現代美語語料庫（COCA）與語言資料庫（LDC）簡介

【研究紀要】

【文／編譯發展中心主任　林慶隆、專案助理　陳怡臻】

語料庫（corpus）為具語言研究價值的文字資料庫。最早建置在電腦中的語料庫是西元 1967 年由 Henry Kucera 與 W. Nelson Francis 所創建的 Brown Corpus，收錄美國文章 500 篇，約 100 萬字（林武聰，2003），該語料庫是一個平衡語料庫，平衡指語料蒐集盡量做到平衡分配在不同的主題和語式上。平衡語料庫在語言學研究上有重要價值，建構一個平衡帶詞類標記的語料庫，收集語料是初步工作，其次是語料整理，包括語料清潔、為語料分類、加詞類標記等（陳克健，1994）。隨著科技發展，現在所稱的語料庫有 3 點特徵，第一，語言資料的整合是根據某項原則或是規定，使得資料庫具代表性。例如現代學習者語料庫常與學習者中間語（inter language）分析連結並做比對，將學習者語言看成是一種規則系統；第二，這些語料通常以資料庫形式存於電腦；第三，研究者可以利用這資料庫作各種量化及質性的分析（洪千惠，2009）。

Brown Corpus 雖然以文體單一特徵來界定語料庫是不足的，但後來新建立的語料庫如 LOB（Lancaster-Oslo/Bergen，英國英文）及 London-Lund（英語口語），都還遵循 Brown Corpus 的架構。而且，Hsu and Huang（1995）為了突破語料過於單純化的線性描述，利用五個不同特徵軸（文類、文體、語式、主題、媒體）的多重分類，以增加語料庫提供研究的活用性（中研院平衡語料庫構建技術手冊，2006）。

現代美語語料庫（Corpus of Contemporary American English）與語言資料庫（Linguistic Data Consortium）是兩個常被使用的語料庫，其內容及特色如下，

一、現代美語語料庫（Corpus of Contemporary American English，簡稱 COCA）

COCA 由美國楊百翰大學語言學教授 Mark Davies 在 2008 年建立，是全球最大免費英語語料庫，收錄 16 萬筆文本、多達 4.25 億則字彙，自 1990 年至 2011 年，以每年收錄 2 千萬字的速度成長。每個月大約有 4 萬人使用，大多是語言學家、教師、翻譯人員及研究人員。COCA 具有五項特色：

1.選取不同文本類型來查詢與比較。COCA 的文本來自 5 項內容：

（1）口說（spoken），包括 150 個電視或廣播節目對話，約 8 千 5 百萬字。

（2）小說（fiction），包括短篇故事、戲劇／電影劇本，約 8 千 1 百萬字。

（3）雜誌（popular magazines），包括新聞、健康、家庭園藝、女性話題、財經、宗教、運動等將近 100 種不同領域的雜誌，約 8 千 6 百萬字。

（4）報紙（newspapers），包括 10 家報紙不同版面（地方新聞、評論、運動、財經）的文章，約 8 千 1 百萬字。

（5）學術期刊（academic journals），包括近 100 種不同種類的期刊，約 8 千 1 百萬字。

2.隨時間推移，比較不同時間點出現的同一詞彙。

3.提供詞彙出現頻率與相關字比較的功能。

4.使用者自行訂定同一類別（服裝、食物、情緒）的字彙表列，便於日後查詢。

5.涵蓋西班牙文與葡萄牙文語料庫。

二、語言資料庫（Linguistic Data Consortium，簡稱 LDC）

LDC 由美國高等研究計畫機構（Advanced Research Projects Agency, ARPA）與美國國科會資訊智慧系統處於 1992 年建立，現在由賓州大學（University of Pennsylvania）主辦。網站營運基金來自公司、大學及網站會員使用費。網站內容包括阿拉伯文、中文、英文等新聞電報文本，Brown Corpus 全文，教育、研究及科技發展相關的語料資源，並且歡迎使用者分享資源。LDC 具有四項特色：

1.語料內容包含「中英翻譯辭彙版本 3.0」，資料來源為字典與網路。

2.「中英新聞雜誌對照文本」語料源於 1976 年至 2004 年臺灣光華雜誌的新聞報導。

3.採付費使用原則，費用介於美金 250 元至 2500 元之間。

4.提供西班牙文、德文、日文、韓文、法文、波斯文、北印度文、坦米爾文與越南文等語料資源。

【參考文獻】

中研院（2006）。**平衡語料庫構建技術手冊**。2012 年 4 月 9 日，取自 http://godel.iis.sinica.edu.tw/contest/CorpusIntroduction.htm

林武聰（2003）。**線上英語學習環境**。雲林科技大學電子與資訊工程研究所，未出版，雲林縣。

洪千惠（2009）。**英譯中：譯文西化分析—語料庫為本的翻譯研究**。輔仁大學翻譯學研究所碩士論文，未出版，新北市。

陳克健（1994）。**素材語言學與文本處理**，2012 年 4 月 9 日，取自 http://rocling.iis.sinica.edu.tw/CKIP/20corpus.htm

Corpus of Contemporary American English (2012). 2012 年 4 月 9 日，取自 http://corpus.byu.edu/coca/

Linguistic Data Consortium (2012). 2012 年 4 月 9 日，取自 http://ldc.upenn.edu/