信息處理用短語功能標記體系的設計暨語料庫標注實踐*

詹卫东
澳門大學社會科學及人文學院中文系
北京大学中文系汉语语言学研究中心
zwd1972@gmail.com

短語結構作為漢語的核心語法單位，其分類體系以及屬性特徵描述無疑是面向信息處理構建漢語語言知識庫的重要組成內容。從目前國內外中文信息處理的發展現狀來看，詞匯一級已經有較大規模可用的電子知識庫，短語一級則還缺乏通用的、大規模的資源；中文短語知識的形式化表示雖已經有一些探討，但相對詞匯而言，短語的分類體系以及短語結構的分析模式還存在不少爭議。

本文在結合前人有關漢語短語結構以及句型研究的基礎上，對漢語短語結構標記體系的制訂原則做了分析，提出了一套短語的功能分類和標記體系。並對基於這套標記體系進行中文語料的短語結構標注實踐做概要的介紹。

1 相關研究簡述

漢語短語和句子構造模式大體一致。語言學界對短語構造的認識，除了語法教科書中羅列的各種基本短語結構類型（如偏正、述賓、述補、主謂、連謂等）外，主要就體現在對漢語句型的描寫和歸納中。上世紀 80 年代以前，關於漢語句型的論述主要是各家語法著作對漢語句型所做的定性的分類，如《漢語知識》、《中學教學語法系統提要》、《現代漢語八百詞》、《現代漢語》（黃廖本、胡裕樹本）等；上世紀 90 年代，開始出現對漢語句型的詳細歸納和定量的研究。其中比較有代表性的是北京語言學院句型研究小組在《世界漢語教學》上發表的《現代漢語基本句型》和社科院語言研究所李臨定先生在《現代漢語句型》一書中對漢語單句句型的歸納（參見文後附錄 1）。前者歸納了漢語中 267 種單句句型；後者列舉了 711 種（數量的差異是因為分類描寫的顆粒度粗細不同）。在此基礎上，清華大學的計算語言學研究人員利用語料庫對北京語言學院的句型系統（選取了其中 209 個句型）進行了分佈統計研究。在定量考察後給出了漢語句型的頻度排序。上述這些句型系統研究有兩個顯著特點：（1）採用“扁平”線性序列的描述模式，不考慮句型的內部構造層次；（2）以句子功能成分（如主語、謂語、賓語等）作為主要的句型構造成分，同時也包含詞類成分（比如“動詞”）和一些特殊詞語（比如“是”“有”等）。也就是說，句型“標記”成分並不是勻質的。這一階段研究的應用目的主要是為了語言教學，對中文信息處理中的句型和短語分析處理問題的關注只是剛剛開始；

1990 年代中後期以後，為了幫助計算機對漢語句子的結構進行全自動的分析，人們開始探索對大規模實際語料進行句法結構標注，即在某種句法理論體系的支撐下，對實際語料中的漢語句子進行結構層次分析和結構類型標注。這種帶有句法結構信息標記的句子組成的語料庫通常稱為“樹庫”（Treebank）。這方面的研究以美國賓州大學中文樹庫為典型的代表。下面是一個標注了句法結構層次信息和短語類別信息的例子：

IP（ NP-SBJ（ -NONE-<*pro*>）VP（ v<讚揚> NP-OBJ（ n<僑胞> n<臺胞>）IP（ NP-SBJ（ -NONE-<*PRO*>）VP（ PP-PRP（ p<為> IP（ NP-SBJ（ -NONE-<*pro*>）VP（ v<支援> NP-OBJ（ DNP（ NP（ n<祖國>）dec<的>）NP（ n<社會主義> vn<建設>））））VP（ v<做出> as<了> NP-OBJ（ ADJP（ a<重要>）NP（ vn<貢獻>））））））））w。<。>）

* 本研究工作得到國家語言文字應用研究“十五”科研項目“信息處理用現代漢語短語及句型標記規範”（項目編號：YB105-49）以及教育部人文社科重大項目“大規模中文樹庫建設及其應用研究”（項目編號：06JJD740001）資助。

其中括弧表示了短語結構的層次和邊界，“IP”“NP-TPC”等表示了短語成分的類型和特定的句法語義性質。這種帶括弧和標記的一維線性句法結構表達形式（適合計算機識別）可以轉換為下面這樣的“樹圖”形式來表示（適合人理解）。

