語料庫為本的語義訊息抽取與辨析
以近義詞研究為例

蔡美智* 黃居仁* 陳克健**
*中研院史語所
**中研院資訊所

電子郵件：tsmei@hp.iis.sinica.edu.tw

摘要

本文以近義詞研究為例，說明如何利用語料庫進行語意抽取與辨析。傳統的研究方法在舉證過程中，因為缺少豐富語料作參考，難免遺漏一些有趣的語言現象，同時也無法反應語言事實。反觀語料庫研究，雖然能巨細靡遺地記錄語言現象，提供各種數據，然而往往會為繁複的語料所困，掌握不到現象背後的眞正導因。這裡我們配合語料庫進行語意方面的研究，一方面詳細觀察詞項之間的句法功能，計算其使用頻率；另一方面從各種差異中找出基本原由，證明詞彙的句法表現取決於自身的語意特性。

1. 緒論：「熱心」與「熱情」

一般而言，近義詞的定義並不明確，而被認定為近義詞的詞項之間彼此到底有什麼差別，也沒有一個嚴格的界定方法。

以 Teng (1994)主編的中文近義詞詞典為例，書中針對漢英翻譯需要，收錄一些英文譯文相同的中文詞，視為近義詞，然後利用各自不同的用法對這些詞加以區分。例如「熱心」和「熱情」這對詞雖然都表示對事物的全心投入，可是彼此的用法並不完全相同，在名詞、形容詞、副詞、動詞四種用法[1]當中，「熱心」沒有名詞功能，無法接受定語修飾(例1)；反之，「熱情」沒有動詞功能，不能後接賓語(例4)。至於兩者都適合的用法，只有形容詞和副詞兩種(例2-3)。

(1) 懷著滿腔(*熱心+熱情)去前方勞軍。

(2) 他十分(熱心+熱情)。

(3) 他總是(熱心+熱情)地幫助別人。

(4) 她一向(熱心+*熱情)慈善事業。

由下表，我們可以清楚看出這兩個詞項在用法上的異同。

(5)

用法＼詞項	熱心	熱情
名詞	-	+
形容詞	+	+
副詞	+	+
動詞	+	-

這種比較詞彙功能的方法，確實提供了一套辨識近義詞的具體標準，可是限於例句的局部性，觀察結果難免有所偏失，例如Teng (1994)根據例(6)，判定「熱情」比「熱心」適合接上定語標誌「的」修飾後面的名詞組，事實上我們發現，只要將句中內容稍作修改，「熱心」一樣可以擔任定語功能（例7）。

(6) (*熱心+熱情)的觀眾為精彩的表演熱烈鼓掌。

(7) 熱心的觀眾為這次的表演四處奔走。

其次，Teng (1994)作了一個有趣的觀察，那就是當介詞「對」引出的對象是人的時候，搭配的謂語以「熱情」為宜(例8a)；反之，如果引出的對象是物的話，像「朋友的事」，那麼「熱心」就較適合(例8b)。

(8) a. 他對人很(*熱心+熱情)。

b. 他一向對朋友的事都很(熱心+*熱情)。

不過，只要我們觀察更多的語料便可以發現，像「人」與「物」這樣的對立關係，不僅只存在於事件所涉及的對象上面，從句子的主語成份，也看得到類似的現象。下面三組例句顯示，「熱心」和「熱情」都可以自由搭配像「民眾、男人、各位貴賓」一類的屬人主語，可是只有「熱情」能搭配「氣氛、陽光、太陽」等非屬人主語。

(9) a. (熱心+熱情)民眾送兩箱草莓慰問。

b. (*熱心+熱情)的氣氛備受世界各地的遊客青睞。

(10) a. 男人不可太(熱心+熱情)。

b. 陽光不再像以前那麼(*熱心+熱情)。

(11) a. 感謝各位貴賓遠道而來，(熱心+熱情)參與。

b. 太陽還是衝破烏雲，(*熱心+熱情)的放出光和熱。

另一個問題是，這種針對幾個例句進行觀察比較的研究方法，可以得到像表(5)的結果，知道一個詞項具備哪些功能，不具備哪些功能，然而該詞項最重要的功能是什麼，以及各項功能間的使用比例為何，卻無從知曉。

諸如此類的問題，唯有配合語料庫的應用，才可以獲致圓滿的解答(參見黃居仁1995)。我們利用中央研究院詞庫小組所建立的兩百萬語料庫(黃居仁、陳克健等1995)進行搜尋，總計取得用例「熱心」71條，「熱情」77條，下表記載兩者各項功能的使用情形。

