詞彙語意與句子閱讀難易度計量*

中央研究院語言學研究所
鄭錦全

摘 要：詞語學習需要大量的文本，讓學習者在大量閱讀中理解詞語的用法。“針對一詞廣泛閱讀”的學習模式所需要研究的問題包括如何給學習者提示和輔導，如何安排文本讓讀者由簡入難，又如何讓人不分難易隨機查閱詞語的共現語境等等。本文討論決定句子閱讀難易度計算的三個因素，一是句子的長短，二是句中詞語出現的頻率，三是詞彙語意的相關性。我們以這三個因素來排比句子的難易度，讓學習者先閱讀比較簡單的句子，逐漸進入比較難的句子，達到學習的目的。
關鍵詞：詞語學習，泛讀，閱讀難易度，可讀性，詞彙語意。

中央研究院的平衡語料庫有五百萬詞的數位文本，現在還在繼續擴展中，很快會達到一千萬詞。可是，一般社會大衆和學生很難用來當作學習資源。我們在行政院國家科學委員會的資助下，彙整這類文本語料，建立語言教學與學習的資源中心，其中一個研究項目是“針對一詞廣泛閱讀”，以簡潔明快的網路介面，幫助學生針對個別詞語廣泛閱讀語料庫文本句子，熟悉詞語的用法。

1．一詞泛讀的理念

多年來我們提出“針對一詞廣泛閱讀”的學習模式（鄭錦全 1998a, 1998b, 1998c; Cheng 2004, Cheng et. al 2004），語言學習者在這個模式中，可以利用電腦所收集的文本，針對一個詞語，閱讀該詞語出現的許多句子。閱讀了各種該詞語和其他詞語共同出現的情形，也就是語言環境，讀者就更能掌握該詞語的用法。人在成長的過程中，閱讀許多文字，在社會交際中又有許多語言互動，因此不自覺地學會詞語的用法，有了一詞泛讀的電腦輔助，更能加速詞語學習。對成人的外語學習來說，需要在兩三年內學好一個語言，並沒有十幾年的成長時期來閱讀許多文本，因此，一詞泛讀對外語學習更有促進的作用。

一詞泛讀的理念源出語料庫。我們在中文和英文的學習上都提出具體的文本資料庫和適當的輔導，現在彙整中央研究院五百萬詞的平衡語料庫以及其他語料庫，建構“語文數位教與學資源中心”。一詞泛讀網頁呈現的形式如圖一。網址是 http://elearning.ling.sinica.edu.tw。

在這個圖示中，可以看到一詞泛讀的網頁有三個區塊，第一區是由學習者輸入一個或多個詞，在這裡是‘以為’兩個字，輸入後程式對輸入字串加以判讀，如果是平衡語料庫的詞語，就顯示該詞的詞類和詞頻，讓讀者對該詞有比較深入的了解。在這一區域裡讀者如果按“閱讀”，在第三區塊就顯示該詞語出現的句子，每次三句，讓讀者可以閱讀該詞語在句中的用法，按“繼續閱讀”就可以繼續閱讀資料庫中該詞語出現的句子。

第三區塊呈現句子。對於句子呈現的先後次序，我們有兩個考量。一是如果學習者是比較初階的學生，那麼呈現的句子應該是從簡單到困難的排列；如果是比較高階的閱讀者，需要對某個特定的詞檢索其用法，那麼文本的呈現就不需要按難易度排列。

關於句子閱讀難易度的計算，我們提出三個因素來考慮，一是句子的長短，二是句中所有詞語在文本中出現頻率的高低，三是詞語語意類別的多少。

2. 句子閱讀難易度計算

如果在一詞泛讀的學習模式中，初階的學習者要了解‘以為’這個詞的用法，而電腦所出現的次序隨文本原來的編排提取，就會看到以下的句子次序：

(1) 該 館 鄰近 的 承昇 汽車 保養廠 工人 說 ， 黃 老板 的 轎車 都 在 該 店 保養 ， 業者 當時 還 以為 有 人 放 鞭炮 ， 歹徒 行動 非常 俐落 不慌不忙 ， 但 有 人 記下 機車 車號 。

圖一 一詞泛讀網頁

(2) 我 以為 喝 這 酒 得 小心 一點 呢 ！
(3) 我 以為 我們 去 唸 其它 學科 不是 從 自己 的 本位 出發 ， 而是 開放 自己 ， 去 學習 其它 學科 的 思考 方式 ， 它們 自己 看待 世界 的 觀點 ， 從而 和 我們 自己 的 本行 做 比較 ， 這 是 一 種 開放 的 學習 方式 ， 套用 電腦界 的 典故 ， 開放 架構 比 封閉 架構 更 能 促進 本身 的 發展 ， 網路 的 開放 系統 也 逐漸 能 取代 封閉 的 大型 電腦 。
(4) 妳 以為 我 傻 ？

