國語中的複合詞和語言剖析

陳克健(Keh-Jiann Chen)*,張麗麗(Li-li Chang)**,張莉萍(Li-ping Chang)**,謝清俊(Ching-chun Hsieh)*,**

* 中央研究院資訊科學研究所

** 中央研究院計算中心

論文摘要

本論文指出複合詞在電腦剖析國語時所產生的問題，並提出一套解決複合詞判定的模式，以語法律及有限的語意關係和詞典的設計相互配合，來判定複合詞的存在及其詞性並解決一些含混的情形。

電腦分析文句時所處理的單位是詞 (word) 而不是字 (character)。因此一般而言，中文的剖析分成兩個步驟，第一個步驟是斷詞，第二個步驟才是句子的分析 [陳 86, chen 86]。斷詞的工作是利用電腦中儲存的詞庫和輸入的句子做匹配 (matching)，找出句子中的詞及相關詞類，以利下一步的剖句 (parsing)，我們知道詞集是一個具有衍生性的開集合，我們無法窮舉所有的詞列入詞庫中。例如：1.重疊 2.數詞 3.複合詞。其中重疊和數詞的合成具規則性，因此可以由語法 (Grammar) 產生各種不同的組合[趙 80, 湯 80] 。但是複合詞的產生就比較複雜，根據[趙 80]有下列六種的產生方式：

(a)主謂式，例：頭疼
(b)並列式，例：書報
(c)主從式，例：牛肉
(d)動賓式，例：排隊
(e)動補式，例：打破
(f)複雜類，例：國民大會代表

§1. 複合詞影響句子分析的複雜性

從詞組律的方向來看，如果複合詞的組合律和語法規律用同等的地位來看，剖句的過程中複合詞的組合律和語法規律 (grammar rules) 就沒有區別。例如 N→NN 代表兩個名詞可以組合成一個複合名詞。可惜此種處理方式根本行不通。根據 [Huang 86]，見表一，以名詞、動詞、形容詞，任何兩種組合可以構成名詞或動詞或形容詞，也就是以下詞組律所表示的 X→XX，X：{N，V，A}。這樣的詞組律根本無法用來分析句子。因此我們必須研究更精確更有效的方式解決複合詞的問題。

表一、主要複合詞組合方式及例子

	N	V	A
NN	木工	根據	矛盾
NV	牙刷	規定	國有
NA	口紅	中飽	性急
VV	動作	分析	保守
VN	領事	鬧鬼	得意
VA	陳舊	挖苦	鎮靜
AN	空位	小心	狠心
AV	小說	公布	好看
AA	空白	歪曲	奇怪
φ φ	幽默	蹉跎	參差

[Huang 86] 列出了複合詞的分佈資料，以國語日報詞典為準的統計結果如表二，可以發現大多數為名詞性複合詞，其次為動詞，再其次才是形容詞。從詞的衍生性這個觀點來看，也以名詞性複合詞的結合方式最多也最複雜，動詞性的動補複合詞次之，但最常使用。形容詞類的複合詞除了動補結構，例如氣死，樂瘋 ... 等，幾乎很少出現［趙 80，第六章］。因此我們可以把重心放在這些具有衍生性的複合詞上面，至於其它形式的複合詞由於較不具衍生性，希望在詞典中可以窮舉盡列。即使無法完全列出時，我們在下一節也提出利用附著詞素 (bound morpheme) 的特性，找出複合詞的方法。

表二、複合詞的分佈圖表 [Huang 86]

category / structure	N	V	A	Total
NN	6910	21	90	7021
NV	306	434	72	812
NA	168	12	209	389
VV	276	3692	103	4071
VN	1559	2922	378	4859
VA	22	434	126	582
AN	2961	18	198	3177
AV	116	707	173	996
AA	163	38	1483	1684
(ϕ ϕ)	257	72	66	395
Total	12738	8350	2898	23986
	(12481)	(8278)	(2832)	(23591)

