基於大規模中文樹庫的漢語句法知識獲取研究*

基於已經標注好的百萬字級漢語樹庫，可以抽取不同層面上的句法知識，爲中文信息處理、漢語句法結構的深入研究、以及對外漢語語法教學提供參考信息。對此，本文主要分爲三個層次從點到面地展開探討。一是考察具體的一類句法結構的內部結構特徵和分布特徵；二是對具有共性的一大類結構（以非中心擴展結構和非同類成分並列結構爲例）的考察；三是對全部句法規則集中右部同型左部根節點不同的潛在歧義結構的考察。

關鍵詞：中文樹庫，句法結構，歧義，非中心擴展結構，非同類並列結構，分布

1 引言：樹庫加工與利用概述

自 1990 年代以來，樹庫加工及應用在語料庫語言學和自然語言處理領域一直是受到相當重視的研究方向（Marcus 1993, Abeillé 2003, Xue 等，2000，周强 2004，Huang 等 2000）。除用於信息處理技術外，從樹庫中獲取句法知識，也可以爲句法本體研究以及語言教學提供參考。本文介紹近年來北京大學中文系樹庫研究小組在這方面所做的一些工作。北京大學現代漢語樹庫加工采用人機結合的方式：先由程序對原始語料進行斷句、分詞、詞性標注、句法結構分析等處理（這些環節都會碰到不少如何界定及規範的問題，限於篇幅，本文對此不展開討論），然後由人在樹圖編輯軟件環境中逐句進行檢查，修改程序自動分析的錯誤，得到最終的標注了正確語法信息的樹庫。流程可簡要示意如下：

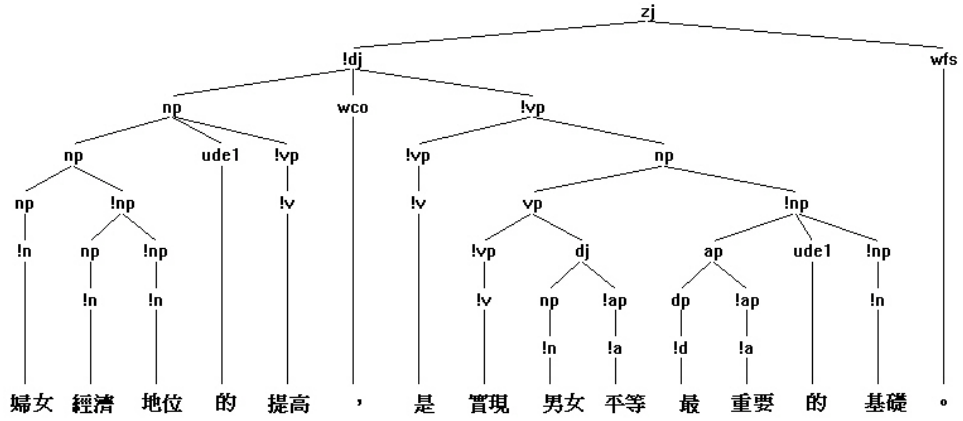
* 本文的研究工作得到霍英東基金項目“大規模中文樹庫構建及其在對外漢語教學中的應用”（項目號：111098）和國家社科基金項目“語言知識資源的可視化技術研究”（項目號：12BYY061）資助。匿名審稿人對本文初稿提出了中肯的修改意見。在此一併致謝。

流程	語料示例
（1）原始語料	三、在經濟領域中的平等權利與重要作用 婦女經濟地位的提高，是實現男女平等最重要的基礎。中國政府爲改善和提高婦女的經濟地位作出了卓有成效的努力。……
（2）斷句處理	1. 三、在經濟領域中的平等權利與重要作用 2. 婦女經濟地位的提高，是實現男女平等最重要的基礎。 3. 中國政府爲改善和提高婦女的經濟地位作出了卓有成效的努力。……
（3）分詞和詞性標注	1. …… 2. 婦女/n 經濟/n 地位/n 的/ude1 提高/v ，/wco 是/v 實現/v 男女/n 平等/a 最/d 重要/a 的/ude1 基礎/n 。/wfs 3. ……
（4）句法結構標注	1. …… 2. (zj (!dj (np (np (np (!n (婦女)) !np (np (!n (經濟)) !np (!n (地位)))) ude1 (的) !vp (!v (提高))) wco (，) !vp (!vp (!v (是)) np (vp (!vp (!v (實現)) dj (np (!n (男女)) !ap (!a (平等)))) !np (ap (dp (!d (最)) !ap (!a (重要))) ude1 (的) !np (!n (基礎)))))) wfs (。))) 3. ……

句法結構標注的形式是在計算機中以括號方式標記[1]在原始句子字符串上

[1] 具體的短語類標記和詞類標記可訪問 http://ccl.pku.edu.cn/doubtfire/Projects/Treebank_Tags.pdf 查詢。有關樹庫部份查詢功能的示例可訪問 http://ccl.pku.edu.cn:8080/WebTreebank/

進行存儲的。上面流程中加工完成的語料示例第 2 句對應的直觀的樹結構圖如下：



（語料來源：中國政府白皮書·1994·《中國婦女的狀況》）

〈圖 1〉句法結構樹示例

目前北大中文樹庫已經標注的語料規模爲 55,742 句，1,309,719 字，899,365 詞。語料類型包括語文課本（56.85%）、句型例句（13.29%）、新聞語料（13.00%）、科技語料（9.43%）、政府白皮書（7.43%）。

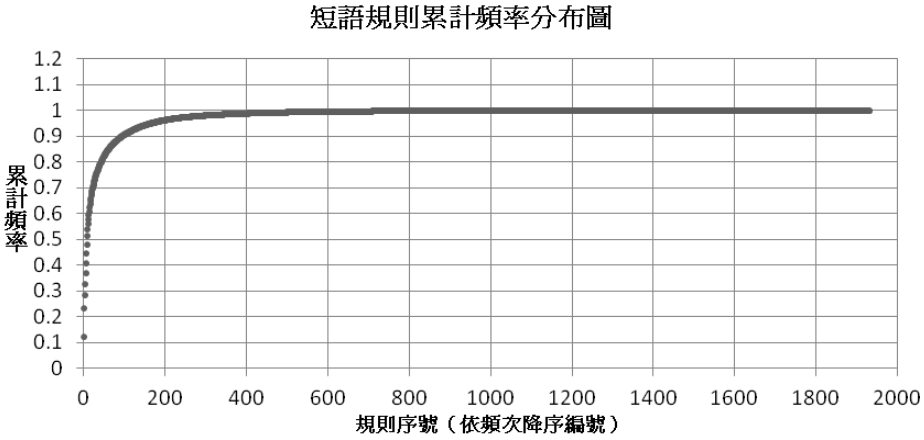
從樹庫中可以很容易抽取短語規則。比如圖 1 中“最重要的基礎”對應的規則是 np → ap ude1 !np。其中“最重要”的對應的規則是 ap → dp !ap。規則中的“！”標記了其後的成份是該短語結構規則的中心成份。把樹庫中像這樣的規則全部抽取出來，按頻次降序排列，就可以得到如下表所示的現代漢語短語結構規則集：

