LANGUAGE AND LINGUISTICS 9.2:359-379, 2008
2008-0-009-002-000064-1

漢語詞彙語義研究及詞彙知識庫建設*

俞士汶　朱學鋒　段慧明　吳雲芳　劉　揚
北京大學

信息技術，特別是 Internet 上搜索引擎技術的需求爲漢語詞彙語義的研究拓展了廣闊的空間。本文論述詞彙語義研究在自然語言處理技術中的意義以及自然語言處理技術對詞彙語義研究所提出的新要求；本文重點介紹北京大學計算語言學研究所面向自然語言處理研製中文詞彙知識庫的成果：(1) 現代漢語語法信息詞典，(2) 現代漢語語義詞典，(3) 中文概念詞典，(4) 現代漢語詞義標注語料庫。本文最後介紹將這些獨立存在的詞彙知識庫集成爲綜合型語言知識庫的構想、技術方案和已經取得的階段性進展。

關鍵詞：漢語詞彙語義學，詞彙知識庫，現代漢語語法信息詞典，現代漢語語義詞典，中文概念詞典，現代漢語詞義標注語料庫，綜合型語言知識庫

1. 引言

自第一屆中文詞彙語義學研討會於 2000 年舉辦以來，中文詞彙語義學研究的發展確實令人鼓舞。六年的歷程證實了鄭錦全院士的高瞻遠矚。信息技術，特別是 Internet 上搜索引擎技術的需求爲中文詞彙語義學的研究拓展了廣闊的空間。這也是本系列研討會得以不間斷召開並且成果逐年彰顯的原動力所在。

本文概述作者對面向語言信息處理的詞彙語義學的基本認識，重點介紹北京大學計算語言學研究所（Institute of Computational Linguistics, Peking University，簡稱 ICL/PKU）在詞彙知識庫建設方面所做的工作以及所積累的一些經驗。本文最後介紹將幾個獨立存在的詞彙知識庫集成爲綜合型語言知識庫的構想、技術方案和已經取得的階段性進展。

* 本文相關研究得到中國國家 973 項目 (2004CB318102)、國家“八六三”高技術研究發展計畫基金項目 (2007AA01Z198) 和國家自然科學基金項目 (60503071, 60775031) 的支持。

2. 詞彙語義學研究與自然語言處理

在面向人的語言學本體研究與語言教學研究領域，詞彙語義研究的歷史源遠流長。訓詁學的研究重點就是對古書的字句（包括現代意義的“詞”）的含義進行解釋，《說文解字》、《廣雅》、《爾雅》等經典著作就包含了豐富的詞彙語義研究成果。現代漢語詞彙語義學研究在中國大陸也有諸多論著問世（蘇新春 2001，蘇寶榮 2000，符淮青 2004，張志毅、張慶雲 2001）。語言學家的成果爲面向信息處理的詞彙語義研究，提供了理論基礎。

在面向語言信息自動處理的計算語言學領域，詞彙語義研究也是最早發展的分支之一。無論早期的機器翻譯採用“詞對詞”的直接模式，還是後來一度盛行的轉換模式，乃至當今大行其道的統計方法，詞義消歧 (Word Sense Disambiguation) 總是這個實用領域首當其衝的基礎研究課題。在 Internet 迅速擴張的今天，詞義消歧也是提高信息檢索效率的最重要的技術手段。詞彙語義學是詞義消歧的理論基礎，而詞義消歧正是詞彙語義學研究成果的用武之地。

本文只聚焦於面向自然語言處理的詞彙語義研究。在開始這項研究的時候以及在這項研究的全過程中，都要清醒地認識和妥善處理以下一些基本問題。

自然語言處理需要的詞彙語義知識有什麼樣的特點？這些知識的源頭在哪裡？如何把計算機需要的詞彙語義知識組織、表示爲計算機程序可以運用的形式？

計算語言學研究者都了解語言知識庫是自然語言處理系統不可或缺的組成部分，語言知識庫的規模和質量在很大程度上決定了自然語言處理系統的成敗。把處理程序和語言規則分開是機器翻譯技術進步中的一個里程碑。不過，在早期採用這種技術的系統中，語言知識庫基本上就是語言規則庫。語言規則主要是以基於上下文無關語法 (Context Free Grammar) 的產生式規則 (Production Rule) 或其擴展形式表述的，在這些把規則中也融入了大量的有關詞語的個性知識。規則庫無可避免地急速膨脹，龐雜無序，難以管理。1980 年代，隨著以複雜特徵集 (Complex Feature Set) 和合一運算 (unification) 爲技術特徵的一批新型計算語言學語法理論（如詞彙功能語法 Lexical Functional Grammar，中心語驅動的短語結構語法 Head-Driven Phrase Structure Grammar 等等）的問世，自然語言處理技術對詞彙知識的要求越來越精細，在應用系統中，便把詞彙知識從規則庫中分離出來，構造獨立的詞彙知識庫。這樣，語言知識庫就可劃分爲句法規則庫和詞彙知識庫兩大部分，而詞彙知識庫是語言知識庫的重心。

