特集
〈徹底研究〉
語彙習得のメカニズム

コーパス言語学がもたらした新たな語彙指導

投野由紀夫
Tono Yukio

はじめに

本稿ではコーパス言語学 (corpus linguistics) の進展により得られた新たな知見が，語彙指導や語彙習得研究の分野にどのような示唆を与えうるかを考察する。コーパス言語学そのものの概要については本稿では触れない。興味のある方は文献解題に紹介した概説書を参照されたい。

コーパス言語学は特定の目的で収集された大量の言語テキストをコンピュータ処理することにより種々の言語統計を抽出して言語研究を行う分野である。最近20年間のコーパス言語学の成果で言語教育に最も関連のあるものとして，(1)母語話者の英語使用頻度データ，(2)コロケーション強度，(3)学習者コーパス研究の3点を取り上げて主要な成果を概観してみよう。

母語話者の英語使用頻度データ

言語教育において言語材料の選定は教える側にとっては最重要課題の1つである。言語の文法を形作る最も基本的な要素をそれに含める，という考え方以外に，よく用いるものを含める，という「使用頻度」の観点は常に関心事であった。言語教育では20世紀初頭の教育測定運動を背景としたThorndikeの研究（特にThorndike & Lorge 1944が代表的なもの）が人手による頻度集計の作業としては最も大規模なものであった。

1964年に米国ブラウン大学において世界で最初の科学的な標本抽出によるブラウン・コーパス (Brown Corpus) が完成した。その後，一連の100万語規模の英語変種コーパス（Hofland et al. 1999）が作成され，それらに基づく研究書（Johansson & Hofland 1989など）により100万語規模でのジャンル別，品詞別などの語彙統計が明らかになった。さらに80年代後半からはCOBUILD Projectを中心にした大規模コーパス構築の動きが盛んになり，1994年にはBritish National Corpus（1億語のイギリス英語コーパス）が完成。1995年には主要な英英学習辞典（*LDOCE*, *COBUILD*, *OALD*, *CIDE*）のすべてが大改訂を行い，ほとんどの辞典がコーパスの語彙頻度をもとにした重要語ランク表示や定義文の制限語彙 (defining vocabulary) を設けるようになった。1億語のBNCをもとにした頻度リストが公開され（Kilgarriffのリストや Mike Scottのリストが有名），BNCによる言語統計の研究書も出版された（Leech et al. 2001）。

コーパス利用により，母語話者の書き言葉，話し言葉の語彙使用頻度に関してさまざまなことがわかってきた。まずは基礎語彙の占める割合が圧倒的に多いという事実である。表1はBNCの話し言葉コーパス1000万語における高頻度語の占める割合をまとめたものである。

基本語ランク	全体に占める割合
1～ 100位	67.2%
1～ 500位	83.5%
1～1000位	89.0%
1～3000位	95.9%

表1：BNC spokenにみる上位語の割合

驚くべきことに上位100語の単語があれば1000万

語のほぼ7割弱（つまり700万語近く）をカバーしてしまう。最初の1000語があれば約9割を占める。このように話し言葉ではごく限られた単語がコアとして用いられていることが分かる。だからといって，1000語を単純に丸暗記していれば会話ができるようになるわけではない。これらの単語は実は基本的であればあるほどその意味や用法が多岐にわたる単語群なのである。そしてこういった基本語の十分な使いこなしをしっかり身につけることが日常会話では重要ということを示唆している。

これらコーパス言語学からの最新の語彙分析データをもとに日本でも語彙表の改訂が盛んになっている。JACET8000 (2003) はその中でも欧米の大規模コーパスに日本人英語学習者の接するべき英文コーパスを照合して作成した語彙表として，コーパス言語学の応用の好例だといえよう。

コロケーション情報

コーパスから得られる情報として2つ目に重要な点は，単語同士の共起 (collocation) 頻度が得られるということである。この共起情報は特に「名詞＋名詞」といった複合語のパタン，「形容詞＋名詞」，「副詞＋形容詞」といった修飾関係や，「動詞＋名詞」，「動詞＋前置詞・副詞」といった動詞補部のパタンを把握するのに威力を発揮する。

例えば，sweet という形容詞がどのような名詞を修飾するのか BNC で検索してみると表2のようになる（数字は単純頻度で sweet の直後1～3語のスパンで計算）。

sweet	smell	73
	pea	63
	shop	57
	potato	46
	thing	46
	tooth	45
	wine	40
	tea	37
	taste	36
	smile	35