〈表 1〉現代漢語短語結構規則及其頻次示例

編號	結構規則	結構類型	頻次
1	np -> !n	名詞結構	159778
2	vp -> !v	動詞結構	146250
3	vp -> !vp np	述賓式動詞結構	65751

4	dj -> np !vp	主謂式小句	58724
……	……	……	……
1929	zj -> !dp wfs	獨詞型整句	1
1930	zj -> dj !fj	複句型整句	1

按規則在樹庫中的頻率降序排列，計算規則的累計頻率，可以瞭解短語規則對語料的覆蓋率情況，下圖是短語規則累計頻率分布圖，按累計頻率降序排列後，前 95 條規則（占 4.9%）覆蓋了樹庫中 90%的語料，前 446 條規則（占 23.1%）覆蓋了樹庫中 99%的語料。剩下的 1484 條規則（占 76.9%）覆蓋剩下的 1%的語料[2]。



[2] 我們也統計了美國賓州中文樹庫 5.0 版語料（Xue 等 2000, 2005）的情況。樹庫的規模是 43,385 個詞型（type），508,385 個詞例（token）。規則種數（type）爲 5220 條，例數（token）爲 537,504 條。其中前 116 條規則（占 2.24%）覆蓋 90%的語料，前 1379 條規則（占 26.42%）覆蓋 99%的語料。剩下的 3841 條規則（占 73.58%）覆蓋剩下 1%的語料。賓州樹庫的句法結構標注中很多是多分支樹結構，因此規則數目比較多。不過，儘管規則體系差別較大，但按規則百分比來看規則對語料的覆蓋率，兩個樹庫語料統計反映的情況是大致相當的。

〈圖 2〉短語規則累計頻率分布圖

上面這樣的統計數據，可以爲語言教學提供宏觀層面的參考。比如對句法結構可以像對詞的常用性分級一樣，區分不同的層級，安排更爲合理的教學順序。結構規則中近 80%的低頻規則，有不少涉及到省略、轉指等相對複雜的語言現象[3]，在語法教學的項目安排中可能就需要考慮更有針對性的策略。

以上扼要介紹了構建樹庫的流程，以及從樹庫中獲取短語結構規則知識。從樹庫中可以獲取的句法結構知識是多維立體的，除短語結構規則及其頻次外，還包括抽取帶詞性和頻度信息的詞表，兼類詞的分布統計，短語分布環境及其頻次，歧義短語結構及其頻次等等。在這些數據基礎上，還可以就研究者感興趣的問題，做專項信息提取和歸納。限於篇幅，本文分三個層次按照從點到面的順序介紹我們從樹庫中獲取不同類型的句法結構知識的情況。下文第 2 節是考察某一類短語在不同句法位置的差異（以名詞性短語 np 爲例）以及特定句法結構的內部構造特點和外部環境特點（以“把”字結構爲例）；第 3 節是考察漢語中違反中心擴展規約和並列條件的短語結構的情況；第 4 節是定量考察短語結構歧義的情況。

2 基於樹庫的漢語短語結構個案考察

2.1 np 在主、賓、定、中等句法位置的差異考察

名詞性短語 np 在不同句法結構位置對應的結構規則如下（xp 代表任意短語）：

[3] 比如“看你把閨女嚇得那個樣子”，“那個樣子”是 np，但出現在了補語位置上，違反了一般的句法結構規則要求。再如“他在他父親的公司裏一直呆到他父親去世”。“他父親去世”是一個小句（dj），這裏做“呆到”的賓語，實際上是轉指一個時間，起到了相當於 tp（時間詞性短語）的作用。下文還有一些出現頻次爲 1 的短語結構例子。這些例子相對於高頻的結構規則來說，某種程度上可以看做是“特例”。儘管母語者對這些組合例子一般來說是感到“習以爲常”的，但從系統的角度來講，它們屬於應該特殊對待的研究和教學對象。

主語位置　　dj → **np** !xp　(規則中 xp 通常爲 vp，ap，np，dj 等)

賓語位置　　vp → !vp **np**（動賓）　　pp → !p **np**（介賓）

定語位置　　np → **np** !xp（定語 1）　　np → **np** ude1 xp（定語 2）

中心語位置　np → xp !**np**（中心語 1）　np → xp ude1 !**np**（中心語 2）

上面規則中加粗標記了 np 所在的相應的句法位置。樹庫中 np 在"主、賓、定、中"句法位置出現的總頻次爲 325,747，占全部 np（392,925）的 82.9%。np 在這些不同句法位置的寬度（按詞數計寬度）統計結果見表 2。

〈表 2〉np 在不同句法位置的寬度統計

句法位置	主語	介賓	動賓	定語 1	定語 2	中心語 1	中心語 2
實例數	68279	15247	65751	42291	15372	84491	34316
寬度種數	37	40	60	46	29	36	26
最小寬度	1	1	1	1	1	1	1
最大寬度	54	60	113	112	54	63	53
均值	1.96	2.85	3.23	1.52	1.83	1.69	1.64
方差	3.93	9.65	12.18	4.22	3.74	2.97	2.17

說明：我們把定中結構分爲兩種情況，甲："定-中"；乙："定–的–中"。定語 1 和中心語 1 爲甲類中的"定"和"中"位置。定語 2 和中心語 2 爲乙類中的"定"和"中"位置。

如果把結構內包含詞數作爲評價結構複雜性的一個指標的話，np 在定語和中心語位置上的複雜性差異不是很大。np 充當甲類定中結構的定語和中心語的頻次顯著多於乙類定中結構的定語和中心語。原因是乙類定中結構的定語和中心語都可以由非 np 類短語（如 vp，ap）充當，而甲類定中結構的定語和中心語則主要由 np 充當。

np 在主語位置和賓語位置的複雜度差異較明顯，主語位置上的 np 平均詞長不超過 2 個詞，顯著低於介賓和動賓位置上的 np。此外，主語位置上 np 寬

度的方差值也顯著低於介賓和動賓位置上的 np，說明主語位置上 np 比賓語位置上 np 的寬度相對更集中。

下面進一步考察 np 在主、賓位置上的內部結構是否存在顯著差異。

〈表 3〉主、賓語位置上的 np 的內部結構及其頻次

語結構規則	主語位置			介賓位置			動後賓語位置		
	序號	頻次	頻率	序號	頻次	頻率	序號	頻次	頻率
np -> !rn	**1**	**23906**	**35.01%**	**2**	**2436**	**15.98%**	**5**	**4788**	**7.28%**
np -> !n	2	15561	22.79%	1	4218	27.66%	1	18852	28.67%
np -> np !np	3	7638	11.19%	3	2227	14.61%	3	7188	10.93%
np -> np ude1 !np	4	3449	5.05%	4	1180	7.7%	4	5313	8.08%
np -> qp !np	**5**	**2630**	**3.85%**	**5**	**775**	**5.08%**	**2**	**8313**	**12.64%**
…	…	…	…	…	…	…	…	…	…
總計	type 數：172 token 數：68279			type 數：128 token 數：15247			type 數：224 token 數：65751		

