《永遠的 POLA：王士元先生七秩壽慶論文集》57-69
2005-8-003-004-000048-1

基本詞彙的預測與驗證：由分佈均勻度激發的研究構想

黃居仁　　張化瑞　　俞士汶
中央研究院　　北京大學

本文嘗試爲基本詞彙提出一個可驗證的定義。斯瓦迪士 Swadesh (1952) 提出的基本詞彙表是基本詞彙庫的原型。這個詞表以概念上的必然性爲其定義基準，但卻無法以規範方法來定義或驗證。

本文的研究奠基在對兩個概念的澄清。基本詞彙庫 (basic lexicon) 由概念定義，是能夠表達這些基本概念的最小詞彙庫。核心詞彙庫 (core lexicon) 由分佈上的可靠性定義，是在不同語境中都能被預測出現的最大詞彙庫。在這個定義下，基本詞彙庫可以視爲所有語言的核心詞彙庫的交集。

本文提出以「詞彙通用度」(lexical usuality, Zhang et al. 2004) 來解釋預測基本詞彙庫。「詞彙通用度」是以張化瑞提出的「分佈均勻度」Distributional Consistency (DC) 爲測量基礎。分佈均勻度可找出在不同類文本中可靠出現的語言單位。本文的初步研究顯示「詞彙通用度」用於驗證與預測基本詞彙，有優異的成績。

關鍵詞：分佈均勻度，詞彙通用度，基本詞彙庫，核心詞彙庫，斯瓦迪士詞彙表

1. 背景：基本詞彙與核心詞彙

基本詞彙是比較語言學與語言演化理論上一個非常重要的指標工具。因爲詞 (word) 是語言中不可分割的最小表意單位。因此語言的比較研究，不能避免的要以詞爲基本單位。這也是因爲語言的變遷與歧異 (language changes and language variations)，最直接反映在詞彙上。可是，正由於詞彙是語言演變中最基本，最迅速的環節，各語言的詞彙也可能有相當大的出入。比如說，某個特定的詞彙，可能不在研究的對象語言中出現。如何能保證詞彙爲本的比較研究，不會因找不到可比對的資料而失敗呢？

斯瓦迪士 Swadesh (1952) 提出的基本詞彙表，就是要解決上述的研究問題。

基本詞彙表中的詞彙，是他認爲在所有語言中都應該會使用的。在這個前提下，語言的對比研究可以有可靠比對標準與確實的事實基礎。這亦是詞源統計方法的基本假設 (Wang 1994)。

相對於基本詞彙，核心詞彙 (core lexicon) 的概念比較晚起。核心詞彙的概念，主要用在語言處理與語言學習的研究上。他的假設，是說每個語言中都有些詞彙是比較重要的，是使用與理解語言不可或缺的。因此，在（母語）語言習得 (acquisition) 上，核心詞彙是應該較早學會的。在語言教學理論上，核心詞彙也是應該及早學習的。在自然語言處理上，核心詞彙則是不受主題，應用等的影響；即使用與意義都最穩定的詞彙，在應用上也最重要。

不論基本詞彙或核心詞彙，雖然都是各相關領域經常使用的資料，也都各有概念上爲各領域所共同接受的定義；卻都還沒有一個嚴謹的規範定義。也就是說，沒有一個可驗證的 (falsifiable) 科學定義。因爲這兩項概念都是根據語料產生的，我們將探討利用語料庫語言學研究成果，提出基本詞彙庫定義的研究方向。

2. 過去的研究方法：經驗法則的定義

基本詞彙與核心詞彙過去的定義法，多是憑藉不同的經驗法則。由經驗的規律整合演繹，得到的結果有相當的可靠性，但由經驗推出的規律，並無法直接驗證。

2.1 由專家知識出發的系統化經驗法則

斯瓦迪士 Swadesh (1952, 1955) 提出的基本詞彙表：利用語言學家的經驗，與基本概念結構的論證，提出的詞彙表，已有比較語言學研究多年的資料支持。最近的研究包括了苗瑤語族及藏緬語族關係的計量研究（王士元與鄧曉華 2003a, b）。其結果都廣爲學界接受。也就是說，雖然斯瓦迪士的基本詞表，建立在主觀的概念立論上，沒有科學上可驗證的客觀定義。但他的使用上已成爲學界的預設 (default) 標準，而且其內容已應用於上千種世界不同的語言，也就是說經由經驗法則通過了跨語言可行性的基本檢驗。[1]