(12)

用法＼詞項	熱心 (71)	熱情 (77)
名物化	5 7.0%	39 50.6%
定語 ∅	7 9.9%	4 5.2%
的	7 9.9%	14 18.2%
狀語 ∅	20 28.2%	4 5.2%
地	10 14.1%	2 2.6%
謂語 ∅	14 19.7%	14 18.2%
賓	7 9.9%	
介	1 1.3%	

由表中提供的數據可以看出，這兩個詞項除了謂語及物用法[2]及名物化用法[3]有明顯差異之外，其他功能的分佈情形也頗有距離。儘管兩者都具備定語和狀語功能，但事實上「熱心」用作狀語的次數是定語用法的兩倍，而「熱情」恰恰相反，狀語用法僅及定語用法的三分之一。兩者在用法上唯一差不多的地方，只有謂語不及物功能一種。另外值得注意的是，「熱心」仍然有少數幾個名物化用例，並不像預期中的絕對禁止。

由此可見，語料庫應用能幫助我們進一步了解近義詞之間的關係。這些詞事實上只有在某些用法底下表現類似，整體而言仍是很不相同。接下來我們將利用語料庫提供的訊息，觀察另一對近義詞「高興」和「快樂」句法行為的異同，並探究其原因。

2. 語料庫應用實例

在兩百萬詞中央研究院平衡語料庫中，「高興」和「快樂」出現的次數分別為280與365，都是使用頻率相當高的動詞，由中文詞知識小組(1994)編制的新聞語料詞頻來看，兩者都是收錄的28326目詞中，前四千個最常用的詞彙。這對詞無論按 Vendler (1967)、Teng (1975)或 Smith (1991)提出的分類原則，都是狀態動詞，具有可以接受程度副詞「很」修飾的語法特徵(例13)。

(13) a. 她很快樂。

b. 她很高興。

2.1 「高興」不等於「快樂」

儘管語意和用法頗類似，兩者的句法表現並不盡相同，如例(14)所示，「高興」可以後接賓語子句，「快樂」就不行。

(14) 她很(高興+*快樂)張三來了。

所以依照中文詞知識庫小組(1993)採用的論元結構標準，這對詞分別列入兩個不同的類：「高興」屬狀態句賓述詞，而「快樂」屬狀態不及物述詞。

可是我們發現除了後接賓語子句這項功能以外，這對詞之間還有其它不同的用法，譬如例(15)的名物化功能，例(16)的定語功能，例(17)的結果補語功能，以及例(18)與動貌標誌「了」搭配，兩者的表現一概相反。

(15) 人有追求(*高興+快樂)的本能。

(16) 如何做個(*高興+快樂)的上班族。

(17) 他看得很(高興+*快樂)。

(18) 客人(高興+*快樂)了會賞你錢。

這些對比並非「狀態句賓」和「狀態不及物」兩個次類劃分所能預測的，因此我們決定先從語料庫中觀察這對詞的實際使用情形，再比較兩者各自特有的句法功能，從中抽離出導致差異的語意因素。

2.2 句法功能與範例

我們將使用平衡語料庫的多項配備，包括詞類標記、關鍵詞檢索、排序、過濾、詞類統計、列印等多種功能，比較這對詞的句法功能，包括謂語、補語、狀語、定語及名物化等用法。

首先，我們利用關鍵詞搜尋(KWIC)功能找出所有包含關鍵詞的例句，接著利用詞類標記統計各種功能的出現次數。例如標有<+NOM>的詞項具名物化功能，出現於<DE>後面的則具補語功能。其他僅由關鍵詞標記無法辨識的功能，如狀語和定語，則利用語料庫可預設視窗範圍的功能，以及詞類檢視、濾除等功能來判斷。各種功能的分佈情形簡要敘述如下：

2.2.1 謂語功能

在語料庫中，「高興」的標記為VK，擔任謂語的情形計有224次，是該詞數項功能中最重要的一項，使用率為百分之八十；標記為VH的「快樂」則有119次充當謂語，使用率約百分之三十。兩者用作謂語的時候，絕大部分都呈現不及物用法(例19)。可是「高興」卻獨自擁有及物用法，後面可以接上子句(例20)。