按頻次降序對不同的結構排序後得到上表的統計結果，從高頻結構的情況可以看到，主語位置的 np 跟介賓位置的 np 性質更爲接近，動後賓語位置上的 np 跟二者相差較大。這在一定程度上印證了以往漢語研究中所觀察到的現象，即舊信息傾向居動詞前位置，新信息傾向居動詞後位置。表 3 中“np → qp !np”規則對應的主要是漢語中的一般的“數+量+名”結構，這種結構的 np 在主語和介賓位置的出現頻率都排在第五位，而在動詞後賓語位置則居第二位。這一點，通過表 4 統計的數據可以更清楚地看出。表 4 是詞長爲 3 的 np 在主語、動賓、介賓位置上的內部組成情況。根據表 2 的數據，主語、動賓、介賓三個位置的 np 詞長均值比較接近 3，因此我們重點統計了詞長爲 3 的 np 在這三個位置的內部構成情況。

〈表 4〉詞長爲 3 的 np 在主語、動賓、介賓位置上的內部組成情況及其頻次

詞結構規則	主語位置			介賓位置			動後賓語位置		
	序號	頻次	頻率	序號	頻次	頻率	序號	頻次	頻率
np -> m q n	**3**	**760**	**10.05%**	**4**	**180**	**8.26%**	**1**	**2771**	**23.83%**
np -> rn ude1 n	1	1091	14.43%	1	260	11.93%	2	1169	10.05%
np -> n ude1 n	4	730	9.66%	3	217	9.96%	3	853	7.33%
np -> rb q n	**2**	**977**	**12.92%**	**2**	**221**	**10.14%**	**4**	**689**	**5.92%**
np -> a ude1 n	7	161	2.13%	5	117	5.37%	5	666	5.73%

結構規則“np → m q n”是“數 + 量 + 名”組合，“np → rb q n”是“指示詞 + 量 + 名”組合。前者一般對應語義上的不定指成分，後者則對應定指成分。在動後賓語位置上，不定指性 np 遠多於定指性 np。而在動前的主語位置和介賓位置，情況則顛倒過來，不過，在動前位置，兩類 np 的數量差異沒有在動後賓語位置相差得那麼大，這主要有兩方面的原因，一是漢語允許“無定 np 主語”（魏紅、儲澤祥，2007），二是形式上的無定 np，在語義上也可以表達定指義或者類指義，如“一個人毀壞了別人的東西，應不應該賠償？”中的“一個人”是無定形式的 np，用於主語位置，語義上並不是表達非定指，而是表達類指。總的來說，從樹庫中獲得的句法數據跟以往漢語研究中從語用角度所觀察到的現象：漢語中舊信息傾向居動詞前位置（主語位置 np 和介賓位置 np 都在謂語動詞前），新信息傾向居動詞後位置（LaPolla 1995）是非常吻合的。

2.2 vp 在“把”字結構中的內部構造以及“把”字結構整體分布環境考察

漢語語法學中傳統上關於“把”字結構，即“把 xp vp”中的 vp 的認識主要是它由複雜動詞詞組充任，不能僅僅是動詞的簡單形式。這樣才能滿足整個結構表達“處置”或“致使”語法意義的需要（比如北京大學中文系現代漢語教研室編寫的《現代漢語》教材“虛詞”章在介紹“把”字結構時就是這樣說明的）。下面我

們通過從樹庫中抽取“把”後 vp 實例以及 vp 內部結構規則的方式，進一步來考察這個結構中的 vp 具有哪些結構上的具體特點。

〈表 5〉“把+xp+vp”中 vp 的寬度考察

寬度	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**
頻次	864	551	463	235	155	85	58	28	17	13
寬度	**11**	**1**	**13**	**14**	**15**	**17**	**16**	**19**	**22**	**42**
頻次	12	12	8	4	3	2	1	1	1	1

表 5 對“把”字結構中 vp 的寬度進行統計的結果顯示，“把”後 vp 主要由複雜動詞詞組構成，平均寬度（詞長）爲 3.71 個詞，多於全部介詞性短語（pp）後的 vp 的平均詞長（3.50 個詞）。這個統計結果佐證了以往人們對“把”後 vp 結構要求具有一定複雜性的語感描寫。同時，表 5 的統計結果也顯示，“把”後 vp 也有單個動詞（詞長爲 1）的情況，北大樹庫中有 12 個這樣的例子，具體的動詞是“公開、分解、抽象化、形式化、平分、消滅、發揚光大、神化、相加、除外、置之度外、還原”。

在考察了“把”後 vp 的寬度之後，下面表 6 給出了“把”後 vp 的具體規則分類情況。

〈表 6〉“把+xp+vp”結構中 vp 的構造類型及示例

構造類型	數量	結構規則	示例
述賓式 vp	999 (39.74%)	vp → !vp np vp → !vp sp vp → !vp qp vp → !vp mp ……	(把 x) 交給 新幹部 放在 桌子上 放在 第一位 砍去 一半 …
述補式 vp	895 (35.60%)	vp → !v v	(把 x) 扔 掉

		vp → !v a vp → !v ude3 ap ….	清理 乾淨 布置 得 非常漂亮 …
狀中式 vp	354 (14.08%)	vp → dp !vp vp → pp !vp vp → ap !vp …	(把 x) 也 拋出來 在幾個工作人員中 分配一下 直接 倒到喉嚨裏去 …
附加式 vp	187 (7.44%)	vp → !vp ule vp → !v uzhe …	(把 x) 摔壞 了 珍藏 著
連謂式 vp	58 (2.31%)	vp → !vp vp vp → !vp wco vp …	(把 x) 帶回家 放好 變成電信號，再加以放大 …
其他	21 (0.84%)	vp → !v vp → c !vp …	(把 x) 公開 一 剝 …
合計	2514 (100%)	48 種	

從表 6 可以看出，“把”後的 vp 以述賓式構造類型爲最多，這個特點以往語法學中討論“把”字句時注意得不够。在討論漢語“把”字句的文獻中，有不少是以“把”後的所謂“保留賓語”爲題展開研究的，即從某種程度上來說認爲“把”後 vp 再帶賓語是一種特殊現象。儘管“保留賓語”確實有其自身的特點，但語料調查的結果也顯示，“把”後 vp 主要的結構類型就是述賓結構（包括帶準賓語的情況）。“把”字結構後出現賓語是該結構的用法特點之一。

對一個結構，除考察其內部構成外，還可以看它所處的上下文環境的特點。表 7 列出了“把+xp+vp”結構的分布環境的類型。本文關於“分布環境”的定義是：一個結構體（S）的分布環境是一個三元組。設樹（T）的根節點是 S 的父

節點，則 S 的分布環境由 S 的父節點、S 的左鄰節點、右鄰節點三個項目構成。其中左鄰節點和右鄰節點都可以為空。

〈表 7〉“把+xp+vp”的分布環境的類型統計（按照父節點的類型不同分組）

父節點	數量	左鄰節點	右鄰節點	示例
vp	1504 (59.86%)	dp vp vp wco - …	- - - vp …	連忙 **把它拾起來** 走過去 **把口琴還給錫海** 爬上樹去，**把小鳥放回窩裏** **把門打開** 放狗出去 …
dj	676 (26.89%)	np np wco …	- - …	你 **把它吃了** 古代的埃及人和中國人，**把它用做藥物** …
fj	237 (9.43%)	dj wco …	- …	他一隻手抓住繩子，**把另一隻手伸給水中的孩子**。 …
zj	36 (1.43%)	- …	wfs …	**把瓶子放在桌上** 。 …
np	29 (1.15%)	- - …	ude1 !np ude1 …	**把人生溶進偉大事業** 的 人 **把咖啡喝光** 的 …
# [4]	20 (0.80%)	-	-	**把桌子拿出去**
tp	7 (0.28%)	- …	f …	**把羊肉和羊骨粉碎** 後 …