[1] 斯瓦迪士的基本詞彙已成功適用於世界上的多種語言。以「羅賽塔計畫」(Rosetta Project, http://www.rosettaproject.org/) 爲例，其中收了超過三千個語言的斯瓦迪士基本詞彙表。可是，若深入觀察，又發現很大部分語言中的詞彙並沒有全填入。這也驗證了王士元先生 (1994) 對斯瓦迪士基本

2.2 由分佈差異出發的對比經驗法則

相對於斯瓦迪士的理性辯證，語料庫語言學興起後，統計式的理論模式蔚爲風氣。但對於基本詞彙庫的定義，並無直接幫助。從語料庫得到最直接的統計就是詞頻，因此經常用來做核心詞彙定義的指標。研究者卻很快發覺，影響詞頻的因素太多。比如說，文本的主題一定會提高文本內相關主題詞的頻率。談生命科學，生命科學詞彙的詞頻就應該比其他詞高。這對於這些主題詞彙，是否是語言中的基本或核心詞彙，並未提供可驗證的證據。

黃居仁等 (Huang et al. 1998) 利用了詞彙庫資源的特性，提出了另一種定義核心詞彙庫的經驗法則。他們的方法的創新之處，在於並不直接用個人的經驗推斷，而是利用由詞彙庫資源做客觀的運算。他們的理論基礎，在於一個很有意思的觀察。也就是說，任兩個不同的詞彙庫，一定有相當大的差異性。換句話說，每個詞彙庫中都有一些別的詞彙庫中不收的特色詞彙。更有意思的是，這個現象並不受詞彙庫大小的影響。他們研究的詞彙庫中最大的超過十五萬六千詞，最小的只有不到四萬詞。但即使是這個最小的詞彙庫，其中還有將近百分之十二的詞（事實上是 11.80%，相當於約 4,400 個詞），是最大的詞彙庫中沒有的。因爲詞彙庫的大小會影響互相的覆蓋率。比如說，上述的四萬詞詞彙庫，即使完全吻合，也只能覆蓋十六萬詞詞彙庫的 25%。因此 Huang et al. (1998) 提出了相互覆蓋率（mutual coverage，簡稱互蓋率）的概念，是甲詞彙庫對乙詞彙庫的覆蓋率，與乙詞彙庫對甲詞彙庫的覆蓋率的平均值。他們所計算五個不同詞彙庫，共十組不同的互蓋率，從 49.01% 到 67.77%。換句話說，詞彙庫之間的相似度，大約是一半到三分之二。每個詞彙庫收詞的原則不盡相同。如果把這個原則當個別詞彙學家提出，該語言中何謂必要詞彙的定義；則這些詞彙庫的交集，便是不同必要詞彙定義下，所共同認爲是基本的詞彙。也就是說，可以用詞彙庫的資料比對，得到專家經驗法則的交集，因而定義出核心詞彙庫來。

這個方法，雖然是不受限於語言或時代，可重複使用的方法。但基本上還是經驗法則的延伸。更重要的，這個方法還是要依靠專家事先準備好的詞彙庫資料，無法直接由語料產生。也就是說，他是建在詞彙庫資料正確的前提下，無法由一手資料中歸納或驗證的。

這個觀察到的事實，也會導致對「基本詞彙」這個概念不同的想法。如果專家在描述同一個語言的詞彙時，都會產生哪些詞該收錄的大規模差異。如何能找

詞彙表的批評。就是各語言不見得會有完全相對應的詞彙語意範圍；極可能有找不到對應詞彙，或對應詞彙的意義有相當差距。

出一個一致的原則，定義所有不同語言中都需要收錄的「基本詞彙」呢？

2.3 區分「基本詞彙」與「核心詞彙」的暫行定義

由上面的討論，我們看到「基本詞彙」與「核心詞彙」雖然在概念上有相通之處，但其出發點與應用上的分歧甚大。因此，我們先嘗試在定義上試行予以區分。希望這樣的區分，有助於更進一步的研究：

基本詞彙 (basic lexicon)：（概念上的）最小充分詞彙
——涵蓋表達所有必要概念的詞彙。也就是說，它是由一個事先定義好的概念集出發的。這個概念集的必然性，需要另行驗證。從另一個觀點看，基本詞彙接近於所有可能詞彙庫的交集。