第一句比第二句長，第三句更長，而第四句最短。如果要排出閱讀可讀性的話，我們的感覺是如下的排序：

(5) 妳 以為 我 傻 ？
(6) 我 以為 喝 這 酒 得 小心 一點 呢 ！
(7) 該 館 鄰近 的 承昇 汽車 保養廠 工人 說 ， 黃 老板 的 轎車 都 在 該 店 保養 ， 業者 當時 還 以為 有 人 放 鞭炮 ， 歹徒 行動 非常 俐落 不慌不忙 ， 但 有 人 記下 機車 車號 。
(8) 我 以為 我們 去 唸 其它 學科 不是 從 自己 的 本位 出發 ， 而是 開放 自己 ， 去 學習 其它 學科 的 思考 方式 ， 它們 自己 看待 世界 的 觀點 ， 從而 和 我們 自己 的 本行 做 比較 ， 這 是 一 種 開放 的 學習 方式 ， 套用 電腦界 的 典故 ， 開放 架構 比 封閉 架構 更 能 促進 本身 的 發展 ， 網路 的 開放 系統 也 逐漸 能 取代 封閉 的 大型 電腦

這樣的排序感覺上是合理的，但是要如何量化我們難易度的感覺，顯然這裡是以句子裡的詞語數量多寡的為標準。我們的看法是，一般說來，短的句子比長的句子容易讀，這也是在研究英文文章的可讀性所用的計算的一個因素(McLaughlin 1969, Zakaluk and Samuels 1988)。平衡語料庫的文本已經做了分詞的工作，詞語之間以全形空格鍵隔開，每個詞語含有詞類的標記，如下列的例子：

(9) 妳(Nh) 以為(VE) 我(Nh) 傻(VH) ?(QUESTIONCATEGORY)
(10) 我(Nh) 以為(VE) 喝(VC) 這(Nep) 酒(Na) 得(D) 小心(VK) 一點(Dfb) 呢(T) ! (EXCLAMATIONCATEGORY)

這樣一來詞語數目的計算就很簡單，但是下面兩個例句的詞語數目相同，兩者比較起來就完全一樣。不過，從詞頻來說，有些詞的出現頻率高，有些詞的詞頻低。詞頻高的詞通常是在一般生活中出現比較多的詞語，因而也比較容易閱讀，詞頻低的，讀者接觸比較少，因此閱讀難度較高。基於這樣的理念，我們對句子難易度的計算，考量點包括詞語的多少和詞語出現的頻率。

在計量上，我們希望可讀性的指數高的代表可閱讀性高，指數低的代表可閱讀性低。現在所用的平衡語料庫中最長的句子是 225 個詞，從 225 減掉每個句子中詞語的數目，就是可讀性的指數。這樣，詞語少的句子可讀性的指數就高。例如下面的句子一共有四個詞，可讀性就是：225-4=221，列在句子的前頭。

(11) (221) 妳(Nh) 以為(VE) 我(Nh) 傻(VH) ?(QUESTIONCATEGORY)

可是，這樣一來，所有同樣長度的句子的可讀性就完全一樣，例如下面的兩個句子：

(12) (220) 我(Nh) 的(DE) 爸爸(Na) 是(SHI) 軍人(Na) 。(PERIODCATEGORY)
(13) (220) 我(Nh) 的(DE) 爸爸(Na) 是(SHI) 廟祝(Na) 。(PERIODCATEGORY)

這樣不很合理，這兩個句子的可讀性應該有別，前者高，後者低。要做區隔，需要加入第二個因素。這兩個句子不同的地方是‘軍人’和‘廟祝’，‘軍人’多見，‘廟祝’少說，兩者的詞頻不一樣。因此，我們加入詞頻的因素來考量可讀性。這兩個句子的詞語的頻率如下：

(14)
我(Nh)	40,312
的(DE)	284,586
爸爸(Na)	1,112
是(SHI)	83,665
軍人(Na)	165
廟祝(Na)	3

同一個詞語在不同的語境中可能有不同的詞類和詞頻，我們從詞庫提取詞頻時也考慮詞類。把句子中的詞頻的總和除以詞語的數目，就可以得到一個平均詞頻，我們用平均詞頻來當作可讀性考量的第二個因素。上文說過，第一個因素是句子的詞語的數目。上面兩個句子各有 5 個詞，第一個可讀性指數是 225-5=220，第二個指數，不記小數點，分別是 81,968 ((40,312 +284,586 + 1,112 + 83,665 + 165)/5)和 81,935 ((40,312 + 284,586 + 1,112 + 83,665 + 3)/5)：