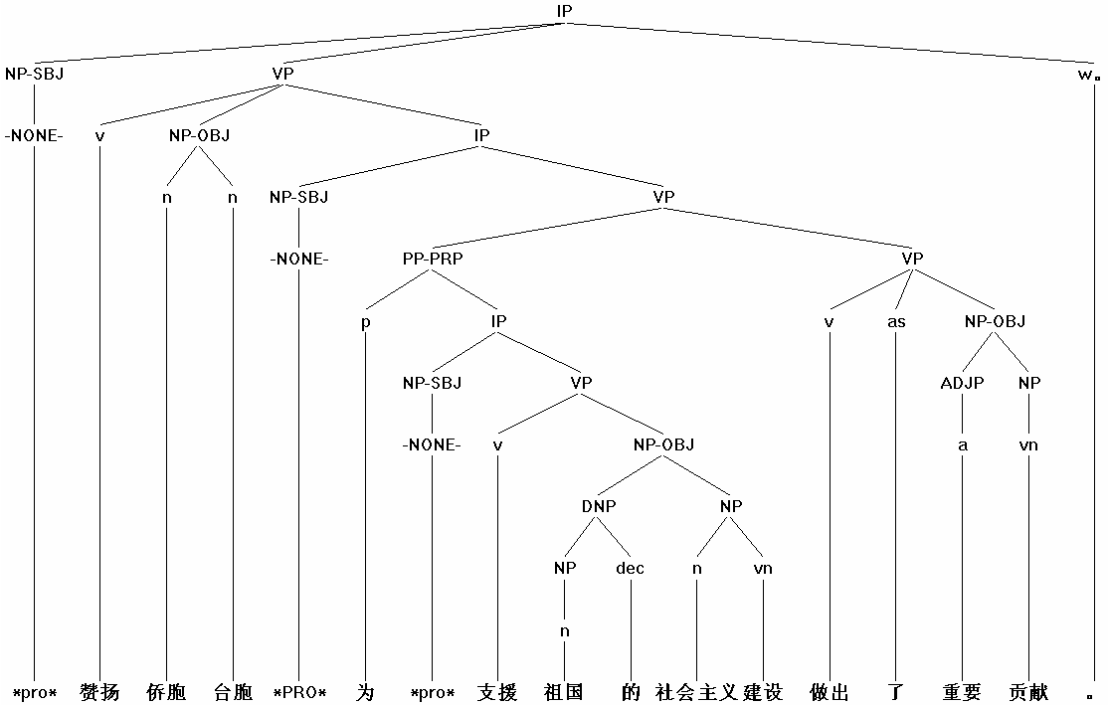


圖 1：賓州大學中文樹庫句法樹結構標注示例

賓州樹庫對漢語短語構造的分析主要特點是：（1）採用生成語法的管約理論（GB）為句法結構分析模式對漢語句子結構進行層級樹描寫（圖 1 中，*pro*，*PRO*等標記都是 GB 理論假設的深層結構成分標記）；（2）以詞和短語的分佈（功能）類為主要句法標記，來刻畫句子的組成類型，同時短語標記中也附加了句子結構成分標記（如 NP-SBJ 中 SBJ 表示該 NP 充當主語）。

這一時期國內北京大學和清華大學也開始了對漢語句子進行結構分析和標注的工作，並提出了初步的標記體系，主要是以結構主義語法理論為背景，參照漢語的詞類體系，對漢語的短語進行分類。

本文認為，現有的對漢語短語的分類還需要在理論上做進一步的思考，對於分類的原則還有深入分析的必要。只有在原則問題上有了更清晰的認識，在操作層面上，設計的短語分類體系才會更為合理。

2 信息處理用短語功能標記體系的設計

2.1 短語結構的分級與分類原則

對句子和短語進行結構分析，實質上是要把分析對象劃分為大大小小的語言單位。一般語法教科書上列出的基本語言單位包括句群（段落）、句子、短語、詞、語素。其中句群和語素一般來說都不是目前作信息處理所要處理的對象。而對於句子、短語、詞三級單位，其性質的差異可以用下表來說明。

表 1：語法單位的性質及其分級

層級	語法單位	說明
一級	整句	不被其他任何單位包含，只能包含第二級單位。

二級	短語	既可以包含第二級或第三級單位，又可以被第二級或第一級單位包含。
三級	詞	不包含其他任何單位，只能被第二級單位包含。

認識到這三級語法單位的性質差異，就不難推論：整句沒有分佈環境可言，因而不能做功能分類，只有結構分類，對於目前樹庫句法結構標注的需要來說，整句沒有進一步分類的必要。詞沒有內部結構，只能做功能分類。現有的詞類體系也正是功能分類的體系。短語是可以自我嵌套的語法單位，既有外部功能（分佈環境）特徵，又有內部結構特徵，因而可以有功能分類和結構分類兩種選擇。從面向計算機構造語法規則的角度講，首先應該選擇功能分類，其次可以將結構特徵作為短語的一項屬性信息來描述。