[4] # 表示父節點爲空，這裏意味著“把”字結構獨立成句，占據一行，且末尾沒有標點。

pp	4 (0.16%)	P …	- …	從 **把水放在爐上** 到水開 …
合計	2514 (100%)	82 種		

表 7 反映了“把”字結構的主要用法中直接做謂語是排在第二位的。排在第一位的是“把”字結構跟其他成分組合成更大的 vp，占到近 60%，頻率是第二位的兩倍多。也就是說，現實中的“把”字結構，其前後往往會有其他的謂詞性成分共現。這個特點，在有關“把”字句的對外漢語教學中應引起注意。當以“把”字句爲視點去看“把”字結構時，往往容易把“把”字結構 vp 直接放在謂語位置上，同時把整個“把”字句跟被字句、主動賓句式放在一個層次上關聯起來，但如果以短語結構的視點去看“把”字結構，會更加全面地看到該結構所在的不同句法位置以及頻率上的差異。

3 現代漢語非中心擴展結構與非同類並列結構考察

通常情況下，一個短語結構的功能類跟其中心成分的功能類是相同的，這樣的短語規則是所謂的符合“（中心）擴展條件”的規則（記作 HE 規則）。兩個成分構成並列結構，則兩個成分應屬同類短語，這樣形成的並列結構是所謂符合“並列條件”的規則（記作 CC 規則）。當代形式語法理論一般也都强調短語結構規則從形式上應該符合 HE 規則和 CC 規則的要求。沈家煊(2007)引用 Lyons（1968: 331）的論述：“N 和 NP 之間，V 和 VP 之間都存在一種必不可少的（essential）的聯繫，對哪種語言都一樣。..... NP 和 VP 不僅僅是幫助記憶的符號，而是分別表示句法成分 NP 必定是名詞性的，VP 必定是動詞性的，因爲兩者分別以 N 和 V 作爲其必需的主要成分。”他接著說，如果有哪位語言學家提出諸如“NP→V+VP，NP→V，VP→T（冠詞）+N”的規則，“那不僅是有悖常情的，在理論上也是站不住的。”這些話是就“擴展條件”而言的，但是也適用於“並列條件”，提出有“NP 和 VP”這樣的並列結構也是有悖常情的，理論上站不住的。

但從樹庫標注的情況來看，我們認爲，實際語料中也有少量的短語結構，

其功能類跟中心成分的功能類是不同的，同時也有少量的並列結構，並列的兩項屬於不同功能類的短語，至少在表層結構形式上是如此。這樣的結構規則我們分別稱爲非中心擴展規則（記作 NHE 規則）和非同類並列規則（記作 NCC 規則）。下面是樹庫中抽取的 HE 規則和 NHE 規則、CC 規則和 NCC 規則各自所占的比例情況。

〈表 8〉樹庫中 HE 規則、NHE 規則、CC 規則、NCC 規則的數量統計

規則類別	規則種數（type）	規則例數（token）	結構示例	示例
全部規則	1,930	1,318,488	dj np !ap	人 多
HE 規則	1672(86.63%)	1,048,669 (97.20%)	ap dp !ap	最 冷
NHE 規則	258(13.37%)	**30,252 (2.80%)**	np sp !vp	體內 分布
CC 規則	61(52.59%)	26,220 (96.37%)	ap !ap c ap	光榮 而 艱巨
NCC 規則	55(47.41%)	**987 (3.63%)**	ap !ap c vp	無知 與 疏忽

說明：上面表中 NHE 及 NCC 規則的統計數據是程序根據規則形式自動判別的，數據會有一定誤差。不過，我們的目的並不是統計出精確的數據做量化分析，而是通過這種方式從實際語料中發現 NHE 規則和 NCC 規則的類型和實例。很顯然，從實例頻次（token）的對比來說，NHE 規則和 NCC 規則相對於 HE 規則和 CC 規則（常規情況）來說，都是絕對少數。換言之，真實語料中的大部分短語組合都是符合“中心擴展條件”和“並列結構條件”的，但是我們想强調的是，也確實存在不符合的實例，儘管比例不高，但違反中心擴展條件和並列結構條件的實例也並非特例。下文即通過具體實例的展示和分析來說明語言使用中存在這樣的組合是合理的。

〈表 9〉NHE 規則的內部成分、中心成分考察

序號	內部構成/中心成分	NHE 規則	示例
1.	跟“的”相關的 NHE	np → np 的 !vp	時間 的 推移

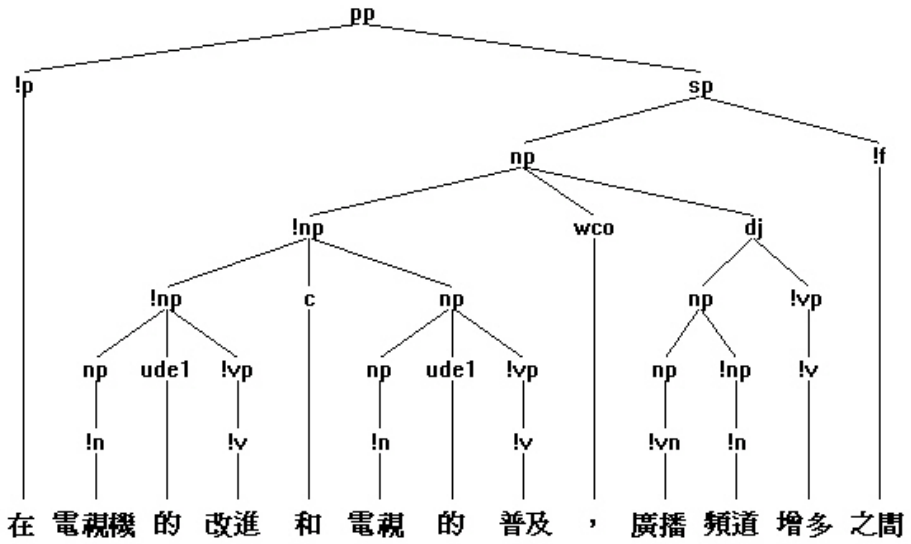
		np → pp 的 !vp np → sp 的 !qp	在電子產品 可靠性方面 的 應用 他們中 的 三個
2.	跟其他助詞(似的、地)相關的 NHE	ap → !np 似的 ap → !dj 似的 dp → !qp 地 dp → !vp 地	雪片 似的 他是這個地方的主人 似的 一寸一寸 地 有秩序 地
3.	ap 擴展爲 np	np → qp !ap	一點 清凉
4.	dj 擴展爲 np	np → qp !dj	這種 再狹窄發生率降低
5.	qp 擴展爲 np	np → np !qp	這 三 本