(19)　出爐了，大家吃得津津有味，高興又快樂。

(20)　我們很高興創刊號終於發行了。

再者，兩個詞後方都可以直接加上補語成份，如例(21)中的「萬分、無比」，或是形成由「得」字帶出的補語結構(例22)。

(21) a.　他眞是高興萬分。

b.　全家一定快樂無比。

(22) a.　我和妹妹都高興得大聲叫好。

b.　一聽到披薩，眞是快樂得不得了。

2.2.2 補語功能

「高興」和「快樂」也都可以出現在「得」字後面，發揮補語功能(例23)。不過，兩者的使用頻率都不高，一個是百分之三，另一個也只有百分之五。

(23) a. 他看得很高興。

b. 爸爸忙得很快樂。

2.2.3 狀語功能

這兩個動詞扮演狀語的次數都比扮演補語的情形來得多，總計「高興」有47條用例，「快樂」有30條，兩者的差別在於「高興」作狀語時一定帶有狀語標誌「的」或「地」(例24)，因為不帶標誌的話,就會被理解成謂語的句賓用法。至於「快樂」帶標誌也可以(例25a)，不帶標誌，直接出現在謂語前方進行修飾也可以(例25b)，帶標誌的情形大約是不帶標誌的兩倍。

(24) 小明高興的跳起來。

(25) a. 她快樂地叫出來。

b. 快樂過新年！

2.2.4 定語功能

在定語功能方面，「高興」的用例為零，「快樂」則有116次之多，使用率達百分之三十，和謂語一樣重要。「快樂」扮演定語功能的時候，如下則例句所示，定語標誌「的」可有可無，但是和「的」一起出現的頻率較高，約為不帶「的」的兩倍。

(26) a. 露出了快樂的笑容

b. 在一起的種種快樂情景

2.2.5 名物化功能

與上一項定語功能的情況類似，兩個動詞名物化的頻率相差十分懸殊，「高興」全部只有一個名物化的用例(例27)，出現率還不到百分之一，至於「快樂」名物化的情形則相當普遍，既可以擔任主語(例28a)，也可以擔任賓語(例28b-c)，出現率超過百分之二十。

(27) 不過卻帶有一些高興，因為終於可以回到自己的家了。

(28) a. 快樂在哪裡?

b. 人有追求快樂，逃避痛苦的本能。

c. 祕密組織以快樂和暴力為尚。

由以上的觀察，我們可以得到一個初步的認識，那就是「高興」和「快樂」這兩個詞雖然意義相近，而且都具有謂語、補語、狀語等用法，但是分佈比例卻大不相同。以下將語料庫研究所得的數據製成圖表：

(29)

功能＼詞項	高興 (280)	快樂 (365)
謂語	224 (80%)	119 (32%)
補語	8 (3%)	17 (5%)
狀語 地	47 (17%)	19 (5%)
狀語 ∅		11 (3%)
定語 的		77 (21%)
定語 ∅		39 (11%)
名物化	1 (0.3%)	83 (23%)

從表中可以清楚看出，除了能否後接子句之外，「高興」的諸多功能當中，以謂語的用法最具代表性，而定語和名物化的用例則微乎其微。相反地，「快樂」除了謂語以外，也經常有擔任定語或名物化的情形。這些句法上的差異，不是「狀態句賓」和「狀態不及物」兩個次類劃分所能預見的，因此我們將在下一節進一步探索，是什麼因素致使「高興」和「快樂」句法結構互異。

3. 詞彙語意特徵研究

我們再次將收集到的語料依照語言現象分門別類，諸如詞項扮演的句法功能，表述的事件動貌，搭配的主語屬性，及建構的句型，以便從中抽離出導致上述許多語法差異的語意特性。

3.1 狀態有無變化 <±change of state>

首先，我們利用句法功能分佈與事件動貌，證明「高興」和「快樂」表述不同的事件類型，前者隱含某種狀態變化，後者屬於沒有變化、均質的狀態。

3.1.1 句法功能

綜觀第二節的功能分佈介紹，「高興」和「快樂」在用法上有以下三大不同點：第一，「快樂」名物化的情形相當普遍，「高興」則有困難(例30)。第二，「快樂」可以扮演定語功能修飾名詞組，「高興」則不行(例31)。第三，「高興」後面可以接句子，「快樂」則無此用法(例32)。

(30) 人有追求(*高興+快樂)，逃避痛苦的本能。

(31) 如何做個(*高興+快樂)的上班族。

(32) 我們很(高興+*快樂)創刊號終於出來了。

這些對比顯示「高興」和「快樂」屬於不同的事件型態，「高興」的動作性較強，後面接的句子雖然不是眞正的受事賓語，但也是促使某種狀態發生的導因(例32)，「快樂」則名詞性較強，可以指涉認知世界中的某種特性(例30)，也可以次類劃分一個集群(例31)。因此就語意觀點而言，我們可以說「高興」含有狀態變化(change- of-state)的語意特性，「快樂」則是穩定、均質的狀態(homogeneous state)。

