LANGUAGE AND LINGUISTICS 9.2:435-447, 2008
2008-0-009-002-000262-1

# 漢語複合詞理解難易度的計算*

葉文曦　　邱立坤
北京大學　　北京城市學院

本文依據字的使用頻率和字的構詞頻率給出了語義範疇重要性層級的劃分標準，考察了複合詞整體所表語義範疇在相關語義場中的位置及重要性等級，列出了語義場關聯模式的優先度列表。在上述工作的基礎上，最後給出了複合詞理解難易度的計算公式及名詞性複合詞語義模式的計算方法。本項研究有助於語言教學中的詞表編製和句子理解難易度的計算。本文的研究方法和成果還有助於加深對語義場內部結構以及複合詞內部語義結構的認識。

關鍵詞：詞彙語義學，複合構詞，語義場，語義理解難易度，計算語言學

## 1. 引言

漢語中的複合詞是一個有爭議的概念，有時被視爲與單純詞相對應的範疇，有時則又可以涵蓋所有的粘著式名詞短語，本文的複合詞包括兩者在內，是詞與詞組之間的交叉地帶。目前語言學界對於複合詞的研究主要集中於複合詞的語法地位，自然語言處理學界則主要從新詞識別的角度進行考察。事實上，複合詞作爲一個交叉地帶，有著承上啓下的重要地位。通過考察複合詞，既可以了解詞語的構造方式，又可以了解短語的構造方式。當前對於短語的研究都是在一個句子的範圍內進行，把短語看成句子的一部分，所看到的短語都是很複雜的，因而面臨許多無法回避的難題。通過考察複合詞，事實上也就是在考察最簡單的短語，在一定程度上可以將研究對象簡化，從而可以更方便地抓住短語研究中最本質的東西。

* 本文使用了董振東先生開發的知網 2000 版和北京大學計算語言學研究所開發的人民日報標注語料庫，在此謹致謝意。
本文在「第七屆漢語詞彙語義學討論會」（2006 年 5 月，台灣新竹交通大學）上宣讀。本項研究得到了北京市教育委員會科技發展計畫面上項目 (No.KM200600006002) 的資助。
感謝匿名審稿人提出重要的修改意見。

本文的研究受到鄭錦全 (2005) 研究的啓發，句子難易度的計算與複合詞難易度的計算是密切相關的。在句子當中，單字詞難易度與複合詞的難易度在一定程度上也影響著句子理解的難易度。

對複合詞的研究可以從多個角度進行，本文的研究目的主要是爲應用服務的。在語言學教學過程中，需要判斷哪些複合詞具有相同的理解模式，具有相同理解模式的複合詞中，又需要區別哪些容易理解，哪些較難理解，等等，從而可以科學地制定分級詞表。在計算機信息處理過程中，計算機發現一個新詞之後，爲了進一步進行語法和語義分析，則需要判斷複合詞的整體功能進而推知其語義。在所有的複合詞中，名詞佔有最重要的地位，新生詞語中也多數屬於名詞。因此，本文在語義範疇重要性等級和語義場關聯分析的基礎上，嘗試對漢語中的名詞性複合詞理解的模式及難易度進行計算。

本文使用的資源主要包括：北京大學計算語言學研究所人民日報（2000 年上半年）標注語料庫和董振東先生的知網（2000 版）。人民日報標注語料庫主要進行了詞語切分和詞性標注兩項工作，其中的切分單位既包括嚴格的語言學意義上的詞，也包括一些固定短語。我們所使用的人民日報標注語料庫文本爲 2000 年上半年人民日報，計 1,400 餘萬字。從中可以統計出單字的字頻和構詞頻率，並取得一個實際的詞表。知網 2000 版則對六萬多個詞語進行了語義描述，從中可以獲知單字詞的語義類別和部分雙字詞的語義類別，在此基礎上，可以通過一定的規則推斷出語料庫詞表中所有名詞的語義類別。

在標注語料庫和知網的基礎上，首先進行以下準備工作：獲取所有單字詞的構詞頻率，並藉此劃分各語義場中單字詞的語義範疇重要性等級。語料庫中的詞，有一些已爲語義詞典所描述，還有一些則未得到描述。已描述詞語的語義類可直接繼承，未描述詞語的語義類則需要通過詞性及構成成分的語法功能和語義類別來計算。在這一步驟中，將會總結出名詞性複合詞語義關聯模式的計算方法。

