Computational Linguistics and Chinese Language Processing
Vol. 18, No. 2, June 2013, pp. 35-56

基於字典釋義關聯方法的同義詞概念擷取：以《同義詞詞林（擴展版）》爲例[1]

A Definition-based Shared-concept Extraction within Groups of Chinese Synonyms: A Study Utilizing the Extended Chinese Synonym Forest

趙逢毅*、鍾曉芳+

F. Y. August Chao and Siaw-Fong Chung

摘要

同義詞在資訊擷取與語義分類上是很重要的語料資訊，但將兩詞歸納爲同義其原由則值得令人探討。從語義(sense)的觀點來說，多義詞組歸到特定同義組合中，其語義中應有與該類字詞同義集合。此類型的代表爲《同義詞詞林》(梅家駒、竺一鳴、高蘊琦與殷鴻翔，1983)，將漢語同義字詞區分成具結構類別。而從計算語言學方法來說，同義詞關聯需要參考語料庫中詞組的出現頻率，輔以機器學習方法來計算同義詞相似度。然而前者專家分類原則是透過語感進行，若沒有對同義詞的類別原則加以定義，則後人便會產生對同義詞的混淆。後者機器學習方法使用統計方法來辨別相似詞彙，則會缺乏語義的辨別。爲了瞭解同義詞組的概念內涵，本研究提出基於辭典釋義文字的關聯計算原則，試透過計算共同擁有的釋義文字出現比率，以解析兩詞彙間所包涵之釋義概念。並且以《同義詞詞林（擴展版）》爲例，從釋義義涵的角度列舉出適合詮釋該詞組

[1]本研究第二作者蒙國科會計畫專案補助編號 101-2410-H-004-176-MY2，初稿亦獲匿名審查委員指正，謹此誌謝。

*國立政治大學資訊管理學系
Department of Management Information Systems, National Chengchi University
E-mail: fychao.tw@gmail.com
+國立政治大學英國語文學系
Department of English, National Chengchi University
E-mail: sfchung@nccu.edu.tw

的詞彙，突顯該類別所包涵的語義。最後，比較 SketchEngine (Kilgarriff *et al.*, 2004)中所取得的同義詞(similar words)之間的差異。本研究計算結果雖然會受辭典釋義內容影響，但辭典釋義內容相較於人工分類原則與統計語料庫所得的數值資料，較能從詞義上詮釋詞彙之間的共有概念。我們希望能透過釋義關聯方法更瞭解詞彙間的交集概念，亦希望能在同義詞的語義計算上，提供辭典釋義與詞條編寫上的思考。

關鍵字:同義概念, 同義詞詞林, 釋義, 辭典

Abstract

Synonym groups can serve as resourceful linguistic metadata for information extraction and word sense disambiguation. Nevertheless, the reasons two words can be categorized into a particular synonym group need further study, especially when no explanation is available as to why any two words are synonymous. Lexical resources, such as the Chinese Synonym Forest (or Tongyici Cilin) (Mei *et al*. 1983), assemble lexical items into hierarchical categories via manual categorization. Other than this, statistical measures, such as co-existing probability, have been adopted widely to verify synonymous relationships. Nevertheless, a purely statistical method does not provide description that can help interpret why such a synonymous relationship occurs. We propose a novel method for the study of shared concepts within any synonym group by comparing co-existing words in the dictionary definition of each member in the group. The co-existing words are seen as the representatives of shared concepts that can be used for interpretating any hidden meaning among members of a synonym group. We also compare our results with the thesaurus function in the Sketch Engine (Kilgarriff *et al.* 2004), which uses statistical data in the form of Sketch scores. The results show that our method can produce concept words according to dictionary definitions, but this method also has its limitations, as it works only with a finite number of synonyms and under limited computing resources.

Keywords: Shared Concept, Synonym, Chinese Synonym Forest, Dictionary Definition

1. 簡介

同義詞語料在自然語言處理與資訊擷取技術領域上是很重要的參考資源。透過同義詞語料不僅能讓機器學習方法更瞭解使用者所闡述的文字內容概念，亦可以對提問文字舉一反三。在詞彙同義的歸納上，大概可區分爲兩種進行方向：詞義訓詁與機器學習。第一，“義訓者，觀念相同，界說相同，特不說兩字之製造及其發音”(黃侃述與黃悼，1983)。

「同義相訓」是在詞義訓詁上的主要工作。這項工作往往需要花費許多人力與時間，才能清楚地分析出詞彙之間的同義內涵。此類型的成果眾多，如梅家駒等(1983)所編撰的中文《同義詞詞林》(以下稱《詞林》)，收錄來自詞素、詞組、成語、方言詞與古語等詞，共五萬三千多詞彙，並且依照同義詞分類義涵有系統地分成不同的類別。其後經由哈爾濱工業大學信息檢索研究室(HIT IR Lab)刪除舊詞與罕用詞，並依新聞語料加入常用新詞，使擴展版詞彙量增加到七萬多。《詞林》中，詞語分類原則是「相對、比較」(梅家駒等，1983)，詞語所屬類別與列舉位置哲學有作者們不可言喻巧思。第二，機器學習方法進行同義詞辨析近年來不斷發展，但著重在使用同義字詞進行上下文之中詞義的消歧。在進行同義消歧的處理過程中，又可區分爲監督式與非監督式兩種學習方法，來進行辨別是否可歸爲同義字群。上述的兩類機器學習方法都需要參考語料庫的詞類頻率計算後，才能得到字詞之間的相似程度。在進行此類方法時，常見的問題是缺乏大型語料庫與同義詞資料稀疏的通則(劉挺與車萬翔，網頁資料擷取於 2012)。

本研究使用辭典釋義內容，首先對詞彙之間的共有概念計算原則進行討論，再對《同義詞詞林（擴展版）》(下稱《擴展版》)之分類進行釋義關聯計算。本研究試以共同使用的釋義用詞，擷取能表達該分類的共有概念詞組。除了計算《擴展版》中的同義類別義涵，並透過釋義涵蓋數與最大平均釋義關聯詞值比較同義類別中的詞彙，標記較適合用以表達該類別的詞彙。最後，我們將所使用的釋義關聯比較原則與 Sketch Engine(Kilgarriff *et al.*, 2004)的語料庫統計方法進行比較，對比兩種方法在詞彙共同義涵計算上異同。Chinese Sketch Engine 語料統計方法基於大量中文語料進行語法統計與共同出現詞頻計算取出同近義詞(Huang *et al.* 2004)。雖然此語料庫平台可以呈現語言行爲相似的詞彙，但近義詞語義關係無法透過此類統計方法的結果明確說明；相反的，釋義關聯能計算兩多義詞彙之間較合適的同義釋義，需依賴辭典釋義進行關聯計算。另外，由於《教育部重編國語辭典修訂本》(下稱《國語辭典》)的釋義屬非專業領域辭典，因此仍有涵蓋面欠缺或不全，此乃是本方法的限制。

2. 釋義關聯

詞彙是「語言中表達意義的最小獨立單位」(黃居仁，2005)，辭典的釋義文字則是組合查詢者瞭解的詞彙，以表達詞條的概念涵義。釋義說明所用的單一詞語包括被解釋字(詞)的部份或完全的涵義，而且在釋義的詞語都屬於知識層級上較爲通俗的語句或概念。例如辭典之中，在解釋“*鱗波*”詞條時，會使用“*魚鱗*”、“*波紋*”等詞彙來說明詞彙的義涵。基於釋義字彙的共同出現頻率原則，提出基於釋義詞彙共同出現的涵蓋比例 (Percentage of co-appearance of relations, CoAP) 的語義關聯程度 (Semantic Relation Degree, SRD) 的比較方法 (趙逢毅與鍾曉芳，2011)，比較詞彙間在釋義文字的涵蓋關聯。計算方式如下：

$$CoAP(x) = \frac{X \bigcap Y}{X} \tag{1}$$

$$SRD_{xy} = 2\frac{CoAP(y)CoAP(x)}{CoAP(y)+CoAP(x)},\ 0 \le SRD_{xy} \le 1 \quad (2)$$

其中 *x, y* 是兩個待計算的詞條字/詞彙， *X* 與 *Y* 分別是 *x, y* 兩字的釋義字彙組合。*CoAP* 是釋義詞彙共同出現的涵蓋比例，即是計算共同出現的釋義字彙在原有字彙中所佔之比例。語義關聯程度 (SRD) 即是將兩詞語 *CoAP*(*x*), *CoAP*(*y*)相互之間概念比例的方向性消除，所以使用平均數以表示兩詞彙之間共同釋義詞彙語義關聯程度。在進行釋義關聯計算中，會使用釋義詞彙之中僅以包涵概念義涵較多的動詞與名詞詞類進行計算。計算過程在此以「*漣漪*」與「*鱗波*」兩詞爲例：

在《教育部重編國語辭典修訂本》釋義中

「*漣漪*」爲 *"水面上細微的波紋。"*

「*鱗波*」爲 *"水面似魚鱗狀的波紋。"*

計算釋義詞彙經中研院斷詞系統[2]取得釋義文字的詞性之後，僅使用動詞與名詞進行計算。計算過程如下：

「*漣漪*」爲 *"**水面(Nc)** 上(Ncd)細微(VH)**波紋(Na)**"*

「*鱗波*」爲 *"**水面(Nc)** 似(VG) 魚鱗狀(Na)**波紋(Na)**"*

因此，

$$CoAP(\text{漣漪}) = \frac{(\text{水面 波紋})}{(\text{水面 上 細微 波紋})} = \frac{2}{4}$$ ，同理

$$CoAP(\text{鱗波}) = \frac{(\text{水面 波紋})}{(\text{水面 似 魚鱗狀 波紋})} = \frac{2}{4}$$

$$SRD_{\text{漣漪,鱗波}} = 2\frac{CoAP(\text{漣漪})CoAP(\text{鱗波})}{CoAP(\text{漣漪})+CoAP(\text{鱗波})} = 2\frac{\frac{2}{4}*\frac{2}{4}}{\frac{2}{4}+\frac{2}{4}} = 0.5$$