3.1.2 事件動貌

一旦弄清楚這兩個近義詞間不同的語意特性，下面的語法現象都可以得到合理的解釋。例(33)中「高興」因為詞彙語意本身即涉及狀態改變，所以能夠帶上

完成貌標誌「了」，並搭配表瞬間動作的時間子句「聽了」。反之，同樣的搭配完全不適合「快樂」這種穩定均質的狀態。

(33) a. 客人(高興+*快樂)了會賞你錢。

b. 父親聽了很(高興+*快樂)。

有關時間副詞的搭配方面，也出現類似的對比情形，表瞬間狀態的副詞「正」適合修飾「高興」，而表恆久狀態的副詞「永遠」則適合修飾「快樂」。

(34) a. 我們談得正(高興+*快樂)，突然…

b. 永遠(快樂+*高興)

在補語的搭配上，兩者都可以接表程度的補語，如例(35a)中的「不得了」，但只有「高興」容許後面接上由句子組成的結果補語，如例(35b)中的「大聲叫好」，表示某事件發生之後所產生的新事件。

(35) a. 一聽到披薩，眞是(快樂+高興)得不得了。

b. 我和妹妹都(高興+*快樂)得大聲叫好。

擔任狀語的時候，兩者和被修飾的謂語呈現不同的選擇限制，「高興」很適合搭配由動作動詞組成的謂語，如例(36a)中的「叫出來」，但遇到像例(36b)中動作性低的事件「成長」就不太自然。至於「快樂」則沒有這層限制，修飾的事件無論是動態或靜態都可以。

(36) a. 她(快樂+高興)地叫出來。

b. 全家人(快樂+*高興)地一起成長。

擔任補語的時候，兩者和前面的謂語同樣表現出不同的選擇限制，「高興」搭配狀態動詞或動作動詞都可以，分別表示狀態的程度和動作的結果，如(37a)中的「忙」和(37b)中的「看」。「快樂」則只適合搭配狀態動詞表程度，搭配動作動詞的話，詮釋起來會很奇怪。

(37) a. 爸爸忙得(很快樂+很高興)。

b. 他看得(很高興+*很快樂)。

(38) a. 玩/2、跳/1、吃/1、看/1、說/1、談/1、欣賞/1 + 得 + 高興

b. 過/8、活/6、玩/2、忙/1 + 得 + 快樂

換句話說，只有「高興」可以表示一個事件過後產生的新狀態。事實上，根據我們對實際語料的統計，如例(38)所示，「高興」當補語時搭配的謂語對象多是「跳、吃、看、說」一類的動作動詞，相反地，「快樂」搭配的對象則絕大部分是像「過、活」這種描寫持續狀態的動詞。

上述各種現象從不同的角度證明，「高興」和「快樂」之間的許多句法差異，都與兩者本身的事件型態有關，「高興」意味著狀態有所變化，而「快樂」則屬穩定均質的狀態，無涉任何變化。

3.2 意志能否控制 <±volition>

接下來我們從語料中透過主語屬性以及句型類別，歸納出另一項可以區別「高興」和「快樂」的語意特性，即其本身所表述的狀態能否由主觀意志控制。

3.2.1 有生主語

「高興」和「快樂」與句中主語也表現出不同的選擇限制，如例(39)所示，「高興」可以接受屬人代名詞「她們」擔任主語，但無法接受無生名詞「生活」當主語。至於「快樂」搭配兩種屬性的主語都沒有問題。

(39) a. 她們(快樂+高興)嗎？

b. 心胸開闊，生活(快樂+*高興)最重要。

3.2.2 句型

在搭配句型方面，兩者也呈現出以下兩點差異：第一，只有「高興」可以形成命令句(例40)；第二，只有「快樂」才適合祝願的句型(例41)。

(40) a. （高興+*快樂）一點！

b. 別(高興+*快樂)！

(41) a. 祝你(快樂+*高興)！

b. 媽媽過節(快樂+*高興)！

由於命令句是要求聽話者作出相關的反應，聽話者顯然具有執行能力，而祝願語

係純屬說話者的願望，無關聽話者的執行能力。因此，既然「高興」能與命令句相容，本身應該是個主觀意識可以控制的狀態。反之，「快樂」適合祝願語而排斥命令句，所以應為主觀意識所無法操控。

這項推論可以解釋前一小節有關主語屬性的觀察，「高興」之所以必須搭配屬人主語，便是因為那才具有控制的能力。「快樂」因為屬自然生成的狀態，不受外力左右，所以搭配的對象是有生主語也好，是無生主語也行。

