Computational Linguistics and Chinese Language Processing
Vol. 10, No. 4, December 2005, pp. 543-552

双向考察和驗證：并列成分中心語的語義關係和 CCD 的名詞語義分類体系[1]

Bidirectional Investigation: The Semantic Relations between the Conjuncts and the Noun Taxonomy in CCD

吳云芳*、李素建*、李芸*、俞士汶*

Yunfang Wu, Sujian Li, Yun Li and Shiwen Yu

摘要

并列結構是語言信息處理中的難點。本文一方面基于中文概念詞典 CCD 的語義分類体系來考察名詞性并列結構并列成分中心語的語義關係，大部分并列結構呈現出語義相似的特性，少部分并列結構呈現出語義相關、語義相對的特性；另一方面透過并列成分中心語的語義關係來審視 CCD 的語義分類体系，對語義分類体系中的一些語義類、以及語義類之間的關係進行了深入思考。這是語言形式（并列成分中心語）和語言意義（語義類和語義關係）的双向考察和驗證。

關鍵字: 并列結構、語義關係、語義相似、分類体系

---

[1] 本文研究得到了中國國家 973 項目(2004CB318102)、中國博士后科學基金(2004035029)和 863 項目(2001AA114210)的支持。

* 北京大學計算語言學研究所，北京 100871
Institute of Computational Linguistics, Peking University, Beijing, 100871
E-mail: {wuyf, yusw}@pku.edu.cn

[Received April 10, 2005; Revised June 16, 2005; Accepted August 15, 2005]

Abstract

This paper presents a bidirectional investigation on linguistic form and meaning. On the one hand, based on the Chinese Concept Dictionary (CCD), this paper examines the semantic relations between the heads of the conjuncts of nominal coordination. Most of the conjuncts show semantic similarity, a few of them show semantic association, and a few of others show semantic opposition. On the other hand, the semantic relations between the conjuncts provide a new perspective on the noun taxonomy and suggest ways to improve the taxonomy.

Keywords: Coordinate Structure, Semantic Relation, Semantic Similarity, Taxonomy

1. 引言

并列結構（coordinate structure）是語言信息處理中的難點。一般認爲并列成分是相似的，并列結構的自動識別研究幾乎全是圍繞并列成分的相似性來進行。[Okumura and Muraki 1994] 和 [Agarwal and Boggess 1992] 對英語并列結構的研究，[Kurohashi and Nagao 1994] 對日語并列結構的研究，[周強 1996] 和 [孫宏林 2001]對漢語并列結構的研究，都是基于“并列成分相似”這樣的語言學假設，在此前提下設計規則和演算法。漢語語言研究同樣認爲并列成分是相似的，[吳競存，梁伯楓 1992] 指出，詞性相同、結構相同、語義類相同、音節數相同的項并列是最理想、最嚴格的并列。

中心語（head）是當代句法理論中的一個核心概念，擴展的短語結構文法（GPSG）、中心語驅動的短語結構文法（HPSG）都把中心語擺在了重要的位置。中心語是其父親節點句法語義特徵的集中体現者，那麼，并列成分的相似也應該集中体現在各并列成分的中心語上。CCD（Chinese Concept Dictionary，中文概念詞典）是北京大學計算語言學研究所研製開發的漢語語義詞典[于江生、俞士汶 2002]，基本上沿襲了 WordNet 的語義分類体系。

本文一方面基于 CCD 的語義分類体系來考察名詞性并列結構并列成分中心語的語義關係，一方面透過并列成分中心語的語義關係來審視 CCD 的語義分類体系，這是一個双向考察和驗證的過程。[Resnik 1993] 基于早期版本的 WordNet 的名詞語義分類体系研究表明，動詞 drink 的直接賓語是 beverage 的下位詞語；在新版本的 WordNet 的語義分類体系中，[Miller 1999] 引用 Resnik 的研究成果來證明上下位語義關係存在的合理性，這是語言現象和語義分類体系的双向驗證過程。