複合詞從意義上或語法分析上來看應該是一個單元。從表一我們了解複合詞的內部結構很複雜，要由電腦分析找到正確的解答十分困難，有時候必須依靠語意及常識 (real world knowledge) 才能解決。例如 (1) 的正確組合是 (2)。

(1) 國民 大會 代表 人民 行使 政權
N N N or V N V N

(2) 國民大會 代表 人民 行使 政權
N V N V N

但 "國民大會代表人民" 形成一個名詞複合詞為句子的主語，語法上也說得通。因此要得到正確的解答，嚴格說起來和常識有關。在沒有太多語意資料的情況下，也許電腦只能提供兩組可能的分析。下面也是一個類似的例子，(3) 可以分析為 (4)。

(3) 張三 打 破 碗。
N V A N

(4) a. 張三 打破 碗
N V N

b. 張三 打 破碗。
N V N

有些含混的情況可以由語法及動詞的格框及格位限制 (case frame and case restriction) 幫忙解決。有些可以用一些經驗法則 (heuristic rules) 來幫忙解決。雖然經驗法則不是絕對可靠，但不失為一個可行的辦法。因此論文內提出的解決方法雖然不是百分之百有效，但是在語意及常識缺乏的電腦系統下還是具有可行性。另外詞典建立的方法也和斷詞及尋找複合詞的成敗休戚相關，因此我們也提出了詞典建立的原則。

§2. 複合詞的結構及判定方式

中文的詞是由詞素 (morphemes) 組合而成

的，詞素可分為自由詞素 (free morpheme) 和附著詞素 (bound morpheme) [趙 80]。所謂自由詞素是可以獨立成詞的詞素，例如山，打，巧克力，(註：詞素不一定是單字)；附著詞素不能單獨成詞必須和其它的詞素結合成複合詞，例如：「第」一，「初」三，桃「子」，可能「性」，「公」「共」。而複合詞的組合成分可能有附著詞素、自由詞素或複合詞。自由詞素組合的複合詞比較難找，但是含有附著成分的複合詞就容易多了。因此句子當中如果斷詞後有附著詞素存在，就等於找到了複合詞。只是還不知道這個附著成分是和它左邊的成分或右邊的成分組合，我們可以用下面的原則來解決：

原則一：(a) 詞首字 [何 83] 和右邊成分結合。
例如：「拆」除，「爬」行

(b) 詞尾字 [何 83] 和左邊成分結合。
例如：物理「學」，經濟「化」

(c) 附著詞素和附著詞素優先結合 [張 86]。

(d) 單字和單字優先結合，但是獨用單字詞 [何 83] 是不和其它詞成分結合的。

詞首字及詞尾字在 [何 83] 附錄 E-1，2中列出，但本文所指的詞首字及詞尾字和 [何 83] 文中詞首字及詞尾字不盡相同。[何 83] 詞首字及詞尾字意義較狹窄，像例子中 "學","化"，皆非詞尾字，因為 "學業","化學" 中，"學" 及 "化" 皆在詞首，應非詞尾字。我們對詞首字及詞尾字的定義較廣：

定義：所謂詞首字 (詞尾字) 為一附著詞素，除了詞典中列出的詞以外，對包含此一詞素的複合詞它都出現在詞首 (詞尾) 的就叫詞首字 (詞尾字)。

因此一些衍生性很強的字，像 "學","化"，"家"，... 都可以列為詞尾字。只要我們在詞典中窮舉，由這些字組合而又非詞尾 (詞首) 的複合詞，例如"學者"，"學習"，" 學術"，"學報"，"學分"，"學生"，... 列入詞典後，"學" 就可以當做詞尾字看待。