如果使用的釋義詞彙完全相同，則透過上述得到的 SRD 值會較高，即爲釋義相同義涵詞彙；反之，若共有釋義詞彙佔有的比例較低，則因沒有共同的釋義內容，而使兩個詞彙在直接的釋義內容上，無法找到共同詞彙交集。當使用這釋義詞彙進行計算時，有三項主要的缺點：(1)共有釋義詞彙數目的計算是以斷詞後的字詞組爲基準，並以字型比較原則進行計算，無關詞彙本身意義。如：「*魚鱗*」與「*龍鱗*」，雖然兩者都是以鱗片進行比喻，但在以字型爲基礎的字/詞組比較時，卻無法將兩者進行關聯；(2)多數詞彙的完整釋義句子都不長，從而使共有釋義詞彙數目對 SRD 值的影響十分敏感，如前述「*鱗波*」一詞。若將釋義「*鱗波*」改爲 *"水面上似魚鱗狀的波紋。"*，則經斷詞之後爲 *"水面 (Nc) 上 (Ncd) 似 (VG) 魚鱗狀 (Na) 波紋 (Na)。"*，則共有釋義詞彙由原本的兩詞

[2] 中研院斷詞系統，Chinese Knowledge and Information Processing (CKIP) Word Segmentation System, http://ckipsvr.iis.sinica.edu.tw/

變成三詞 *"水面 (Nc)上 (Ncd)波紋 (Na)。"* 並使 $CoAP(\text{漣漪}) = \frac{(\text{水面 上 波紋})}{(\text{水面 上 細微 波紋})} = \frac{3}{4}$ 且 $CoAP(\text{鱗波}) = \frac{(\text{水面 上 波紋})}{(\text{水面 似 魚鱗狀 波紋})} = \frac{3}{4}$ 最後兩個的 SRD 值則會是 $SRD_{\text{漣漪,鱗波}} = 2\frac{\frac{3}{4}*\frac{3}{4}}{\frac{3}{4}+\frac{3}{4}} = 0.75$ 而提高 25%； (3)所有共同釋義詞彙的權重皆相同，不會區別字詞在釋義句中的語法角色，因此釋義各字與詞彙的詮釋比重皆相同。如經斷詞之後的「*鱗波*」爲 *"水面(Nc) 上(Ncd) 似(VG) 魚鱗狀(Na) 波紋(Na)。"*，依句型文法與詮釋的角度解讀，其概念表達僅需要 *"魚鱗狀"*、*"波紋"*、*"水面"* 三詞彙，但上述釋義關聯方法則無法把功能詞去除。基於釋義文字的字詞組爲表達辭條概念意涵的原則之下，我們則將釋義文字的釋義也加入兩詞彙之間的關聯比較，因此用來計算的共用釋義詞彙能夠不僅局限於單一釋義階層的詞彙(字型相同)比較上。以使用相同的辭典內容，再次解釋出現的共有釋義詞彙意義，並納入關聯計算函式中，除了以增加共有釋義詞彙廣度，同時也依重覆出現的釋義詞彙計次來決定參與計算權重，以改善前述的三項缺點。計算方式如下：令 X^1 爲條目 x 所有符合過濾條件的辭典釋義詞彙，即**第一階(直接)釋義詞彙組合**；X^2 爲 X^1 釋義詞彙再經過辭典釋義文字擴充並符合過濾條件詞彙，即**第二階釋義詞彙組合**(釋義文字的釋義)；同理，Y^2 爲 Y^1 符合的釋義詞彙，再經辭典釋義擴充後且符合過濾條件詞彙集合。而 $X^{(1+2)}$ 爲 X^1 第一階(直接)與 X^2 第二階(釋義文字的釋義)釋義詞彙的集合總合；同理，$Y^{(1+2)}$ 爲 Y^1 第一階與 Y^2 第二階釋義詞彙的集合總合。而 X^1、X^2 與 Y^1、Y^2 共同擁有的釋義文字的集合則表示爲 $X^{(1+2)} \bigcap Y^{(1+2)}$，使前述的第二階共有釋義關聯值計算則修改成 $CoAP^2_{x,y}(x) = \frac{(X^1 + X^2) \bigcap (Y^1 + Y^2)}{(X^1 + X^2)}$ 以表示所有用來釋義 *x* 詞條的第一階與第二階文字集合，與所有用來釋義 *y* 的第一階與第二階文字集合，兩者之間交集共同出現的擴充釋義字詞的比例。以此類推，則第 n 階層釋義關聯的一般式即可寫成：

$$CoAP^n_{x,y}(x) = \frac{\sum_{i=n} X^i \bigcap \sum_{i=n} Y^i}{\sum_{i=n} X^i} \tag{3}$$

從而用來計算兩者之間的第 n 階釋義關聯數值則修改成爲下列一般式：

$$SRD^n_{x,y} = 2\frac{CoAP^n_{x,y}(x) CoAP^n_{x,y}(y)}{CoAP^n_{x,y}(x) + CoAP^n_{x,y}(y)},\ 0 \le SRD^n_{x,y} \le 1 \tag{4}$$

透過前述多階層反覆釋義並計算共同出現的釋義文字比例，在先前 SRD 的比較方法研究中已討論過，除了可以發掘詞彙深層的釋義詞彙，亦可使釋義關聯值不受釋義文字

些微修改而大幅影響釋義關聯值。在計算共同擁有的釋義文字在反覆的階層釋義過程中，也可利用反覆釋義而出現的詞彙累計，突顯出權重高的佔有詞彙並加入計算。在多階層反覆釋義的過程，能找出兩詞彙之間深層的共有釋義詞彙，並且因爲能被擴充釋義的詞彙都已經透過辭典釋義過，所以反覆釋義有益於穩定計算關聯的結果(即所有釋義成份所佔百分比會傾向一定的組成比率)。最後在前先的研究中亦建議多階層釋義關聯值，可以取用第四階層的語義關聯程度進行討論(即反覆進行釋義處理至 X^1、X^2、X^3、X^4，並將所有結果加總計次)。

3. 《同義詞詞林（擴展版）》

哈爾濱工業大學信息檢索研究室 (HIT IR Lab) 所提供的公開版本《擴展版》，是整理自《詞林》(梅家駒等，1983)，除了刪除舊詞與罕用詞外，並依新聞語料加入常用新詞。在梅版的《詞林》中，收錄來自詞素、詞組、成語、方言詞與古語等詞共五萬三千多詞彙數，並且依照同義詞分類涵義有系統地區分爲人、物、時間、空間、抽象事物、特徵、動作、心理活動、現象與狀態、關聯、語助、敬語等十二組大類(以 A 到 L 標記類別的第一位英文字)以及若干中類與小類。同類型詞語依照「相對、比較」的排序原則(梅家駒等，1983)依同義/近義程度在同類型中，單行詞語由左自右排列，詞語所屬類別與列舉位置則隱含有作者們的巧思。《擴展版》對原始的分類也擴展到五層，其中加入「相等、同義」(=)、「不等、同類」(#)及「自我封閉、獨立」(@)等相關涵義。

表 1.《同義詞詞林（擴展版）》例

Bp20B03=招子**幌子**市招
Dd15A09=**幌子**招牌牌子旗號金字招牌
Aa01B03#良民順民
Bg02B07#超聲波低聲波聲波
Aa01C05@眾學生
Bg03A01@火

在表 1 中可看得出，《擴展版》都保留了分類類別、字彙及同義詞彙，並沒有針對該類別給予明確的類別涵義定義，亦沒有對所類別中的詞彙給予明確定義。以 Bp20B03= 來說明，雖依《詞林》編撰原則—同義詞在前、近義詞在後—說明“*招子*”與 Bp20B03= 應屬同義，在《國語辭典》中分爲六個釋義，其分別爲*「招牌、廣告、海報」*、*「門票。」*、*「亦稱爲花招、招兒」*、*「死刑犯就刑時，插於背後的紙標，用來揭示犯人的罪狀、姓名。」*、*「隨風招展的長布簾子。」*、*「眼睛。多用於江湖人物間。」*。而排序在第二個詞*幌子*，在《國語辭典》中分爲二個釋義，分別爲*「掛在店鋪門外，用來招徠顧客的招牌。」*與*「表現在外用以蒙蔽他人的言行。」*；*市招*的釋義僅只有*「商店門外標示其名號及所賣貨物的招牌或標誌。」*。雖然分類類別是屬 B 類(物)，但從分類架構或比較同類型詞組，亦無法從《擴展版》的分類之中得知在 Bp20B03=是屬“*招子*”的六個釋義中那一個定義。此外，一詞多義(Homographs)在《擴展版》之中則會分列在不同的類

別之中，如前述的“*招子*”則分別被歸在 Bp20B03=(B 類，物)與 Dk15A03= (D 類，抽象事物)之中。此分類結果不僅進行同義詞歸類時會造成模糊，亦有可能會在分析文本時造成所表達的義涵辨別錯誤。由於《詞林》僅將詞彙高度概化後，再將詞彙置放於相對的類別之中(鮑克怡，1983)，雖然經過增、刪、修改之後已較符合時下用語，但這樣人爲的相對分類原則很難讓後人把研究的語料歸納到《詞林》分類之中。

