

## 【補助資料 7】

### 3 Web As Corpusの利用

BNC, Cobuild, Michigan Corpus of Academic Spoken English, Virtual Language Center 等, 実に様々なものが利用できる。

しかし最も便利なのは Web 自体ではなかろうか。

Adam Kilgarriff 氏(2003)は次のように書く。

"The web is immense, free and available by mouse-click. It contains hundreds of billions of words of text and can be used for all manners of language speech."

"Language scientists and technologists are increasingly turning to it as a source of language data, because it is so big, because it is the only available source for the type of language they are interested in, or simply because it is free and instantly available. The mode of work has increased dramatically from a standing start seven years ago with the web being used as a data source in a wide range of research activities: the papers in the Special Issue form a sample of the best of it."

そして、この論の先で彼はこう断定する。

"The answer to the question 'Is the web a corpus?' is yes."

Zurich 大学の Martin Volk 氏(2002)は次のように述べる。

"Corpus linguistics, in the sense of using natural language samples for linguistics, is much older than computer science. This dictionary makers of the 19th century can be considered Corpus Linguistics pioneers (e.g. James Murray for the Oxford English Dictionary or the Grimm brothers for the Deutsches Wörterbuch). But the advent of computers has changed the field completely."

"Linguists started compiling collections of raw text for ease of searching. In a next step, the texts were semi-automatically annotated with lemmas and recently with syntactic structures. First, corpora were considered large when they exceeded one million words. Nowadays, large corpora comprise more than 100 million words. And the World Wide Web(WWW) can be seen as the largest corpus ever with more than one billion documents."

Volk 氏は standard な corpus は不便であるとさえ言う。

"Corpora distributed on tape or CD-ROM have some disadvantages. They are limited in size, their availability is restricted by their means of distribution and they do no longer represent the current language by the time they are published. The use of the web as corpus avoids these problems. It is ubiquitously available and due to its dynamic nature represents up to date language use."

さらに次のように言い切る。

"Using the web for Corpus Linguistics is a very recent trend. The number of approaches that are relevant to Computational Linguistics is still rather small. But already the web has been tried for tasks on various linguistic levels: lexicography, syntax, semantics and translation."

ここで考えられる便利なものは Internet の search engine ではなかろうか。これはすでに Mike

Russell が "The Biggest corpus of all" と題して "Yahoo!" や "Altavista" を検索に使用することを 2000 年に提案していることでもある。

また、Thomas Rob 氏(2003)も Google をツールとして利用することを論じている。

こうした search engine, 特に Google の発展・成長は著しいものがある。コーパスとしてのデータも驚くほど膨大で、しかも日々増大している。

このプラス面も実に大きく、Rob 氏は次の 3 点を挙げている。

(1) It is much more accessible than any corpus.

(2) The database is huge compared to any existing corpus.

(3) The Index sites include blogs and discussions which come very close to spoken language whereas much of the data in formal corpora are from more corpora

しかし、彼も指摘するように決して理想的なものではなく、問題もある。

(1) You can only search for specific words, not word categories or inflected forms.

(2) There is no control over the educational level, nationality, or other characteristics of the creators of the utterances found.

(3) The results are not in a easy-to-read format.

やはり問題となるのはデータとしての信頼性である。英語を第 1 言語としない国々の資料も多い。言語資料としてその「正しさ」の面で心配なサイトも多い。

Google の場合、サイト指定検索、サイト・ドメイン指定検索等を方法で信頼できる情報のみを取得可能である。エクスクルーディング（マイナス検索）の機能もあり、排除したいサイト等を情報を除くことも可能である。ワイルドカード検索等の機能とともに活用することにより、有用な資料が十分得られよう。

WWW 上の膨大なデータは近年注目を集めており、Web データを 1 つの巨大なコーパスと考え、コンコーダンスを作成されるようになった。代表されるものは次の図の WebCorp であろう。

(図 1) Webcorp の検索画面

WebCorp

WebCorp Advanced Wordlist Generator Guide Publications Feedback

**NEWS:** Significant improvements in the speed of WebCorp and range of processing options available, as it becomes part of a fully-tailored linguistic search engine... [more]

Search term:  
Enter a word, phrase (no quotes necessary) or [pattern](#)

Search engine: Google

Concordance Span: 5 word(s) to left and right

Case options: Case Sensitive

Send Results by Email:  
Option temporarily unavailable

[Advanced Search Options](#)

Submit

By using the WebCorp tools you are agreeing to be bound by the [Terms of Use](#).

© 1999-2006 Research and Development Unit for English Studies Privacy Policy

このツールでは Google だけでなく AltaVista 等, 他の search engine を使った検索が可能である。Advanced の利用でもっと細かな設定をすることにより, さらに便利な利用が期待される役立つツールである。

この WebCorp を論じている論文の中で Andrew Kehoe & Antoinette Renouf(2002)は Web を利用する利点を論じている。

"Corpus linguistics is the study of a body of electronic text to discover facts about the language which are not observable or quantifiable by manual means. However, the design and creation of text corpora can be expensive and corpora are fixed at a point in time; they do not provide access to up-to-date information on language use or the changes which are occurring.

"An obvious source of such language data is the Web. As a text-based information source, the Web also has tremendous potential value as a linguistic resource. It is orders of magnitude larger than any finite corpora, constantly updated and expanded, broad in domain coverage and potentially available without cost to the research community."