2. 從 CCD 看并列成分中心語的語義關係

沿襲 WordNet 的分類体系，CCD 的名詞分爲了 25 個基本語義類[2]。在經過了詞語切分和

[2] 這 25 個語類是：1) 動物(animal), 2) 人(person), 3) 植物(plant), 4) 人工物(artifact), 5) 自然物(natural object), 6) 身体(body), 7) 物質(substance), 8) 食物(food), 9) 屬性(attribute), 10) 數量

詞性標注的《人民日報》1998 年 1 月 1－10 日語料基礎上，作者手工標注了語料中出現的有標記的短語層面的并列結構[3]，從中抽取了 2101 個名詞性并列結構，基于 CCD 對并列成分中心語語義關係進行了定量考察。本文的例句均取自于此。待試驗的并列結構都是兩項的，多項并列結構可看作是多個兩項并列結構的疊加，和兩項并列結構應該具有相同的語義約束。待試驗的并列結構僅包括并列成分中心語是名詞的并列結構，如“被[習慣勢力和陳舊觀念]所束縛”，“全部是由[國家、集体]投資”。對名詞性并列結構，各并列成分的最右端一個詞默認爲是中心語；當并列成分是光杆詞語時，其自身也就是中心語。考察兩個并列成分中心語的語義關係，其計算機操作過程可概要地敍述爲：1）提取兩個并列成分的中心語，并列標記之前一個詞是前并列成分的中心語，并列結構結尾處最后一個詞是后并列成分的中心語；2）在 CCD 名詞語義知識庫中對應各中心語的語義類，當詞語是多義詞有多個語義類歸屬時，由人工甄別選擇正確的語義類；3）生成并列成分中心語語義類同現列表。考察結果如表 1 所示。

表 1：名詞性并列結構并列成分中心語語義關係考察

有共同祖先節點（屬于同一語義類）:	1639	78%
無共同祖先節點（不屬于同一語義類）:	462	22%
總計：	2101	100%

表 1 顯示，78%的名詞性并列結構其并列成分的中心語屬于同一語義類，是“同類并列”，呈現出語義上的相似性（semantic similarity）；而有 22%的名詞性并列結構其并列成分的中心語不屬于同一語義類，是“非同類并列”。這麼大比例的“非同類并列”與我們的先驗期待不相符合，可能存在兩個原因：1）或是 CCD 的 25 個名詞基本語義類的設定不合適，至少從并列結構的角度來看不合適；2）或是并列結構本來就不是我們所想像的那樣完全遵從“同類并列”的原則。前一種可能爲我們提供了一個新的視角來審視 CCD 的語義分類体系；后一種可能要求我們重新分析并列成分中心語的語義關係。

2.1 并列成分中心語語義相似

大多數名詞性并列結構并列成分中心語呈現出語義相似的特性。CCD 是用標記樹來表示語義關係的，我們就用“樹”的術語來描述這些語義關係：同一初始語義類下并列的概念稱爲兄弟節點；同一初始語義類下的上下位概念，不論距離遠近稱爲祖孫節點；同一初始語義類下的其他概念，如果不屬于同一同義詞集合（synset）、不形成兄弟節點和祖孫節點，則稱爲遠距離節點。根據并列成分中心語在語義分類樹上的相互位置，語義相似又可分爲 5 種情況。1) 詞形相同。如“競爭機制和激勵機制”，這約占并列結構總數

(quantity), 11) 關系(relation), 12) 通信(communication), 13) 時間(time), 14) 認知(cognition), 15) 情感(feeling), 16) 動機(motivation), 17) 自然現象(natural phenomenon), 18) 過程(process), 19) 行爲(activity), 20) 事件(event), 21) 群体(group), 22) 處所位置(location), 23) 所有物(possession), 24) 形狀(shape), 25) 狀態(state)。

[3] 這個標注了并列結構的語料從 www.icl.pku.edu.cn 網址上可自由下載，供研究之用。

的 7%。2) 同一個同義詞集合，如“產品有[20 個大類、1000 多個品種]”，這約占并列結構總數的 3%。3) 兄弟節點。如“促進整個[長江、黃河]流域生態環境的好轉”，這約占并列結構總數的 22%。4) 祖孫節點。如“中國願意加強同[聯合國和其他國際組織]的協調”，這約占并列結構總數的 3%。5) 遠距離節點。如“[專家、各界觀眾]也提出許多修改意見”，這約占并列結構總數的 43%。