核心詞彙 (core lexicon)：（分佈上的）最大必要詞彙
——在所有不同語言環境中，預期都會使用的詞彙。也就是說，它是由語言使用來決定的。是觀察到在不同語境與使用中，可靠出現的詞彙。

在以上的前提上，我們在下文嘗試提出一個定義基本詞彙，以語料庫爲本的統計式理論模式。

3. 分佈均勻度的定義以及其在詞彙統計學上的應用

分佈均勻度 (Distributional Consistency) 用於詞彙統計學，是由北京大學計算語言學研究所的研究群在 2003 年中發展提出 (Yu 2003)。其統計公式與工具的發展，則主要是由張化瑞提出 (Zhang et al. 2004)。

3.1 分佈均勻度的定義與解釋

分佈均勻度（簡稱「勻度」，英文簡稱 DC），是和頻率 (frequency) 相對的概念。直覺來說，頻率求的是在某個範圍內量最大的元素，而勻度求的是在某個範圍內最穩定的元素。這個直覺說法，可以幫助對分佈均勻度意義的瞭解。

分佈均勻度（Distributional Consistency，簡稱 DC）的定義公式如下 (Zhang et al. 2004)：

- **分佈均勻度 DC= *SMR / Mean***
- SMR 及 Mean 分別定義如下

$$SMR = \left(\sum_{i=1}^{n} \sqrt{F_i} \Big/ n \right)^2$$

$$Mean = \sum_{i=1}^{n} F_i \Big/ n$$

上式中，n 表語料庫區分成 n 等分（比如，詞數相等）
F_i 是詞語 *w* 在第 i 等分出現的次數

由統計學的觀點，分佈均勻度與習用的標準差 (standard deviation) 有相當的關係。標準差的定義使用均方根 (RMS: Root Mean Square)，得出的值是分佈差異的預測範圍。更精確的說，是先在一定涵蓋的前提下，預測這些資料點落在與平均數 (mean) 的距離不會超過標準差的範圍內。也就是在單一語料庫中，看語料分佈的集中程度。勻度用的則是均根方 (SMR: Square Mean Root)，得出的值是語料庫等分切分後，語料均勻分散的程度。其取值範圍，可以證明爲 (0,1)（未出現詞不在考慮範圍之內）。分佈越均勻，則越接近於 1。

分佈均勻度可能的值說明如下：在整個語料庫被分成 n 等分的前提下，

(1) 如果一個詞只在其中某一個等分中出現，則其通用度爲 1/n。
(2) 如果一個詞在每一等分中都以相同頻次（非零）出現，則其通用度爲 1。
(3) 如果一個詞在 m 個等分 (1<m<=n) 中以不同頻次（非零）出現，則其通用度小於 m/n。

由這個定義可以看出，分佈均勻度是資料在不同類別取樣中出現分佈的量化。從語料庫來說，它是個別詞彙在不同語料中分佈的量化。我們可以把這個公式的意義解釋如下：

當一份資料，可以因其取樣 (sampling) 的來源或性質不同而分成 n 等分時，我們可以用分佈均勻度來考察，不同對象是否因爲取樣不同而有不同的分佈。在所有不同種類取樣中出現的頻率一致性越高的對象，它的分佈均勻度越高。反

之，若它的出現集中在某些特定取樣中，或在不同取樣中，出現的頻率起伏很大，則分佈均勻度低。換句話說，頻率高的成分，是在整個資料庫中比例最高，最容易觀察到的。而分佈均勻度越高的成分，則是資料庫中各部分出現狀況最穩定，最容易預測的。因此，當我們把資料庫當成一個不可分割的單元時，頻率是其內部成分最直接有效的測量。但如果整個資料庫是有結構的，具有有意義的內部切分時，則分佈均勻度可以提供單純頻率所不能提供的訊息。任何語料庫，一定會有內部的文本單位與結構，因此分佈均勻度在語料與詞彙統計學上，可以有很好的應用潛力。