詞素的自由或附著性資料正由中央研究院詞知識庫計畫整理進行中 [謝 87]。若詳細的分析，一個字可能代表好幾個詞素，因一個字可以有不同的意義，當意義不同時就代表不同的詞素。所以有些字具有不同意義時，其附著或自由的性質亦跟隨改變。例如："偷"，當動詞用時為自由詞素，當名詞用時為附著詞素。詞素的整理工作因此也相當複雜。

中文詞有趨向於多音節的傾向 [趙 80]，因此原則一(4)優先結合單字是有其根據的。[何 83] 附錄 E-2，列出了一百多個純單字詞，這些字是自由詞素，且不和其它詞素結合成複合詞。

具有附著詞素的複合詞比較容易掌握，但是自由詞素或複合詞間的組合就十分複雜。人可以很自然的將它們組合出來，因為它和語意的結合有非常密切的關係。由於語意結合的關係非常複雜，同時又涉及一般常識，無法在電腦中充分表示，因此我們只能依賴語法關係及非常少量的語意關係來解決這個問題。這種複合詞的解決相信和句子的分析是密不可分的，不太可能像具有附著詞素的複合詞可以獨立作業，先把所有的複合詞組織好，再來分析句子。所以由詞組律幫忙決定複合詞將是一個主要的解決方式。在這裏我們不討論句子的分析方法，我們把重點放在幾種最常見最具有衍生性的複合詞類型，加以各別分析

。其它不具衍生性的複合詞我們可以假設這些詞全部納入詞典中，沒有納入的，希望可以從附著詞素來解決。

從前述表一及［趙 80］第六章，我們了解名詞性的複合詞，結合的方式最多也最複雜。［Li and Thompson 84］動詞性的複合詞次之。其它詞類的複合詞幾乎沒有衍生性，可以不加考慮。

名詞性的複合詞具有衍生性的，都符合詞尾中心律 (head final)，細分為三類：

N1類：N N → N

例如：棒球手套，國民大會，...

N2類：V N → N

例如：檢查官，保管箱，戰鬥機，...

N3類：A N → N

例如：安全帽，高級品，...

動詞性的複合詞具有衍生性的類型，多為動補結構及動賓結構，可細分為下面三類：

V1類：V A → V （動補結構）

例：打破，抬高，哭壞，...

V2類：V V → V （動補結構）

例：拿來，提升，收回，...

V3類：V N → V （動賓結構）

例：開玩笑，開刀，加油，...

其它N V結構在表一中亦不在少數，但此結構之複合詞較不具有衍生性，可視為列得完的一類。

形容詞類只有A A → A一類，其意義上為動補形式，A1：A A → A。例如：

(5)高興極了，氣病，氣瘋，...

從以上的例子知道這些具有衍生性的複合詞是無法窮舉的。因此必須依靠電腦程式來判定是否有複合詞存在，而不能直接從詞典中匹配得到。至於其它類型的複合詞都應納入詞典，沒有納入者，希望可由原則一解決。

對以上具有衍生性的複合詞，我們再進一步的分析發現N3類及V3類和詞組律完全相符，因此不把它們當做複合詞也沒有關係。另外A1類的形容詞只能當作述語 (predicate)，不能用來修飾名詞，可以在句子分析時以動補結構解決，因此也不必考慮。如此一來，複合詞的問題已經簡化不少，是不是可以將N1，N2，V1及V2四條規律納入詞組律中，和語法分析用同一種方式解決？由於缺乏語意上的組合限制，這四條詞組律結合性依然太強，可以造成太多的含混情形。因此我們依然認為應該有一套不同的作業，在分析句子的過程當中，將複合詞研判出來。

從以上的討論，我們把一個極為複雜的問題簡化為，如何決定輸入句子中有 N1，N2，V1或V2類型的複合詞。決定複合詞的存在，附著詞素只能幫助解決部份的問題，大部分還是以自由詞素和自由詞素結合的情形較多。因此句子的分析及一些經驗法則，在缺乏語意關係的情況下，就成了主要的決定方法。首先用一個例子說明，假設輸入的句子經過斷詞後得到(6)。