上表的例子中絕大多數情況都是通過結構**助詞**，系統地改變結構的性質，比如"的""地"等結構助詞，可以系統地使得結構整體的功能不同於其中中心成分的功能。此外，漢語中也存在結構功能不需要標記成分的幫助，直接發生功能轉換的情況，比如例 3、4、5 都是這類情況。陳述性成分、修飾性成分都**直**接轉爲指稱性成分（關於指稱、陳述概念，參見朱德熙 1982）。

〈表 10〉NCC 規則的內部成分考察

序號	並列項	NCC 規則	示例
1.	ap – vp	ap → !ap c vp	對朋友誠實 和 幫助老人
2.	vp – dj	vp → !vp c dj	地震 與 火山噴發
3.	np – dj	np → !np wco dj	電視機的改進和電視的普及，廣播頻道增多
4.	np – vp	np → !np wco vp	一間紅瓦灰墻的小屋，一排白漆的大柵欄，或許還有三五個人影（，眨眼就消失了。）
5.	dj – tp	dj → !dj c tp	我應該今天開始 還是 明天
6.	ap – dp	ap → !ap c dp	（失戀以後，會是）頹廢 或 奮力

表中第 3 行是 np 跟 dj 構成並列結構，此例的句法結構如下面圖 3 所示。



〈圖 3〉一個 NCC 規則的結構樹圖

上面例子引出的一個核心問題是：漢語中主謂結構是陳述性成分還是指稱性成分？如果是陳述性成分，則主謂結構跟 np 並列的時候，就違反了“並列結構條件”。如果是指稱性成分，則不違反並列結構條件。但是，當它不違反“並列結構條件”的時候，就要進一步追問，主謂結構的中心語又是什麼呢？如果主謂結構的中心語是謂語 vp，那麼，例中作爲指稱性成分的主謂結構，其功能顯然跟 vp（“增多”）的功能是不一致的，前者是起指稱作用，而後者一般應該是起陳述作用。這樣，就又違反了“中心擴展條件”。看起來，在短語結構的組合過程中，如果在一個層次上要遵守“並列結構條件”，就可能在另一個層次上違反“中心擴展條件”，二者並不總是能兼顧的。實際上，語料中可以觀察到的表層語言現象是：漢語中的主謂結構既可以用於陳述表達功能，也可以用於指稱的表達功能，在表層句法結構上就是表現爲主謂結構可以做主、賓語。

如果違反“並列結構條件”和“中心擴展條件”的 NCC 規則和 NHE 規則都是

不可避免的，那很自然會想到一個問題，爲什麼在常規的符合中心擴展條件和並列結構條件的組合之外，語言系統會有“非中心擴展”“非同類並列”的結構存在呢？我們的看法是，這是語言使用中的“簡約”要求使然。語言使用中的“簡約”（或者說經濟）要求使得人們在使用句法結構規則時，常常省略其中的成分，即以部分結構成分代行整體之功能，其中最突出的例子，莫過於漢語中的“的”字結構了，例如：“成套的書”這個組合是常規 np，“書”是中心成分，整個結構符合中心擴展條件，但實際使用時，人們完全可以用“成套的”來替代“成套的書”，語料中後者往往是已知的背景信息，因而中心成分很容易省略。省略後得到的“成套的”這個結構，其中心成分是哪一個，如何讓中心成分跟結構整體功能保持一致，就要讓理論語言學者大傷腦筋了。它造成的“麻煩”不僅僅停留在短語結構規則集多出了 NHE 規則，一個連帶的後果是，非常規結構規則融入常規組合規則中，由此形成的漢語句法結構規則，用於計算機的自動句法分析，會造成更多的系統性的**潛在結構歧義**。通過考察樹庫中句法組合歧義的分布和數量，可以更深入地瞭解這種影響。

4 現代漢語句法結構歧義情況的定量考察

本節討論如何從樹庫中抽取有歧義的句法結構。句法結構的歧義有不同的類型，歧義程度也有高低之分。這裏關於句法結構歧義情況的統計和考察僅僅是在組合規則層次上看歧義，還沒有深入到具體的語言實例層次。從樹庫中抽取規則時，可以考慮兩個層面的組合規則，一是通常的短語結構規則（如上文表 1 所示），另一種是以詞性標記串來表達組合規則[5]。如 vp → p n v v n。對這兩種組合規則，都可以統計規則右部同型而左部根節點不同類的情況。如下表 11 所示：

[5] 本文只考慮了 2 到 8 個詞形成短語結構的情況。

〈表 11〉規則右部同型、左部根節點不同形成的歧義組合示例

	規則左部根節點	規則右部組合模式	實例
短語組合規則	np	qp !ap	兩個 不同
	dj	qp !ap	兩個 不同
詞類組合規則	dj	n v n	幹部 領導 群衆
	np	n v n	政府 領導 幹部
	vp	n v n	科學 種植 西瓜

表中短語組合規則對應的實例就短語本身來看是確實有歧義的，不過在具體語境中往往因上下文的影響而只有一種理解。比如：

例 1 新版本跟上一個版本相比有**兩個不同**，……

例 2 論學歷**兩個不同**，論能力兩個完全一樣。

例 1 中的“兩個不同”應解作 np，即“兩個不同之處”的簡省說法。例 2 中的“兩個不同”則應解作 dj，“兩個”跟“不同”之間是主謂結構關係，“兩個”是“兩個人”的簡省說法。

樹庫中更多的歧義組合是像表 11 中“詞類組合規則”對應的實例所反映的情況，即實例本身並沒有明顯的歧義，但在詞類範疇（或短語範疇）層面，則可以有不同的結構解讀方式，這是計算機在分析句子時會碰到的主要歧義問題。比如“幹部領導群衆”和“政府領導幹部”理論上是可以有歧解的，但兩例各自都沒有明顯的歧義，前者的結構是主謂，短語類應歸爲 dj，後者的結構是定中，短語類應歸爲 np。把這兩例放在更大語境中，可更好地體會它們的區別。例如：

例 3 上級領導下級，**幹部領導群衆**。

例 4 我縣近期組織對**政府領導幹部**進行群衆滿意度調查。

這類“歧義”例子不是針對人而言的歧義，但是對計算機處理來說，則是貨真價實的歧義現象。下面我們就來統計樹庫中同型組合形成的根節點不同的規則總體是一個什麼情況。這可以從一個側面反映按照目前的短語結構標注體系

進行短語句法結構組合分析時隱含歧義的程度[6]。

〈表 12〉短語組合規則中右部同型、根節點不同的規則統計[7]

	短語組合規則	同型歧義短語組合		百分比
Type 數	1930	同型短語組合數	296	15.34%
		同型短語組合規則數	670	34.72%
Token 數	1,318,488	477,142		36.19%

〈表 13〉詞類組合規則中右部同型、根節點不同的規則統計

	詞類組合規則	同型歧義的詞類組合		百分比
Type 數	124,611	同型詞類組合數	3932	3.16%
		同型詞類組合規則數	8263	6.63%
Token 數	542,153	240,542		44.47%