4. 同義概念的相關研究

在討論詞彙的同義概念的研究上，大概可區分爲兩種面向進行：詞義訓詁與機器學習。詞義訓詁中的同義相訓工作即是在建立同義詞關係，如《爾雅‧釋言》：「宵，夜也。」與《說文》：「麗爾，猶靡麗也。」。然而使用人工方法進行詞義訓詁時，如《爾雅》在訓釋散佈於各處的資料中，難以尋找在同一基礎概念上構築的同義詞彙，使訓詁工作需要專業人士經過長時間累積 (王建莉，2012)。此外，在已由前人歸類完成的同義詞組內涵中，亦是由於沒有明確說明詞組內容，從而使後人會有理解上的差異。如范紅麗 (2011) 對"*拜*"、"*揖*"、"*稽首*"、"*頓首*"、"*稽顙*"、"*拜稽首*"等詞群，以《左傳》爲資料分析在同義詞群中的異同。而在現代詞彙的研究之中，由於無古文可訓，從而使用語料庫方法進行詞彙同義比較。全文奕與郭聖林(2012)對現代漢語中的"*計程車*"、"*出租汽車*"、"*計程車*"、"*德士*"同義詞群的來源義涵與競爭進行討論，並輔以北大 CCL 語料庫驗證，以討論外來譯語存在的分佈情況與其限制。

如前述提及的機器學習方法，基於統計資料分析則可區分爲監督式與非監督式的學習方法兩類。監督式的學習方法以詞語規則、文法規則或概念規則中擷取計算詞彙之間的相似度。曾慧馨、劉昭麟、高照明、陳克健(2002)使用了詞彙中用字的組合規則，和結構與概念之間相似度對動詞進行未知詞的同義分類。而非監督式的學習方法則是將詞彙成對，透過在文本之中的出現字串的『相對頻率』(relative frequency)、『*互*見資訊』(mutual information)和『上下文依附』(context dependency) 等各種統比率(林頌堅，2004)計算統計上的相似程度。舉例來說，使用詞彙出現頻率的餘弦(cosine)關係進行圖書資訊檢索中的詞彙擴展(陳光華與莊雅棻，2001)與*互*斥資訊熵原則 PMI-IR (Pair-wise Mutual Information-Information Retrieval)學習並預測在 TOEFL 考試中的同義詞考題(Turney, 2001)。

5. 研究方法

本研究旨在使用辭典釋義關聯計算同義字群中的共有概念擷取，並提供除了同義詞組外更明確的辭典釋義文字，以利後續的詞義消歧工作。在此我們則使用《教育部重編國語辭典修訂本》釋義作爲計算的基礎，計算在《同義詞詞林（擴展版）》之中的所有分類項後，提取出該分類項中共同有的釋義文字、交叉比對釋義關聯比例與平均最大釋義關聯詞(average dictionary definition relationship)，以呈現在同義詞群之中最能表達該群的概念。

5.1 語料準備

本研究所使用的語料主要有 HIT IR Lab 的《擴展版》與教育部的《國語辭典》。其中《擴展版》的資料是以簡體字碼編寫，因此我們使用了維基百科的繁簡分歧詞表進行繁簡轉換。維基百科的繁簡分歧詞表包括了大陸、台灣、香港與新加坡各地的漢語編碼與詞彙互換原則，例如 hardware 一詞在大陸稱作*硬件*、台灣則稱做*硬體*。另一方面使用《國語辭典》作爲釋義計算的基礎，因此需要將辭典內的詞條去除掉詞目、正形、注音、引證和案語等資訊，只保留釋義的部份再送至中研院 CKIP 斷詞系統進行詞性標記(P.O.S. tagging)。經過處理資訊總計處理條目數爲 156296，共計 278793 種不同的釋義。

5.2 詞彙間釋義關聯計算

爲了解詞彙之間共同有的釋義詞彙，我們使用釋義關聯對詞彙之間的釋義做交互比對，取得到最大釋義關聯值的文字作爲兩詞彙之間的同義概念。在進行釋義關聯計算時，則使用較多語義的動詞詞類 ('VA', 'VAC', 'VB', 'VC', 'Vi', 'Vt', 'VCL', 'VD', 'VE', 'VF', 'VG', 'VH', 'VHC', 'VI', 'VJ', 'VK', 'VL', 'V_2') 與名詞詞類 ('Na', 'Nb', 'Nc', 'Ncc', 'Nd', 'N') 進行釋義關聯的計算。在釋義關聯計算原則上，我們修正前多階層釋義關聯計算原則。由於先前提出的多階層釋義關聯計算原則 (趙逢毅與鍾曉芳，2011)在進行釋義階層擴展時，會局限在特定的部首的概念中，使階層較高的釋義權重與低階層權重相同。不同於先前的計算原則，在本研究中的同義詞釋義關聯計算是以**無特定概念的方向擴展**，因此深階層釋義的概念權重應比較低階層釋義權重低(即釋義權重與階層深度成反比)。爲了使第一階層(直接)釋義文字的權重能高於第二階層(釋義文字的釋義)，我們則將每次的釋義計算過程中，累計低階層中的計次數值，使第一階層中出現的釋義詞彙權重，會較第二階層中出現的釋義詞彙重(不一定成倍數關係。因爲在第一階層中出現詞彙不一定只會出現一次，不一定在每一階層都會出現。)。若 x, y 爲兩待測詞彙，X', Y' 爲包涵 x, y 兩詞彙與其釋義的各別集合，則第 n 階層之釋義詞彙則爲 $X'^{n+1} = X'^{n} + X^{n+1}$ 且 $Y'^{n+1} = Y'^{n} + Y^{n+1}$:

$$CoAP'^{n}_{x,y}(x) = \frac{\sum_{i=n} X'^{i} \bigcap \sum_{i=n} Y'^{i}}{\sum_{i=n} X'^{i}} \tag{5}$$

$$SRD'^{n}_{x,y} = 2\frac{CoAP'^{n}_{x,y}(x)CoAP'^{n}_{x,y}(y)}{CoAP'^{n}_{x,y}(x) + CoAP'^{n}_{x,y}(y)}, \quad 0 \le SRD'^{n}_{x,y} \le 1 \tag{6}$$

由於計算關聯的原則不同，因此我們先觀察在計算不同階層結果以下，是否如先前研究結果一樣，可以取第四層爲基準進行兩詞彙之間的關聯比較。在此則以*「漣漪」*與*「鱗波」*兩詞，計算修改後版本的詞彙間釋義關聯，結果呈現由第一階層到第八階層的結果(詳見下頁表 2)。

從表 2 的結果可以知道，當階層越深則釋義關聯越高，且因爲使用反覆的辭典釋義擴充參與計算的詞彙，因此較深層的義涵在第三、四層之間，共有的釋義文字佔有的比

例產生較明顯的差異。在低階層釋義關聯(淺層釋義)中,因爲詞彙並沒有經過太多解釋(階層數少),因此共有文字相較於高階層釋義關聯(深層釋義)較爲明確具體,如:*水面*、*波紋*、*魚鱗狀*等等。當詞彙經多次釋義之後,共同出現的詞彙則會依知識概念 (concept) 較高的層級堆疊累積而清楚說明,而開始將明確具體的共有詞彙所佔的比例降低,從而出現*表面*、*部份*等文字,直到計算第八階層的計算結果出現共有釋義詞彙如*個體*、*事物*等文字。雖然在本次的實驗之中,我們因爲加入了**無特定概念的趨向**(先前的研究中,局限在屬「目」字部的字與詞彙)的多階層擴充原則,而增加了低階層的共有釋義在計算過程之中的比例。在上述的結果顯示,這樣的計算方法可以擷取出較低階層中(明確具體)與較深階層中(一般性概念)的共有釋義詞彙。在此需要特別說明的是,共有釋義文字*水面*、*波紋*等在第一、二、三、四階層中,這些詞彙擁有高權重的原因是多數釋義文字都共同使用到,而使*水面*、*波紋*詞彙的概念主宰支配(dominate)著該詞條的主要涵義。但在高階層的釋義關聯計算結果中,*水面*、*波紋*並非沒有出現,由於這此詞彙所佔比例太低且落於前 20 名單之後,取而代之的主宰支配共有文字則爲*表面*、*部分*、*事物*、*個體*等一般性概念詞彙。

表 2.漣漪與鱗波兩詞由第一階層到第八階層釋義關聯結果及共用的釋義文字比例 (每階層只取前釋義詞彙所佔權重的 Top 20 列表)

關聯層級深度	第一階層	第二階層	第三階層	第四階層	第五階層	第六階層	第七階層	第八階層
釋義關聯值(SRD)	0.4	0.714	0.9	0.964	0.989	0.997	0.999	≒1
(釋義文字所佔權重)	(釋義文字出現比例)							
水面	**20.00%**	**23.81%**	**20.00%**	**12.97%**	6.97%	3.39%	1.59%	-
波紋	**20.00%**	**14.29%**	**8.00%**	**3.60%**	-	-	-	-
表面	-	4.76%	**9.00%**	**9.73%**	**7.71%**	**5.21%**	**3.29%**	**2.05%**
部分	-	-	2.00%	**4.68%**	**6.14%**	**6.17%**	**5.46%**	**4.55%**
上	10.00%	4.76%	3.00%	2.34%	1.94%	1.65%	1.43%	1.26%
紋理	-	4.76%	6.00%	4.86%	3.14%	1.85%	-	-
稱爲	-	4.76%	6.00%	4.68%	2.94%	1.79%	-	-
水	-	4.76%	6.00%	4.68%	2.86%	1.58%	-	-
水皮兒	-	4.76%	6.00%	4.68%	2.80%	1.46%	-	-
漣漪	10.00%	4.76%	2.00%	-	-	-	-	-
似	10.00%	4.76%	2.00%	-	-	-	-	-
魚鱗狀	10.00%	4.76%	2.00%	-	-	-	-	-
鱗波	10.00%	4.76%	2.00%	-	-	-	-	-
細微	10.00%	4.76%	2.00%	-	-	-	-	-
物體	-	-	2.00%	3.60%	3.71%	3.02%	2.28%	1.70%
形成	-	4.76%	5.00%	3.24%	1.63%	-	-	-