## 2. 複合詞理解難易度的語義基礎和語義範疇重要性層級的劃分標準

關於理解難易度的計算沒有現成的語義理論可供參考，只有在研究問題的過程中加以總結和闡釋。什麼是語義理解？這是研究複合詞理解難易度首先要回答的一個基本問題。僅就複合詞而言，語義理解就是了解和懂得複合詞所具有的意義，它的所指和它的內涵。語義理解的目的是通過弄清複合詞內部的語義結構，

尋求和現實所指進行掛鈎，這是最基本的，還有就是了解複合詞意義的內涵，雖然存在基本的內涵意義，但在不同場合內涵會有多有少，有深有淺。例如“海魚”這一複合詞的意義是“生活在海洋裡的魚”，它的意義可從以下三個方面來認識：空間＋動物，海＋魚；語義類別，[魚]；語義內涵，[海] 提示基本內涵意義“生活在海洋裡”。

複合詞的理解，難有難的原因，容易有容易的原因，都是可以解釋的。

複合詞理解難易度跟人們對現實事物的認識、掌握的程度以及熟悉的程度密切相關，例如：“海魚、海龜”和“海獺、海蛰”，前一組較常見，後一組相對生僻。事物常見必然導致相關詞語常見，事物冷僻必然導致相關詞語冷僻，最終會在單字和詞語的使用頻率上有所表現。

複合詞理解難易度還跟人們對編碼理據的熟悉程度密切相關。所謂理據是指語言社團人們熟知的現實中存在的不同事物之間的真實關聯，例如顏色和動物的關聯、空間和動物的關聯、質料和衣物的關聯等等。常見的關聯人們在理解時容易進行類推，相應地，理解就相對容易；相反，不常見的關聯或罕見的關聯人們理解起來就不容易進行類推，理解就相對困難。

複合詞的內部語義結構反映現實理據，語義的認知理解也遵循一定的軌道。例如可以通過認識與“馬”組配的其他字所表示的語義範疇所在的語義場來考察“馬”可從哪些維度進行認知，形狀、體積、顏色、性質、速度、耐力、性情、功用、空間、舉止和動作是認知“馬”所用的必要的若干維度。上述設定的對“馬”的描述框架似乎帶來有先驗、主觀的色彩，其實不然，它是多少年來漢人形成的關於“馬”的經驗框架，具有極強的慣性，其他動物也大致可以從相同的角度進行認知和描述。爲了計算的方便，我們可以把上述問題轉化爲語義場跟語義場之間的關聯模式，例如形狀、顏色、功用、空間、舉止等語義場分別和動物語義場存在關聯。

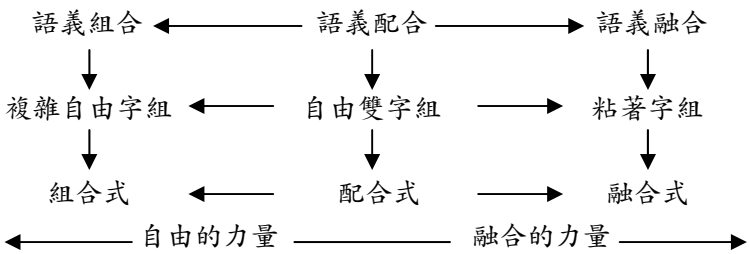
複合詞理解難易度還跟複合詞理據的透明不透明有關係。理據越透明就越容易理解，相反理據越不透明就越不容易理解。關鍵問題是，複合詞前後成分能否直接提示編碼的現實理據，在這一點上不同複合詞有不同表現，例如“戰馬”和“丁香”，前者理據透明，容易理解，後者理據不透明，難於理解。理據不透明的複合詞對人來說，有時需要依賴常識來理解，例如，我們知道“丁香”屬於“花”或“植物”，而計算機則需要依賴現成的分類詞彙集來理解。

近些年來，認知語言學在分析語義結構時多使用“語義範疇”(semantic category) 這一概念，語義範疇就是語義類別，是對現實事物進行概括反映所得到的類別，例如 [人] [馬] [魚] [吃] [聽] [看] [紅] [綠] [大] 等都是常見的語義範疇。