(14) (220:81968) 我(Nh) 的(DE) 爸爸(Na) 是(SHI) 軍人(Na) 。(PERIODCATEGORY)
(15) (220:81935) 我(Nh) 的(DE) 爸爸(Na) 是(SHI) 廟祝(Na) 。(PERIODCATEGORY)

平衡語料庫一共有二十幾萬個句子，根據句子的詞語數目和詞頻兩個因素算出可讀性的指數，再根據這兩個指數從高到低排序，編成由簡入繁的句子資料庫，在一詞泛讀的網頁所顯示的句子就是從可讀性高到底的次序。可是，閱讀的最終目標是理解句子的意思，可讀性應該也要考慮詞彙語意。

3. 詞彙語意可讀性研議

下面的對話中的句子都合語法，理解也沒有問題：

(16) 我 的 爸爸 研究 的 是 數學 。
你 的 爸爸 呢？
我 的 爸爸 是 書法。

那麼，現在來比對下面兩個句子的閱讀難易度：

(17) 我 的 爸爸 是 軍人 。
(18) 我 的 爸爸 是 書法。

直覺上，第一個句子比第二個容易懂，因為第一句的‘我’、‘爸爸’和‘軍人’在語意上都是人這一類。而第二句的‘書法’是繪畫文教的抽象事物，由‘是’把書法當作人聯系起來，需要多一層次的思考，如（16）所說，才能領會。句子的意思要從語法結構來了解，但是我們希望不須要經過自然語言理解的困難過程，就能探討出可以驗證的計量方法，在電腦上自動分析出語意的難易度。中央研究院參照《同義詞詞林》（梅家駒等 1984）的詞的語意分類建立了一個電子檔案，詞語的分類的格式大致如下例：

(19) 我 Aa02 /Aa02 我 我們 /Aa 泛稱 /A 人

這個例子的解釋是，‘我’的編號是 Aa02，Aa02 的詞語是‘我 我們’，其上位詞是 Aa 人的泛稱詞，Aa 的上位是 /A 人。多意詞的語意分類可能有好幾個。上面比對的兩個句子有幾個多意詞，所有詞語的語意層次如下：

(20) 我 Aa02 /Aa02 我 我們 /Aa 泛稱 /A 人
我 Aa021 /Aa02 我 我們 /Aa 泛稱 /A 人
我 Aa021_01/ Aa02 我 我們/ Aa 泛稱 /A 人
的 Bo292_06 /Bo29 弓 箭 矛 盾 劍 /Bo 機具/ B 物
的 Ed013_01 /Ed01 真實 虛假 確實 虛妄 /Ed 性質 /E 特征
的 Kd01 /Kd01 的 然 所 /Kd 輔助 /K 助語
的 Kd011 /Kd01 的 然 所 /Kd 輔助 /K 助語
的 Kd011_01 /Kd01 的 然 所 /Kd 輔助 /K 助語
爸爸 Ah041_01 /Ah04 父 母 父母 父子 /Ah 親人 眷屬 /A 人
是 Ed121_01 /Ed12 正確 準確 錯誤 /Ed 性質 /E 特征
是 Ed611_01 /Ed61 這個 那個 某個 各個 其他 何 /Ed 性質 /E 特征
是 Hg182_02 /Hg18 潤色 修改 修訂/Hg 教衛科研 /H 活動
是 Ja01 /Ja01 是 當做 比作 /Ja 聯系 /J 關聯
是 Ja011 /Ja01 是 當做 比作 /Ja 聯系 /J 關聯
是 Ja011_01 /Ja01 是 當做 比作 /Ja 聯系 /J 關聯
是 Jd011_01 /Jd01 存在 遺留 保持 /Jd 存在 /J 關聯
是 Kc062_01 /Kc06 連 凡是 /Kc 聯接 /K 助語
軍人 Ae10 /Ae10 軍官 將士 軍人 士兵 /Ae 職業 /A 人
軍人 Ae103 /Ae10 軍官 將士 軍人 士兵 /Ae 職業 /A 人
軍人 Ae103_01 /Ae10 軍官 將士 軍人 士兵 /Ae 職業 /A 人
書法 Dk312_01 /Dk31 繪畫 字畫 雕刻 塑像 /Dk 文教 /D 抽象事物
書法 Dk312_01 /Dk31 繪畫 字畫 雕刻 塑像 /Dk 文教 /D 抽象事物