2.2 并列成分中心語語義相關

語義相關（semantic association）是另一種重要的詞語之間的語義聯繫。人腦思維中容易同時激活同一情境（situation）或同一框架（frame）下的不同概念，不同語義類并不能妨礙這種激活，情境或框架足可以成爲激活因子（trigger）。例如，面對“商業事件”這一情境時，人們很容易聯想到買主、賣主、商品、錢，以及買、賣的行爲[Fillmore 1982]。又如，面對“醫療行爲”這一情境時，人們很容易聯想到醫生、護士、醫院、疾病、費用等等相關概念[董振東、董強 2000]。同一情境下的相關概念在一定的語境中就可形成并列。例如：

（1） a 有利于提高[企業和資金]運作效率。

（企業[＋社會團体]，資金[＋所有物]，情境：商）

b 從這裏出發的[車輛和人群]如洪水般流向麥加。

（車輛[＋人工物]，人群[＋人們]，情境：道路交通）

c 促進更多的[中文、中國]信息上網際網路。

（中文[＋通信]，中國[＋處所位置]，情境：中國）

d 環境教育的[師資、教材]都非常缺乏。

（師資[＋人]，教材[＋人工物]，情境：教育）

e 造就出一批批自強不息、直面挑戰的[企業和企業家]。

（企業[＋社會團体]，企業家[＋人]，情境：企業）

這些不同語義類的詞語因在同一個情境下共存而可形成并列，并列的詞語通過不同的方式“指引了（index）”或是“喚起了（evoke）”相同的普遍情境。HowNet 致力于反映概念之間和概念的屬性之間的各種關係，同一情境下的不同概念之間存在著相關聯的描述，并列成分中心語語義相關在 HowNet 的描述中也可得到部分驗證。例如，對（1）a、d 并列成分中心語，HowNet 的描述是[4]：

[4] 此處參考的是 HowNet 2000 版本，向董先生表示謝意。

（2） a 企業：InstitutePlace|場所,*produce|製造,*sell|賣,commercial|商
資金：$spend|花費，#money|貨幣，commercial|商
d 師資：human|人,*teach|教,education|教育,mass|眾
教材：readings|讀物,*teach|教,education|教育

2.3 并列成分中心語語義相對

有時并列成分中心語呈現出語義相對的特性。例如：

（3）a 接收河西醫院全部[人員和資產]。 （人員[＋人們]，資產[＋所有物]）
b 漫漫史河的[許多實事、眾多人物]。 （實事[＋事件]，人物[＋人們]
c [社會心理、人們情感]變化是值得抒寫的。（心理[＋認知]，情感[＋情感]
d 贏得了寶貴的[時間和空間]。 （時間[＋時間]，空間[＋處所位置]）
e 典型本身的[真實事蹟和先進思想]， （事蹟[＋行爲]，思想[＋認知])

（3）中并列的概念在某種意義上是對立的，表示兩個互補的集合，這兩個集合相并就形成一個對語言交際而言完整的集合，對這個完整的集合我們還無法用一個更爲抽象的語詞來指稱。人們常說“人財兩空”，“人”和“財”在漢語言人們的認知世界中是對立的和互補的，因此有了 a 的并列。同樣“人”和“事”、“心理”和“情感”、“時間”和“空間”也是對立互補的。e 是“認知”類名詞和“行爲”類名詞并列。哲學上強調理論和實踐的統一，在人們的思維中同樣注重“認知”和“行爲”的辨正統一，“認知”和“行爲”類名詞在語言中經常形成并列結構：

（4） a 提出了近 15 年內的基本[措施和政策]。
（措施[＋行爲]，政策[＋認知]）
b 以自己的[聰明才智和實際行動]，譜寫青春之歌。
（聰明才智[＋認知]，行動[＋行爲]）