事物	-	-	1.00%	**2.16%**	**2.74%**	**2.86%**	**2.81%**	**2.70%**
外在	-	-	1.00%	2.16%	2.57%	2.30%	1.79%	1.28%
微浪	-	4.76%	4.00%	2.16%	-	-	-	-
外界	-	-	-	2.16%	2.57%	2.31%	1.81%	1.34%
現象	-	-	-	2.16%	2.57%	2.30%	1.79%	1.30%
個體	-	-	-	-	1.97%	**2.50%**	**2.51%**	**2.25%**
實質	-	-	-	1.98%	2.17%	1.80%	1.30%	-
接觸	-	-	-	1.80%	1.83%	1.42%	-	-
中	-	-	-	-	-	1.52%	1.67%	1.67%
構成	-	-	-	1.62%	1.60%	1.28%	-	-
認識	-	-	-	-	-	1.33%	1.52%	1.52%
呈現	-	-	1.00%	1.62%	1.57%	-	-	-
整體	-	-	-	-	-	-	1.46%	1.48%
人	-	-	-	-	-	-	1.30%	1.58%
某些	-	-	-	-	-	-	1.26%	1.19%
部下	-	-	-	-	-	-	1.26%	1.18%
秩序	-	-	-	-	-	-	1.26%	1.18%
局部	-	-	-	-	-	-	1.26%	1.18%
部屬	-	-	-	-	-	-	1.26%	1.18%
空間	-	-	-	-	-	1.29%	-	1.07%
紋路	-	-	-	-	1.60%	-	-	-
指	-	-	-	-	-	-	-	0.95%

本文研究旨在尋找在同義詞彙之中，以其共同擁有的概念爲內容。綜合上述分析，使用多階層釋義關聯若以深階層的計算結果爲主(如第八階層)，則實驗結果會使所有詞彙都落於一般性概念詞彙上，從而無法突顯同義詞彙較具體的概念涵義；反之，若使用淺階層(如第一階層)釋義關聯，則共同釋義詞彙又會因字型不符而無法歸類在一起。故從而則參考先前的研究結果，選定以第四階層釋義關聯進行計算，並取出現比率較高的前 20 釋義文字作爲同義詞彙的概念釋義。

5.3 多義詞處理與最大平均釋義關聯詞

在辭典釋義中，單一詞彙往往會有許多釋義協助辨析該詞條的主要涵義、用法及來源典故。但在關聯計算上，雖然可以透過處理多階層釋義解決多詞同義的問題，但這個方法卻不能解決一詞多義的問題。爲了能確定多義詞彙在同義詞組所適用的涵義，在本研究中，我們則將不同的釋義分開來計算，以一對一的詞條釋義比對原則來尋找能得到最大釋義關聯值的釋義，作爲在該同義詞組的最佳同義說明。以前述在《擴展版》中同義詞組“Bp20B03=*招子 幌子 市招*”爲例(相關在《國語辭典》中的釋義，請參照前述)。經

過釋義處理後，各別可供計算的釋義文字如下表 3。

表 3.經處理過後的同義詞組 Bp20B03=《國語辭典》釋義，及關聯計算結果

同義詞彙	經處理後的《國語辭典》釋義		釋義關聯值 SRD[4]		
			幌子-1	幌子-2	市招
招子	招子-1.	花招 招兒	0.380	0.475	0.475
	招子-2.	死刑犯 刑 時 插於 背後 紙標 用來 揭示 犯人 罪狀 姓名	0.470	0.566	**0.636**
	招子-3.	招牌 廣告 海報	0.549	**0.675**	0.620
	招子-4.	隨風 長布 帘子	0.421	0.541	0.505
	招子-5.	門票	0.398	0.481	0.499
	招子-6.	用於 江湖 人物 間	0.506	0.581	0.597
	招子-7.	眼睛	0.427	0.524	0.499
幌子	幌子-1.	表現 在外 蒙蔽 他人 言行 幌子			0.662
	幌子-2.	掛 店鋪 門 外 用來 招徠 顧客 招牌			**0.813**
市招	商店 門 外 標示 其 名號 賣 貨物 招牌 標誌 市招		N/A	N/A	

在表 3 中同時比較七個詞義的“*招子*”、二個詞義的“*幌子*”與單一詞義的“*市招*”，並透過多階層釋義關聯計算列出對應表並加總求最大值。從表 3 中可以知道最適合詞的詞彙語義關係建立在“*招子*”與“*幌子*”之間，因爲“*市招*”與“*幌子-1*”、“*幌子-2*”之間在第四階層的釋義語義關聯下仍無法產生關係(沒有 *SRD*[4] 數值 N/A)，而釋義關係最大值是建立在“*招子-3*”與“*幌子-2*”之間。在這兩兩成對的關係之中，我們計算各別平均的關係並取其最大值平均值作爲此類別的代表詞彙，則“*招子*”爲(0.675+0.636)/2=0.655、“***幌子***”**爲(0.675+0.813)/2=0.744**、“*市招*”爲(0.813+0.636)/2= 0.724，故“*幌子*”平均值在釋義概念上爲其它詞彙的釋義關聯最高，很適合作爲此分類的代表詞彙。接著，我們將出現在這三個詞彙裡第四階層共同擁有的釋義文字權重中，取最高的前 20 個釋義詞彙列表，作爲該同義詞彙之中主宰支配(dominate)整個同義詞組的主要釋義詞彙。而經計算後選取出各詞彙之間第四階層共有文字釋義，以統計各別文字的分佈(同 5.2 處理原則)，例如前 20 個較高釋義文字，順序如下(由高到低排列)：

*共有釋義文字分佈：**掛=3.12%人=1.69%招牌=1.54%牌子=1.47%***
招子=1.46%有=1.40%上=0.95%標識=0.86%揭示=0.82%
用來=0.81%單位=0.79%表示=0.78%懸吊=0.74%廣告=0.71%
賣=0.70%刑=0.68%物品=0.65%獻=0.61%計算=0.59%團體=0.59%
#總釋義文字：152296

雖然從這些釋義文字中可分析出，在此同義詞之中較多共用釋義的概念爲何。我們能羅列上述 20 個主要主宰支配這個同義詞組的共有釋義詞詞彙，但卻無法將上述的詞彙組合成爲文字，除了尚有約 15 萬個字詞沒有列出外，在使用多階層釋義關聯計算時，我們僅只保留了動詞與名詞詞性，忽略了文法結構。此外我們亦無法將所取得的共有釋義文字組合成爲單一精簡句型，但我們亦可以透過比較相似的同義字群組，幫助我們了解不同的同義詞組之間的差別。分析在《詞林》中擁有相同詞彙“*幌子*”的同義詞組“*Dd15A09=幌子招牌牌子旗號金字招牌*”同義詞組的結果(見表 4)，我們依前述方法，逐一比較同義詞組之中各詞彙的每一項釋義內容，除了使用與其它同組詞彙的第四階釋義語義關聯最大值平均值，與共有釋義詞彙最大涵蓋數來決定最適合代表的詞彙之外，表 4 中亦列出共有釋義詞彙之中的 Top 20 作爲比較。從表 4 中我們可知，“*幌子*”在此群組之中與“*招牌*”的共有釋義詞彙涵蓋度較高，在第四階釋義詞彙語義關聯值 0.975，但是與其它詞彙間的關聯就相對低。再比較表 3 之結果，我們可以說“*幌子*”釋義關聯與“*招牌*”的關係較與“*市招*”的關係較接近，接著才是“*金字招牌*”與“*招子*”。

表 4.同義詞類 Dd15A09=之各詞間釋義關聯表

	幌子	*招牌*	*牌子*	*旗號*	*金字招牌*	平均值
幌子		0.975	0.548	0.583	0.738	0.711
招牌	0.975		0.789	0.527	0.830	0.780
牌子	0.548	0.789		0.381	0.637	0.588
旗號	0.583	0.527	0.381		0.638	0.532
金字招牌	0.738	0.830	0.637	0.638		0.710
平均值	0.711	0.780	0.588	0.532	0.710	

max Average:招牌 0.780
*共有釋義文字分佈：**人=2.87%掛=2.45%**有=1.91%**牌子=1.66%表示=1.08%***
*調子=1.07%上=0.84%**招牌=0.82%**某=0.73%指=0.70%單位=0.64%他人=0.63%*
種=0.62%事物=0.61%一=0.60%懸吊=0.57%名義=0.55% 獻=0.55%具有=0.53%
標識=0.52%
#總釋義文字：426366