我們根據以下的理念計算語意上的可讀性：(一)詞語多意，閱讀時需要排除歧義，會增加難度；（二）詞語的語意類別多的句子，其語意內容比較複雜，閱讀的複雜程度比較高。我們從語意的最低類別到各個層次的上位詞類別做了各種計算，試驗的結果，以最高層的語意上位類別來衡量，比較容易區分難易度。下面一句的詞語的最高上位詞類有 6 種，詞語數目是 5，詞類出現的次數如下：

(21) 我 的 爸爸 是 軍人 。

A	7
B	1
E	3
H	1
J	4
K	4

語意類別出現的次數可以不計算，因為次數多就是表示類別少，用類別種類的多少來決定難易度，計算的方法是：

(22) 詞語數 / 語意類數 * 1000

我們取詞語數與語意類別數的比率作為衡量的標準，兩個句子如果詞語數目相同，語意類別多的句子的分母大，計算出來的比率就比語意類別少的句子小，也就是難度比較大。比率乘上 1000 以便從小數點來擴大辨析，最後刪掉小數點，取其整數作為語意因素的可讀性的指數。上面的例句的可讀性是 5/6*1000：

(23) (833): 我 的 爸爸 是 軍人 。

跟這個句子比對的‘我的爸爸是書法’的語意類別以及計算(5/7*1000=714)的結果如下：

(24) 我 的 爸爸 是 書法 。
A 4
B 1
D 2
E 3
H 1
J 4
K 4
(714): 我 的 爸爸 是 書法 。

這樣，‘我的爸爸是軍人’的語意可讀性高於‘我的爸爸是書法’，我們就這樣計算，在語意上分辨這兩個句子的可讀性。現在一詞泛讀的網頁上所安排的句子次序，詞意是第三個考慮的因素，就是優先排列句子的長短和詞頻，因此，詞意對閱讀難易度的排序作用比較小。這裡還有思考的空間。

4. 結語

一詞泛讀的學習模式提供學生針對一個詞語閱讀大量該詞出現的句子，從閱讀中理解詞語的用法。但是，初階讀者在閱讀時需要從簡單到複雜的學習過程，因此，句子的呈現要從可讀性高的逐步推演到可讀性低的。我們從三個觀點來區分句子的閱讀難易度。一是句中詞語的多寡，句子越長，難度越高。二是句子的詞語的出現頻率，頻率越高越容易閱讀。三是句子的詞語數目與語意類別的比率，比率越低越難讀。這三個因素的排序是以句子的長短優先考慮，其次是詞頻，最後是語意類別。閱讀難易度的感知應該比我們提出的三個觀點更加複雜，例如上下文的語境、概念的難易、句子的結構等等。但是，我們提出的算法已經可以實際應用在網路教學上。學海無涯，本文提出計量的理據，也只能說是一點心思而已。

*本研究在 2004-05 年得到行政院國家科學委員會的資助，計畫是 NSC93-2524-S-001-003 “兼具教學與研究功能的全球華語文數位教與學資源中心”.

引用書目

Cheng, Chin-chuan. 2004. “Word-Focused Extensive Reading with Guidance”. *Selected Papers from the Thirteenth International Symposium on English Teaching* 24-32. Taipei: Crane Publishing Co.

Cheng, Chin-Chuan, Chu-ren Huang, Xiang-yu Chen, Yu-chun Huang, Joyce Ya-Chi Han, and Feng-ju Lo. 2004. “Extensive Reading with Guidance”. 19th Pacific Asia Conference on Language, Information and Computation--Interactive Workshop on Language e-Learning. Tokyo, December 10.

McLaughlin, G. 1969 . “SMOG grading: A new readability formula”. *Journal of Reading* 12.8: 639-646.

Zakaluk, Beverly L. and S. Jay Samuels. Eds. 1988. *Readability: It's Past, Present, & Future*. Newark, Delaware: International Reading Association.

梅家駒等編. 1984.*同義詞詞林*. 上海：上海辭書出版社.

鄭錦全. 1998a. “一詞泛讀：英文詞語用法檢索軟體”. 戴維揚編. *超倍速英語學習年代* S1-S11. 台北: 文鶴出版社.

鄭錦全. 1998b. “針對一詞廣泛閱讀：電腦輔助的詞語學習”. *華文世界* 87:30-44.

鄭錦全. 1998c. *英語用法寶典 (English Word Usage)*. 台北: 文鶴出版社