2.4 并列成分中心語語義既不相似也不相關也不相對

語言中存在少數的并列結構，其并列成分中心語的語義既不相似也不相關也不相對。例如：

（5） a 草案確定了反恐怖活動戰略計畫的執行[機構和辦法]。
b 領導幹部受到[人民和法律]的監督。

對此我們還無法進行有效的描述和解釋。[儲澤祥等 2002]從“語用需要、經濟原則”的角度描述兩個名詞的非常規聯合，但沒有涉及句法語義的解釋。

3. 從并列成分中心語語義相似看 CCD 的名詞語義分類体系

把名詞性并列結構并列成分中心語語義相似看作是一種客觀存在的語言特性，那麼并列結構爲我們提供了一個很好的視角來審視 CCD 的語義分類体系，這種審視對其他的語義分類体系也很具參考價值。

3.1 “人們”、“社會團体”語義類名詞和“人”語義類名詞可形成并列—考慮移動併入

CCD 在“群体（group）”語義類下設有“人們（people）”一小類（記作[＋人們]），表示“任何一群人（any group of human beings）”，而 25 個基本語義類中又設有“人（person）”一類。現代漢語中，[＋人們]和[＋人]名詞經常形成并列，例如：

（6） a 日益被[各層領導和社會公眾]所認識。 （領導[＋人]，公眾[＋人們]）
b 雅俗共賞，極受[專家和人民群眾]喜愛。 （專家[＋人]，群眾[＋人們]）

[＋人們]、[＋人]名詞能自由形成并列，而可以不論是否是“群体”，即數的多少，這是由于漢語沒有數的形態變化而造成的。b 中并列結構若翻譯成英語必得是“expert**s** and common people”。“人們”、“人”這兩個語義類在漢語中是相近的，應該合併在一起。事實上，董振東先生的 HowNet、北京大學的《語義詞典》[王惠等 2003]都是將“公眾”、“群眾”這樣的詞置于“人”語義類下。因此，在 CCD 中可以將“人們”小類從“群体”類中移出併入“人”語義類。

CCD 在“群体（group）”語義類下設有“社會團体（social_group）”一小類（記作[＋社會團体]），[＋社會團体]名詞經常和[＋人]名詞形成并列，例如：

（7） a [求職者和用人單位]反映最爲強烈的，（求職者[＋人]，單位[＋社會團体]）
b [旅客和航空公司]都受到損失， （旅客[＋人]，公司[＋社會團体]）

[＋人]名詞和[＋社會團体]名詞都具有“施事”功能，很多動詞對它們具有相同的選擇限制（selectional restriction）。例如，“反映”的主体可以是“求職者”，也可以是“單位”，因此可形成 a 的并列；“受到損失”的主体可以是“旅客”，也可以是“公司”，因此可形成 b 的并列。除了人所特有的一些生理動作[＋社會團体]名詞不能勝任，例如不能說“**單位吃”，“**公司跑”，[＋社會團体]名詞可以充當大多數動詞的施動者，如“單位贈送錦旗”，“公司轉讓債權”，“航空公司請求延期”等等。雖然表面上[＋

社會團体]名詞不具有生命，但它由具有生命的人所組成，并且由其中的代表法人來實施某種行爲，因此句法功能上[＋社會團体]名詞和[＋人]名詞有很多相似之處。《語義詞典》將“社會團体”置于“人”語義類下作爲一個次類[5]，這是比較合適的。由此，在 CCD 中可以將“社會團体”小類從“群体”類中移出併入“人”語義類。

同爲“群体”語義類的[＋人們]名詞和[＋社會團体]名詞可形成并列。例如：

（8） a 要求[各國政府和全人類]採取緊急行動，(政府[＋社會團体]，人類[＋人們])
b [地方政府和人民群眾]積極支持部隊， (政府[＋社會團体]，群眾[＋人們])