然而表 3 與表 4 之間亦可從與其它同組詞彙的第四階釋義語義關聯最大的平均值，與共有釋義詞彙最大涵蓋數來決定最適合代表的詞彙。在表 3 中“*幌子*”釋義爲“(1) 表現在外用以蒙蔽他人的言行。 (2) 掛在店鋪門外，用來招徠顧客的招牌。”在表 4 中“*招牌*”釋義爲“(1) 商店機構作爲標識的牌子。 (2) 演藝人員或團體揭示其所獻技藝有關事項的牌子。(3) 拿手的，可作爲標識的。(4) 比喻騙人的幌子。”從釋義之中，我們可以理解共有釋義的處理原則是將“*招牌-4*”中所出現的“*幌子*”釋義文字納入計算而得的最大值來關聯，但“*招牌*”對其它 Dd15A09=同義詞組中的詞彙關聯卻不從“*招牌-4*”而來，而是“*招牌-1*”或“*招牌-3*”。另一方面，“*幌子*”能與 Dd15A09=同義詞組中的

詞彙產生釋義關聯的，僅僅只有“*幌子-2*”釋義中，較爲明確的“*招牌*”釋義而來，因此在此一同義群組之中得到的平均數值就會低許多。透過上述的說明我們可以知道，雖然最適合同義詞組的代表詞彙必需是滿足釋義語義關聯最大值平均值與共有釋義詞彙最大涵蓋數。但要同時達到這兩項條件，則該詞彙必須包涵的各階層釋義文字集合，且大多都要能出現在同一辭組之中其它詞彙的釋義文字裡，也就是要「被用來說明辭典裡的某個辭條」時，才會可以成爲該辭組的最合適同義詞彙。

在計算上，並非以深層釋義關聯(如第八階層)所能取得的一般性概念詞彙進行計算，而是詞層概念層級在有限度(本研究以第四階層進行計算)的概化擴充之後，仍要滿足前述的條件。此外，詞彙間是 part-and-whole 的隱喻關係的“*金字招牌*”與“*招牌*”兩詞，在《詞林》的依同義/近義程度在同類型中由左至右排列原則之下，我們也可以得出相對於“*招牌*”一詞，“*金字招牌*”可視爲近義詞而非同義關係。這項結果也可以從表 4 中看出“*招牌*”一詞可以完全可主宰此同義詞組共有釋義，而“*金字招牌*”卻無法主宰看出 (因爲兩詞彙在高度釋義相關的條件下，應同時能主宰該同義詞組。若否，在一詞多義的條件之下，則可推論兩者相關聯的釋義並非此同義詞組的主宰釋義)。 最後，從共同擁有的釋義權重 Top 20 中，兩同義詞組所表現的釋義詞彙亦是不同的。可以看出，Bp20B03=之中從較具體的“*招牌*”，而在 Dd15A09=則從抽象的“*牌子*、*表示*”等概念涵意，此與《詞林》之分類規則上(B 類爲物、D 類是抽象事物)是相同的。

雖然同義詞組經過上述計算可以得到詞組之中最適合用來表達的同義詞彙，與共同擁有釋義文字的權重比例(在此僅列出 Top 20)，但此方法亦有限制。由於此方法需透過辭典反覆對詞彙進行釋義、斷詞再擴充釋義文字，所以當無法取得釋義內容時，則會造成無法計算釋義關聯的窘境。以下參考維基百科爲例。描述「訓詁學」定義一組同義詞“*訓詁訓故故訓古訓解故解詁*”，其計算結果如下表 5。從表中可以看出“*訓故*”與“*故訓*”三詞在所使用的教育部《國語詞典》中找不到釋義，因此無法計算釋義關聯，此爲缺點一。而從其它三個字的四階層釋義關聯計算之後的結果，最大釋義關聯文字可以使用“*解詁*"代表，因爲經過計算之後最大平均值是 0.741。這項數值與“*解故*”的 0.740 之間雖然只差千分之一，但仍無法直接將“*解詁*”與"*解故*”視爲相同詞彙或概念義涵相同(因爲本研究所用方法僅止於計閱釋義而非詞彙概念)。最後釋義文字比重前二十個字詞亦能表現出此同義字組的共有釋義文字，如：“*指*”、“*解釋*”、“*古代*”、“*文字*”、“*說明*”、“*分析*”等，但卻不能組織並架構成一句精簡的同義詞釋義(前文已討論過)。儘管此方法還有許多尚待改善的缺陷，但能提到詞彙之間客觀的比較基準與可供參考的釋義文字權重，相對人工訓詁方法，仍能在眾多的資料之中提供快速同義關聯參考依據。

表 5. 自建同義詞組"訓詁"及各詞間釋義關聯表

	訓詁	*訓故*	*故訓*	*古訓*	*解故*	*解詁*	平均值
訓詁		N/A	N/A	0.814	0.575	0.569	0.652
訓故	N/A		N/A	N/A	N/A	N/A	N/A
故訓	N/A	N/A		N/A	N/A	N/A	N/A
古訓	0.814	N/A	N/A		0.677	0.684	0.725
解故	0.575	N/A	N/A	0.677		0.971	0.740
解詁	0.569	N/A	N/A	0.684	0.971		0.741
平均值	0.652	N/A	N/A	0.725	0.740	0.741	

max Average:解詁 0.741

共有釋義文字分佈：指=4.96%解釋=2.73%手=1.98%某=1.97%古代=1.79%

個=1.72%文字=1.51%說明=1.47%人=1.40%分析=1.26%部分=1.15%

一=1.07%原因=1.04%事=0.94%希望=0.92%指示=0.92%中=0.91%

理由=0.84%消除=0.83%直立=0.82%

#總釋義文字：93720

5.4《擴展版》詞類義涵分析

在《擴展版》中將同義詞類區分爲「相等、同義」(=)共計 9995 類 55844 字/詞數、「不等、同類」(#)爲 3445 類 29893 字/詞數與「自我封閉、獨立」(@)有 4377 類 4377 字/詞數。從資料上可以看出，在「自我封閉、獨立」分類下的同義詞均爲單一字/詞彙分類，而「不等、同類」的詞彙，如「*Bg02B08# 麥浪松濤煙波*」，不僅難從辭典釋義之中找到關聯 (“*麥浪*”釋義爲“*麥田中的麥苗遭風吹拂時起伏如浪的樣子*”；“*松濤*”釋義爲“*風吹松樹所發出像波濤般的聲音*”；“*煙波*”釋義爲“*雲煙瀰漫的水面。*”)，甚至在詞意概念關聯上仍需要人爲的語感協助才能分別。

因此，在本研究中僅以「相等、同義」(=)分類別進行分析。總計處理詞類 8311 類，其中同類詞組中找不到任何《國語辭典》釋義詞組計有 502 組、僅能找到單一詞彙之同義詞組爲 1182 組，共計處理的字/詞數爲 44872，釋義總計爲 126605 組。處理《擴展版》的方式如前所述，單一合適的「相等、同義」詞組取得詞彙的釋義之後，並將釋義內容送至中研院斷詞系統處理後留取合適的動詞與名詞，並反覆進行四次以計算第四階層釋義語義關聯。最後將釋義內容一對一交互比對，取得最合適的代表詞組詞彙與 Top 20 共有釋義詞彙權重，再依表 1 中《擴展版》的資料羅列如下列表 6 排版。表 6 僅列出各大項(A 組到 L 組之中的取樣結果)。