下面幾個例句分別從不同的角度來支持這項論點。首先，兩個近義詞之間只有含意志控制特性的詞項「高興」才能搭配情態動詞「應該、要」(例42)。

(42) a. 你應該(高興+*快樂)。

b. 要我(高興+*快樂)就陪我打麻將！

其次，也只有「高興」這種心理狀態，當事人能夠清楚認知，進而以言行來表達(例43)。

(43) a. 打從心裡（高興+*快樂）。

b. 研究人員表示(高興+*快樂)。

再者，可以憑常理判斷，作出適當情緒反應的，同樣也只有「高興」（例44）。

(44) a. 陳菊出獄不值得(高興+*快樂)。

b. 為陳老師(高興+*快樂)。

最後，只有「快樂」這種心理的自然反應，意志所無法控制的情緒，才可能連當事人都沒有辦法掌握，進而發生自問或者弄錯的情形（例45）。

(45) a. 可是，我(快樂+*高興)嗎？

b. 他以為自己非常(快樂+*高興)。

根據上述的考量和比較，我們得以從語料中複雜的語法現象，歸納出兩項基本的詞彙語意特性，一為狀態有無發生變化<±change-of-state>，二為意志能否自由控制<±volition>，而兩者都與動詞本身所表達的事件類型有關。這一點印證了Pustejovsky 1991，Levin 1993 等所採用的詞彙-句法互動原則，並肯定了詞彙語意在句法中的主導地位。

4. 結論

以上我們利用語料庫的多項功能配備，成功地剖析出致使近義詞之間彼此句法行為迥異的詞彙語意特性。這個示範說明語料庫應用是當前語言學研究的利器，無論是語法現象的考察，語意因素的探索，或是語用效應的評估，我們都能夠賴以進行最詳盡的比對和觀察，並且由統計數據了解語言現象的全貌，這一點是傳統方法中檢驗有限幾條例句所無法探究的。

附註

[1] Teng (1994)顯然依照英文文法的觀點，將「熱心、熱情」當作形容詞，它們除了本身的用法以外，還兼具其他詞類的功能。我們則主張將詞類和句法功能的界限劃分清楚：一種詞類可以扮演多種句法功能，如狀態動詞「熱情」可以同時具有名物化、狀語、謂語等多種功能；同理，一種句法功能可以由不同的詞類來扮演，如「昨天觀眾一直熱情地鼓掌」一句中時間名詞「昨天」、副詞「一直」、動詞「熱情」等都可以勝任狀語功能。

[2] 即Teng (1994)所謂的動詞用法。

[3] 「熱情」名詞性用法相當突出，佔總出現率一半以上，在詞庫制定的詞類分析裡面便以多重詞類看待，詞類標記可能是Na名詞或是VH狀態不及物述詞。另外也有人主張將之視為名詞，因為就構詞率而言，「熱情」屬偏正結構複合詞，理應保有中心語「情」名詞的特性。可是反觀「熱心」，雖然中心語「心」也是名詞，但屬性卻毫無疑問是動詞。因此我們採用單一詞類觀點，將這個詞視為狀態不及物述詞，至於這三種作法哪一種比較妥當，這裡暫時不予討論。

參考書目

中文詞知識庫小組. 1993. *詞庫小組技術報告93-05 中文詞類分析.* 台北南港: 中央研究院.

中文詞知識庫小組. 1994. *詞庫小組技術報告94-01 中文書面語頻率詞點(新聞語料詞頻統計).* 台北南港: 中央研究院.

黃居仁 1995. 科技整合與整合科技—談計算語言學與語料庫語言學之角色與發展. 語言學門現況與發展研討會. 台北: 國立師範大學.

黃居仁, 陳克健, 張莉萍, 許蕙麗 1995. 中央研究院平衡語料庫簡介. *第八屆計算語言學研討會論文集.* pp. 81-99. 內壢: 元智工學院.

LEVIN, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation.* Chicago: University of Chicago Press.

PUSTEJOVSKY, J. 1991. The Generative Lexicon. *Computational Linguistics* 17.4.
SMITH, C. 1991. *The Parameter of Aspect.* Dordrecht: Kluwer.
TENG, S.-H. 1975. *A semantic study in transitivity relations in Chinese.* Ph.D dissertation. University of California at Berkeley.
TENG, S.-H. 1994. *Chinese Synonyms Usage Dictionary.* Taipei: Crane Publishing.
VENDLER, Z. 1967. *Linguistics in Philosophy.* Ithaca: Cornell University Press.