(6)老師　教　學生　物理　化學
　　N　　V　N　　N　　N

從句子分析的過程當中知道"教"為雙賓動詞，當句中似乎有三個賓語"學生"，"物理"，"化學"，因此首先確定其中存在複合詞。第二步再決定"學生"，"物理"，"化學"，應為"學生物理""化學"還是"學生""物理化學"，從動

詞 "教" 的格框知道，第一個賓語是有生命的，因此 "學生" "物理化學" 的組合較為正確。

決定複合詞有三件事 1.決定複合詞存在 2.決定組合成分 3.結合後的複合詞的詞類。

從以上的例子知道如何決定複合詞和句子的剖析有密切的關係。我們必須依賴語法結構來判定是否應該結合其中的成分形成一個複合詞。因此我們必須假設：

假設一：輸入的句子符合文法。

這一個假設並不表示，不合文法的句子都可以剖析為合法的句子。這個假設只是說剖析句子的程式是以句子為正確的前題下，去考慮複合詞的問題。

至於如何決定複合詞的詞類，我們採用下面的假設。

假設二：除了動補形式的複合詞（包含V1，V2類），複合詞的詞類都以詞尾中心律決定。

也就是說複合詞的詞類是由組合個體中最後一個詞的詞類為其詞類。這一個假設應該可以實現，因為從表二可以發現如果除去動補形式的複合詞，具有詞尾中心特性的複合詞佔了百分之九十（包含具有衍生性的N1，N2，N3及A1類）。因此將剩餘百分之十不符詞尾中心律的複合詞，完全納入詞典，其它非N1，N2，N3，V1，V2，V3 及A1類而符合詞尾中心律的複合詞，大多可由附著詞素判定，組合結果可依照詞尾中心原則決定詞類。

詞尾中心律的引用是我們假設複合詞組成成分詞類沒有含混（ambiguous）的情形。事實上許多中文詞具有多重詞類，例如 "游泳" 是名詞也是動詞，"幽默" 是名詞也是形容詞。因此對於有含混的詞尾，我們是以詞組律來幫助決定詞類。如果詞組律無法幫助判定，這時以名詞優先。畢竟名詞性的複合詞佔了大多數。

原則二：詞組律可以用來判定複合詞存在與否及輔助解決具有含混詞尾的複合詞的詞類。

以下列舉一些可以幫助判別複合詞的詞組律：

詞組律：(a)定量詞＋名詞

例如：一趟　中　距離　游泳　→
　　定量　A　N　N or V

一趟　中距離游泳
定量　N

(b)介詞＋名詞片語

例如：用　旅行 ... →用　旅行 ...
　　介　N or V　　介　N

(c)形容詞＋("的")＋名詞

副詞＋"的"＋動詞或形容詞

(d)動詞＋"著","了","得","過", ...

(e)動詞的格框及格限

§2.1 名詞類複合詞的決定方式

由於語意關係及一般知識並未儲存在電腦中，因此名詞性複合詞的決定，除了依賴附著詞素詞尾字之外，只能由動詞格框及名詞片語結構的部份成分來判定，亦即詞組律所列出的一些語法。例如(7)是利用詞組律(e)的例子，(8)是利用詞組律(a)的例子，(9)是利用(b)的例子，(10)是利用(c)的例子。

(7)老師教學生物理化學。
(8)打破一個瓷瓶。
(9)利用大眾心理來...。
(10)破爛的公共汽車站。

§2.2 動詞類複合詞的決定方式

動補結構的複合詞有一個特徵可用來幫助判定。

原則三：動補結構的複合詞其成分多為單音節詞素。例如：打破，打爛，打輸，...。

對V1類的動補結構，所用的補語可以參考常見之補語表［趙 80］見附錄 1，只有少數幾個補語為雙音節。V2 類的補語除了趙元任所提的方向補語：來，去，上，下，進，出，起，回，過，開，攏，上來，下來，上去，下去，進來，出來，起來，回來，過來，過去，還有走，跑，升。例如：趕走，逃走，嚇跑，趕跑，提升，回升，... 。

