# 中央研究院中文句結構樹資料庫 3.1 版 (Sinica Treebank Version 3.1)說明

## （一） 簡介

中央研究院中文樹圖資料庫 3.1 版(Sinica Treebank Version 3.1)，從 86 年起到 95 年 6 月份止，是中央研究院詞庫小組從中央研究院平衡語料庫（Sinica Corpus）中，抽取句子，經由電腦剖析成結構樹，並加以人工修正、檢驗後的所得的成果。在結構樹圖中，我們標示了中文句語意角色和語法結構的訊息。它的目的在於提供中文自然語言處理研究一個具有標記句結構的研究素材，可以從這個中文句結構樹資料庫中抽取語法知識，也藉由語法知識的抽取與瞭解使剖析系統功能更趨完善。

## （二） 內容說明

### 1. 數量

中文句結構樹資料庫 3.1 中，包含了 6 個檔案，65,434 個中文樹圖，392,237 個詞。

### 2. 檔案內容及來源說明

本資料庫中的檔案，是從平衡語料庫中，直接抽取文章，經由電腦剖析即人工修正後的結構樹，文章中的句子並沒有刻意的挑選過，因此，有些句子會有形式不完整或是語意有誤的情形，而這些情形，我們都保持原貌，沒有刪除，所以在完成的檔案中，會看到標示“%”符號的句子，代表的即是句子不完整無法分析，或是語意錯誤不合文法。舉例來說，檔案中一般的句結構樹為：

# S(agent:NP(Head:Nca:觀光局)|evaluation:Dbb:還|quantity:Daa:另|Head:VE12:安排|aspect:Di:了|theme:NP(property:NP(quantifier:DM:幾處|Head:Ncb:市郊)|property:Nv4:遊覽|Head:Nac:活動))# 。(PERIODCATEGORY)

而句子本身有問題的，則以"%"標示，為：

# %(contrast:Cbca: 但 |reason:PP(Head:P03: 為 了 |DUMMY:VP(Head:VC2: 重 展 |goal:NP(property:Ndda: 昔 日 |property:VH11: 燦 爛 |Head:Nac: 風 貌 ))))# ，(COMMACATEGORY)

在資料庫中的檔案，共有 6 個。news.check 和 travel.check 是抽取自中研院平衡語料庫，因此來源為報紙期刊、網路或書籍等等中的文章。ko.check、ev.check 內容為國小國語課本，ko.check 為國立編譯館出版之國小國語課本一至十二冊，ev.check 為南一書局出版之國小國語課本第七冊與第八冊。oral.check 是和中研院語言所合作的語音平衡檔案，因此不包含完整的文章，此檔案除了剖析成中文句結構樹外，並有語音的錄音資料，然錄音資料不包含在此一資料庫中。sino.check 為光華雜誌的句結構語料，此部分語料亦包含在平衡語料庫中；Sinica Treebank 3.1 版的 sino.check 增加了 4,347 個中文樹圖，30,403 個詞。

六個句結構樹檔案裡還含有原始文本的標頭，用以說明文章本身的屬性，包含了文章的文類、文體、語式、主題和媒體（oral.check 檔除外）。

在 Sinica Teebank 3.1 的版本內，除了較 3.0 版本增加了光華雜誌語料 4,347 個中文樹圖，30,403 個詞之外，在名物化的結構上也做了大幅度的修正，將所有修飾語的名物化詞類改為其原動詞詞類，僅名詞組的中心語仍為名物化。例如，動詞當修飾語 property 時，在 3.0 版本之前的詞類為名物化詞類，

#NP(property:NP ・ 的 (head:NP(property:Nv4: 下 塌 |Head:Ncb: 飯 店 )|Head:DE: 的)|property:Nab:套房|Head:Nab:設施)#