要對短語進行功能分類，跟詞類劃分的原則一樣，也應遵循功能（分佈）原則，即根據分佈上的顯著差異，將短語劃分為不同的類別。不過，由於短語的分佈情況遠遠比詞要複雜，對短語進行分類，在依據分佈特徵的同時，還需要參照短語的結構類型，以及構成成分，特別是中心詞的情況，來決定短語的類別。舉例來說，“有水準”可以受“很”的修飾，同時也不能帶賓語。類似於詞類劃分中以“很”和“帶賓語”兩項特徵作為鑒別標準，將能受“很”修飾，同時“不能帶賓語”的詞定為形容詞的做法，“有水準”就要劃歸為形容詞性短語。但是，“有水準”中的中心成分“有”是動詞，“有”和“水準”的關係是述賓結構關係，而形容詞不帶賓語，不能形成述賓結構，這樣，“有水準”就又不適合歸入形容詞短語，應歸入動詞性短語。為了綜合反映短語的上述複雜情況，就應該把分類視作一個基本的特徵，它只是描述了一個短語的基本信息，但並沒有反映全部屬性，在分類的基礎上，還需要設置多項特徵來刻畫一個短語的性質，比如對動詞性短語來說，需要描寫一個動詞性短語受程度副詞“很”修飾的情況，即雖然很多動詞性短語不能受“很”的修飾（如“* 很看書、*很喝醉”），但也有的動詞性短語可以受“很”的修飾（如“很有水準、很看得開”）。從這個角度說，“很”對於區分動詞性短語和形容詞性短語，不是一個具有分界性的區別特徵，但是是一項需要標記的分佈特徵。從以上所舉的簡單例子不難看出，漢語語法學界在給詞進行分類時，尚且很難找到嚴格的單一的具有鑒別力的分佈標準，而往往是代之以多項分佈標準共同來作為一個詞類的判別依據，在給短語進行分類時，選取什麼樣的分佈標準，就更需要多項特徵綜合考慮了。根據漢語語法特點的實際情況，本文提出短語功能分佈的基本原則如下：

（1）劃分短語類別，應根據短語的分佈差異；

（2）在單項分佈差異難以區分一類短語的情況下，可以多項分佈差異作為劃分依據；

（3）短語的功能分類要兼顧短語的結構類型，不應跟組成成分的詞類性質發生矛盾；

（4）短語的功能類別（分佈特徵）應跟短語中心成分的功能類別（分佈特徵）盡可能多地保持一致。

（5）短語的分類應該是有層級性的。換言之，分佈差異是相對的。分佈差異大的短語更可能歸入不同的短語類；分佈有差異，但差異小的短語，則可歸入同一個短語類，它們之間的差異不在短語類別層次上體現，可以在短語的語法或語義屬性特徵層次上體現。

下面的例子說明來短語分類的層級性（或相對性）。

功能類別	結構類別	分佈	
		數量 + _____	動詞 + _____
名詞性短語	定中結構	三雙皮涼鞋	住花園洋房
	聯合結構	*三雙皮鞋和涼鞋 *三雙皮鞋涼鞋	*住花園和洋房

例中“皮涼鞋”是定中型的名詞性短語，“皮鞋和涼鞋”是聯合型名詞性短語。它們雖都屬名詞性短語，但在“受數量成分修飾”這一分佈特徵上，卻表現出一定的差異。如果需

要，可以根據這項分佈差異，將這兩個短語分別歸入名詞性短語不同的下位次類。

2.2 現代漢語短語的功能分類及其標記體系

基於上一節對短語功能分類的原則性的認識，下面首先給出用來描述語言成分的分佈特徵的基礎框架，即漢語抽象的句法結構關係，以及句法結構關係所定義的句法位置（表 2），然後在這個框架中，給出區別不同類型短語的分佈標準（表 3）。確定一個具體短語的歸屬，主要就根據該短語是否可以有某項（些）分佈或沒有某項（些）分佈。比如，動詞性短語跟形容詞性短語具有一些共同的分佈，如作謂語、作狀中結構的中心語等等，但在一些分佈上有較大差異，如動詞性短語可以進入連謂結構的前項或後項位置，形容詞性短語則不佔據這些位置。動詞性短語（部分）可以出現在“所”後位置，形容詞性短語則不出現在這個位置。通過這些分佈特徵上的差異，就基本可以把動詞性短語跟形容詞性短語區別開。值得強調的是，表 2 給出的結構類型還可以繼續擴充。我們認為，跟詞類劃分具有相對性一樣，短語的劃分也是相對的。分多少類，分到多細的程度，取決於分佈框架的設置。此外，表 3 給出的分佈標準也具有相對性。應理解為典型範疇的分佈描述。在判別一個具體短語的功能類型時，還需要考慮該短語的結構類型和中心詞情況。表 4 列出了短語的功能類型與結構類型及其中心詞之間的基本對應關係。