A1類的補語後面通常加 "了"。例如：死，病，極，透，多。

呂叔湘［呂 86］和李臨定［李 80］對動補結構的句型有詳細的研究，研究詞組律時可以參考。

§3. 詞組及複合詞成分含混問題

因N1，N2，N3，V1，V2，V3複合而造成含混有下面三種情形。解決這幾種含混的情形，有時必須依賴語意，有些情形可以用語法來解決。

(a) V A N → V A + N　or　V + A N

例 "他打破碗"，為必然含混，兩種結構"打破碗" 及 "打破碗" 都說得通，因此必須有兩種不同的分析。但 "他打破一個碗" 必然是 "打破一個碗" 的結構，因為名詞修飾語的順序是 1. 表領屬的名詞或代詞，2. 數量詞，3. 形容詞，4. 表性質的名詞［朱 57］。

其它可以幫助判別的特徵為

(i) 動補結構中間可能帶有 marker "得"，"不"。

(11) a. 他打得破木板。

b. *他打得破木板。

(ii) 有些形容詞必須加"的"才能修飾名詞。

例如："清楚"。

(12) a. 他看清楚書本。

b. *他看清楚書本。

(b) V N1 N2 ： N1為N2的修飾語或為V的賓語之一

前面看過的例子 "他教學生物理化學" 可以解決，但是 "他教生物化學"，就有些困難，因為 "教" 可以是雙賓動詞，也有單賓的結構，這時候就必須依賴語意來判定，可用的語意資料就是動詞的格限，"教" 的間接賓語必須是有生命的，"生物" 是不是有生命的呢？這時候必須依靠常識來幫助判斷，可是目前電腦沒有這種能力。

(c) V V N → V + V N　或　V V + N

以下兩個例子，可以了解這個問題

(11) 法官　列席　偵察　犯人
　　　　V　　V　　N

(12) 大家　爭取　代表　資格
　　　　V　　N or V　　N

有一些法則可以幫助解決以上例子，例 1中 "列席"為及物動詞，且其賓語為動詞片語或地點若把 "偵察犯人" 結合為複合詞，則與動詞 "列席" 之格框不合。例 2中 "代表" 同時具名詞性及動詞性，如果一定要得到唯一的分析結果應優先選擇名詞複合。又例如 "法官列席偵察庭" 的例子，"庭" 為附著詞素，因此就沒有什困難了。

從以上的討論，我們了解動詞的格框及格位限制，在句子分析中扮演一個非常重要的角色，和複合詞的判定息息相關。至於詞及詞類特性也都是分析句子不可缺少的資訊，因此剖句的成敗和詞典的設計有密切的關係，下一節我們就討論詞典的架構及選詞原則。

§4. 詞典的架構及選詞原則

為了要滿足以上所討論的分析方法，電腦中必須儲存有一份詞典，這個詞典能夠提供必要的詞，及詞素和它們的屬性資料。在這裏我們不談詞典的資料結構，我們只列出選詞的原則及必要的屬性資料。

選詞原則：詞典中應包含

(a) 所有單字及詞素。大部分的詞素都是單字，只有少數詞素是多音節的，如：玻璃，葡萄。大部分的字都是詞素只有極少數非詞素，如："玻"，"葡"，"鸚"。一個字意義不同時，可能代表幾個不同的詞素。

(b) 意義上不具結合性或結合性不明顯的複合詞。例如：黑板，熱心 ...，這些複合詞可以視為詞素看待。

(c) 意義上具有結合性，但非詞尾中心的複合詞，但V1及V2類的複合詞除外。

屬性資料：

(a) 自由詞素及複合詞：

語法分類，若為動詞應加所支配格之語意限制，若為名詞應加基本語意分類 (semantic category)。

(b) 附著詞素：應標明為附著詞素，並給予名詞、動詞或形容詞的特性資料。例如 "盜"為附著詞素，具名詞性及動詞性的特性。"偷" 為附著詞素時，具名詞性 。為自由詞素時歸為動詞類。