現在的 3.1 版本改為：

# NP(property:NP ・ 的 (head:NP(property:VC1: 下 塌 |Head:Ncb: 飯 店 )|Head:DE: 的)|property:Nab:套房|Head:Nab:設施)#

而名詞組的中心語仍保留名物化，即 3.0 版和現在的 3.1 版都是：

# NP(theme:NP(Head:Nab:書)|nominal:DE:的|Head:Nv4:誕生)#

3. 原則、符號及其它

（1）原則：**中心語主導原則**

中文句子的語法結構常常有歧義的現象。如果只是表達詞和詞的語法結

構，不足以解決中文這種相當程度的歧義現象。因此為了解決這種歧義性，我們採取中心語主導原則（Head-Driven Principle）。中心語主導原則是指每一個句子或詞組結構皆有一中心語（Head），詞組結構由中心語與其論元(argument)或附加成分(adjunct)組成，並由中心語決定詞組的詞類，例如:句子（S）和述詞詞組（VP）的中心語皆為述詞（V）；名詞詞組（NP）的中心語為名詞（N）；介詞詞組（PP）的中心語為介詞（P）；方位詞詞組（GP）的中心語為方位詞（Ng）。藉由中心語指導原則，在剖析中文句子的時候，不但可以決定每一個中心語的詞組類型，並且利用中心語和其他成分所記載的語法和語意訊息，將句子中詞和詞之間的語法和語意的限制關係清楚的表達出來。

以例句(a)和(b)說明如下: 動詞「刊登」為(a)的中心語，它指派其語意角色，「他」為「刊登」的主事者(agent)，語法形式為名詞組(NP)，「一則廣告」為「刊登」的客體(theme)，語法形式為包含一定量詞組(DM)的名詞組，而「在報紙上」是由表地點(location)的介詞「在」所引介的介詞組(PP)，句子剖析的結果如(b)。

(a) 他刊登一則廣告在報紙上。

(b) S(agent:NP(Head:Nhaa:*他*) | Head:VC33:*刊登* | theme: NP (quantifier: DM:*一則* | Head: Nac: *廣告*) | location: PP (Head:P21:*在* | DUMMY: GP(DUMMY:NP(Head:Nab:*報紙*) | Head:Ng:*上*)))#。(PERIODCATEGORY)

（2）符號說明

#：以“#”置於前後，做為一結構樹段落。

( )：詞組的組合成分為複雜結構，以"("、")"標示其詞組結構左右邊界。

|：分隔在同一層次上的成分結構。

（3）語意角色：

Head：語法上的中心語。表示句子或任何詞組結構的中心語（註：大多數的語法中心語也是語義上的中心語）。

head：語義上的中心語。中文有些結構，如“VP 的”形式，其中心語我們不容易決定，因為在形式上它的中心語為“的”，但實際上，它的語意是由“的”前面的 VP 所決定；因此，我們以 Head 表示形式上也就是語法上的中心語，以 head 表示實際載有意義的語義中心語。

DUMMY：未定的語意角色。

除以上的語義角色，其他語義角色有：agent、addition、alternative、

apposition、aspect、avoidance、benefactor、causer、companion、comparison、complement、condition、conjunction、concession、conclusion、contrast、conversion、degree、deixis、deontics、duration、epistemics、evaluation、exclusion、experiencer、frequency、goal、hypothesis、imperative、inclusion、instrument、interjection、listing、location、manner、negation、nominal、particle、possessor、predication、property、purpose、quantifier、quantity、range、reason、recipient、rejection、restriction、result、selection、source、standard、target、theme、time、topic、uncondition、whatever 等等。

（4）詞組結構：

S：表示結構樹為句子（S）；以述詞為中心語；此外當主詞和述詞的賓語或補語的型式為句子或子句的時候，詞組結構標記為S，不為NP。

VP：述詞詞組；中心語為述詞（V）。

NP：名詞詞組；中心語為名詞（N）。

GP：方位詞詞組；中心語為方位詞（Ng）；所帶論元角色為 DUMMY。

PP：介詞詞組；中心語為介詞（P）；所帶論元角色亦為 DUMMY。

DM：定量詞詞組。

其它如結構標記原則，細節請參見陳鳳儀、蔡碧芳、陳克健、黃居仁《中文句結構樹資料庫 (Sinica Treebank)的構建》。