表 2：漢語短語的結構類型

序號	結構類型	句法結構位置		實　例
1	主謂結構	主語	謂語	樹葉黃了;小明喜歡看電視;感冒傳染
2	述賓結構	述語 1	賓語	喝了三杯酒;學了三年;企圖逃跑;送他香煙
3	述補結構	述語 2	補語	洗乾淨;做得非常好;好得很;吃得完;拿出來
4	定中結構	定語	中心語 1	一斤白菜;老師的眼淚;大紅燈籠;削梨的刀
5	狀中結構	狀語	中心語 2	快跑;認真地學習;把飯吃完;明天見;屋裏坐
6	連謂結構	前項	後項	開著窗戶睡覺;打電話請醫生;派助手辦理;請他來
7	聯合結構	前項	後項	小說和戲劇;又高興又難過;批評教育
8	附加結構	中心語 3	附加語	紅著;吃了;砍光了;努力奮鬥過
9	的字結構	中心語 4	附加語	買菜的;老師表揚了的;冰涼的;慢性的
10	所字結構	附加語	中心語 5	所知道;所瞭解

表 3: 漢語短語的功能類型（sp 和 tp 是根據語義從 np 中分出的次類，並非嚴格的句法分類）

序號	標記	功能類名稱	典型功能														
			a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p
1.	ap	形容詞性短語		+			+	+	+		+	+		+	+	+		+
2.	dj	單句型短語		+		+										+		+
3.	dp	副詞性短語									+							
4.	fj	複句型短語		+		+										+		−
5.	mp	數詞短語														+		+
6.	np	名詞性短語	+			+			+	+				+		+		+
7.	pp	介詞性短語									+		+					
8.	qp	數量短語							+									+
9.	sp	處所詞性短語				+										+		+
10.	tp	時間詞性短語				+										+		+
11.	vp	動詞性短語		+	+		+	+				+	+	+	+	+	+	+

說明：a:作主語；b:作謂語；c:作述賓結構的述語；d:作賓語；e:作述補結構的述語；f:作補語；g:作定語；h:作定中結構的中心語；i:作狀語；j:作狀中結構的中心語；k:作連謂結構前後項；l:作聯合結構前後項；m:帶“著了過”等時體標記；n:“的”前位置；o:“所”後位置；p:中心語 3 位置。

表 4：短語功能類型、結構關係類型和中心詞之間的對應關係（詞性標記說明見附錄）

短語類	結構關係	中心詞
ap	ZZ LH SBU DC	a z b
dj	ZW ZZ FJ DC	v n q m z a ~p ~b
dp	FJ LH DC	d a v n
mp	DZ LH DC	m
np	DZ DE LH TW DC	n v a
pp	JB LH	p
qp	DZ LH DC	q
sp	DZ LH DC	s f n
tp	DZ LH DC	t f n
vp	SB SBU LW ZZ LH DC	v

SB：述賓關係；SBU：述補關係；LW：連謂關係；ZZ：狀中關係；LH：聯合關係；DZ：定中關係；DE：的字結構；LH：聯合關係；TW：同位關係；ZW：主謂關係；FJ：附加結構；JB：介賓關係。DC：單詞

表 3 中，複句型短語（fj）的分佈特徵只有從負面給出的“不能附加成分”特徵具有區別性。對比單句型短語（dj）來看，dj 可以附加語氣詞。而 fj 不能附加語氣詞，fj 中包含的語氣詞是附著在其中的組成成分上，而不是附著在 fj 整體上的。fj 的其他一些分佈特徵，如可以作謂語、賓語等，都跟 dj 相同，沒有區別意義。表 4 未列出 fj 的結構類型和中心詞，因為表 2 所給出的結構關係類型實質上沒有考慮 fj 的結構情況。由於 fj 的長度在短語中最長，已經接近整句的性質，目前還很難從結構類型上系統給出 fj 的分佈框架。這是有待進一步研究的課題。此外，由表 4 也可以看出，短語的結構關係跟短語的功能類型之間有明顯的對應，但又不是簡單的一一對應關係。比如“主謂關係”對應著單句，但反過來並不意味著單句一定是主謂關係，因為單句中也可以有“狀中關係”的（比如“在家裏他不停地鬧”就是狀語性成分“在家裏”加上單句型短語“他不停地鬧”形成的一個 dj）。多數短語的功能類型與其中心詞的功能類（詞類）是一致的，比如動詞性短語的中心成分一般也是動詞，名詞性短語的中心詞一般也是名詞。但並非所有的短語類型都如此，比如短語功能類型中的 dj，就沒有對應的詞類。dp 經常由某個詞類的詞附加“地”構成，因而其中心成分可以有除了副詞（d）之外的形容詞（a），動詞（v），甚至還有名詞（n）。

