

コーパスに基づいた EST 教育

English for Science and Technology Using Corpus Based Approach

小山 由紀江

(名古屋工業大学)

koyama@center.nitech.ac.jp

1. 始めに

ESP においては多面的なニーズ分析が不可欠であるが、その基本となるのが実際の言語運用を解明するテキスト分析である。工業系大学ではさらに ESP の領域を限定し English for Science and Technology (EST) の立場に基づいた英語教育が要請され、テキスト分析については、近年コーパスに基づいた言語分析法が急速に発達している。本研究は、独自に構築した科学技術系英文のコーパスに基づいたテキスト分析の結果を考察し、工業系大学の英語教育 EST の内容を確定するための手法と、今後の英語教材作成の方向性を提示することを目的とする。

2. EST コーパスの構築

本研究では、EST の観点から工学系学術論文(以下 J コーパス)と一般科学雑誌の記事(以下 M コーパス)の2種類のコーパスを独自に構築した。J コーパスについては対象を電気関連の分野に限定した。内容のバランス、分野を代表するものであるか等考慮し、電気関連の異なる分野から5種類の論文誌をデータソースとした。データは web と CD-ROM から採取したもので、最終的にはそれぞれ約 10 万語のサブコーパスから成る全体としては約 50 万語のコーパスとなった。M コーパスは Scientific America の web 上の記事(1997 年から 2001 年)を採取したもので、これも全体として約 50 万語から成る。

3. 分析結果と考察

本研究の分析にはコンコーダンサー: WordSmith Tools Version 6.0 と品詞タグ付けソフト: Brill's Tagger Version 1.14 を使用した。WordSmith によるコーパス J と M の基本統計は以下の通りである。

	J コーパス	M コーパス		J コーパス	M コーパス
Tokens	521,829	593,714	Ave. Word Length	4.88	5.02
Types	15,417	29,566	Sentences	19,353	24,862
Type/Token Ratio	2.95	4.98	Sent. Length	26.52	23.77

3.1. 頻度表から見る 2 分野の共通性と差異

3.1.1. 高頻度語彙

WordSmith Tools で得られた頻度表を精製し、さらに中学既習語 1000 語を除き、J コーパスと M コーパスそれぞれの頻度表を作成した。これらを比較した結果、以下の特性が見られた。100 位までの語彙の中で両コーパスに共通している語彙は以下の 21 語 (base, cell, change, control, cost, data, design, energy, form, however, line, model, problem, process, set, speed, state, surface, system, thus, while) である。注目すべきなのは system が両コーパスにおいて頻度第一位で、学術論文、雑誌の違いを問わず広く使用されていることである。又、data も両者の上位に共通して出現している。このように両コーパスに共通して高頻度で出現する語彙は、EST の教材作成に際し最重要語彙として採用すべきであろう。M コーパスの高

9月5日(金) 研究発表1 第8室(521)

頻度の語彙が一般的な語彙 (e.g. computer, university, space, human, technology, percent, information, drug) であるのに対し、J コーパスにおいてはやはり専門用語 (e.g. value, voltage, circuit, flow, load, input, output, variable, current, algorithm, parameter, signal, thermal, combustion) が中心であり、電気関連の専門分野の学生には、これらの語彙を中心とした EST 教材を作成することが必要であろう。

3.1.2. 過去分詞形の頻度

頻度表の精製作業をする前段階では動詞は現在形、過去形などが別々に表示されるが、J コーパスでは、単純頻度 100 位までの動詞の場合、過去形 (あるいは過去分詞形) の出現数が現在形よりはるかに多いことが解った。そこで、過去分詞の出現頻度を調べるために両コーパスの全テキストにタグ付けを行い、動詞の種類別の出現率を調べたところ、過去分詞形は J コーパスでは 4.37% と動詞の変化形のうち最も高いことが確認された。又、M コーパスにおいても *-ing* 形 (2.82%) に次いで高い率 (2.64%) であった。工学部の学生が英文読解の際犯す誤りに、形容詞的用法の過去分詞と主動詞の取り違えがあることはよく指摘されるが、テキスト分析の点からも、過去分詞形の理解が専門論文の読解に重要であることが確認された。

3.1.3. shown の用法

上記のように過去分詞形の使用頻度が高いことが確認されたが、実際に過去分詞形が文章の中でどのように使われているかを *shown* を例として調べてみた。*shown* は J コーパスにおいて 808 回出現するが、さらに WordSmith の clustering 機能を使い *shown* に関連する連鎖の頻度を検索したところ、1 番頻度が高かったのは *shown in fig* で 382 回であった。

<過去・過去分詞形と現在形の出現頻度比較>

単語原形	過去・過去分詞	現在形
show	shown 808 / showed 97	245
base	816	139
propose	534	63
obtain	502	115
find	287	76
compare	277	67

<shown の clustering による頻度>

順位	cluster	Freq.
1	shown in fig	382
2	as shown in	206
3	is shown in	192
4	are shown in	133
5	shown in figure	110

cluster の頻度では第 2 位の形 *as shown in* をさらに調べると *as shown in* の頻度 206 のうち、やはり一番多かったのは *as shown in fig* あるいは *figure* で 178、*as shown in table* 17、その他のパターンが 11 である。他方、M コーパスにおける *as shown in* の出現回数は 1 回のみ *as shown in three days of demonstrations* という形であり、一般的な記事の文章では非常にまれにしか使用されないことが解った。

4. まとめと今後の課題

以上が、頻度表の考察を通した本研究で得られた知見であるが、これらはそれぞれ 50 万語を越える 2 つのコーパスの言語データの語るところのほんの一片に過ぎない。今後は、この言語材料をより精確に分析し、EST に適切な教材作成を実現することが課題である。語法の特徴をより正確に把握するためにはタガーとコンコーダナーを組み合わせる品詞情報を含む分析をすること(これはより精確な単語頻度表の作成を可能にし、分野別語彙リストの教材化の可能性につながる)、構文解析ソフトを使用することにより構文の特徴を解明することの 2 点も、今後、着手すべき課題である。いずれの場合も、基本は分析の元となるデータであり、今回構築した 2 つのコーパスを量と質の両面からより充実させてゆくことを常に念頭に置かなければならない。そのような専門分野に特化した良質なデータを用いることによって初めて、専門研究に直結した効率的な英語 web 教材の作成が明確に視座に据えられるわけである。