從更廣泛的角度考慮，我們可以把複合詞前後兩個成分的關係看作是兩個語義範疇之間的關係。因此我們可以從雙字各自所表示的語義範疇的性質入手來探討判定一個雙字複合詞的語義理解難易度。

在以往的研究中（葉文曦 2004a, b），我們發現不同的語義範疇進行組配有三種不同的情況，即配合式、融合式跟組合式，如下圖所示：

漢語語義組配層次和語形實現




兩個基本級語義範疇直接相鄰組配的結果有兩種，一種爲語義配合，一種爲語義融合，原因是直接相鄰在語義組配上是非常敏感的，直接相鄰的兩個語義範疇容易發生配合或融合。例如 [騎]–[馬] 和 [駿]–[馬]，前者是配合式，後者是融合式。配合和融合的區別在於，配合內部的語義關係是可以類推的，是不同語義域之間的無標記關聯，而融合內部的語義關係不能類推或者較少能夠類推。能夠類推的語義關係是常見的，無標記的，詞典不必收錄相關字組，而不能夠類推的語義關係是少見的，是有標記的，詞典應該收錄相關字組。

具體語義場內部是不平衡的，即屬於同一語義場內部的不同語義範疇存在重要性等級的差別。這裡以“動物”和“植物”兩個語義場內部的層級爲例加以說明，確定層級的標準是綜合了構詞條數、充當字組後字的條數以及單用情況：

漢語“動物”義場語義範疇的重要性等級

I [鳥 獸 蟲 魚 畜]

II [馬 牛 龍 豬 雞 狗 羊 鼠 虎 狼 雀 熊 鷹 猴 蛇 鴨 貓 鵝 猿 鯨 蝶 禽 燕 鴉 蝦 蟻 豺 兔 蝗 驢 蛙 蠅 蚊 蛾 龜 駒 騾 犀 雁 鹿 狐 豹 鶴 獅 犬 鯽 雛 鶯/鵲 蝟 鱷 鷗 鴿 犢 羔 蟹]

III [蛐蛐 虱 鼠 獺 螻蛄 蚌 鴻 貂 雉 螯 鱉 鷂 梟 蟋蟀 鵪鶉 蜻蜓 蚯蚓 蝙蝠 蝌蚪 鴛鴦 蚱蜢 鸚鵡 麒麟 螟蛉 鷓鴣 鯉 鳩 鰍 猹 鮭 鴿 鰻 蚤 羚 鯰 鱒 蠹 獾 獐 鱸 鰱]

漢語“植物”義場語義範疇的重要性等級

I [花 草 樹 木 果 菜 瓜 苗 葉]

II [桃 稻 蔥 蓮 蘭 麥 梅 芽 栗 梨 楊 柳 橘 桑 杏 薯 藤 葵 棗 蕉 蒲 楓 柿 椿 桂 李 柏 樟 禾 槐 茄 蒜 椒 杉]

III [蔗 菊 蘑菇 菠 榆 藕 檸檬 樺 豌 苔蘚 棠 卉 萵苣 桉 椰 橄欖 芋 榛 榕 稞 茉莉 苜蓿 楂 枇杷 荸薺 檳榔 稗 棕櫚 芫荽 芙蓉 茯苓]

在每個語義場中高級和中級都是最常用的級別，但性質有差別。高級比較概括，反映的是語言中最重要的事物大類，例如 [鳥 獸 蟲 魚 畜] 和 [花 草 樹 果 菜 瓜 苗 葉]。而語義場裡最重要的語義範疇是中級，例如 [馬 牛 豬 雞 狗 貓 羊 鼠 虎 狼 雀 熊 鷹 猴 蛇 鴨 鵝] 和 [桃 稻 蔥 蓮 蘭 麥 梅 梨 楊 柳 橘 桑 杏] 等，在認知語義上這一類既概括，又比較具體，意象明確，識別方便，理解容易，和動作、性狀類語義範疇的組配最爲典型和活躍，因此是基本級語義範疇裡最重要的一類。從有無標記的角度看，等級越高，越傾向於無標記，越容易認知理解，相反，等級越低，越傾向於有標記，越不容易認知理解。