我們把單一詞彙的分佈均勻度稱之爲該詞彙通用度，並簡稱爲通用度（暫譯爲 Lexical Usuality[2]）。而用於通用度測量所根據的資料切分原則，則稱之爲「基於 X 的通用度」，簡稱爲「X 通用度」。以張化瑞等在北京大學的初步研究爲例，他們以一年份的人民日報爲基礎，依十二個月份分成十二等分。測量出在十二個月份中，分佈均勻的詞彙。我們就把這樣的通用度稱爲「基於月份的通用度」，簡稱爲「月份通用度」。由此看來，分佈均勻度的解釋關鍵，在於其資料如何切分。我們也可以按每星期的七天，或按幾個主題，來測量分佈均勻度。每有不同的等分切分，詞彙的分佈均勻度可能會不同，而有不同的解釋。也就是說，只要有適當的語料，我們將來也可能測量「主題通用度」,「年代通用度」,「方言通用度」等。

4. 由分佈均勻度出發的研究構想

4.1 用通用度預測與驗證基本詞彙的構想

這個研究構想的提出，有兩項假設：

一、在個別語言中，基本詞彙的通用度較高。

因爲語言演進與變異，各個不同的語言，會有各自不同的核心詞彙庫，因此不可能以任何單一語言爲基礎，而把基本詞彙定義爲通用度最高的 N 個詞

[2] Usuality 的概念由 Zadeh (1985) 在模糊邏輯 (Fuzzy logic) 的理論中提出。他所要描述的是某一個模式 (pattern) 出現時的固定傾向 (disposition)。比如說，夏季雷陣雨通常在午後下，這是事情的通常狀態。但如何在規範理論中表達與測量呢？這就是 Zadeh 提出的 Usuality 理論。這個理論描述的對象與數學模式，都和我們現在討論的不同，但是在描述分佈出現傾向的觀點上是一致的。我們採用同樣的名詞，以表達同樣的精神。但爲了區分我們描述的是詞彙的出現傾向，以及數學模式的不同，我們的定義正式的名稱是「詞彙通用度」(Lexical Usuality)，有別於 Zadeh 的 Usuality。

彙。但是，如果承襲我們早先把基本詞彙定義爲最小充分詞彙的想法；就是當對這個語言的語料內容作所有的可能分類切分，並對每種切分取通用度最高的詞彙之後，基本詞彙應該落在所有高通用度的交集裡。這樣的假設，有一個可被驗證的前提，就是說在只用單一種分佈均勻度測量時，也應該可以在高通用度的詞彙群中，找到個別的基本詞彙。這是我們目前已可以檢測的一個假設。

二、在跨語言的比對中，基本詞彙的通用度最穩定。而且相對於其他詞彙，有較高的值。

我們因此可以大膽假設，當有足夠的語言有分佈均勻度的數據時，我們可以把相對當的詞彙的通用度建成一個 n- 維資料庫。當我們再計算這 n- 維的分佈均勻度時，這個第二階段得到的通用度最高的詞，就應該是所謂的「基本詞彙」。這是一個仍待確認的假設。

在上述的前提下，我們不但可以在理論上驗證斯瓦迪士的基本詞彙表，我們也可以因著語言分類，訂出相關的語言群（如漢語，如南島語系等等），而計算出這些以定義語言群的基本詞彙庫。

4.2 初步檢測

我們利用北京大學人民日報語料庫的月份通用度資料與斯瓦迪士的基本詞表，對上述的假設做一驗證。爲了取材的廣度，也因爲即使最小的基本詞彙集也尚無嚴謹的定義支持，我們採用的是 207 個詞彙的原始詞表。正如王士元與鄧曉華 (2003a) 的觀察，漢語研究中尚未建立標準的斯瓦迪士詞彙表；因此我們是參考了羅賽塔計畫 (www.rosettaproject.org) 與維基辭典 (wiktionary.org) 中的不完全的部分詞彙表後，補齊了斯瓦迪士基本詞彙的中文版。在附表中，除了斯瓦迪士的原編號，英文詞彙，與中文詞彙外，最後一欄是根據月份通用度排序時，這個中文詞的編號。我們只取了月份通用度最高的 2,966 個詞。

4.3 預測與驗證

如 Wang (1994) 所評，斯瓦迪士基本詞彙是由概念出發，也的確包含了個別語言中的重要詞彙。但概念的必然，並不一定完全等同於詞彙的約定俗成。更何況斯瓦迪士詞彙表中，有些概念是否必然，仍待商榷。因此斯瓦迪士詞彙表並不