[6] 這種統計同時也可以作爲檢查語料標注一致性的一種手段。限於篇幅，本文對此不展開討論。需要說明的是，下文給出的統計數據中因此也可能存在一定偏差，即有的同型規則有可能是標注錯誤造成的不一致問題，而非真正的同型歧義組合。

[7] 這裏的右部同型規則只計算了兩分支以上的規則，沒有計算單分支規則（形如 np → !n 這樣的規則）。如果把單分支規則算在內，則同型短語組合數爲 327（占比 16.93%），同型短語組合規則數爲 741（占比 38.43%）。以規則實例（token）計，共 819,440 個組合涉及同型組合歧義（占比 62.15%）。我們按照同樣方式統計了賓州大學中文樹庫的同型短語組合歧義情況，賓州樹庫短語規則（type）數爲 5220 條。其中兩分支以上的同型短語 316 個（占比 6.05%），同型短語組合規則數爲 724（占比 13.85%）。以規則實例（token）計，共 253774 個組合涉及同型組合歧義（占比 47.21%）。從 type 數來看，賓州中文樹庫的同型歧義情況要顯著低於北大中文樹庫。從 token 數來看，同型歧義組合的比例則高於北大中文樹庫。這大體反映了一方面賓州中文樹庫標注具有更好的內部一致性，另一方面賓州樹庫的短語標記的區分度要更大一些（賓州樹庫的短語標記共 25 個，北大中文樹庫是 17 個）。 從這樣的對比來看，北大中文樹庫的短語標記體系還有進一步細分的必要，可以通過短語類的細分來降低同型短語組合規則的比例。此外，語料標注的內部一致性也還需要提高。

下面就進一步從不同的角度來看同型歧義組合中歧義程度相對比較高的情形。這裏主要考慮了三個角度，一是看一個同型組合能形成幾種不同的短語類，即統計同型組合構成的不同根節點數量（以下簡稱"根數"）的多少；二是看同型組合的頻次高低；三是把一個同型組合形成不同短語類看做是一個隨機事件，計算這個隨機事件的信息熵值，比較熵值的大小。

（一）從同型組合形成的根節點個數多少來看歧義程度

〈表 14〉同型短語組合的不同根數頻次分布

根數	頻次	百分比
2	229	77.36%
3	57	19.26%
4	9	3.04%
5	1	0.34%
合計	**296**	100%

〈表 15〉同型詞類組合的不同根數頻次分布

根數	頻次	百分比
2	3558	90.49%
3	350	8.90%
4	24	0.61%
合計	**3932**	100%

vp 跟 ap 短語組合可能形成根節點數最多達到 5 個，具體每種組合的頻次分布如下表所示。

〈表 16〉根數爲 5 的短語組合及其頻次分布示例：vp+ap 組合

規則左部根節點	規則右部組合模式	頻次	示例
dj	vp !ap	462	發展 很快
ap	vp !ap	85	看著 非常舒服
fj	vp !ap	6	不是星期日 還不著急呢
np	vp !ap	1	（聯繫） 教學 實際
sp	vp !ap	1	過橋 不遠
合計		**555**	

〈表 17〉根數爲 4 的短語組合及其頻次分布示例：ap+vp 組合

規則左部根節點	規則右部組合模式	頻次	示例
vp	ap !vp	3061	認真 學習
dj	ap !vp	117	快樂 在等待我們
np	ap !vp	33	不同 解釋
fj	ap !vp	1	由於恐懼 而逆來順受
	合計	**3212**	

〈表 18〉根數爲 4 的詞類組合及其頻次分布示例： a + m + q 組合

規則左部根節點	規則右部組合模式	頻次	示例
qp	a m q	181	近 一千億 元
ap	a m q	93	少 三 票
tp	a m q	7	近 幾十 年
dj	a m q	6	寬 九 米
	合計	**287**	

（二）從同型組合的頻次高低看歧義程度

同型短語組合中有 96 種頻次超過 200。同型詞類組合中有 139 種頻次超過 200。下面分別列出同型短語組合和同型詞類組合中頻次前 5 位的組合，包括它們能構成的根節點數量，具體是哪些短語類，頻次信息以及示例。

〈表 19〉同型短語組合頻次最高的前 5 個組合

短語類組合	根數	根節點	頻次	合計	示例
!vp np	2	vp	65752	65756	有 天大的困難
		dj	4		是我 聽見的
np !vp	3	dj	58724	60842	我們 正在嘗試
		vp	1448		科學 種植西瓜
		np	670		科學 研究

np !np	2	np	41452	41818	中國 國民經濟
		dj	366		總人口 一千萬人
!vp vp	2	vp	25310	25340	打算 研製新產品
		fj	30		沒有革命的理論 就沒有革命的運動
mp !q	2	qp	19584	19613	兩千多 個
		tp	29		二〇〇六 年

〈表 20〉同型詞類組合頻次最高的前 5 個組合

詞類組合	根數	根節點	頻次	合計	示例
m q	2	qp	18487	18516	三 輛
		tp	29		第五 年
v n	2	vp	14005	16941	送 朋友
		np	2936		輔導 教材
v v	2	vp	16507	16538	推 出去
		dj	31		會談 擱淺
n n	2	np	12979	13015	人民 群衆
		dj	36		小名 鐵蛋
n f	2	sp	5059	5133	樹 後
		tp	74		晚飯 後

表 16-20 分別給出了根數最多和頻次最高的同型短語類組合和同型詞類組合及其實例。從這兩個角度評價同型組合的歧義程度高，有一個明顯的問題，就是同型組合形成的不同短語類頻次分布可能並不均勻，比如表 16 的“vp !ap”的各種組合的頻次就相差很大，表 19 的“!vp np”兩種組合的頻次相差更是懸殊。很顯然，這樣的歧義組合，其歧義程度並不算高。爲了描述同型歧義組合形成不同短語類的頻次分布的均勻程度，可以引入信息熵的概念。

（三）從同型組合的信息熵值大小看歧義程度

如果把像“vp !ap”這樣的組合形成不同的根節點看做一個隨機事件，就可以

用隨機變量的信息熵值來度量一種同型組合形成不同根節點的分布均勻程度，熵越大，分布越均勻，相應的，歧義程度也越高。反之，則分布不均勻，歧義程度也就低一些。熵值計算公式爲：

$$H(\mathcal{S}) = -\sum p_i \log_2 p_i$$

公式中 p_i 表示隨機變量 S 可能的取值中第 i 個值出現的概率。對於同型歧義組合規則來說，它組成爲不同短語類的概率可以用各個組合的頻次來估計，比如“vp ap”短語組合爲 dj 的概率爲 462/555，即 0.83。依此計算出各組合規則的概率後，帶入上面的公式，就可以求得“vp ap”短語組合的熵值爲：0.7383。類似地，可以計算得到表 17 中的“ap !vp”短語組合的熵值爲：0.3118。