因此，從并列結構形成的角度考慮，“人們”、“社會團体”可適當併入“人”語義類。

3.2 “社會團体”語義類名詞和“行政區”語義類名詞可形成并列—考慮移動靠近

“群体”語義類下的“社會團体”名詞和“處所位置”語義類下的“行政區(district)”名詞（記作[＋行政區]）可以形成并列。[＋社會團体]名詞既可以指稱共用某些社會關係的人們，也可以表示這些人們所在的處所位置，例如“銀行”，在（9）a 中表示“銀行的領導者或人們”，而在（9）b 中表示“處所位置”之義。反之，[＋行政區]名詞既可以表示占有一定空間的處所位置，又可以指稱相關的社會團体，例如“北京”，在（10）a 中表示“處所位置”之義，而在（10）b 中表示“北京的領導者或人們”。表現在并列結構上，[＋社會團体]名詞和[＋行政區]名詞可以自由形成并列，如（11）中的例子。

（9） a 這 13 家銀行也作出了積極反應。
b 他走進了那家銀行。

（10）a 1921 年 9 月 26 日生于北京。
b 北京按照國際奧會的要求，如期將申辦報告文本送交國際奧會審閱。

（11）[俄羅斯和北約]建立戰略夥伴關係。(俄羅斯[＋行政區]，北約[＋社會團体])

由此可見，“社會團体”和“行政區”這兩個語義類在漢語中是相近的，語義類設置中應使兩者靠近。

[＋社會團体]名詞可以和[＋人]名詞形成并列，[＋社會團体]名詞也可以和[＋行政區]

[5] 《語義詞典》將此語義記作“團体(group)”。

名詞形成并列，但[＋人]名詞卻很少能夠和[＋行政區]名詞形成并列。可見，詞語之間的并列關係是不可傳遞的。

3.3 “抽象物”類語義類更易形成并列—分類宜粗

“實体（entity）”類語義類形成并列結構時，其語義類相同要求更細，而“抽象物（abstraction）”類語義類形成并列結構時，其語義類相同要求略粗。事實上，同屬于“抽象物”類的“屬性”、“關係”、“通信”、“認知”、“群体”、“狀態”6 個語義類之下的詞語可以相當自由地彼此形成并列。例如：

（12）a 要堅持不懈改善[生態環境和生產條件]。（環境[＋狀態]，條件[＋屬性]）
b 可持續發展的[定義和內容]。（定義[＋通信]，內容[＋認知]）
c 被[習慣勢力和陳舊觀念]所束縛。（勢力[＋屬性]，觀念[＋認知]）
d 有著[悠久的文明和豐富的文獻傳統]。（文明[＋群体]，傳統[＋認知]）
e 一國兩制事業的[可行性和輝煌前景]。（可行性[＋屬性]，前景[＋狀態]）

人們對抽象事物的認識其實并沒有那麼清晰的分類意識。假如問一個人“環境”、“傳統”、“可行性”的語義類分別是什麼，他或許會回答“環境”的語義類是“認知”，“傳統”的語義類是“通信”，“可行性”的語義類是“狀態”。各家語義分類体系對抽象詞語的歸類也存在諸多的不一致性。《語義詞典》將“環境”和“傳統”籠統地歸入“抽象事物”，將“可行性”歸入“屬性”。HowNet 將“環境”和“可行性”歸入“屬性”，將“傳統”歸入“規矩”（相當于 CCD 中的“認知”類）。而對具体事物（實体）就是另一番光景了，沒有人會認爲“桌子”是“食品”，或者“狗”是“人工物”。表現在語言上，具体事物（實体）的語義類之間不能隨意并列，偶爾并列也依賴于一定語境的支撐。現代漢語并列結構的形成啓示我們，對路徑長度相同的兩對節點，“具体事物”類下的兩個節點語義距離較大，而“抽象事物”類下的兩個節點語義距離較小。由此可知，一方面在并列結構的自動識別過程中，對抽象名詞的語義相似性要求可適當放寬，而對具体名詞的語義相似性要求需適當加嚴；另一方面在語義類設置過程中，對抽象類名詞的分類宜粗而不宜過細。

4. 從并列成分中心語語義相對看 CCD 的名詞語義分類体系

4.1 高度抽象的詞語容易形成并列—宜從一個新的角度進行語義歸類

上文 2.3 談到，表示相對意義的詞語經常形成并列。例如：

（13） a [社會心理、人們情感]變化是值得抒寫的。（心理[＋認知]，情感[＋情感]）
b 贏得了寶貴的[時間和空間]。 （時間[＋時間]，空間[＋處所位置]）