依據字頻和構詞頻率可以定性並定量地給出語義範疇重要性層級的劃分標準。字頻和構詞頻率能夠反映出同一語義場的不同語義範疇之間重要性的不同。單字反映的語義範疇重要性層級的難度係數可以設定爲 0、1、2、3 四級，數目越大，越容易理解，數目越小，越不容易理解。在語料庫中未出現的爲 0 級，出現過的依照構詞頻率由高到低分別爲 3、2、1 級。作爲示例，下面給出“走獸”語義場的語義範疇重要性層級列表。

〈表 1〉語義範疇重要性層級示例

| 字 | 共計 | 等級 |
|---|---|---|
| 躡 | 0 | 0 |
| 鼢 | 0 | 0 |
| 貉 | 0 | 0 |
| 虺 | 0 | 0 |
| 麋 | 0 | 0 |
| 貊 | 0 | 0 |
| 羆 | 0 | 0 |
| 貔 | 0 | 0 |
| 鼉 | 0 | 0 |
| 蝠 | 0 | 0 |
| 鼯 | 0 | 0 |

| 字 | 共計 | 等級 |
|---|---|---|
| 鼷 | 0 | 0 |
| 魑 | 0 | 0 |
| 貅 | 0 | 0 |
| 鼬 | 0 | 0 |
| 獘 | 0 | 0 |
| 螭 | 1 | 1 |
| 蝠 | 1 | 1 |
| 貘 | 1 | 1 |
| 麅 | 1 | 1 |
| 麝 | 1 | 1 |
| 鼴 | 1 | 1 |

| 字 | 共計 | 等級 |
|---|---|---|
| 獐 | 1 | 1 |
| 貂 | 2 | 1 |
| 獾 | 2 | 1 |
| 麋 | 2 | 1 |
| 猩 | 2 | 1 |
| 豺 | 3 | 1 |
| 麂 | 3 | 1 |
| 虬 | 3 | 1 |
| 獺 | 3 | 1 |
| 蜥 | 3 | 1 |
| 蟾 | 4 | 2 |

| 字 | 共計 | 等級 |
|---|---|---|
| 狸 | 4 | 2 |
| 鬣 | 4 | 2 |
| 麒 | 4 | 2 |
| 犀 | 4 | 2 |
| 鱷 | 6 | 2 |
| 鯨 | 7 | 2 |
| 蛟 | 8 | 2 |
| 蟒 | 8 | 2 |
| 羚 | 10 | 2 |
| 猿 | 12 | 2 |

| 字 | 共計 | 等級 |
|---|---|---|
| 蛙 | 13 | 2 |
| 狐 | 14 | 2 |
| 猴 | 16 | 2 |
| 豹 | 19 | 2 |
| 蹄 | 19 | 3 |
| 駝 | 21 | 3 |
| 仔 | 27 | 3 |
| 彪 | 28 | 3 |
| 鼠 | 29 | 3 |
| 狼 | 30 | 3 |

| 字 | 共計 | 等級 |
|---|---|---|
| 蛇 | 32 | 3 |
| 獸 | 32 | 3 |
| 麟 | 34 | 3 |
| 獅 | 39 | 3 |
| 熊 | 42 | 3 |
| 鹿 | 44 | 3 |
| 象 | 102 | 3 |
| 虎 | 132 | 3 |
| 龍 | 555 | 3 |

雙字複合詞整體所表示的語義範疇在不同語義場中也存在重要性等級問題，使用頻率越高，越容易理解，重要性等級也越高；相反，使用頻率越低，越不容易理解，重要性等級也越低。下面給出“走獸”類雙字複合詞整體在語義場中重要性層級的表現作爲示例。