按照這種方式，可以計算全部同型短語組合和同型詞類組合的信息熵值[8]。考慮到從詞類組合上升到短語組合的過程中，會減少同型區別的數量（表 12 和表 13 中 token 數的對比），下面就以同型詞類組合的熵值爲例來看歧義程度的差異。因爲是以頻率估算概率，這樣就要求頻率足够大，才能得到相對比較準確的概率值，但限於目前樹庫的規模，大量的組合都是低頻組合。在 3932 個同型詞類組合中，頻次在 100 以下的占 94.15%。對這樣的低頻組合來說，算出來的熵值是並不可靠的。爲兼顧頻次和熵值，我們在計算出全部同型詞類組合的熵值後，取了頻次在 1000 以上，熵值在 0.5 以上的組合，共得到 6 個這樣的同型詞類組合。下面是這 6 個組合形成的不同短語類、頻次值、熵值及實例。

〈表 21〉同型詞類組合中頻次及熵值均較高的 6 個組合

詞類組合	根數	根節點	頻次	合計	熵值	示例
n v	3	dj	1838	2777	1.26	前人 開路
		vp	514			全綫 崩潰
		np	425			燃料 供應

[8] 通過求全部同型歧義組合的平均熵值，可以在一定程度上評價整個樹庫標注體系的不確定性程度。北大樹庫跟賓州樹庫的同型短語歧義組合的平均熵值都約爲 0.57。但考慮到北大樹庫抽取的規則總數是 1930 條，而賓州樹庫規則總數爲 5220 條，據此估計，賓州樹庫標注體系的確定性程度更高，這在一定程度上反映了賓州樹庫標注內部一致性（尤其是低頻規則的內部一致性）可能優於北大樹庫。

n v n	3	np	685	1176	1.04	電子 發射 裝置
		dj	482			麥子 需要 春雨
		vp	9			重金 獎勵 發明人
v n ude1 n	2	vp	590	1161	1.00	解釋 工廠 的 困難
		np	571			劃分 句型 的 標準
v n n	2	vp	1291	1622	0.74	解决 技術 問題
		np	331			有 問題 農藥
v n	2	vp	14005	16941	0.67	符合 國情
		np	2936			煉鋼 工人
a v	2	vp	1173	1353	0.67	努力 學習
		ap	149			難 住

上面 6 個組合中涉及的詞類正是名、動、形三大類實詞。這說明目前采用的詞類體系對於句法組合的制約能力有限，對計算機來說，可能造成較嚴重的歧義問題，因此，面向計算機句法分析的需要，詞類的劃分還應加細。這 6 個組合中恰恰包含了漢語中比較經典的句法結構歧義組合“v n ude1 n”（實例“咬死獵人的狗”）。其熵值也基本爲 1。這意味著，如果讓計算機來猜測這個組合該分析爲 vp，還是 np，則命中率就如同扔硬幣猜正反面一樣，只有 50%。

需要說明的是，本文采用計算熵值的方式來評估一個同型組合的歧義程序，只是初步的探索，還不够成熟。一是如上文已經指出的，受語料規模的限制，用低頻現象去估計隨機事件的概率值，是不可靠的。另一方面，對同型組合的分析深度也是影響熵值的重要因素。比如“v a n”這個同型詞類組合，其實例數爲 743，根節點有 vp，np 兩種，由此計算得到的
“v a n”的熵值爲 0.15。但如果考慮“v a n”形成的 vp 和 np 各自內部都有不同的結構情况，再來計算熵值，結果就可能顯著提高。下面兩個表對比了不同分析深度下，對同一個同型詞類組合的熵值計算的差異。

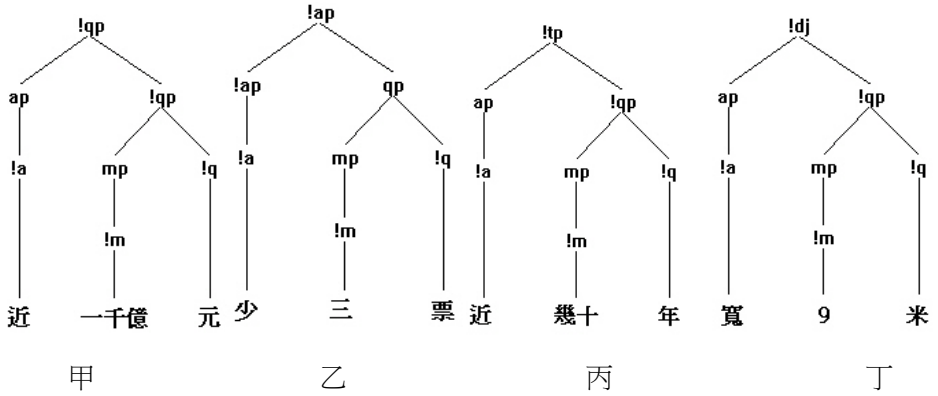
〈表 22〉只考慮根節點差異計算同型詞類組合的熵值

根	詞類組合	頻次	合計	熵
vp	v a n	727	743	0.15
np	v a n	16		

〈表 23〉考慮詞類組合的內部結構差異後計算的熵值

根	詞類組合	內部結構	頻次	合計	熵
vp	v a n	vp (!vp (!v ( )) np (ap (!a ( )) !np (!n ( ))))	608	743	0.80
		vp (!vp (!v ( ) a ( )) np (!n ( )))	119		
np	v a n	np (vp (!v ( )) !np (ap (!a ( )) !np (!n ( ))))	12		
		np (!vp (!v ( ) a ( )) !np (!n ( )))	4		

此外，值得指出的是，按照上述考察方式得出的歧義程度高的組合，也並不一定意味著歧義消解就更困難。以表 18 中介紹的“a m q”的 4 種結構爲例。一方面，這一同型詞類組合的熵值雖然達到 1.25，但該組合形成的 4 類短語有比較清楚的區分條件，因而排除歧義並不困難。下面是“a m q”的 4 種結構。



〈圖 4〉“a m q”的句法結構樹圖

“a m q”的上述 4 種分析方式中，甲和丙可分爲一組來對比，共性爲其中的“a”都由“近”來充當，如果要細分，則甲中的“近”是“接近”，丙中的“近”是“最近”。此外，甲中的量詞 q 爲度量衡單位量詞“元、噸”等，丙中的量詞 q 只能是時間量詞“年、天”等。乙中的 a 只能由“多、少”來充當，丁中的 a 只能由“長、寬、高、重”等少數詞語充當。顯然，“a m q”同型組合中的 a 和 q 都有一定的限制，而且範圍很窄，其結構分析的難度並不大。

5 結語

本文嘗試基於樹庫語料獲取現代漢語句法結構知識。具體內容分三個層面展開。

第一個層面是對特定的句法結構進行考察。本文選取了兩個對象，一個是 np，考察了 np 在主、賓、定、中等不同句法位置的寬度（詞長）以及內部結構的差異。所得統計結果印證了以往語法研究的定性分析。另一個考察對象是“把”字結構，考察了“把”字結構中 vp 的內部構成情況，以及“把”字結構在句中的分布環境。有兩點發現值得注意：一是“把”字結構中的 vp 以帶賓語的情況爲最