日本にも専修大学の佐藤弘明氏が作成した GuggleFormatter があったが, 現在は一般の人の使用はできない形になっている。しかし佐藤氏のホームページから ggleGrab を使わせていただいて一部の機能はまだ利用可能であり, サンプル資料として活用できるようである。

コンピュータで検索するコーパスの出現により, "Data-driven learning"(DDL)という新しい英語教育法が生まれた。学習者がコンコーダンスーを利用して, 多くの実例の触れ, 語法や Collocation 等に関する知識を学習者自身が身につける手法であり, 主に大学で利用されはじめているという。

最後に, "KMT's"のサイト (<http://www001.upp.so-net.ne.jp/google/>) に掲載されている「英語学習に Google をいかす ― コーパスとしての活用」を引用しておきたい。

「KMT's 英語学習者のサイト」を開設してから寄せられた質問を整理すると, 学校英文法や学習参考書だけで学んでいるだけでは身につかないが, 英語を母語として日常的に使っている人 (以下ネイティブ) なら即答できるものが多かった。ネイティブには全く気にならない英語の言い回しについても, 日本人学習者にとっては厄介で悩みの種となっているようである。

英語学習者がそういった問題に直面した時, Google(グーグル) が手助けとなる。ある表現について疑問を感じた時, 直接ネイティブにたずねることのできる環境に恵まれているならともかく, 一般的には英英辞典・英和辞典・和英辞典・語法辞典……を駆使して調べるしかない。また, 仮にネイティブに訊くことができたとしても, 答えに納得がいかなかったり, 訊いた相手に答えが変わってしまったりして疑問が残ってしまうということも頻繁にある。その人がネイティブであったとしても, 生活してきた国や地域, さらに受けてきた教育で使われる英語は驚くほど変わり, その人がどこの国の人か, どういう教育を受けてきたのかという問題も考慮しなくてはならないからである。

Google(グーグル) の機能をいかすと, その英語が実際にどのような場面で, どのように使われているのか, 日常の表現から学術的な分野にいたるまで, 簡単に調べることができる。Google をコーパスとして英語学習に活用するにあたっての有益な機能は次の通りである。

Google は検索語が含まれているページを検索結果として表示するだけでなく, 検索語を太字で表示する。また, 検索結果画面の"Cached"をクリックすると, 検索語がそれぞれハイライト表示されるため, 目的のことばを容易に探せる。

サーチフィールドに検索語を入力するだけで, 英英辞典・同義語辞典・類義語辞典・百科事典とし

て利用できる。綴りが曖昧な語については、スペルチェック機能で確かめることができる。綴りが違っている場合は正しい単語を提案してくれる。

Google フレーズ検索を活用すると、これまで膨大な時間を費やして調べた“ある英語表現が「実際どのように使われているか」を瞬時に調べることができる。気になる英語表現の実例が簡単に入手できるのである（フレーズ検索・ストップワーズ参照）。単語だけでなく、イディオム、構文、文の実用例まで手軽に探し出すことができる。

検索結果画面に表示されるページ数は、検索をかけた表現が、どの程度、一般的に使われているか把握する目安となる。また URL は、そのサイトに述べられている内容の信憑性を判断する手助けとなる（ドメイン資料）。例えば、英米の“政府機関の公式サイト”や“大学の公式サイト”で頻繁に使われている表現は語法上、正しいと推測できる。

サイトやドメインを指定すると、特定のサイト内だけで検索できる。よって、日頃から信頼しているサイト内で検索語を確かめることが可能となる。例えば、CNN、Newsweek、Time、The New York Times、The Washington Post 等、“報道機関の公式サイト”を限定して検索できる。

小説やエッセイから学術論文に至るまで、Web 上にある情報ならファイル形式にかかわらず検索できる。また PDF ファイルをはじめ、様々なファイル形式を指定して検索することも可能である。

以上、英語学習に役立つという観点から、Google の機能をあげた。辞書や参考書を数冊調べただけで、ある表現を「言う」、「言わない」と日本人が論ずるのはナンセンス、時間の無駄である。これまで満足いく解答が見つからなかった学習者にとって、Google で得られる生きた英語は価値あるコーパス（言語資料）となる。」

役立つ機能が多いので英語科教員として利用したいものである。

## 4 新しい辞書の活用

近年、新しい形の辞書が次々と出版された。井上氏(1997)の言葉を引用しよう。

「1995 年は英国の学習辞典界にとって記念すべき年であった。Longman Dictionary of Contemporary English (LDOCE), Collins COBUILD English Dictionary (COBUILD), Oxford Advanced Learner's Dictionary (OALD), Cambridge International Dictionary of English (CIDE), Harrap's Essential English Dictionary (Harrap) など。主要な出版社から売れ筋の辞典の改訂版や注目すべき新刊の学習辞典が申し合わせたようにいっせいに発売された。そればかりか、それらはすべてがそれぞれにコーパスを辞書編集に活用したことを主張している。」

国内でもコーパスを使った、「ウイズダム英和辞典」(三省堂)、「ユースプログレッシブ英和辞典」が生まれた。「ルミナス英和辞典」や「レクスス英和辞典」あるいはこれまで高校現場で人気のあった辞書も改訂版が出され、コーパスの利用を前面に打ち出している。

こうした新しい辞書を活用して、生徒にコーパスからの資料(説明)に注目させる指導も必要になってくる。語法は時と共に変わっている。最近とくに変化が著しい感じがある。古い辞書や情報などに頼ってはいられない状況である。

しかし、問題となるのは、やはり「新しい」用法が、「認められる」表現としてどのくらい社会に認知されているかである。少数の native check だけでは不安である。新しい辞書や Web から情報を得る教員の日々の努力も必要なことは言うまでもない。

新しい辞書の活用も期待したい。