〈表 2〉複合詞整體重要性層級示例

| 詞 | 整體語義場 | 前字語義場 | 後字語義場 | 詞頻 | 等級 |
|---|---|---|---|---|---|
| 猛虎 | 走獸 | 舉止 | 走獸 | 16 | 3 |
| 海豹 | 走獸 | 水域 | 走獸 | 15 | 3 |
| 青蛙 | 走獸 | 顏色 | 走獸 | 11 | 3 |
| 野獸 | 走獸 | 舉止 | 走獸 | 10 | 3 |
| 海龍 | 走獸 | 水域 | 走獸 | 8 | 3 |
| 狐狸 | 走獸 | 走獸 | 走獸 | 5 | 3 |
| 蛟龍 | 走獸 | 走獸 | 走獸 | 5 | 2 |
| 猩猩 | 走獸 | 走獸 | 走獸 | 4 | 2 |
| 麒麟 | 走獸 | 走獸 | 走獸 | 4 | 2 |
| 猛獸 | 走獸 | 舉止 | 走獸 | 4 | 2 |
| 蟒蛇 | 走獸 | 走獸 | 走獸 | 3 | 2 |
| 虬龍 | 走獸 | 走獸 | 走獸 | 2 | 2 |
| 水獺 | 走獸 | 水域 | 走獸 | 2 | 1 |
| 牛蛙 | 走獸 | 舉止 | 走獸 | 2 | 1 |
| 毒蛇 | 走獸 | 舉止 | 走獸 | 2 | 1 |
| 豺狼 | 走獸 | 走獸 | 走獸 | 1 | 1 |
| 白狐 | 走獸 | 顏色 | 走獸 | 1 | 1 |
| 黑熊 | 走獸 | 顏色 | 走獸 | 1 | 1 |

| 詞 | 整體語義場 | 前字語義場 | 後字語義場 | 詞頻 | 等級 |
|---|---|---|---|---|---|
| 鼷鼠 | 走獸 | 走獸 | 走獸 | | 0 |
| 駝鹿 | 走獸 | 走獸 | 走獸 | | 0 |
| 鼯鼠 | 走獸 | 走獸 | 走獸 | | 0 |
| 鼷鼠 | 走獸 | 走獸 | 走獸 | | 0 |
| 貂熊 | 走獸 | 走獸 | 走獸 | | 0 |
| 鼢鼠 | 走獸 | 走獸 | 走獸 | | 0 |
| 猿猴 | 走獸 | 走獸 | 走獸 | | 0 |
| 狼獾 | 走獸 | 走獸 | 走獸 | | 0 |
| 鼴鼠 | 走獸 | 走獸 | 走獸 | | 0 |
| 龍虎 | 走獸 | 走獸 | 走獸 | | 0 |
| 貔貅 | 走獸 | 走獸 | 走獸 | | 0 |
| 麋鹿 | 走獸 | 走獸 | 走獸 | | 0 |
| 貔虎 | 走獸 | 走獸 | 走獸 | | 0 |
| 鼬獾 | 走獸 | 走獸 | 走獸 | | 0 |
| 紫貂 | 走獸 | 顏色 | 走獸 | | 0 |
| 紅狐 | 走獸 | 顏色 | 走獸 | | 0 |
| 黃鼠 | 走獸 | 顏色 | 走獸 | | 0 |

## 3. 語義場關聯度的計算

從微觀上看，語義組配是具體的不同語義範疇之間發生關聯；而從宏觀上看則是具體語義範疇所屬的大的語義領域之間發生關聯。通過統計考察範圍內的複合詞的語義關聯模式，可以給出語義關聯模式的優先性列表，從而可以確定哪些語義場之間經常關聯，哪些語義場之間很少關聯，哪些語義場基本上沒有關聯。我們可以將優先性高的關聯稱爲正常關聯，將優先性低的關聯稱爲異常關聯，並參照語義範疇重要性等級的劃分方法，給出語義場關聯模式的優先性等級。

語料庫中的詞，有一些已爲語義詞典所描述，還有一些則未得到描述。已描述詞語的語義類可直接繼承，未描述詞語的語義類則需要通過詞性及構成成分的語法功能和語義類別來計算。在這一步驟中，將會總結出名詞性複合詞語義關聯模式的計算方法。複合詞的構成模式其實也就是複合詞的理解模式，比如“赤狐”是“顏色”語義場與“走獸”語義場組合而成的。通過統計複合詞的構成模式可以計算語義場之間的關聯度，經常在一起構造複合詞的兩個語義場之間的關聯度高，反之則低。

知網中的每一個詞均標注了語義類，比如“蒼白、赤紅、白、藍”爲顏色類，“豺、狼、豹、豺狼”爲走獸類，我們將各詞的語義類作爲語義場。我們的計算基於以下假設：(1) 雙字詞各成分的語義場必然屬於知網所描述的語義場中的一個，比如“后”（簡體字）在知網中屬於四個語義場，即“次序、時間、位置、人”，則“后漢”中“后”的語義場必然是這四個語義場中的一個；(2) 同一語義場內的兩個雙字詞，作爲其成分的單字詞可能分別屬於多個語義場，如果兩者的前後字分別屬於一個相同的語義場，則這一相同的語義場即應該是這一雙字詞的語義場關聯模式。比如“紅狐、黑熊”均屬於走獸語義場，其中的“紅、黑”均屬於多個語義場，但它們有一個共同的語義場爲“顏色”，則“紅、黑”在兩詞中的語義場應該是“顏色”。