大多數；二是“把”字結構用得最多的並不是直接做謂語，而是跟其他謂詞性成分組合成更複雜的謂詞性結構。

第二個層面是對一類句法結構現象進行考察。本文考察的對象是非中心擴展結構和非同類並列結構。通過分析語料中這兩類結構的實例，我們認爲，這兩類結構是語言中實際存在的，是言語交際中人們出於經濟高效的需要，省略成分或通過借用成分製造出的結構，語法理論設計中關於中心擴展條件和並列結構條件的假設適用於大多數常規情況，但並不能否認實際語言中非常規結構的存在。同時因爲這類結構規則加入到常規結構規則集合中，由此形成的句法結構系統用於計算機分析時，就可能帶來更多的潛在歧義。

第三個層面是對句法結構中的潛在歧義情況做宏觀的定量考察。我們嘗試從同型短語組合和同型詞類組合可能形成的短語類個數、同型組合的頻次、同型組合的熵值等不同角度來衡量一個組合的歧義程度的大小。目前的探討雖對語法研究，特別是計算機自動句法結構分析有一定參考價值，但還是比較初步的，對於句法結構的系統性歧義的考察，還有待在樹庫語料規模擴大，標注信息更爲豐富的基礎上，做出更加可靠的分析。

樹庫資源以往通常是用於數據驅動的計算機自動句法分析（DOP）的模型參數訓練，從樹庫中自動抽取可計算的形式化語法模型（如 LFG、HPSG 語法等）。本文則面向語法本體研究，探討從樹庫中獲取不同層次的語法知識。一個樹庫中蘊含的語法知識一部分來自它的標注體系，還有一部分來自基於該標注體系對語言實際材料的標注結果。通過後者獲取的知識，還可以反過來評價前者的設計是否合理，比如如果一個標注體系在標注實際語料後得到的歧義結構的平均熵值較高，就有可能需要回過頭去審視最初的標注體系中各功能標記的設置是否合理。對此，本文還只是做了一些初步的探索，我們希望得到讀者的寶貴意見和建議。

參考文獻

Anne Abeillé, ed. 2003. *Treebanks: Building and Using Parsed Corpora*, Text, Speech and Language Technology Volume 20. Dordrecht: Kluwer Academic Publishers.

Aoife Cahill, Michael Burke, Martin Forst, Ruth Odonovan, Christian Rohrer, Josef Genabith and Andy Way. 2005. Treebank-Based Acquisition of Multilingual Unification Grammar Resources. *Research on Language and Computation*, 3.2:247–279.

Chou, Qiang (周强). 2004. Hanyu jufashuku bianzhu tixi 漢語句法樹庫標注體系 (Annotation Scheme for Chinese Treebank). *Zhongwen xinxi xuebao 中文信息學報* [*Journal of Chinese Information Processing*], 4:1-8.

Chu, Dexi (朱德熙). 1982. *Yufa Jianyi 語法講義* [*Lectures on Syntax*]. Beijing: The Commercial Press (商務印書館).

Huang, Chu-Ren, Feng-Yi Chen, Keh-Jiann Chen, Zhao-ming Gao & Kuang-Yu Chen. 2000. Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface, In Martha Palmer, Mitch Marcus, Aravind Joshi and Fei Xia (eds). *Proceedings of the Second Chinese Language Processing Workshop*, 29-37. HongKong: Hong Kong University of Science and Technology.

LaPolla, Randy J. 1995. Pragmatic relations and word order in Chinese. In Pamela Downing, Michael Noonam (eds), *Word Order in Discourse*, 299-331. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Lyons, John. 1968. An Introduction to Theoretical Linguistics. Cambridge: Cambridge University Press.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 9.2:313-330.

Shen, Jiaxuan (沈家煊). 2007. Hanyu li de mingci han dongci 漢語裏的名詞和動詞 [Nouns and verbs in Chinese]. *Han Zangyu xuebao 漢藏語學報* [*Essays on Sino-Tibetan languages*], 1:1-16.

Shen, Jiaxuan (沈家煊). 2009a. Wo kan hanyu de cilei 我看漢語的詞類 (My view of word classes in Chinese). *Linguistic Sciences* [*語言科學*], 1:1-12.

Shen, Jiaxuan (沈家煊), 2009b. Wo zhishi jiezhe xiangqian kuale banbu – zaitan hanyu de mingci han dongci 我只是接著向前跨了半步——再談漢語的名詞

和動詞 [Just a Small Step Forward——Further Remarks on Nouns and Verbs in Chinese] *Yuyanxue Luncong 語言學論叢* [*Essays on Linguistics*], 40:3-22.

Wei, Hong (魏紅) and Zexiang Chu (儲澤祥). 2007. 'You ding ju hou' yu xianshixing de wuding NP zhuyuju "有定居後"與現實性的無定 NP 主語句 [On 'postposition of definite subject' and the realistic sentence with indefinite NP as subject]. *Shijie Hanyu Jiaoxue 世界漢語教學* [*Chinese Teaching in the World*], 3:38-50.

Xue, Nianwen and Fei Xia. 2000. The Bracketing Guidelines for the Penn Chinese Treebank (3.0), http://www.cis.upenn.edu/~chinese/parseguide.3rd.ch.pdf

Xue, Nianwen, Fei Xia, Fu-Dong Chiou and Marta Palmer. 2005. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*. 11.2:207-238. Cambridge: Cambridge University Press.

Zhan, Weidong (詹衛東), Baobao Chan (常寶寶), and Shiwen Yu (俞士汶). 1999. Hanyu duanyu jiegou dingjie qiyi leixing fenxi ji fenbu tongji 漢語短語結構定界歧義類型分析及分布統計 [Analysis on types of phrase boundary ambiguity in contemporary Chinese]. *Zhongwen Xinxi Xuebao 中文信息學報* [*Journal of Chinese Information Processing*], 3:9-17.

Zhan, Weidong (詹衛東). 2000. Yuyan chengfen de zuhe yu gongneng chuandi 語言成分的組合與功能傳遞 [Combination of Linguistic Units and Inheritance of Syntactic Features]. In 《面臨新世紀挑戰的現代漢語語法研究》[*Modern Chinese Grammar Studies Meeting Challenges of the New Century: Papers from 1998 International Conference on Modern Chinese Grammar*], ed. by Jianming Lu (陸儉明) . 823-833. Jinan (濟南): Shandong Education Press (山東教育出版社).

Extracting Syntactic Knowledge from Large-scale Chinese Treebank

Zhan Weidong

Dept. of Chinese Language & Literature, Peking University

Abstract: This paper illustrates three cases of syntactic knowledge extraction on different levels of linguistic analysis from a Chinese treebank which contains more than 1 million Chinese characters. The first case involves the extraction of the inner structures and the distributions of specified phrases, for instance, noun phrases that used as subject, object and modifier, and verb phrases that used in Chinese *Ba*-construction. The second case concerns the extraction of the non-endocentric constructions and the coordinate constructions whose constituents are asymmetric syntactically. The third case lies in the extraction of ambiguous structures from the treebank and the measure of the degree of ambiguity. In this paper, an ambiguous structure is confined to a construction in which the immediate constituents correspond to two or more reduction rules, which have same components in the right hand side and different categories reduced in the left hand side. The entropy of each ambiguous structure can be calculated by counting its probability. And the entropy value of an ambiguous structure indicates the degree of a structure's ambiguity.

Keywords: Chinese Treebank, Syntactic Annotation, Linguistic Knowledge Extraction