表 6.分析《擴展版》同義詞類範例結果：摘錄 A 組到 L 組各組一筆

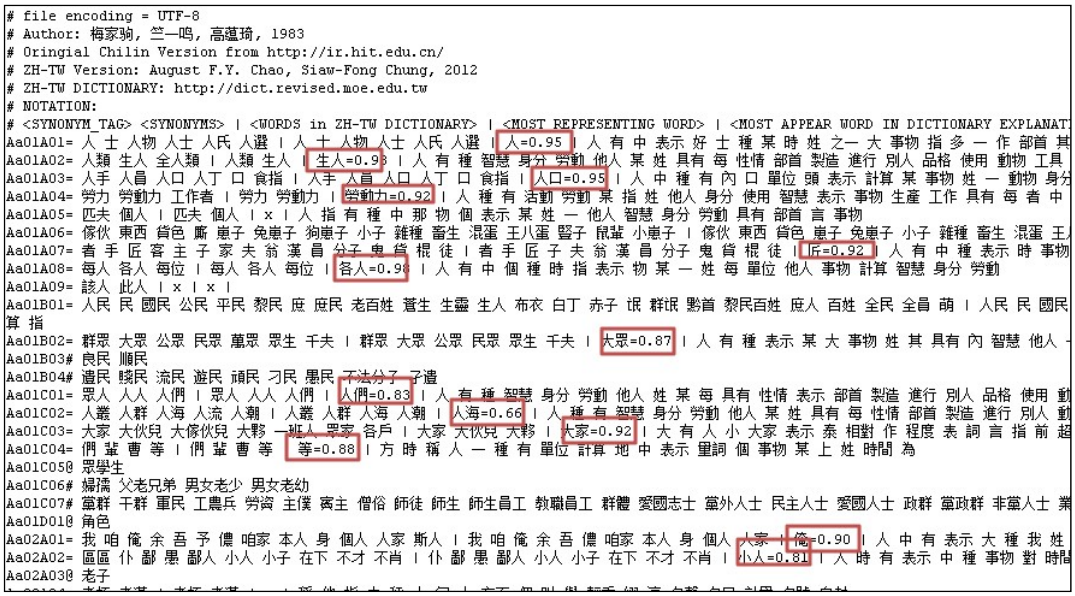
<SYNONYM_TAG><SYNONYMS> | <WORDS in ZH-TW DICTIONARY> | <MOST REPRESENTING WORD> | <MOST APPEAR WORD IN DICTIONARY EXPLANATION>
Ab01C01=男女 士女 兒女 紅男綠女 男男女女 少男少女|男女 士女 兒女 紅男綠女 男男女女|男女=0.87 |有 女 表示 大 男女 中 人 作 者 色 相對 前 多 男 指 某 用於 一 表 事物
Bf06B01=閃電 電 銀線 電閃|閃電 電|電=0.90 |種 有 中 人 表示 其 時 事物 一 電 內 兩 單位 計算 姓 多 帶 使 指 某
Bq03C03=絨褲 衛生褲| x | x |
Cb06B01=附近 就近 鄰近 近處 一帶 內外 左右 左近 前後 近水樓台 就地 近旁 跟前 不遠處|附近 就近 鄰近 近處 一帶 內外 左右 左近 前後 就地 近旁 跟前|附近=0.82 |中 表示 其 有 言 大 稱 人 左右 處 某 一 指 姓 句 個 地方 屬 時 單位
Dk16B01=電報 電 報|電報 電 報|電報=0.94 |種 有 人 地 表示 其 時 事物 電 一 單位 內 得 中 計算 指 兩 使 上 種子
Df01B01=感想 感 感觸 感受|感想 感 感觸 感受|感觸=0.90 |中 心 內 有 人 感 種 事物 一 表示 某 影響 三 稱爲 之一 思想 姓 內心 上 指
Ef06C01=空閑 悠閑 幽閑 安閑 清閑 輕閑 閑暇 空暇 閑空 悠然 閑 空 逸 暇 空餘 得空 有空 沒事|清閑 閑暇 空暇 悠然 閑 空 暇 得空 有空 沒事|空=0.61 |有 表示 養 閑 一 某 時間 時候 閒 事物 前 樣子 指 人 存在 正面 餘數 地方 相對 用於
Fa04B01=捧 掬 端 端面 端平|捧 掬 端|捧=0.94 |手 兩 人 單位 計算 一 某 等於 種 公斤 量詞 斤 事物 詞 雙 公制 臺 表 做 部首
Ga17B01=感動 感 觸 動容 感觸 百感叢生 催人淚下 令人感動 動感情 動人心魄|感動 感 觸 動容 感觸 動人心魄|感動=0.89 |人 指 有 使 事物 物 感應 一 某 種 稱爲 表示 影響 物體 姓 個 現象 中 電 感動
Hb02C03=開戰 開仗 開火 動武 用武 動干戈 宣戰|開戰 開仗 開火 動武 用武 動干戈 宣戰|開火=0.78 |指 一 一點 中 個 開始 開戰 雙方 開火 手 稱爲 某 兩 事物 人 爭鬥 開仗 部分 有 指示
Ig04B01=循環 輪迴 循環往複 周而復始 大循環 巡迴|循環 輪迴 周而復始 大循環|循環=0.76 |有 指 人 事物 一 中 一切 承 組織 血液 動物 表示 各 佛教 種 物 稱爲 個 內 眾生
Jd08A03=失去 失掉 失卻 失 去 奪 錯過 錯開|失去 失掉 失卻 失 奪 錯過|失掉=0.47 |奪 失 錯過 遺落 做 到 失去 決定 挫 機會 脫漏 失掉 丟掉 眩目 耀眼 漏掉 爭取 人 強取 衝過
Ka21A01=借故 託故 假託 推託 借口 假說 託詞 託辭|借故 託故 推託 借口 託詞 託辭|託詞=0.83 |藉口 理由 某 人 推託 假借 有 借口 借用 字 話 自己 語言 事 別人 我 爲 事物 作爲 論說
La04C01=抱歉 對不起 對不住|抱歉 對不起 對不住 |對不住=0.98 |中 心 人 有 表示 之一 姓 內 種 某 一 端 進行 性情 宿 部首 兩 名 星 事物

在表 6 之中，我們依表 1《同義詞詞林-擴展版》的原始資料保留在每行的前兩個欄位，標記爲<SYNONYM_TAG><SYNONYMS>，隨後我們將能在《國語辭典》中找到的詞彙列於<WORDS in ZH-TW DICTIONARY>欄位、最合適的代表該詞組的詞彙列於<MOST REPRESENTING WORD>，最後擷取用於計算的共有釋義詞彙權重最高的 Top

20 一併依序羅列於最後欄位之中<MOST APPEAR WORD IN DICTIONARY EXPLANATION>。從表 6 觀察在列表的資料，其中<MOST REPRESENTING WORD>一欄結果爲詞彙在表達該同義詞組字彙 ci 涵蓋率高，且其平均值也是最高的結果中，Ab01C01= 是以“*男女*”第四階釋義語義關聯值爲 0.87、Bf06B01=爲“*電*”0.90、Cb06B01=爲“*附近*”0.82、Dk16B01=爲“*電報*”0.94、Fa04B01=爲“*捧*”0.94、Ga17B01=爲“*感動*”0.89 、Hb02C03=爲“*開火*”0.78、Ig04B01=爲“*循環*”0.76、Ka21A01=爲“*託詞*”0.83、 La04C01=爲“*對不住*” 0.98 等，這些詞彙在表達該詞組上從字面上即可以了解該同義詞組的主要包涵的概念。至於無法得到較高釋義關聯數值的同義詞組，如 Ef06C01=“*空*”0.61 與 Jd08A03=“*失掉*”0.47 兩辭組，主因爲該同義詞組中包括了概念層級很上位的詞彙(Ef06C01=裡的“*空*”使其該詞彙在釋義的涵蓋度很高)，且在詞組之中加入詞彙的釋義與其它詞彙釋義的概念領域之間交集程度較低的近義詞彙(如 Jd08A03= “*奪*”有八組釋義爲“*強取*”、“*削除、使失去。*”、“*爭取。*”、“*做決定。*”、“*錯過。*”、“*衝過。*”、“*耀眼、眩目。*”、“*脫漏、漏掉。*”，此八組釋義與 Jd08A03=同義詞彙能建立較高釋義關聯的共有釋義詞彙是“*錯過。*”；從而使 Jd08A03=同義詞組中“*失去*”、“*失掉*”、“*失卻*”、“*失*”與“*奪*”、“*錯過*”可區分爲兩個子辭組，後者爲前者的近義詞。)。另外要說明的是，在詞典之中，同義詞找不到任何詞條釋義就無法進行關聯運算，因此在資料則會如上表 6 中的 Bq03C03= 同義詞類以"| *x* | *x* |"表示；同理，若僅只找一個詞彙則會無法計算釋義關聯，則亦無法決定最大平均釋義關聯詞，故以“| *x* |“表示。而在最後一欄的<MOST APPEAR WORD IN DICTIONARY EXPLANATION>詞組是羅列共同擁有的釋義文字權重 Top20 的詞彙，並依權重高低順序排列，雖然不見得就能構成爲可讀的句子，如表 6 之中的 Bf06B01=“*電*”與 Dk16B01= “*電報*”兩詞組在共有釋義詞彙 Top20 的重覆程度太高，因而難以區分。但可提供輔助詮釋該同義詞組更多資訊，在共有釋義詞彙 Top20 的中就可以很明顯的區別出。Ga17B01= “*感動*”及其共有釋義詞彙 Top20 中獨有的八個詞彙“*使*”、“*物*”、“***感應***”、“***物體***”、“*個*”、“*現象*”、“*電*”、“*感動*”，相對比較 Df01B01=“*感觸*”及其獨有的八個共有釋義詞彙：“***心***”、“*內*”、“*感*”、“*三*”、“*之一*”、“***思想***”、“***內心***”、“*上*”，亦可以說明《詞林》的大分類中 D 大類指是抽象事物，而 G 大類是心理的結果。

綜合前述的資料結果，我們將《同義詞詞林-擴展版》的原始資料，附加本次研究的內容結果放在 Google Code 的 tw-synonyms-chilin 的專案之中，網址爲 http://code.google.com/p/tw-synonyms-chilin/(或使用 http://goo.gl/H6YRK 直接下載處理後的文本)供其它研究人員在網站的軟體庫存專案中下載。爲了處理上述資料，我們使用了 2 台 Unix 電腦(Unbuntu: IntelCore2 Duo 2.80GHz; FreeBSD: AMD Sempron 1.8GHz)，輔以 NLTK(Loper & Bird, 2002)工具開發計算工具，進行同義詞組的多階層釋義關聯計算，與辭組之間多釋詞彙之間的交互比較，總處理時間大約 32 小時，總計處理同義詞類 8311 類，結果文件約爲 2.2MB，內容列於下表 7。