3 現代漢語短語結構標注語料庫的構建

基於上述短語分類體系，我們近年來以現代漢語真實語料為加工對象，進行了構建樹庫的實踐。從操作上講，加工樹庫的主要工作包括兩方面：（1）給短語定界，即劃分短語結構的層次；（2）給短語定性，即給一個具體的短語確定其短語功能類別。此外，還可以根據研究或應用目的的需要，標注短語的內部結構類型、語義角色、篇章功能等等。從資料的易維護性、易擴展性角度考慮，我們主張分層分級來標注上述信息，而不是像賓州樹庫那樣將這些信息都以樹節點標籤的形式來反映。事實上，像 LFG、HPSG 等當代句法結構理論中反映短語屬性也都是才用短語結構樹加複雜特徵集合的方式來描述短語的各類各層級的屬性。

針對真實語料的情況，在標注短語類別時，除上一節中的主要功能分類外，還需要增加一些標記。下表是目前所有標記體系中除主要短語外其他的一些短語標記。

表 5：樹庫加工用短語標記（“例子”欄中的對應例句部分以黑體突出顯示）

標記	說明	例子
zj	整句	**讚揚僑胞臺胞為支援祖國的社會主義建設做出了重要貢獻。**
yj	引句	**“我買的東西，為什麼讓給你？”**小芬不滿地說。
ypc	語篇插入性成分	平均每百戶居民擁有彩電 86 台**（比 1990 年增加 27 台）**
yph	語篇呼語性成分	**喂，喂，**你是誰？

上一節表 3 的 11 個短語功能標記加上這裏表 5 的 4 個標記，共 15 個標記，就是目前樹庫標記短語結構的全部標記。對於作中心成分的短語或詞，則在該標記前加西文嘆號!來標記。按照這一模式，上文賓州樹庫的例句在我們的樹庫中分析結果如下：