按照這一處理方法，從知網中可以得到 1,101 個語義場。“走獸”語義場做後字時最常見的關聯模式爲“走獸｜顏色｜水域｜舉止＋走獸”。

語義場關聯度的計算涉及到一些複雜的問題，比如如何判斷一個新詞或未登錄詞內部的語義場關聯模式，我們將另文討論。

基於上述假設，我們計算出下列模式爲常見的語義場關聯模式：

〈表 3〉常見語義場關聯模式

| 雙字詞<br>語義場 | 前字<br>語義場 | 後字<br>語義場 | 詞頻 |
|---|---|---|---|
| 人 | 人 | 人 | 388 |
| 時間 | 時間 | 時間 | 225 |
| 用具 | 用具 | 用具 | 149 |
| 動物 | 動物 | 動物 | 143 |
| 地方 | 地方 | 地方 | 119 |
| 人 | 舉止 | 人 | 100 |
| 人 | 地方 | 人 | 82 |
| 材料 | 材料 | 材料 | 55 |
| 人 | 時間 | 人 | 44 |
| 設施 | 設施 | 設施 | 41 |
| 時間 | 數量值 | 時間 | 41 |
| 人 | 數量值 | 人 | 40 |
| 時間 | 舉止 | 時間 | 40 |
| 人 | 年齡 | 人 | 39 |
| 用具 | 舉止 | 用具 | 38 |
| 動物 | 動物 | 物形 | 36 |
| 人 | 場所 | 人 | 35 |
| 事情 | 舉止 | 事情 | 35 |
| 材料 | 舉止 | 材料 | 35 |
| 用具 | 材料 | 用具 | 34 |
| 地方 | 人 | 地方 | 34 |
| 事務 | 事務 | 事務 | 32 |
| 魚 | 魚 | 魚 | 32 |
| 陸地 | 陸地 | 陸地 | 32 |
| 衣物 | 衣物 | 衣物 | 31 |
| 事情 | 事情 | 事情 | 29 |
| 用具 | 動物 | 用具 | 29 |

| 雙字詞<br>語義場 | 前字<br>語義場 | 後字<br>語義場 | 詞頻 |
|---|---|---|---|
| 疾病 | 動物 | 疾病 | 29 |
| 設施 | 人 | 設施 | 28 |
| 事情 | 時間 | 事情 | 28 |
| 位置 | 位置 | 位置 | 28 |
| 土石 | 土石 | 土石 | 28 |
| 語文 | 語文 | 語文 | 26 |
| 用具 | 用具 | 物形 | 26 |
| 樹 | 樹 | 樹 | 24 |
| 境況 | 境況 | 境況 | 24 |
| 人 | 人 | 動物 | 23 |
| 情感 | 情感 | 情感 | 23 |
| 設施 | 舉止 | 設施 | 22 |
| 人 | 用具 | 人 | 22 |
| 牲畜 | 人 | 牲畜 | 22 |
| 團體 | 人 | 團體 | 22 |
| 衣物 | 時間 | 衣物 | 22 |
| 衣物 | 舉止 | 衣物 | 22 |
| 用具 | 用具 | 動物 | 22 |
| 語言 | 地方 | 語言 | 22 |
| 地方 | 數量值 | 地方 | 22 |
| 時間 | 人 | 時間 | 21 |
| 情感 | 舉止 | 情感 | 21 |
| 人 | 位置 | 人 | 21 |
| 走獸 | 走獸 | 走獸 | 21 |
| 動物 | 人 | 動物 | 20 |
| 場所 | 時間 | 場所 | 20 |

## 4. 複合詞理解難易度的計算

在上述工作的基礎上計算複合詞理解的難易度，可以綜合考慮以下三個方面的因素：

一、單字表示的語義範疇重要性等級。同一語義場的詞所表示的語義範疇可以依據其重要性劃分爲不同層級，這主要通過構詞頻率來體現。

二、語義場關聯度。複合詞的構成模式其實也就是複合詞的理解模式，比如“赤狐”是因“顏色”語義場與“走獸”語義場之間存在關聯而成詞的。通過統計複合詞的構成模式可以計算語義場之間的關聯度，經常在一起構造複合詞的兩個語義場之間的關聯度高，反之則低。