能保證所收入的每個特定概念，在所有語言中一定有詞彙表達。因此語言學者在使用斯瓦迪士詞彙表時的基本態度，是預期可在任一語言中找到大部分的詞彙，但不會有必須找到所有詞彙的先入爲主概念。此外，斯瓦迪士詞表是根據語意概念訂的，並未直接預測這個概念在語言中的使用分佈。因此在語料爲本的實際計量研究中，可能有頻率或通用度上的分佈不一致。作爲比較的基準，我們的假設是通用度對基本詞彙的預測優於頻率的預測。

在月份通用度最高的 2,966 個詞彙，共出現了 120 個斯瓦迪士基本詞彙，覆蓋率是 57.97%。這 120 個詞的通用度值的範圍由 0.99985（在）到 0.89165（女）。也就是說，這些詞屬於月份通用度多峰分佈中最高的一峰。相對而言，這 120 個詞的頻率分佈則甚爲離散，從最高的 147,835 次（在）到最低的 454 次（水果），超過 325 倍。事實上，以頻率分，月份通用度最高的這 2,966 個詞中，頻率高於 1,000 與低於 1,000 的詞數差不多相等。而斯瓦迪士詞彙在這兩組詞彙中的分佈，也是幾乎相等的。換句話說，正如學界一向懷疑的，基本詞彙表與頻率間，並無絕對相關性。反倒是一如我們早先的懷疑，基本詞彙與高通用度間，似乎有一定的關係。

月份通用度與基本詞彙間有什麼關係呢？我們在解釋分佈均勻度時提到，分佈均勻度的解釋，取決於對資料取樣的分割。月份通用度的資料，是建立在把語料分成出版月份的 12 個等份上。因此，月份通用度高的詞，就是不受取樣切分的影響的詞彙。因此，我們看到斯瓦迪士詞表中，指稱詞，數詞，上下文（脈絡）連接詞等幾組不受時序影響的詞彙，可以很可靠的由月份通用度所預測。而明顯會受季節時序影響的，如天氣現象則有較多的例外。另外，值得注意的是：斯瓦迪士詞表是按概念排序的。而我們也可很明顯看出，通用度與基本詞彙間的關係，往往是整組的。也就是說，某一組概念集相關詞的通用度高低，似乎有其內部的一致性。這也是我們認爲通用度可以預測基本詞彙的佐證之一。也就是說，通用度測量時一定要對取樣分組。而這些分組基準的改變，一定會抑制某些概念的通用度，也會相對凸顯某些組的概念。正好回到我們原先的假設，以多重基準作高通用度詞彙的統計，而這些多重基準的交叉，很可能可以面向基本詞彙的真正定義。

4.4 理論意義的探討與反思

分佈均勻度這個統計特性用於詞彙通用度的初步研究，得到了兩個似乎並不相容的解釋性描述。我們剛在上節中觀察到詞彙通用度具有概念類聚性。也就是

說，通用度與概念有相關性。但是，另一方面，我們也預測並觀察到通用度則是功能詞優先的。這原來是通用度非常重要的動機，就是它不同於詞頻，可以與主題詞脫鉤。因而也預測到通用度高的應該包括語言結構上必然的功能詞。

概念類聚以及主題脫鉤，這是否是兩個相矛盾的特性呢？仔細考察之後，實則不然。主題是受限於文本範圍的概念。也就是說，如果概念的重要性，只限於少數幾個文本；我們就會把這些概念，稱之爲主題。如果概念的重要性，不受文本的限制，那就可能是基本概念了。另一方面，功能詞的顯著性，是否與概念類聚有衝突呢？特別是當我們把功能詞稱之爲虛詞，有別於表達內容的實詞。其實，若光由概念出發，並無實虛之分。功能詞表達的事件分類，事物遠近，因果關係等，都是知識表達中不能避免的概念。因此，我們觀察到的，通用度概念類聚及主題脫鉤這兩個看似矛盾的特性，反而是它與基本詞彙有高相關性的重要指標。

事實上，基本詞彙與斯瓦迪士詞表的設計，就是要找出在不同語言中，一定會使用的一組詞彙。而目前對詞彙通用度的研究，找出的是，在同一語言，不同環境中，都會可靠使用的一組詞彙。從這個觀點看來，基本詞彙的定義是跨語言的，而核心詞彙是針對單一語言的。單一語言的核心詞彙與跨語言的基本詞彙間，是否有一定程度的相關性呢？

