はじめに

前回は学習者が英語で表現する際に、英語にしにくい部分を日本語で書かせた作文コーパスをもとに、英語に出来なかった部分の分析を試みた。今回は、学習者コーパスによる日本語の影響をより精密に調べるためにパラレル学習者コーパス (parallel learner corpora) を作る試みを紹介しよう。

パラレル・コーパスとは?

パラレル・コーパスとは対訳テクストをコーパス化したものである。特に機械翻訳や人工知能の分野で現在大変大きな注目を集めている。以前はこういった分野では、理論言語学の手法を利用して、文法理論をベースに限定的なルールを記述することでなるべく人間に近い処理の出来るプログラムを書こうとしていたのであるが、最近はそういったものと並行して、大量の言語データをコーパス化することにより対訳データの翻訳パタンの確率を計算し、そこから新しいデータに確率的に高い翻訳を当てはめる方法が取られている。そういった土壌の中から、パラレル・テクストの利用のノウハウが確立されてきている。

日英の場合にはNHKが大量の外電の対訳データを持っており、また郵政省やNTTでも研究プロジェクトが行なわれているようだ。しかし、大学などの研究機関でパラレル・テクストの大きなデータを持っているところはあまりないだろう。ましてや英語教育の分野では、今後の研究を待たねばならない。しかし近い将来、言語学や言語教育の分野でもパラレル・テクストをもとにした研究はますます盛んになるはずである。

英文和訳・和文英訳研究との違い

対訳のテクストは昔から数多くあった。また、英文和訳や和文英訳の研究は対訳テクストの分析という意味では、パラレル・コーパスの起源と言えなくもない。例えば、和文英訳の訳例をコーパス化している好例としては、広島大学の三浦省五氏のコーパスがある (http://www.ipc.hiroshima-u.ac.jp/~d052121/eigo1.html)。この資料は中学教科書に出てくる新出事項のターゲットセンテンスをもとに和文英訳をさせた訳出例をすべてその頻度とともにリストしている労作である。氏は教科書語彙の研究などにも功績があり、コーパスということをあまり話題にする前からこのようなデータの採取などを積極的に行なっていた。まさに先見の明があったと言えるだろう。

ただ、我々がここでパラレル・コーパスと言っているデータは、やはりいくつかの点でこのような単なる和文英訳データとは異なる。その特徴をいくつか挙げておこう:

① 全文が1対1に対応した対訳を持っている

parallelという名の通り、コーパスのすべての文がそれに対応する翻訳テクストを別ファイルとして持っている。これにより、双方向で柔軟な文字列検索が可能になる。

② パラレル・データを同時並行して検索可能

単純な対訳のリストではなく、ダイナミックに語彙検索が可能になる。例えば、図1・2は筆者がプロトタイプで作成中の学習者コーパスのパラレル版をParaConcというソフトで検索している例である。図1では、英作文データのほうでangryを検索している例が出ている。これに対して、図2では日本語の表現のほうから、「怒って...」という部分を検索した例である。このように日本語のある表現が英語でどのように表現されているかが、柔軟に検索可能になる。

③ 高度なデータ分析の可能性を持つ

コーパスは、テクストの内部に統語情報や意味情報を盛り込んでいくことで豊かなリソースとして活用出来るようになる。パラレル・データも文レベルの対応だけでなく単語レベルの対応のタグなどをつける技術も進んできており、学習者データの場合には母語の干渉に関する情報をタグにして埋め込んでみるなどの試みが可能だ。