三、雙字複合詞整體所表語義範疇在相關語義場中的位置及重要性等級。如果該詞在語料庫中未出現，說明該詞不常見，設其等級係數爲 0。

以上三個係數，各分四個層級，依次爲 0、1、2、3，難易度係數值域爲 0-9，值越大越容易理解。這一係數是相對的，在一個連續數值域中分段可粗可細，本文主要考慮在同一個語義場範圍內的各詞之間理解難易度的區別。如果在不同語義場成員之間進行比較的話，則可以以此爲基礎，進一步考慮不同語義場之間的區別。

以“銀狐”爲例，“銀”和“狐”單字語義範疇重要性等級係數分別爲 2、2；其語義場關聯模式爲“顏色＋走獸”，等級係數爲 3；“銀狐”在語料庫中未出現，整體的重要性等級係數爲 0。則其難易度係數爲 (2+2)/2+3+0=5。

三個係數事實上是不平衡的，對於理解而言，最重要的還是語義場關聯係數，這是語義理解的核心。

根據以上研究可以把複合詞難易度的計算公式確定如下：

**$(X_1+X_2)/2+Y+Z$。**

以上公式中，$X_1$ 表示前字所表語義範疇重要性等級係數；$X_2$ 表示後字所表語義範疇重要性等級係數；Y 表示語義場關聯模式的等級係數；Z 表示複合詞整體的重要性等級係數。得分高的表示較容易理解，得分低的表示較難於理解。下面給出動物義場複合詞難易度計算的一組實例：

〈表 4〉複合詞理解難易度示例

| 詞語 | 語義場關聯模式 | 計算 | 得分和分級 |
|---|---|---|---|
| 青蛙 | 顏色＋走獸 | (3+2)/2+3+3 | 8.5 |
| 狐狸 | 走獸＋走獸 | (2+2)/2+3+3 | 8 |
| 蛟龍 | 走獸＋走獸 | (2+3)/2+3+2 | 7.5 |
| 蟒蛇 | 走獸＋走獸 | (2+3)/2+3+2 | 7.5 |
| 黑熊 | 顏色＋走獸 | (3+3)/2+3+1 | 7 |
| 水獺 | 水域＋走獸 | (3+1)/2+3+1 | 6 |
| 貂熊 | 走獸＋走獸 | (1+3)/2+3+0 | 5 |
| 麋鹿 | 走獸＋走獸 | (0+3)/2+3+0 | 4.5 |
| 鼬獾 | 走獸＋走獸 | (0+1)/2+3+0 | 3.5 |
| 貔貅 | 走獸＋走獸 | (0+0)/2+3+0 | 3 |

顯然，上述例子顯現出的難易度分級是初步的、粗略的，也是相對的，隨著研究的進展還有一定的調整的必要。

## 5. 結語：問題、應用價值和理論價值

跟問題的高度複雜性相比較，本文對漢語複合詞難易度的計算無疑是初步的，雖然複雜問題得到了一定程度的簡化，但是遺留的問題還有不少，有些問題給計算帶來了很大的麻煩。例如，“后”表示不同意義，但對計算機來說形式上一樣的，在計算時必須考慮怎麼樣將它們區分開來。又例如，一個字出現在複合詞的前後位置時語義功能可以有大的差別，如“狐”在“狐媚”和“銀狐”中有不同的意義。有時還會遇到同一個字所表示的各種意義有很大差別的情況，例如“商”所表示的意義分屬人、音樂和時間三個不同的語義場。上述幾種情況是比較多見的，要進行準確歸類和計算還需要依賴基礎材料的整理，但現在整理還是很粗略的。

還有兩類給計算帶來困難的問題是跟複合詞相關的意義的轉化問題和意義的比喻引申問題，這兩方面的問題在基礎理論研究上才剛剛起步，計算還一時難以處理。例如：在後字位置上由性狀類轉化爲名物的，如“幼兒”和“老幼”。又例如兩個屬於身體語義場的單字並列表示人的抽象隸屬物的實例（王洪君 2005），如“肺腑、肝膽、血汗、臉面、嘴臉”等。要確定此類複合詞的難度係數會困難一些。