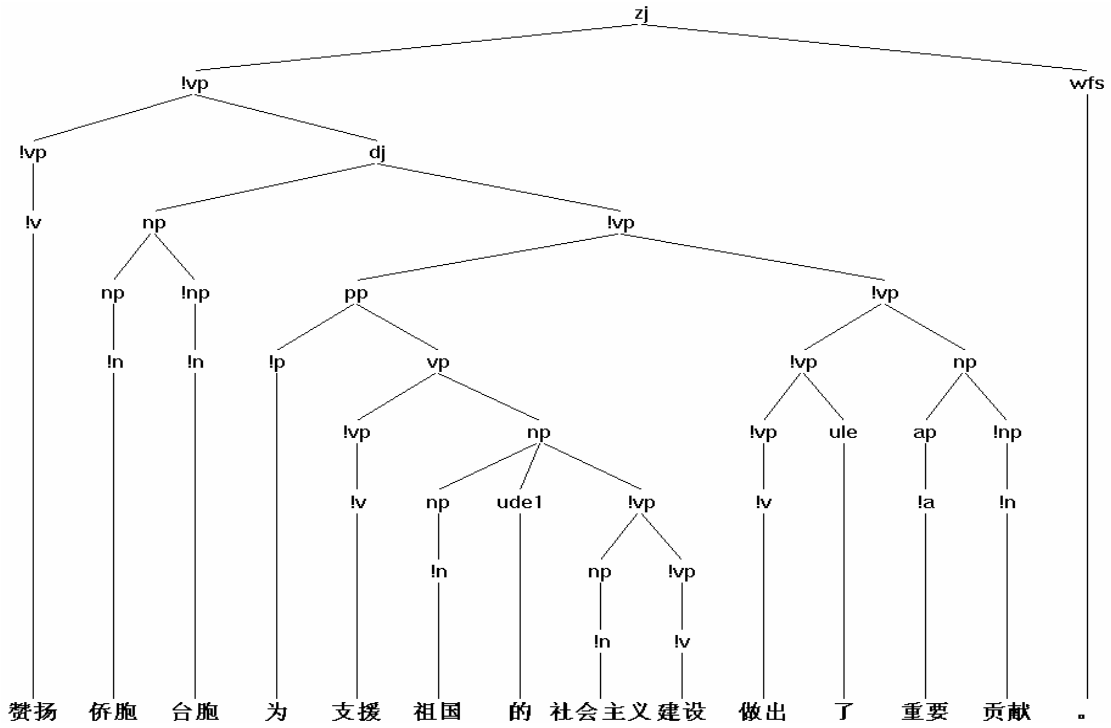


圖 2：北京大學中文樹庫句法結構標注示例

對比圖 2 與上文圖 1，不難看出，圖 2 所示的句法結構分析中不假設深層結構，因而也沒有深層結構成分的標記，短語節點標籤只體現短語的基本功能類別，不含句法成分、語義、語篇等信息（這些信息可以另外以特徵結構形式給出）。每個短語中都含有且僅含有一個中心成分。中心成分的功能信息與短語整體的功能信息保持一致關係。在很大程度上，漢語短語結構中中心成分的確定可以標記該短語的結構類型，比如對於一個 np 短語來說，中心成分在後，表明它是一個定中型短語，中心成分在前，表明它是一個聯合型短語。

目前已加工的樹庫語料分為四部分：（T-I）中國政府白皮書；（T-II）新聞報導類語料；（T-III）語文課本。（T-IV）機器翻譯系統評測語料。語料庫的規模如下：

名稱	字數 (type)	字數 (token)	詞數 (type)	詞數 (token)	句數	平均句長 (按詞數計)	短語	規則數
T-I	1553	51295	4917	35480	1268	27.981	26583	985
T-II	2415	151033	11763	93984	3553	26.452	69846	2778
T-III	1983	63499	5695	52202	4108	12.707	40887	1147
T-IV	2610	111494	9957	89794	10631	8.446	70223	1875
合計	3205	377321	22911	271460	19560	13.878	207539	4853

樹庫加工中首先利用分詞和詞性標注軟件進行處理，然後再由句法分析器進行自動分析得到初始樹庫，再借助輔助校對工具（TreeEditor）對機器分析的結果進行人工修改，經過質量檢查合格後，即得到正確標注句法結構信息的樹庫。TreeEditor 實現了可視化的樹庫編輯功能。包括（1）顯示直觀的樹狀圖；（2）可以在樹上拖動、刪除、合併節點、進行修改標籤等操作；（3）具有樹結構查找、替換功能，方便對樹庫中的結構進行批量修改。

從語言知識挖掘的角度說，基於樹庫可以提取很多有用的統計信息。比如：

（1）抽取帶頻度信息的短語規則、詞頻資料等。

從目前 27 萬詞的樹庫中可抽取 4853 條規則（不計中心成分），超過 5 次的規則 1180 條，只占 24.31%，超過 75%的規則出現次數在 5 次以下。由此可見大多數組合模式在真實語料中是低頻存在的（跟字頻、詞頻統計結果類似），這也可以說明，一般基於規則的系統，由於規則都是人工寫的，而人工規則往往集中在高頻部分，對那些出現頻度很低的組合模式，通常不大涉及到，這就造成一般人工規則對真實語料的覆蓋率低。對於樹庫標注來說，低頻的規則組合模式有時候意味著人工標注的錯誤，因此自動抽取規則後檢查低頻規則是否合乎語感，也是檢查樹庫質量的一種方法。

（2）按照樹結構、寬度、深度等作為查詢條件來提取短語結構。

比如可以提取樹庫中所有的介詞結構，考察一個介詞結構可以覆蓋的詞數。下表是從 T-I 樹庫中抽取的介詞短語 pp 的覆蓋詞數分佈情況：

跨度	2	3	4	5	6	7	8	9	10	11	12	13	14	15	…	45
頻次	231	189	135	102	88	76	54	37	35	30	18	16	15	13	…	1

T-I 樹庫中共有 1093 個 pp 短語，最多覆蓋 45 個詞。覆蓋詞數少於 8 個的 pp 占 80%。

（3）考察短語的功能分佈。

除了考察一個短語的內部構成情況，還可以考察短語的外部環境（分佈）情況，比如下表就是從 T-I 樹庫中抽取的介詞短語 pp 的分佈情況表（頻次排在前 10 位的 pp 的分佈）

序號	短語類	父節點	左鄰	右鄰	頻次	頻率	累計頻率
1	pp	vp	##	!vp	580	0.530650	0.530650
2	pp	dj	##	wco !dj	136	0.124428	0.655078
3	pp	fj	##	wco !fj	91	0.083257	0.738335
4	pp	vp	##	wco !vp	86	0.078683	0.817017
5	pp	np	##	ude1 !np	57	0.052150	0.869167
6	pp	ap	##	!ap	31	0.028362	0.897530
7	pp	vp	!vp	##	22	0.020128	0.917658
8	pp	np	##	ude1 !vp	17	0.015554	0.933211
9	pp	dj	##	!dj	9	0.008234	0.941446
10	pp	dj	##	wco !vp	7	0.006404	0.947850