需要注意的是，（13）中各并列成分多是概括的、抽象的詞語，它們不指稱具体的概念，一旦將其中一個抽象概念換作具体概念，并列就不能成立。這種抽象概念往往就是某個語義類的“標籤”，其本身的意義和用法與語義類內部具体詞語差別很大。例如“時間”的語義類是[＋時間]，但和具体的時間詞“今天”、“明年”用法語義差別很大，可以說“今天學習”，但不能說“**時間學習”。“時間”能和別的語義類的詞語形成并列，如“時間和精力”，“時間和空間”，但具体的時間詞卻只能跟時間詞自己并列，不能跟別的語義類的詞語形成并列，“**今天和精力”，“**明年和空間”這樣的并列在語言中是不存在的。又例如“情感”的語義類是[＋情感]，但它和具体表示情感的詞語（例如“溫情”、“恐慌”）在用法語義上差別很大，“情感”和“溫情”、“恐慌”在形成并列結構時鮮有共同點。像語義類標籤的這些高度抽象的詞語可看作是廣義的屬性（attribute），語義類之下的具体實例 （instances）可看作是屬性值（attribute values），我們有 VALUE(時間) = 今天|明年，VALUE(情感)=溫情|恐慌。我們懷疑，類似語義類標籤的這些高度抽象的詞語，是否應該從一個新的角度進行語義歸類。

5. 結語

本文一方面基于 CCD 的語義分類体系，考察了現代漢語名詞性并列結構并列成分中心語的語義關係，呈現出四種：語義相似、語義相關、語義相對、語義既不相似也不相關也不相對。語義相似是并列成分之間最主要的語義關係，但并不是所有的并列成分都呈現出語義相似的特性。另一方面，透過并列成分之間的語義關係其中主要是語義相似的關係，我們重新審視了 CCD 的語義分類体系，對詞語之間的語義關係進行了深入的思考，爲語義類的設置提供了一些有價值的參照座標。本文對并列成分中心語的語義關係和 CCD 的名詞語義分類体系進行了双向考察和驗證，這其實也就是形式和意義的相互驗證。

參考文獻

儲澤祥等，《漢語聯合短語研究》，長沙，湖南大學出版社，2002。

董振東、董強，知網. 見：http://www.keenage.com. 2000

孫宏林，《現代漢語非受限文本的實語塊分析》，北京大學計算機系博士學位論文，2001。

王惠、詹衛東、俞士汶，“現代漢語語義詞典規格說明書”，《漢語語言與計算學報》,13(2), 2003, pp.159-174.

吳競存、梁伯樞，《現代漢語句法結構與分析》， 北京，語文出版社，1992。

于江生、俞士汶， “中文概念詞典的結構”，《中文信息學報》，16(4), 2002, pp.12-20 轉 44 頁.

周強，《漢語語料庫的短語自動劃分和標注研究》，北京大學計算機系博士學位論文，1996。

Agarwal, R., and L. Boggess, "A simple but useful approach to conjunct identification," In *Proceedings of 30th Annual Meeting of Association for Computational Linguistics*, 1992, Newark, Delaware, pp. 15-21.

Kurohashi, S., and M. Nagao,1994. "A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures," *Computational Linguistics* 20,(4), 1994, pp. 507-34.

Miller, A., "Nouns in WordNet", In Fellbaum, C., (ed.), *Wordnet: An Electronic Lexical Database*, Cambridge: MIT Press, pp. 23-46, 1999.

Okumura, A., and K. Muraki, "Symmetric pattern matching analysis for English coordinate structures," In *Proceedings of the 4th Conference on Applied Natural Language Processing*, 1994, University of Stuttgart, pp. 41-46.

Resnik, P., *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. Dissertation, University of Pennsylvania, 1993.

Fillmore, C. J., "Frame semantics," In *Linguistics in the Morning Calm*. Seoul: Hanshin Publishing Corporation, 1982, pp.111-137. 中譯本：詹衛東譯，2003，框架語義學。《語言學論從》第 27 輯，北京：商務印書館。382-412 頁。