表2：sweet ＋名詞の top 10

語彙学習の際に sweet ＝「甘い」という1対1対応で覚えさせるだけでは active vocabulary としては不十分であろう。リストを見ただけでも，smell, taste, tea, wine のように「(味覚・臭覚として) 甘い」という場合は比較的単語の意味だけで応用が利きそうだが，sweet smile のような派生的な意味を使いこなせる必要があるし，sweet pea (スィートピー)，sweet shop (売店)，sweet tooth (甘党) などは高頻度なのにあまり我々は sweet からの連想としてぱっと浮かばないのではないだろうか。逆に日本語からの発想でいけば「甘い誘惑」という表現はよくするが，sweet のコロケーションとして temptation は1件しかない。逆に temptation のコロケーションを検索すると，great, strong のような形容詞で強調する方がずっと一般的であることがわかる。こういった語彙知識の「盲点」的なものをコーパスは我々に教えてくれる。それを活かした教材開発の可能性は今後ますます重要になるはずだ。

表2は単純頻度でリストを集計しているが，実際は単純頻度では結びつきの強度は十分には分からない。具体例を挙げてみよう。先ほどの sweet＋名詞のリストの第1番目は smell であるが，smell が73回出てくるという事実と pea が63回出てくるという事実から sweet smell の方が sweet との結びつきが強いと断言できるだろうか？　実際はそうはいかない。というのは，コーパス全体における smell と pea の単体での発生頻度が異なるからである。単純頻度のみを比べていては，標準化された共起頻度比較ができない。実際に smell と pea のそれぞれの単体での頻度を調べてみると smell が2148回，それに対して pea の方は複数形とあわせても783回であった。つまり割合的に言えば，smell と pea の個別の確率と，sweet smell, sweet pea などの共起確率を厳密に勘案しながら共起強度を算出しなければならないのである。

自然言語処理 (Natural Language Processing: NLP) の分野ではこれらの共起統計は類似度測定 (similarity measure) として情報抽出，機械翻訳などの領域で広く活用されている (Matsumoto and Utsuro 2000)。そして単純頻度以外のさまざまな結合強度を示す統計値が考案されている。代表的なものとしては相互情報量 (MI)

(Church and Hanks 1990), t-スコア (ibid.), Φ^2 統計 (Gale and Church 1991), Dice 係数 (Salton and McGill 1983), 対数尤度 (Dunning 1993), log-log 値 (Kilgarriff の考案) などが挙げられる。これらより高度な統計値を組み入れたコーパス検索ソフト (例えば小学館コーパスネットワーク (http://www.corpora.jp), BNCWeb, Qwic, XARA など) を用いればコロケーション情報はより有効に抽出可能だ。例えば, log-log 値という MI にさらに対数をかけた指標によれば, sweet peas が最も log-log 値が高く第1位, 逆に sweet smell は第4位であった。共起頻度と絶対頻度がずれた場合にはこれらのより精密に頻度の意味づけを行う統計を活用するべきである。

学習者コーパス研究

最後にコーパス言語学のもう1つの新しい波として学習者コーパスの動向を紹介しておこう。学習者コーパス (learner corpus) は, コーパス言語学の手法を用いて学習者の発話・作文データを大量に収集・分析する研究分野である (最新情報は http://leo.meikai.ac.jp/~tono/ 参照)。この10年ほどでコーパス言語学の言語教育への応用分野として脚光を浴びてきている (Granger 1998 参照)。ここでは最新の研究成果の一端として Tono (2002) を紹介する。Tono (2002) では学習者コーパスを用いた動詞下位範疇化情報の習得を研究した。動詞型は従来からさまざまな分類が提唱され, また言語習得の分野でも項構造 (argument structure) の獲得として意味・統語の両分野が交差する興味深い領域として知られている。またコンピュータ言語学では格フレーム (case frame) の自動獲得という分野の問題として, 自然言語から自動的に動詞型パタンを抽出する技法が試みられている。

Tono (2002) ではこれらの複合領域的な分野の知見を用いながら, (1)学年別の英作文コーパス (約31万語) にみる高頻度動詞の下位範疇化パタンの習得度 (Apple Pie Parser を用いた構文解析および正規表現検索によるパタン抽出), (2)中高英語教科書 (約250万語) におけるそれらの動詞下位範疇化パタンの頻度 (上記同様の手法), (3)日本語コーパス (自作の新聞データを中心に1100万語程度) における日本語動詞の (形態素解析ツール「茶筅」とパタン検索・目視による) 下位範疇化パタンの頻度, (4)日英の主要動詞の下位範疇化情報の類似度 (IPAL および COMLEX Lexicon による類似度データベースを作成), (5) Beth Levin の動詞分類にもとづく意味特性および交替文 (alteration) の分類, といった多面的要素を15の英語動詞すべてに関してコーパスおよび lexical database より抽出し下位範疇化フレーム・データベース (Subcategorization Frame Database) を構築した。

このデータベースを用いながら, 第二言語における英語動詞の下位範疇化情報の獲得が次の要素のどの影響を最も強く受けるのか, およびその交互作用の大きさを検証した:

(a) 英語習得独自の発達的要素
(b) (英語教科書に見る) input の多寡
(c) 英語・日本語動詞のパタン類似度
(d) 英語特有の動詞意味特性
(e) 英語下位範疇化パタンの使用頻度

これら多変量の複合モデルを作成し, それを最も経済的 (parsimonious) なモデルに絞り込むための統計手法として対数線形分析 (log-linear analysis) を用い, 上記の変数間の最も強い因子とその関連モデルを15の動詞に関して特定した。

分析の結果, 以下のような点が明らかになった:

① 下位範疇化情報の習得度とインプットの量は無関係であった。むしろインプットの量は動詞そのものの overuse, underuse と関係が深かった。
② 習得度と顕著な関係にあったのは, 日英動詞の下位範疇化情報のずれの度合い, および頻度のずれといった日英の差によるもの, また動詞の意味特性のようなより普遍的な要素であった。ただしこれらの影響の強弱は動詞によって異なる結果を示した。
③ 学習段階が上がっていくにつれ, 下位範疇化情報を無理なく習得する動詞とそうでない動詞とに分かれた。これらの動詞間のグルーピングは Beth Levin の分類でコレスポンデンス分

析を行って要約しようとしたが十分に意味あるグルーピングが得られなかった。今後の研究に待たねばならない。

この研究の特徴は，従来教室環境での外国語習得研究で十分に捉えきれなかった母語の影響，主要なインプット源である教科書の影響，それらと発達的・普遍的な習得要因との相対的関係をコーパス言語学と多変量解析の手法を用いて因果関係モデルとして検証を試みている点である。そのため単一の学習者コーパスのみを用いて行われていた従来の研究に比べると，格段に複雑な習得モデルを扱うことが可能になる。

Tono (2002) ではこのように単一の学習者コーパスのみの比較を行うだけでなく，学習者コーパスと母語コーパス，目標言語のコーパスなどを多重比較するアプローチ (multiple comparison approach) を提唱している。またそれにふさわしい多変量統計手法を絡めることにより，より有意味な理論モデル構築や先行研究のより精密なコーパスによる再検証などが可能になると述べている。

学習者コーパスは近年新しい研究方法として注目を集めてきている。今回は紹介できないが，学習者の発話・作文の語彙的特徴を習得段階別や母語話者のものと比較すると，さまざまな知見が得られる可能性を秘めている。現在 ICLE (International Corpus of Learner English) の1部データが公開されているほか，私が中心で構築している JEFLL，アルクのスピーキング・テスト (SST) の音声データをコーパス化した SST Corpus などの日本人学習者のデータも公開に向けて整備中である。

まとめ

以上，本稿ではコーパス言語学が可能にする語彙学習や語彙習得の可能性に関してまとめてみた。コーパス言語学が教育分野に応用され始めてまだ10年足らずである。今後は教材開発，第二言語習得研究の分野でコーパスはさまざまな応用可能性が探られるに違いない。コーパスは「道具」であるから，各研究分野でその使い道を活発に議論してみてはいかがだろうか。

【参考文献】
Church, K. and Hanks, P. (1990). "Word association norm, mutual information, and lexicography." *Computational Linguistics* 16:22-29.
大学英語教育学会基本語改訂委員会 (編) (2003).『大学英語教育学会基本語リスト JACET List of 8000 Basic Words』.
Dunning, T. (1993). "Accurate methods for statistics of surprise and coincidence." *Computational Linguistics* 19:61-74.
Granger, S. (1998). *Learner English on Computer*. London & New York: Addison Wesley Longman.
Gale, W. and Church, K. (1991). "Identifying word correspondences in parallel texts." *Proceedings of Speech and Natural Language Workshop*, Orange Grove, CA, 152-157.
Hofland, K., Lindebjerg, A. & Thunestvedt, J. (1999). *ICAME Collection of English Language Corpora, CD-ROM*. Bergen: HIT-senteret.
Johansson, S. & Hofland, K. (1989). *Frequency Analysis of English Vocabulary and Grammar. Vol. 1-2*. Oxford: Clarendon Press.
Leech, G., Rayson, P. & Wilson, A. (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Longman.
Matsumoto, Y. and Utsuro, T. (2000). "Lexical knowledge acquisition." In Dale, R. et al. (eds.) *Handbook of Natural Language Processing*. New York: Marcel Dekker, Inc.
Salton, G. and McGill, M.J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.
Thorndike, E.L. and Lorge, I. (1944). *The Teacher's Word Book of 30,000 Words*. New York: Columbia University Press.
Tono, Y. (2002). *The Role of Learner Corpora in SLA Research and Foreign Language Teaching: the Multiple Comparison Approach*. Unpublished PhD dissertation. Lancaster University.

(明海大学外国語学部助教授)