④ データの共有

電子化されており、プログラムなどを使って検索

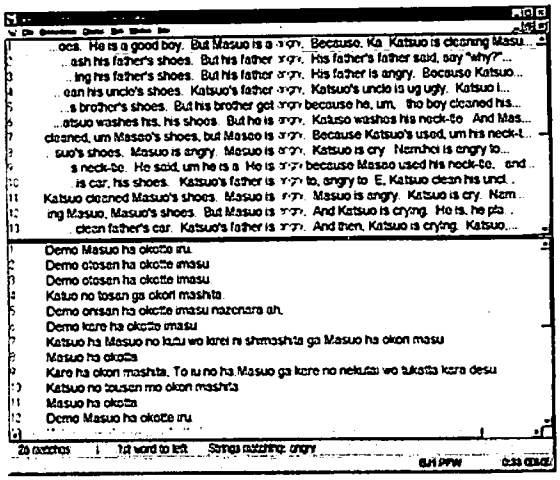

図 1

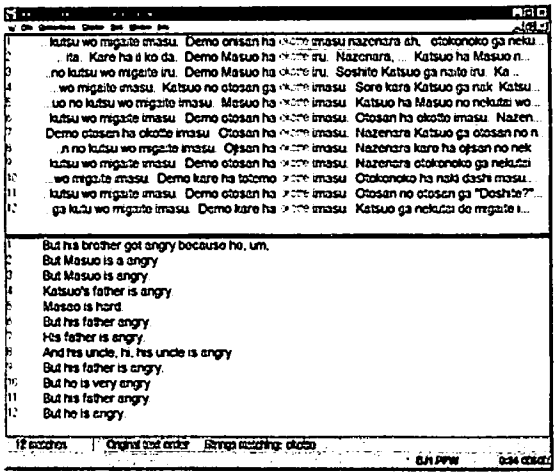

図 2

可能なようにして提供できる。共有資産として研究用に公開していけば、その利用価値は計り知れない。

パラレル・コーパスのプログラム

さて、それでは現在利用可能なパラレル・コーパスのプログラムにはどのようなものがあるのだろうか？ 普通のコンコーダンサーに比べると、あまり数は多くない。Mac または Windows で動くものでは、先ほど紹介した ParaConc がある。これは MonoConc の作者である Rice University の Michael Barlow 氏の開発したソフトである。きわめて操作性に優れて使いやすいソフトだ。Windows 版はベータ版が彼のホームページからダウンロード可能だ。

また ICAME (ノルウェーにあるコーパスや関連資料を配布する研究機関) の中心人物である Knut Hofland が TALC98 (オックスフォードで開かれたコーパスの学会)でのパラレル・コーパスのワークショップで紹介した TCE というソフトもある。私もそれに参加したが、日本語のような 2 バイト文字でも UNICODE (世界中の文字をコンピューター上で扱えるようにコード化したもの)で対応可能になると言っていた。

日本では、現在のところ日英対訳コーパスを扱っているのは、NHK と NTT、郵政省くらいしか私は知らないのだが、これらのプログラムは一般に利用可能なものではない。日本語表示が可能なパラレル・コーパス処理プログラムのよいものが 1 日も早く出来て欲しいものである。

フリーの辞書ソフトでパラレル処理

最後に、現在来年 2 月のベルギー Louvain 大学での学会に備えて開発中の学習者パラレル・コーパスのプロトタイプについて紹介しよう。

このプログラムは、日本で現在パラレル処理のコンコーダンサーに利用可能なよいものがないので、フリーの辞書ソフトである Personal Dictionary for Win 32 (開発者: TaN 氏) を使ってデータを加工して作っている。このソフトは単語登録機能が実に充実しており、ユーザーの手によってさまざまな辞書が公開されている (NIFTY-Serve FENGC (英会話フォーラム・Communication 館) で取得可能)。また単語登録と言いながらフレーズや短文も登録できるので、作文データを短く切っていけば、あたかもパラレル・データのように作文を登録していくことが出来るのである。

基本的にやっていることは非常にシンプルなことで、例えば中学生の書いた英作文があれば、その 1 文 1 文に対して日本語訳をつけていき、辞書にそれを登録すると、文頭の文字を入れれば対応する訳の日本語を表示してくれる(図 3)。