上表中，##表示空標記。pp 最常見的分佈位置是在 vp 前作狀語（排在第 1 位）。在實際語料中，如果 pp 修飾的是 dj，fj 成分，則 pp 跟中心語之間有逗號隔開的情況比較常見（上表中排在第 2、3 位的 pp，可對比排在第 9 位的 pp）。

從樹庫中抽取的上述資料，無論是對構造句法分析器的統計模型，還是對漢語教學特別是漢語句型教學，都可以提供直接的參考。

4 結語

短語結構的分類是一項基礎研究。將理論上的分類結果應用於樹庫加工實踐時，會碰到更多具體的問題。比如短語的定界（層次分析）問題，是否對所有的短語都堅持二分支的分析原則？如何處理非連續結構（如“他老發我的脾氣”）？等等，都還需要做深入地考察和

討論。此外，短語的功能分類在理論層面是靜態的分類，在樹庫加工的操作層面也會碰到跟詞有兼類以及活用等類似的問題，即具體語境中標注時該處理為兼類，還是像堅持“詞有定類”那樣，堅持“短語有定類”。比如“三個人”靜態地看是數量成分加名詞形成的名詞性短語，但在“三個人睡一個房間”中，“三個人”是否標記為 np？還是像有的學者主張的那樣，分析為“數目短語”，就值得討論。再比如，“結婚三天了”是分析為 vp（看作述賓結構或述補結構），還是分析為 dj（看作主謂結構）。這些都需要結合語言學理論分析和真實語料的資料來尋找合理的處理辦法。總的來說，語言學界在理論層面對漢語短語的許多具體類型都做了分析和討論，如何將這些認識結合到面向中心信息處理的短語結構分析中，還需要做很多工作。

參考文獻：

[1] 北京語言學院句型研究小組,1989，現代漢語基本句型，載《世界漢語教學》1989 年第 1 期- 1991 年第 1 期（分 5 次刊出）

[2] 范繼淹,1986《範繼淹語言學論文集》，語文出版社。

[3] 李臨定,1986《現代漢語句型》，商務印書館 1986 年版。

[4] 李豔惠、陸丙甫，2002，數目短語，《中國語文》2002 年第 4 期。

[5] 羅振聲、鄭碧霞,1994，漢語句型自動分析與分佈統計演算法與策略的研究，載《中文信息學報》1994 年第 2 期。

[6] 邵敬敏,1985，漢語句型研究述評，載《語文導報》1985 年第 4 期。

[7] 石定栩，2006，動詞後數量短語的句法地位，《漢語學報》2006 年第 1 期。

[8] 吳蔚天、羅建林,1994《漢語計算語言學》，電子工業出版社 1994 年版。

[9] 俞士汶等，2003，《現代漢語語法信息詞典詳解》（第二版），清華大學出版社 2003 年。

[10] 詹衛東,2000《面向中文信息處理的現代漢語短語結構規則研究》，清華大學出版社 2000 年版。

[11] 趙淑華,1991《談 80 年代和 90 年代的漢語句型研究》，載《語言教學與研究》1991 年第 4 期。

[12] 周強，1997，《漢語短語的自動劃分和標注》，《中文信息學報》1997 年第 1 期。

[13] 周強，1997，《漢語樹庫的構建》，《中文信息學報》1997 年第 4 期。

[14] 周強、俞士汶，1996，《漢語短語標注標記集的確定》，《中文信息學報》1996 年第 4 期。

[15] 周強、詹衛東、任海波,2001《構建大規模的漢語語塊庫》，載黃昌甯、張普主編（2001）《自然語言理解與機器翻譯》，清華大學出版社 2001 年版，pp102-107。

[16] Mitchell P. Marcus, et al, 1993, Building a large annotated corpus of English: the Penn Treebank, Computational Linguistics, Vol.19.

[17] Robert D. Borsley, 1996, *Modern Phrase Structure Grammar*, No. 11 in Blackwell textbooks in Linguistics, Blackwell Publishers Inc..

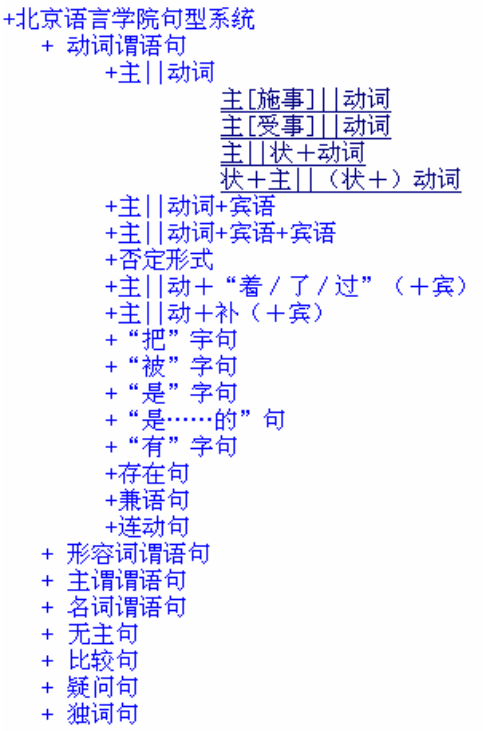
[18] Sag, Ivan A. & Thomas Wasow, 1999, *Syntactic Theory: A Formal Introduction*, CSLI Publications, Stanford, California.