§5. 結論

複合詞的問題以往都不太重視，直覺認為只要建立一個完整的詞典就不會有複合詞的問題，可是事實不然，許多複合詞類是列舉不完的。人處理一個語句，從來不會想到複合詞，從語意的組合過程很自然的組合了所謂的複合詞，人工智慧的研究目前對知識的表達及使用能力還很薄弱，因此對複合詞的處理無法模擬人的行為，以常識做為複合詞的判斷方法。本論文才提出一個比較可行的模式及法則，雖然這些法則不是百分之百的有效，相信可以解決絕大部分的問題。

附錄1　常見的補語表

[趙 80] pp.227-228

好	勻	散（去聲）	煳
壞	齊	散（上聲）	清
對	正	鬆	渾
錯	反	緊	生
早	翻	輕	熟（「煮熟」）
遲	掉	重	熟（不陌生）
晚	倒（去聲）	脆	紅
快（銳利）	倒（上聲）	佃	綠
慢	斜	硬	黃
久（像：「等久了」）	歪	軟	青
遠	方	結實	藍
近	圓	强	紫
長	扁	弱	粉
短	鈍	破	黑
高	空	斷	白
低	滿	碎	甜
矮	粗	爛	酸
大	細	乾	苦
小	厚	濕	辣
寬	薄	潮，溻濕	鹹
窄	光	熱	鮮
深	毛	冷	香
淺	透	涼（快）	臭
彎	穿	暖和	肥
直	通	凍	胖
平	塞	化	瘦

餓	少	明白	窮
飽	够	明	濶
渴	足	定	慣
疼	沒(有)	疘(復)	迷
撩撩	整齊	動	煩
麻	亂	活動	氣
痠	亮	成	瘋
新	暗	眞	賦
舊	黑	假	惡心
貴	乾淨	老	病
賤	髒	少	死
便宜	清楚	巧	活
多	糊塗	笨	

除此之外還有掉，住，醒，撐，垮，塌，雜

參考書目

1.朱德熙(1957) "定語和狀語",上海教育出版社

2.何容主編(1985) <國語日報辭典> 第十三版，台北：國語日報出版社

3.何文雄(1983) <中文斷詞的研究> 國立台灣工業技術學院工程技術研究所碩士論文。

4.呂叔湘(1986) "漢語句法的靈活性" 中國語文，1986 第一期 pp1-9

5.李臨定(1980) "動補格句式" 中國語文，1980 第二期 pp20-35

6.陳克健，陳正佳，林隆碁(1986) "中文語句分析的研究--斷詞興構詞" Tr-86-004 ，中央研究院資訊所。

7.張麗麗，張莉萍，黃瑞珠，鄭雅霞，魏文真(1986) <國語的詞類分析> 技術報告0002，中央研究院計算中心。

8.湯廷池(1980) "國語形容詞的重疊規律" ，師大學報二十七期，pp279-293

9.趙元任 著 丁邦新 譯(1980) <中國話的文法> 香港：中文大學出版社

10.謝清俊及詞庫小組(1987) 中文詞知識庫計畫

11.Chen, C.G.,K.J.Chen & L.S. Lee (1986) "A model for Lexical Analysis and Parsing of Chinese Sentences", Porceedings of 1986 ICCC, Singapore, PP33-44

12.Li and Thompson 著 黃宣範 譯(1984) <漢語語法> 台北：文鶴書局

13.Sandra Thompson (1973)"Resultative Verb Compounds in Mandarin Chinese" Language : 49:2 361-379

14.Shuanfan Huang (1986), "Chinese Morphology: anatomy of a double-headed language" in the Second International onference on Sinology.