英語指導……(8) 日英比較分析 [2]

words 投野由紀夫

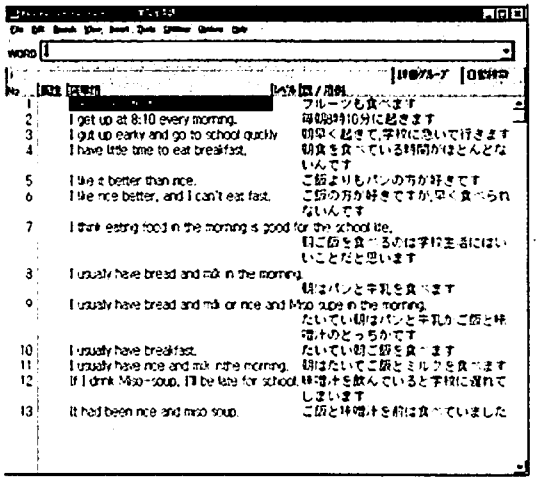

図3

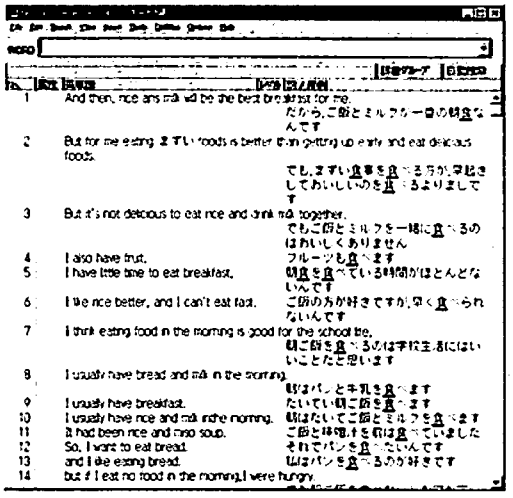

図4

これだけでは使いものにならないが、PDicWinには強力な「正規表現」(文字列検索のための記号)がついている。英語データの部分と日本語データの部分を別々に検索をかけられるので、ちょうどパラレル・コーパスのような感じで単語検索などが出来る。図4は、日本語部分の検索で「食」という漢字を検索した画面である。正規表現で絞り込めば、かなり細かい表現を抽出して、それに対応する英文を検索することが出来る。ある日本語の表現がどのような英語になっているか、そこに学習上の特徴的な誤りやつまずきやすい部分がないかどうか、などの学習困難点の分析に効果があるのではないだろうか。

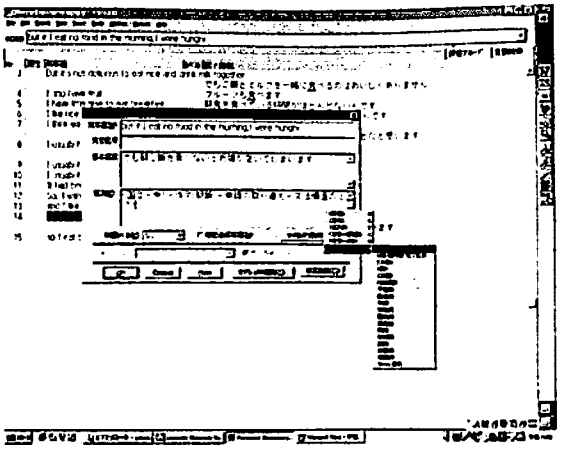

図5

日英比較エラー分析も可能に

このシステムは現在、学習者の書いた英作文データの一部を日本語化して、それをもとに対訳データを作っているが、将来的にデータが増えた段階で先ほど述べた学習困難点の分析をするために、例文にラベリングを出来るようにエラー・カテゴリーを入れられるようにしてある。

図5は、図4の検索結果をもとに発見した誤りに分類ラベルを貼りつけているところである。このようにして、日英表現の対比による誤りを発見したり、それに分類ラベルを付けたりするのを支援するプログラムがあれば、それを共有して、学習者英語がどのような日本語からの影響を受けているかを組織的に研究することが可能になる。

今回はパラレル学習者コーパスのプロトタイプの説明をした。具体的な分析結果などはまだない。今後このようなアプローチの研究がますます盛んになっていくことを期待して、その1つの参考例として現在取り組んでいるプロジェクトの一部を紹介させていただいた。先生方の中でこのような分野に関心のある方は、是非、一緒にデータ収集やノウハウの共有化を進めていきたいものである。

(とうの・ゆきお / 元東京学芸大学講師・ランカスター大学言語学科博士課程在籍: e-mail:y.tono@lancaster.ac.uk)

[訂正] 10月号の本欄 (p. 39) 図1の説明で File Driver とあるのは File Diver の誤りです。 (編集部)

学習者コーパスと英語指導