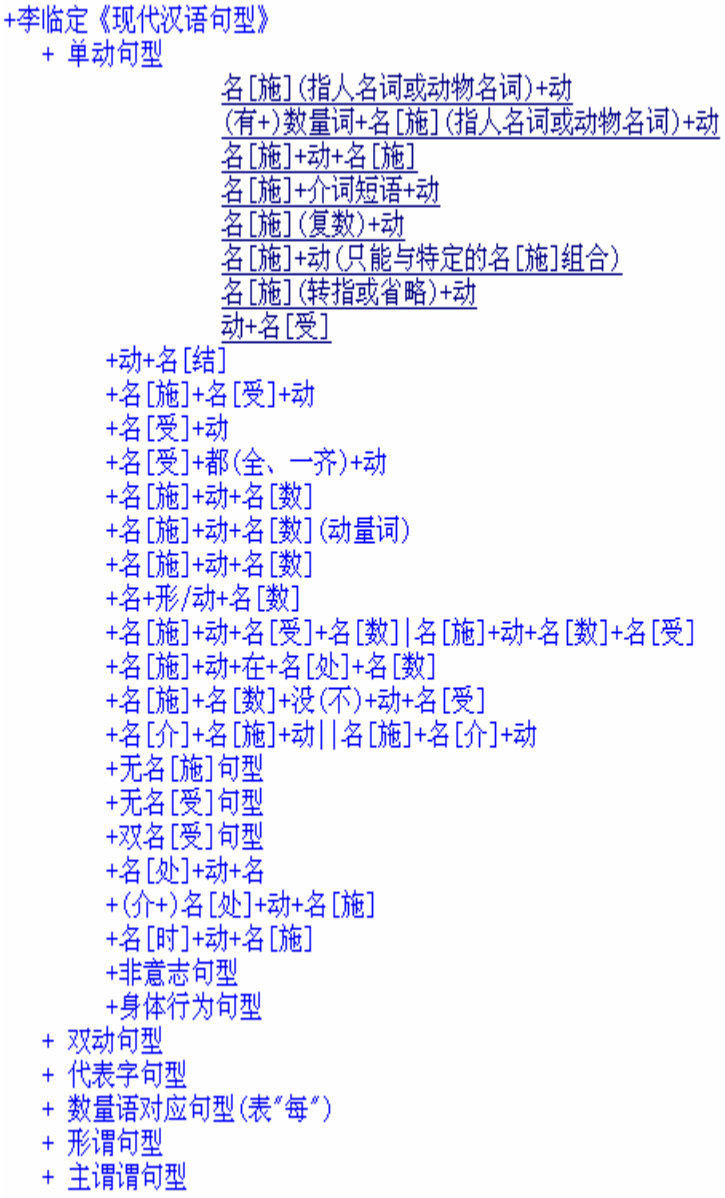
[19] Xue,Nianwen & Xia, Fei, 2000, The Bracketing Guidelinges for the Penn Chinese Treebank (3.0)

[20] http://www.cis.upenn.edu/~treebank/ （美國賓州大學樹庫網址）

[21] http://ccl.pku.edu.cn:8080/webtreebank （北京大學中文樹庫網址）

附錄 1 北京語言學院句型系統和社科院李臨定先生的句型系統





兩個系統都採用了層級分類的模式。北京語言學院的第一層分類為 8 類，最下層的小類為 267 類；李臨定《現代漢語句型》的第一層分類為 6 類，最下層的小類為 711 類。限於篇幅，作為示例，這裏只展開了第一個大類的最下位次類。

附錄 2：詞類及其標記

a	形容詞	h	前綴	o	擬聲詞	v	動詞
b	區別詞	i	成語	p	介詞	w	標點符號
c	連詞	j	縮略語	q	量詞	x	其他成分
d	副詞	k	后綴	r	代詞	y	語氣詞
e	嘆詞	l	慣用語	s	處所詞	z	狀態詞
f	方位詞	m	數詞	t	時間詞		
g	語素	n	名詞	u	助詞		