■ICAME より新 Corpus Collection リリース

ICAME (International Computer Archive of Modern and Medieval English) より新しい Corpus Collection がリリースされたので簡単に紹介する。合計 1700 万語を越える 20 のコーパスが収められている。

【タイトル】

ICAME Collection of English Language Corpora (Version 2, Bergen, June 1999; ISBN 82-7283-091-4) Coordinators: Knut Hofland, Anne Lindebjerg, Jørn Tunestvedt, The HIT Centre, University of Bergen

価格：NOK3,000〔約¥45,000〕; 1998 年に旧 CD-ROM を購入した人には NOK1,500〔約¥22,500〕で提供される。※価格はいずれもシングルユーザーの場合。

【内容】

Written

Brown Corpus *untagged / tagged <WC>
*LOB Corpus untagged / tagged <WC>
Freiburg-LOB (FLOB) <WC>
Freiburg-Brown (Frown) <WC>
*Kolhapur Corpus <WC>
Australian Corpus of English (ACE) <WC>

Spoken

*London Lund Corpus <WC>
Lancaster/IBM Spoken English Corpus (SEC) <WC>
Corpus of London Teenage Language (COLT) <WC>

Written & Spoken

Wellington Corpus (WC) <WC>
Wellington Spoken Corpus (WSC) <WC>
The International Corpus Of English—East African component (ICE-EA)

Historical

*The Helsinki Corpus of English Texts: Diachronic Part <WC>
The Helsinki Corpus of Older Scots <WC>
Corpus of Early English Correspondance, sampler (CEEC, CEECS) <WC>
The Newdigate Newsletters <WC>
The Lampeter Corpus of Early Modern English Tracts <WC>
Innsbruck Computer-Archive of Machine-Readable English Texts (ICAMET)

Parsed

Polytechnic of Wales Corpus (PoW)
Lancaster Parsed Corpus

上記のリストのうち，*印のついたものは 1991 年 2 月にリリースされた初版のときにすでに含まれていたものなので，ここでは新たに加えられたものを中心に紹介してゆく。

書き言葉では，Brown Corpus に LOB Corpus と同様の horizontal 及び vertical の tagged versions が用意された。研究目的によっては，Brown Corpus と LOB Corpus はその内容の古さから敬遠されていたわけであるが，そういった研究者に朗報がある。Albert-Ludwigs-Universitat Freiburg の Christian Mair を中心に，Brown/LOB Corpus と同じ手法で 1990 年代初期のテクストを使って構築された FLOB /flob/ Corpus (Freiburg-LOB Corpus of British English) と Frown /fraun/ Corpus (Freiburg-Brown Corpus of American English) が利用できるようになった〔tagged version はない〕。また新たに加えられたものとして，1986 年に書かれたテクストをもとにしているということを除いて Brown/LOB Corpus とほぼ同じ手法で構築されている The Australian Corpus of English がある。

一方，話し言葉では，従来は CD-ROM に含められていなかった Lancaster/IBM Spoken English Corpus が含められたほか，BNC の手法を取り入れて構築された英国のティーンエイジャーのコーパスである The Bergen Corpus of London Teenage Language〔50 万語; orthographic，prosodic 及び tagged versions を含む〕が新たに加えられた。

書き言葉と話し言葉の両方が用意されているものとしては，東アフリカ英語を集めた The International Corpus of English—East African component〔ケニヤとタンザニアの話し言葉及び書き言葉 100 万語〕のほか，Brown/LOB Corpus とほぼ同じ手法で 1986 年と 1987 年のものを中心に 1986–1990 年にわたるニュージーランド英語の書き言葉を集めた The Wellington Corpus of Written New Zealand English〔100 万語；tagged version を含む〕と，1990–1994 年を中心に，1988–1994 年の範囲でニュージーランド英語の話し言葉を集めた The Wellington Corpus of Spoken New Zealand English〔100 万語; WSC と spoken component of ICE-NZ は 9 カテゴリをシェアしている〕がある。Wellington Corpus については CD-ROM が先行されて発売されており，筆者も含めて，早まって購入したことを後悔している会員の方もいらっしゃるかもしれない。

通時コーパスの方では，新たに University of Helsinki の Anneli Meurman-Solin が中心となって構築された The Helsinki Corpus of Older Scots〔約 80 万語〕，Department of English, University of Helsinki の Sociolinguistics and Language History project team による The Corpus of Early English Correspondence〔270 万語〕及び The Corpus of Early English Correspondence Sampler〔45 万語〕，Newdigate 家〔ほとんどは Sir Richard Newdigate (d. 1710) 宛〕に送られた 3950 通に及ぶ手書きの newsletter のうち，最初の 2100 通 (13 January 1673/4 through 11 June 1692) を収録した The Newdigate Newsletters，1640–1740 年に発行された小冊子・パンフレット類 (tracts and pamphlets)で，University of Wales, Lampeter の Founders' Library に所蔵されているもの全て〔約 110 万語〕を収録した The Lampeter Corpus of Early Modern English Tracts，Dr. Manfred Markus が中心となって，主に Middle English と Early Modern English の散文テクストを集めた Innsbruck Computer-Archive of Machine-Readable English Texts の一部などが追加された。

構文解析用コーパスとして，Polytechnic of Wales Corpus 及び LOB コーパスをもとに作られた Lancaster Parsed Corpus が含められている。

最後に，コーパス分析ツールについて紹介しておく。DOS 版のツールである WordCruncher と TACT は，初版同様 CD-ROM に同梱されているが，特に WordCruncher の方は，今回新たに CD-ROM に含まれるほとんど全てのコーパスに対応したインデック

スファイルが提供されており〔上記リストに<WC>と表記してある〕, 特別なインストール作業なしに直接 CD-ROM から利用できる。すでによく知られている高機能コーパス分析ツール LEXA と，PC 上で発音記号を始めとする非標準的な文字を表示・印刷するためのツール LinguaFont suite (Ver. 5) も同梱された。注目すべきなのは JAVA アプリケーションの Qwick (Ver. 1.01) であろう。すでに BNC Sampler などにも添付されており，ご存じの会員も多いと思うが, コンコーダンスはもちろん, mutual information, t-score，z-score などを始めとする数種の統計値も得ることができる。インストールに若干コツを要するが, JAVA が動く環境であれば OS を選ばないのはうれしい。標準で FLOB が利用できるインデックスファイルがついてくるが，将来的には利用者が自分で集めたデータをインデックス化して利用できるようにする予定という〔作者からの私信より〕。Windows 版のコーパス分析ツールとして有名な WordSmith Tools (Ver. 3.0) も今回フルバージョンが同梱されており〔CD-ROM のジャケットの内側に USER NAME と CODE が印刷されている〕，この CD-ROM 一枚あればコーパス言語学の基本的な研究環境が整うと言っても過言ではない。なお，ほとんどすべてのコーパスに詳細な電子マニュアルが整備されている〔ICAMET は別途注文〕。

【照会先】

The HIT Centre
Allegt. 27
N-5007 Bergen
Norway
Telephone: +47 5558 2954
Telefax: +47 5558 9470
E-mail: icame@hit.uib.no
Website: http://www.hd.uib.no/icame/newcd.htm

井上永幸（島根大学）