如何判斷一個新詞或未登錄詞內部的語義場關聯模式，例如知網中沒有的新出現的複合詞，如“說吧、哭吧、貼吧”等，也是一個需要仔細研究的問題，這需要另文討論。

語料選擇的偏差會影響頻率和語義範疇重要性等級的確定，比如本項研究選擇的是人民日報語料，屬於書面語料，範圍較狹窄，種類較單一，局限是明顯的。以後的研究還需要進一步擴大語料庫的規模和語料的種類，以取得必要的平衡。

本文主要依據字頻和構詞頻率定性並定量地給出語義範疇重要性層級的劃分標準，還考察了雙字複合詞整體在語義場中的重要性層級，在此基礎上統計出了語義場關聯模式的優先度列表。最後給出了複合詞理解難易度的計算公式及名詞性複合詞語義模式的計算方法。本項研究對句子難易度的計算會有一定幫助，在句子當中，單字詞難易度與複合詞的難易度在一定程度上也影響著句子理解的難易度。研究成果對自然語言信息處理中的詞義消歧研究也有一定的參考價值，通過確定新詞或未登錄詞內部的語義場關聯模式，有助於理解其意義。研究成果還可以爲語言教學中編寫有用的較爲科學的詞表提供必要的參考。在理論上，本項研究有助於加深對語義場內部結構以及不同語義場之間關聯模式的認識，最終會有助於加深對漢語複合構詞機制的認識。本項研究工作還需要進一步充實和改進，有不少重要問題值得繼續討論。

## 引用文獻

Lakoff, George. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: University of Chicago Press.

Lakoff, George, and Mark Johnson. 1980. *Metaphors We Live By.* Chicago: University of Chicago Press.

Lyons, John. 1977. *Semantics*. Cambridge & New York: Cambridge University Press.

Ullmann, Stephen. 1957. *The Principles of Semantics*. Oxford: Blackwell.

王洪君. 2005.〈動物、身體兩義場單字及兩字組轉義模式比較〉,《語文研究》2005.1:4-8。

王還等. 1986.《現代漢語頻率詞典》。北京：北京語言學院出版社。

邱立坤. 2005.〈現代漢語動名語串結構關係的判定〉,第六屆漢語詞彙語義學研討會論文。廈門：廈門大學。

俞士汶, 黃居仁主編. 2005.《計算語言學前瞻》。北京：商務印書館。

梅家駒, 竺一鳴, 高蘊琦, 殷鴻翔編. 1984.《同義詞詞林》。上海：上海辭書出版社。

陳保亞. 1999.《20 世紀中國語言學方法論》。濟南：山東教育出版社。

葉文曦. 2004a.〈漢語語義範疇的層級結構和構詞的語義問題〉,《語言學論叢》29:95-109。北京：商務印書館。

葉文曦. 2004b.〈語義範疇的重要性等級和漢語單字的語義功能〉,未刊稿。

鄭錦全. 2005.〈詞彙語義與句子閱讀難易度計算〉,第六屆漢語詞彙語義學研討會論文。廈門：廈門大學。

## 網路資源

《人民日報》(2000 年上半年）標注語料庫：http://icl.pku.edu.cn

知網 2000 版：http://www.keenage.com

[Received 22 December 2006; revised 26 August 2007; accepted 1 October 2007]

葉文曦
北京大學中文系
中國 100871 北京市海淀區
ywx@pku.edu.cn

# Computing the Degree of Difficulty in Understanding Chinese Compounds

Wenxi Ye
*Peking University*

Likun Qiu
*Beijing City University*

This article provides criteria for differentiating the importance of semantic category according to the frequency of character use and the frequency of character use in compound formation. The article also observes the position of the semantic categories expressed by compound as one whole in the related semantic fields and its importance level. Based on the above work, the article finally provides a computational formula of degree of difficulty in compound understanding and the computational method of the semantic pattern in nominal compounds. This study contributes to the compiling of a word list in language teaching and the computation of the degree of difficulty in sentence understanding. This research method and its results may contribute to a deeper understanding of the internal structure of the semantic field and the internal semantic structure of compounds.

Key words: lexical semantics, complex word formation, semantic field, degree of difficulty in semantic understanding, computational linguistics