內的大出版社紛紛建構大型機讀語料庫來編纂英文辭典。以大型機讀語料庫來編纂辭典的好處是可以客觀且方便地檢視詞的頻率，搭配語，及語意，語法，與語用的功能，來判斷詞的用法。以語料庫為主的語言學與計算語言學研究在沈寂了近二十年後才漸漸復甦， 1990 年代從事自然語言處理的研究人員將原先為語音辨認所發展的統計演算法運用到自然語言剖析，詞彙知識自動習得(automatic lexical knowledge acquisition)，機器翻譯等以語料庫為主計算語言學的研究上，獲得豐盛的成果。在加上大型機讀語料因為網際網路的盛行而垂手可得，以及個人電腦功能日益強大，售價卻十分低廉，這些因素使得 1990 年中期以後，以語料庫為主的計算語言學研究成為主流。

貳 、語料庫的資源

料庫種類除了有口語，書面語，與語音資料庫還依是否為平衡語料庫，有否加標記，單語或多語等方式來區分。加標記的語料庫包括加註文章結構標記(如標題，句子，段落等)，詞類標記，語意標記，或語法樹等數種。未加標記的英文語料庫以 Brown Corpus, LOB Lancaster-Oslo-Bergen) Corpus, BNC (British National Corpus), 與 Project Gutenburg 最著名，後者收集了許多英文小說。加詞類標記的英文語料庫包括 Penn Corpus, Sussane Corpus 等，中文加詞類標記的語料庫目前有中研院平衡語料庫(約 500 萬詞)。加註語意標記的語料庫目前尚缺乏。英文語法樹庫則有 Penn Treebank。至於多語語料庫最著名的是加拿大國會以英法文記錄的的 Hansard Corpus。目前國內可以線上取得的中英，中日平行語料庫則有光華雜誌。

目前最常用的兩個句法樹庫資料,分別是中央研究院中文句結構樹資料庫（Sinica Treebank） (http://www.aclclp.org.tw/use_stb_c.php)，以及美國賓州大學

中文句法樹庫（Penn Chinese Treebank ）(http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T05)。兩者在語言，語料來源，語料庫大小，標記集，標記單位，標記訊息，及依據的語言學理論都不相同。

Sinica Treebank 與 Penn Chinese Treebank 最大的差別在於結構樹的語法單位不同。前者以標點符號作為分隔不同結構樹的單位，因此一個結構樹很多時候只是一個詞組（如 PP, NP）而不是一個完整的句子。而後者除小部分結構樹是句子的片段（以 FRAG 標示）大部分的結構樹是完整的句子(sentence)(以 IP 標示)。另外 Sinica Treebank 語法結構採取中心語主導原則 (Head-Driven Principle)，註明中心語(Head)和其他成分（如附加語）的語法和語意訊息，表達出句子中詞和詞之間的語法結構和語意角色關係，而 Penn Chinese Treebank 並沒有中心語與語意角色的訊息，而是在詞組上加註如主詞 SBJ 受詞 OBJ 等語法功能的方式來取代。

由於 Sinica Treebank 有未簡化標記，簡化標記及精簡標記三種標記集，相較於 Penn Treebank 只有一種標記集，Sinica Treebank 的三種不同的標記集可以作為不同的特徵。除此之外只有 Sinica Treebank 有標示語意角色的訊息，Penn Chinese Treebank。由 Linguistic Data Consortium (LDC)所發行，其中標示語意角色的 Penn Chinese Treebank 稱為 Chinese Proposition Bank。

參、語料庫語言學的工具

一、關鍵詞前後文程式(concordancer)：輸入一個關鍵詞或字串，程式自動將語料庫中所有包含這個詞或字串例子找出來置中並顯示前後語境。Antconc 是一個免費軟體，可以計算語料關鍵詞的頻率，並檢索關鍵詞以及搭配語。下面的畫面擷取自 Antconc 關鍵詞上下文檢索程式 concordancer 的功能。