表 7.取自 tw-synonyms-chilin 之摘要結果。(方框爲最合適代表該詞組的詞彙)



```
# file encoding = UTF-8
# Author: 梅家驹, 竺一鸣, 高蕙琦, 1983
# Oringial Chilin Version from http://ir.hit.edu.cn/
# ZH-TW Version: August F.Y. Chao, Siaw-Fong Chung, 2012
# ZH-TW DICTIONARY: http://dict.revised.moe.edu.tw
# NOTATION:
# <SYNONYM_TAG> <SYNONYMS> | <WORDS in ZH-TW DICTIONARY> | <MOST REPRESENTING WORD> | <MOST APPEAR WORD IN DICTIONARY EXPLANAT
Aa01A01= 人 士 人物 人士 人氏 人選 | 人 士 人物 人士 人氏 人選 | 人=0.95 | 人 有 中 表示 好 士 種 某 時 姓 之一 大 事物 指 多 一 作 部首 其
Aa01A02= 人類 生人 全人類 | 人類 生人 | 生人=0.93 | 人 有 種 智慧 身分 勞動 他人 某 姓 具有 每 性情 部首 製造 進行 別人 品格 使用 動物 工具
Aa01A03= 人手 人員 人口 人丁 口 食指 | 人手 人員 人口 人丁 口 食指 | 人口=0.95 | 人 中 種 有 內 口 單位 頭 表示 計算 某 事物 姓 一 動物 身分
Aa01A04= 勞力 勞動力 工作者 | 勞力 勞動力 | 勞動力=0.92 | 人 種 有 活動 勞動 某 指 姓 他人 身分 使用 智慧 表示 事物 生產 工作 具有 每 者 中
Aa01A05= 匹夫 個人 | 匹夫 個人 | x | 人 指 有 種 中 那 物 個 表示 某 姓 一 他人 智慧 身分 勞動 具有 部首 言 事物
Aa01A06= 傢伙 東西 貨色 廝 崽子 兔崽子 狗崽子 小子 雜種 畜生 混蛋 王八蛋 賤子 鼠輩 小崽子 | 傢伙 東西 貨色 崽子 兔崽子 小子 雜種 畜生 混蛋 王八
Aa01A07= 者 手 匠 客 主 子 家 夫 翁 漢 昌 分子 鬼 貨 棍 徒 | 者 手 匠 子 夫 翁 漢 昌 分子 鬼 貨 棍 徒 | 匠=0.92 | 人 有 中 種 表示 時 事物
Aa01A08= 每人 各人 每位 | 每人 各人 每位 | 各人=0.98 | 人 有 中 個 種 時 指 表示 物 某 一 姓 每 單位 他人 事物 計算 智慧 身分 勞動
Aa01A09= 該人 此人 | x | x |
Aa01B01= 人民 民 國民 公民 平民 黎民 庶 庶民 老百姓 蒼生 生靈 生人 布衣 白丁 赤子 氓 群氓 黔首 黎民百姓 庶人 百姓 全民 全員 萌 | 人民 民 國民
算 指
Aa01B02= 群眾 大眾 公眾 民眾 萬眾 眾生 千夫 | 群眾 大眾 公眾 民眾 眾生 千夫 | 大眾=0.87 | 人 有 種 表示 某 大 事物 姓 其 具有 內 智慧 他人
Aa01B03# 良民 順民
Aa01B04# 遺民 賸民 流民 遊民 頑民 刁民 愚民 干法分子 子遺
Aa01C01= 眾人 人人 人們 | 眾人 人人 人們 | 人們=0.83 | 人 有 種 智慧 身分 勞動 他人 姓 某 每 具有 性情 表示 部首 製造 進行 別人 品格 使用 動
Aa01C02= 人叢 人群 人海 人流 人潮 | 人叢 人群 人海 人潮 | 人海=0.66 | 人 種 有 智慧 身分 勞動 他人 某 姓 具有 每 性情 部首 製造 進行 別人 動
Aa01C03= 大家 大伙兒 大傢伙兒 大夥 一般人 眾家 各戶 | 大家 大伙兒 大夥 | 大家=0.92 | 大 有 人 小 大家 表示 泰 相對 作 程度 表 詞 言 指 前 超
Aa01C04= 們 輩 曹 等 | 們 輩 曹 等 | 等=0.88 | 方 時 稱 人 一 種 有 單位 計算 地 中 表示 量詞 個 事物 某 上 姓 時間 為
Aa01C05@ 眾學生
Aa01C06# 婦孺 父老兄弟 男女老少 男女老幼
Aa01C07# 黨群 軍民 工農兵 勞資 主僕 賓主 僧俗 師徒 師生 師生員工 教職員工 群體 愛國志士 黨外人士 民主人士 愛國人士 政群 黨政群 非黨人士 業
Aa01D01@ 角色
Aa02A01= 我 咱 俺 余 吾 予 儂 咱家 本人 身 個人 人家 斯人 | 我 咱 俺 余 吾 儂 咱家 本人 身 個人 人家 斯人 | 俺=0.90 | 人 中 有 表示 大 種 我 姓
Aa02A02= 區區 仆 鄙 愚 鄙人 小人 小子 在下 不才 不肖 | 仆 鄙 愚 鄙人 小人 小子 在下 不才 不肖 | 小人=0.81 | 人 時 有 表示 中 種 事物 對 時間
Aa02A03@ 老子
```

在表中亦將使用第四階釋義語義關聯值，以選出的最合適代表該詞組詞彙並以方框標示，從而可以比較在原始《詞林》編排方式下，同義詞群群首與最適合代表該詞組詞彙之間表達該辭群的義涵。如 Aa01A04=中，群首詞彙爲“*人手*”(有“*他人的手。辦事的人。*”兩釋義)與最適合代表詞彙“*人口*”(有“人。家族或家中的人數。人的嘴巴。指言語議論。一定時間內一地區具有戶籍身分的全部居民。”五釋義)；Aa01A07=中，群首詞彙爲“*者*”(有“*人或事物的代稱。指示形容詞。用於句中，表示停頓。用於句末，表示語氣結束。表比擬。*”六釋義)與最適合代表詞彙“*匠*”(有“*泛稱各種技術工人。尊稱在某方面有特殊造詣的人。技藝靈巧、構思巧妙。*”三釋義)。

5.5 與 Sketch Engine 比較

Sketch Engine (http://the.sketchengine.co.uk) (Kilgarriff *et al.*, 2004)是語料庫處理系統，主要的功能是使用 KWIC (key word in context) 出現頻率及分佈，結合中文語法關聯 (grammatical relations, gramrels) 分析，計算文字共同出現行爲的統計結果，以提供索引 (concordance)、詞彙列表 (word list)、詞彙速描 (word sketch)、同近義詞 (thesaurus) 等功能 (Huang *et al.*, 2005)。我們以“*招牌*”爲例，使用 zhTenTen 語料庫取得其同義詞，並比較本研究所使用的釋義語義關聯原則所表示的關聯程度，與使用 Sketch Engine 中語法模式及文字共同出現行爲所尋找的同義詞彙進行比較。選用的 zhTenTen 語料庫是由程式自動抓取網路上簡體字中文文本後，使用 Stanford Chinese Word Segmenter 及 Chinese Penn Treebank standard 模式所建立的 Stanford Log-linear Part-of-speech Tagger 原則處理的語料庫，目前約有 20 億個字，合計 17 億不重覆字在語料庫中。在 Sketch Engine 之中，我們以“*招牌*”詞彙查詢 Thesaurus 功能中的 Find Similar Words，“*招牌*”在 zhTenTen

之中出現詞頻爲 10185，且得到相似詞彙共 60 個。接著我們使用相同的維基百科的繁簡分歧詞表進行繁簡轉換，將取得的簡體詞彙轉換成繁體詞彙後，使用《國語辭典》進行釋義，並一一與“*招牌*”計算第四階層釋義語義關聯值，區別 Sketch Engine 所得到的相似詞彙與本研究中所指的釋義語義相關之間的差異(結果見下表 8)。

表 8. Sketch Engine 與釋義關聯計算結果比較 (部分刪除)

Sketch Lemma	Sketch Scores	Sketch Freq.	SRD-Scores	SRD- 共有辭典釋義字出現比率 Top 10
牌匾	0.21	5421	0.79	有 字 表示 題 前 單位 提 記錄 人 題目
廣告牌	0.15	4839	0	
標語	0.15	19510	0.43	宣傳 宣布 廣告 宣揚 說明 講解 傳達 大眾 公布 文字
橫幅	0.12	15051	0.35	繪畫 吊掛 懸掛 書法 字畫 筆 作品 文字 藝術 筆畫
喜歡	0.12	247564	0.3	高興 事情 決定 根據 歡喜 快樂 興致 興趣 愉悅 愉快
牌子	0.12	18442	0.79	調子 音 牌子 大調 說話 程度 高低 表示 音調 時
可口	0.12	6800	0.57	美 有 人 使 好 野 表示 變 事物 好看
條幅	0.11	5848	0.66	組 指 單位 組織 物 機關 人事 事物 人 中
喜愛	0.11	47276	0.08	愛好 喜歡 高興 喜好 喜愛 快樂 自愛 事情 事物 根據
櫥窗	0.10	7917	0.62	事物 物 者 人 媒介 指 中 一切 現象 各
名片	0.10	18497	0.3	電影 影片 人 膠片 名聲 供 活動 底版 響亮 人物
標牌	0.10	5443	0	
海報	0.10	15672	0.62	指 大家 人 一 稱爲 有 個 中 上 種
廣告	0.10	212258	0.67	事物 有 人 一 上 觀念 精神 表示 指 意識
門面	0.10	5216	0.65	人 體面 門面 個人 指 身分 有 面子 上 一
餐館	0.09	13639	0.73	人 供 者 有 受 說 別人 表示 他人 智慧
~~過長刪除~~~				
字號	0.07	13197	0.93	人 牌子 招牌 商店 名稱 標識 字號 獻 有 號碼
出名	0.07	10320	0.38	名 具名 出名 出面 人 單位 簽名 計算 稱號 署名
~~過長刪除~~~				
做	0.07	1650329	0.68	前 做 人 某 我 進行 事物 製造 自稱 詞
特色	0.07	566599	0.66	地方 事物 物 當地 人 某 客觀 一切 指 存在
熟悉	0.07	141431	0.22	知道 明白 詳細 道理 仔細 細節 瞭解 明曉 詳情 熟悉

對聯	0.07	6769	0.62	兩 單位 計算 一 人 個 量詞 物 等於 公斤
獨特	0.07	135938	0.07	獨有 占有 特殊 僅有 指 特別 意思 不同於 所有 占據
歡迎	0.07	196262	0.62	他 指 人 方面 個 綴 這 別的 第三人 中
亮點	0.07	64220	0	
促銷	0.07	39704	0.67	有 人 存 使 事物 表示 某 各 意識 令
精緻	0.07	23820	0.28	細密 東西 仔細 文明 精緻 周詳 小史 紅樓夢 精深 東
形象	0.07	274352	0.72	人 指 事物 有 個 中 實體 物 一 表示
吃	0.07	433702	0.65	說話 說 事物 樣子 短 唐 講 事情 今 一