目前可以確定的是，通用度有概念類聚效應，而且在表現一個語言內部的差異，與頻度結合在一起，還可能透射出許多社會、思維與人類的深層潛意識。張化瑞進行中的研究，已用通用度分析過幾種語言（英漢日俄）中星期幾的分佈情況。不但證明了通用度的跨語言適用性，以目前僅僅根據人民日報語料得到的通用度作爲現代漢語詞語通用度的一個初始近似值，也是可行的，甚至不管這個初始值是怎樣的，最後隨著語料的豐富都會逐漸收斂於當下的實在通用度（當然它是會隨著時期的不同而變化的）。至於跨語言的研究，則得到了，雖然不同語言中因通用度而得到的詞表有差異性，其本質在很大程度上與語言無關。

5. 結語

本文提出以「分佈均勻度」這個統計工具來測量詞彙通用度，並探討通用度與基本詞彙間的可能關係。這個研究構想雖然還沒有嚴謹的證明，或大規模的實際結果。我們以現有資料作的初步考察，顯示通用度與學界習用的斯瓦迪士基本詞彙表間，的確有些相當有意思的對應。而這些對應是頻率所無法捕捉的。我們希望這個粗略的構想，能夠在未來的研究中得到驗證與修正。

引用文獻

Huang, Chu-Ren, Zhao-ming Gao, Claude C. C. Shen, and Keh-jiann Chen. 1998. Quantitative Criteria for Computational Chinese Lexicography: a study based on a Standard Reference Lexicon for Chinese NLP. *Proceedings of ROCLING XI*, 87-108.

Swadesh, Morris. 1952. Lexicostatistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96, 152-63.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistical dating. *International Journal of American Linguistics* 21:121-137.

Wang, William S-Y. 1994. Glottochronology, lexicostatistics, and other numerical methods. *Encyclopedia of Language and Linguistics*, 1445-1450. Oxford and New York: Pergamon Press. [中譯本：陳保亞、周政后譯 (2002)〈詞源統計分析法，詞彙統計學和其他數理統計分析法〉,《王士元語言學論文集》，262-279。北京：商務印書館。]

Yu, Shiwen. 2003. Mining of language-data-based knowledge and integration of language resources database. Paper presented at the third Japan-China Natural Language Processing Joint Research Promotion Conference. Kyoto, Japan.

Zadeh, L. A. 1985. Syllogistic Reasoning in Fuzzy Logic and its Application to Usuality and Reasoning with Dispositions. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-15, 6, 754-763.

Zhang, Huarui, Chu-Ren Huang, and Shiwen Yu. 2004. Distributional consistency: a general method for defining a core lexicon. To be presented at the 4th International Conference on Language Resources and Evaluation (LREC2004). Lisbon, Portugal. May 26-28, 2004.

王士元, 鄧曉華. 2003a.〈苗瑤語族語言的親緣關係的計量研究——詞源統計分析方法〉,《中國語文》2003.3:253-263。

王士元, 鄧曉華. 2003b.〈藏緬語族語言的數理分類及其形成過程的分析〉,《民族語文》2003.4:8-18。

附表：斯瓦迪士基本詞彙表與中文月份通用度

字號	英文詞	中文詞	通用度排序/通用度值
1	I	我	758
2	thou	你	635, .99500
3	he	他	514
4	we	我們	139
5	you	你們	1502
6	they	他們	265
7	this	這	60
8	that	那	455
9	here	這裏	389
10	there	那裏	92
11	who	誰	264
12	what	什麼	160
13	where	那裏	2659
14	when	什麼時候	160+139
15	how	怎麼	998
16	not	不	18
17	all	都	15
18	many	許多	49
19	some	幾	95
20	few	一些	87
21	other	其他	330
22	one	一	25
23	two	二，兩	147
24	three	三	88
25	four	四	47
26	five	五	2050, .97900
27	big	大	12
28	long	長	1490
29	wide	寬	1399
30	thick	厚	——
31	heavy	重	868, .99349
32	small	小	460
33	short	短	1287
34	narrow	窄	——
35	thin	薄	——
36	woman	女	2922, .89165
37	man	男	2320
38	person	人	16
39	child	子女	1470
40	wife	妻子	2420
41	husband	丈夫	2135
42	mother	母親	2542
43	father	父親	2124
44	animal	動物	2469
45	fish	魚	1820
46	bird	鳥	——
47	dog	狗	——
48	louse	虱	——
49	snake	蛇	——
50	worm	蟲	——
51	tree	樹	2524
52	forest	森林	2492
53	stick	棍棒	——
54	fruit	水果	2695
55	seed	種子	2718
56	leaf	葉	——
57	root	根	——
58	bark	樹皮	——
59	flower	花	764
60	grass	草	2446
61	rope	繩	——
62	skin	皮	——
63	meat	肉	——
64	blood	血	——
65	bone	骨	——
66	fat (n.)	油	1889
67	egg	蛋	——

