOCE* ではすでに 500 万語(当時)の学習者コーパスからデータを抽出して語法・コラム記事などに活用している。

日本でも英和・和英辞典の作成にコーパスは不可欠になるだろう。その際に自社の現代英語コーパスを作るというのは、もう世界の潮流から見ると時代遅れだ。これからは目的を特化したサブコーパスの作成で特徴を出さないと、後塵を拝している日本はデータの質的には勝てないだろう。その鍵を握るのは学習者データだと言って間違いない。学習者コーパスは研究用で現在開発が進んでいるが、実は商用でも十分魅力的な財産になる可能性を秘めており、現在日本の出版社でも数社、開発の基礎的な研究をしているところがある。

(2) 教科書・教材開発

語彙習得の専門家と話していると、教科書の語彙表はどのような基準で選ばれたのか根拠がはっきりしない、といつも言われる。日本のような環境では段階を追ったシラバスと語彙表の整備は非常に重要だ。学習者コーパスの構築が進めば、日本人の学習者の語彙レベルとネイティブ・スピーカーの語彙レベル、またネイティブの小学生の語彙レベルなどとの比較もできる。また英米で生活していると普通に使う日常生活語彙がどのくらい教科書に載っているかとか、学習者がそれを知っているか、といったこともすぐに調べられる。教科書や付属教材も、学習者のレベル別の語彙表や特に誤りの多い項目のリスト、他のコーパスと比較したデータなどがあれば、もっと教材の選択・配列などに科学的な基準を得ることができる。

おわりに

今回は連載の第 1 回目として、学習者コーパスを用いると何ができるかという点に絞って解説した。大きな新しい研究の潮流として、学習者コーパスは今後絶対に目が離せない研究分野である。

次回は世界の学習者コーパスの現状を解説する予定である。世界で、また国内で現在どのくらい学習者コーパスの開発は進んでいるのか？ 我々英語教師はそれをどのように利用できるのか？ そういった点に関して詳しく解説しよう。

(とうの・ゆきお / 元東京学芸大学講師、
ランカスター大学博士課程在籍)

学習者コーパスと英語指導
(1)学習者コーパスとは何か