在表 8 中，前三欄分別爲 Sketch Engine 所提供語法模式及文字共同出現行爲，找出,與“*招牌*”相似的還原字詞(Sketch Lemma)、分數(Sketch Score)與頻率(Sketch-Freq.)。後兩欄爲使用(第四階層)多階層釋義關聯計算的結果(SRD Score)與 SRD-共有辭典釋義字出現比率。表 8 是依 Sketch Score 由大而小排列，所以可以看到在 Sketch Score 相對較高的詞語中，詞彙行爲雖與「*招牌*」相似，但字義上與“*招牌*”無關的詞彙，如：“*喜歡*、*可口*”等。而 SRD Score 值較高的詞彙(粗體線外框)，則可看出與“*招牌*”義涵上有較高的同義，且共有的辭典釋義字亦羅列於後。如“*招牌*”與“*字號*”產生釋義關係詞彙較高權重值的 Top 10 是“*人牌子招牌商店名稱標識字號獻有號碼*”等，更清楚地知道兩詞彙是透過“*牌子*”、“*招牌*”等進行釋義關聯。

雖然前述兩種方法都能取得兩詞彙相似關聯參考指標，但使用辭典釋義字進行同義關聯探究的方法與 Sketch Engine 方法不同在於:(1) 兩者的相似關係所建立的原則不同：Sketch Engine 中的同義關聯的建立是透過詞彙在許多語料庫之中的出現頻率、文法結構樣式與造句行爲來決定，而本研究所使用的方式是以辭典爲基礎的共用釋義字詞觀點出發，同義詞彙則是建構在使用相同釋義字所佔的比率多寡而決定。但比較下，使用多階層釋義語義關聯計算原則下的詞彙相關性，相較於使用 Sketch Engine 所得到的結果，更能令人直覺理解期涵義。因爲計算過程中釋義不斷進行詞彙擴充，並計算兩者間的共有釋義詞彙，因此尋找出的詞彙也較 Sketch Engine 易於了解。(2) 相似詞表產生方式：Sketch Engine 透過文法模式與統計值產生相似詞表，可以依模式尋找符合的詞彙而產生列表。然而使用多階層語義關聯計算原則，因受階層變數決定所計算的兩詞彙之間概念深淺而定，且需經過詞彙間交互比較計算後才可得到結果，即需要將辭典中的所有詞條進行交互比較。若以目前實驗中所使用的《國語辭典》爲例，是在 15 萬條詞條、27 萬個不同釋義之中交互比較，且在階層計算原則下會產生指數增加的計算負荷(computing loading)，所以這確實一樣大工程。在本實驗中要計算 Dd15A09=第一到第四階層大約都在 1 秒之內、第五階層約在 1.24 分鐘而第六階層則約需要 5 分鐘。雖然釋義語義關聯計算方法不適合用在產生同義字表上，但在計算已知的同義詞或兩生詞之間的關聯，因爲能提供兩詞彙之間較能令人理解的釋義關聯，相較 Sketch Engine 方法是更爲合適的。

6. 結論與討論

在本文中，我們已經討論多階層釋義關聯在計算同義詞彙上的應用，也比較與現有的 Sketch Engine 中 Thesaurus 計算原則上的差別。在本研究中，爲了避免釋義詞彙在多階層的詮釋後而太過發散(無法收斂)，從而使用修正後的多階層釋義關聯計算方法。將較淺階層的釋義詞彙權重增加，以減少在深層釋義之中泛一般性的概念詞彙影響，以突顯兩詞彙之間的共同擁有釋義詞彙的特色，並建立關聯。關聯值的計算，是將兩詞彙間共同擁有的釋義字詞出現佔有比率，來表示共同釋義文字概念的交集，並使用反覆釋義的多階層原則，以減少釋義文字同義不同型的問題，同時利用階層釋義文字比率作爲釋義詞彙參與計算中的權重。此多階層統計中的共有釋義文字權重，可視爲解釋詞彙之間共同擁有的釋義內涵，作爲兩詞彙間關聯描述使用。

在完成多階層釋義關聯原則後定義，我們以處理詞彙間一詞多義的情況，使用多階層釋義關聯的最大值，以及共有釋義文字的涵蓋程度，來決定哪一組釋義內容適合作爲同義詞組間一詞多義代表。並且以此方式將《擴展版》中「相等、同義」詞類進行實驗，其結果雖然受限於《國語辭典》無法完全釋義《擴展版》中的各項詞彙，但我們將現有的資料進行標記，區別出《擴展版》中《國語辭典》所擁有的辭條，並將最適合詮釋同義詞組與共有釋義詞彙權重最高的 Top 20 筆資料，整理出 8311 筆合併於《擴展版》中，並放於網路上供研究者參考。

最後，爲更了解多階層釋使用在同義詞組的計算上有什麼區別，我們亦比較透過計算龐大語料庫 (如：Sketch Engine) 所取得的同義詞彙，與本研究方法的同義關聯之間的差別。雖然多階層釋義語義關聯方法無法如 Sketch Engine 進行大語料庫中詞語的計算後產生相似字表以供查詢，但可以作爲 Sketch Engine 計算取得結果之後，同義詞後的釋義比較。

多階層釋義關聯基於辭典釋義計算，雖然辭典內容的編寫用字會影響計算結果，但因中文釋義能提供詞彙語義中所包括的概念內容，從而使已知的兩詞彙間進行同義概念的探討時，釋義語義關聯相較於文法、詞頻與共現次數，較能取得更好的同義結果。雖然概念表達是由字/詞彙組合而成，而釋義內容所使用的文字雖不必然會與概念組成文字相同，但透過多階層的釋義比較下，亦能對多義詞進行同義歸納。辭典與《詞林》的編撰都是艱巨的文學語料工作，在過去的人工的相互訓詁工作之中，我們希望能應用電腦輔助語料工具方法，協助編撰者能對釋義內容進行整理、校對。亦希望利用人工釋義的內容，能與知識概念，如 HowNet 或 Chinese WordNet，進行交互比對，使釋義關聯計算能直接使用釋義概念進行比較，從而讓詞彙之間的同義關係可以更加清楚。

References

Huang, C.R., Kilgarriff, A., Wu, Y., Chiu, C.M., Smith, S., Rychly, P., Bai, M.H., & Chen, K.J. (2005). Chinese Sketch Engine and the Extraction of Grammatical Collocations, In

Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing, Jeju island, Korea, 48-55.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D.(2004). The Sketch Engine, *Information Technology Research Institute Technical Report*, ITRI-04-08.

Loper, E. & Bird, S. (2002). NLTK: The Natural Language Toolkit, In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics, 1*, 63-70.

Thesaurus Entry，https://trac.sketchengine.co.uk/wiki/SkE/Help/PageSpecificHelp/Thesaurus, last visited 2012/6/30.

Turney, P. (2001). Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL, In *Proceedings of the Twelfth European Conference on Machine Learning*, 491-502.

王建莉(2012)。論爾雅的同義詞詞典性質。*辭書研究*，(02)，60-65。

中研院斷詞系統，http://ckipsvr.iis.sinica.edu.tw/, last visited 2012/6/27.

全文奕，郭聖林(2012)。“淺談的士”及其同義詞群的競爭與選擇。*前沿*，02，153-154。

《同義詞詞林》擴展版，http://ir.hit.edu.cn/phpwebsite/index.php?module=pagemaster&PAGE_user_op=view_page&PAGE_id=162, last visited 2012/6/17.

周亞民，黃居仁(2005)。漢字意符知識結構的建立。*第六屆漢語詞彙語義學研討會論文集*。

范紅麗(2011)。《左傳》中跪拜義同義詞群考察，*西南科技大學學報(哲學社會科學版)*，*28*(5)，93-97。

林頌堅(2004)。基於術語抽取與術語叢集技術的主題。*Computational Linguistics and Chinese Language Processing*, *9*(1)，97-112。

重編國語辭典修訂本，http://dict.revised.moe.edu.tw/, last visited 2012/6/17.

陳光華、莊雅棻(2001)。應用於資訊檢索的中文同義詞之建構。*中國圖書館學會會報*，*67*，93-108。

梅家駒、竺一鳴、高蘊琦與殷鴻翔(1983)。編纂漢語類義詞典的嘗試-《同義詞詞林》簡介。*辭書研究*，*1983*(01)，133-138。

黃侃述、黃焯編(1983)。《文字聲韻訓詁筆記》，*上海古籍出版社*，1983 年 4 月版，190。

曾慧馨、劉昭麟、高照明與陳克健(2002)。以構詞與相似法爲本的中文動詞自動分類研究。*International Journal of Computational Linguistics and Chinese Language Processing*，*7*(1)，1-28。

趙逢毅與鍾曉芳(2011)。基於辭典詞彙釋義之多階層語義關聯程度計量-以「目」字部爲例。*中文計算語言學期刊*，*16*(3-4)，21-40。

維基百科，繁簡分歧詞表，http://zh.wikipedia.org/zh-hant/Wikipedia:繁简分歧词表，last visited 2012/6/27.

劉挺、車萬翔。中文語義處理，http://ir.hit.edu.cn/，last visited 2012/12/06.

鮑克怡(1983)。漢語類義詞典探索《同義詞詞林》編後。*辭書研究*，(02)，64-70。