68	horn	角	——
69	tail	尾	——
70	feather	羽毛	——
71	hair	髮	——
72	head	頭	1422
73	ear	耳	——
74	eye	眼	1220
75	nose	鼻	——
76	mouth	口	2797
77	tooth	牙	——
78	tongue	舌	——
79	fingernail	指甲	——
80	foot	腳	1803
81	leg	腿	——
82	knee	膝	——
83	hand	手	1783
84	wing	翅	——
85	belly	肚	——
86	guts	腸	——
87	neck	頸/脖子	——
88	back	背	——
89	breast	胸	——
90	heart	心	1872
91	liver	肝	——
92	drink	喝	2199
93	eat	吃	1553
94	bite	咬	——
95	suck	吸	——
96	spit	吐	——
97	vomit	吐	——
98	blow	吹	——
99	breathe	呼吸	——
100	laugh	笑	777
101	see	看	122
102	hear	聽	727
103	know	知道	686

104	think	想	855
105	smell	聞	——
106	fear	怕	1150
107	sleep	睡	——
108	live	生存	1702
109	die	死亡	1517
110	kill	殺	——
111	fight	打	152
112	hunt	獵	——
113	hit	打	152
114	cut	切	——
115	split	分	809
116	stab	刺	——
117	scratch	抓	673
118	dig	挖	2009
119	swim	游	——
120	fly (v.)	飛	895
121	walk	走	40
122	come	來	31
123	lie	躺	——
124	sit	坐	937
125	stand	立	2078
126	turn	轉	1385
127	fall	落	766
128	give	給	2657
129	hold	握	——
130	squeeze	壓	2330
131	rub	擦	——
132	wash	洗	——
133	wipe	擦	——
134	pull	拉	1636
135	push	推	1230
136	throw	丟	——
137	tie	結	879
138	sew	縫	——
139	count	算	763

140	say	說	230
141	sing	唱	2214
142	play	玩	——
143	float	浮，飄	——
144	flow	流	1930
145	freeze	凍	——
146	swell	腫	——
147	sun	日	——
148	moon	月	299
149	star	星	——
150	water	水	2250
151	rain	雨	2352
152	river	河	2653
153	lake	湖	——
154	sea	海	1308
155	salt	鹽	——
156	stone	石	——
157	sand	沙	——
158	dust	塵	——
159	earth	地	817
160	cloud	雲	——
161	fog	霧	——
162	sky	天	1588
163	wind	風	2006
164	snow	雪	2915
165	ice	冰	——
166	smoke	煙	——
167	fire	火	——
168	ashes	灰	——
169	burn	燒	——
170	road	路	579
171	mountain	山	2234
172	red	紅	1441
173	green	綠	2161
174	yellow	黃	——

175	white	白	1562
176	black	黑	1810
177	night	晚上	838
178	day	日	1644
179	year	年	2634
180	warm	熱	2385
181	cold	冷	——
182	full	滿	2387
183	new	新	223
184	old	老	575
185	good	好	275
186	bad	壞	——
187	rotten	爛	——
188	dirty	髒	——
189	straight	直	1532
190	round	圓	——
191	sharp	銳	——
192	dull	鈍	——
193	smooth	平滑	——
194	wet	濕	——
195	dry	乾	——
196	correct	對	2319
197	near	近	1643
198	far	遠	863
199	right	右	——
200	left	左	——
201	at	在	3
202	in	內	98
203	with	與	54
204	and	和	13
205	if	如果	767
206	because	因爲	1180
207	name	名字	960