儘管計算機程序面對的只是知識庫中的以二進制編碼形式呈現的字符串，並不理會哪些是詞法知識和句法知識，哪些是語義知識。不過知識庫總是要由人來

建設和管理的，不妨遵循語言學關於語言知識的劃分，可以對詞彙知識庫中的知識作詞法、句法和語義的區分。

面向語言信息處理的詞彙語義研究理所當然地以建設詞彙語義知識庫爲己任。

詞彙語義知識的源頭是人們使用的話語和創作的文本，語言學家的論著和語文詞典是這些知識的集大成者。不過，面向人的論著和詞典，是當代的計算機還理解不了或不便應用的。從事計算語言學研究的學者要善於從這些語料和論著中彙集計算機所需要的知識，並把它們改造成計算機可以應用的形式。計算機可以操作的語言知識的數據結構也是多種多樣的，諸如：關係數據庫中的二維表（每個文件由包含若干字段的諸多記錄組成）、樹 (tree)、框架 (frame)、向量 (vector)、矩陣 (matrix) 乃至表述“條件–動作”的規則等等。計算語言學的工作就是要把各種類型的詞彙語義知識裝配到適當的數據結構中。

計算機沒有舉一反三、觸類旁通的能力。除非受制於存儲空間或計算時間的約束，詞彙語義知識庫必須無一遺漏地描述所收入的每一個詞的詳細語義知識，其重要特徵是規格化和形式化。漢語的詞彙語義研究，像黃居仁教授辨析“聲”與“音”、“看”與“見”的異同（黃居仁、洪嘉馡 2006），並給出了合理的解釋，很深入，這些知識對自然語言理解的研究當然是重要的，但是從當前計算機處理語言信息的整體水平來看，計算機系統更需要建設具有大覆蓋面的知識庫，庫中所彙集的詞彙語義知識的深度要適中，且要考慮每一項具體語言工程的有限投入和有限目標。

經過近 20 年的努力，中文信息處理學界已經建設了若干頗有影響而且也發揮了實際作用的漢語詞彙語義知識庫，如中國大陸的 HowNet[1]，述語動詞詞典（陳群秀 1995），台灣的 BOW[2] 等等。

環視一下書架上琳琅滿目的語文詞典，可以斷言，現在機器可用的漢語詞彙知識庫不是太多了，而是太少了。在漢語詞彙語義研究以及漢語詞彙知識庫的建設中，要提倡多樣化。詞彙知識庫的多樣化表現在規模、深度、專業領域、知識範疇的各有不同。當然，也要避免同水平重複勞動造成的浪費。爲此，加強學術交流就是十分必要的，既交流成果、信息，也交流經驗、心得。漢語詞彙語義學的系列研討會 (CLSW) 提供了這樣一個交流的平台。

[1] http://www.keenage.com/

[2] http://BOW.sinica.edu.tw/

3. 北大詞彙語義知識庫概要

ICL/PKU 於 1986 年成立伊始，便著手語言數據資源的建設。在各種語言知識中，詞彙知識是最基本的。ICL/PKU 把詞彙知識庫看作是語言知識庫的主體，經過近 20 年的努力，在詞彙語義知識庫方面已有如下積累：

一、現代漢語語法信息詞典 (Grammatical Knowledge Base, GKB)（俞士汶、朱學鋒等 2003）
二、現代漢語語義詞典 (Chinese Semantic Dictionary, CSD)（王惠等 2003）
三、中文概念詞典 (Chinese Concept Dictionary, CCD)（于江生等 2003）
四、現代漢語詞義標注語料庫 (Chinese Semantic Corpus, CSC)

建設語言數據資源，首先要選取適當的語言單位作爲著力點。語言單位的選取要服從於應用目標，而應用系統的設計與實現又要受當時可採用的計算機硬軟件的制約。1980 年代中期之後，以“詞”作爲漢語的“輸入–變換”單位逐漸成爲中文鍵盤輸入的主流，基於“分析–轉換–生成”模式的機器翻譯系統採用的對譯單位基本上也是“詞”。縱觀語言信息處理的全局，語言知識庫的建設從“詞”起步是恰當的，也就是首先要建設好詞彙知識庫。

語言學從“聚合關係 (Paradigmatic Relation)”和“組合關係 (Syntagmatic Relation)”兩個視角觀察、彙集、組織詞彙語義知識。組合關係關注於詞與詞如何結合成更大的語言單位。聚合關係主要描述詞語的分類體系。在詞彙知識庫中，每一個詞都要歸於一個適當的類別，屬於同一類的詞具有某些相同的屬性或功能，但分類不宜過細，屬於同一類的詞必然仍有相互區別的屬性或功能，因此更需要逐類分別描述每一個詞的特徵，即各個詞特有的屬性或功能。

GKB，CSD 和 CCD 都是在上述原則指導下建立的詞彙知識庫。GKB，CSD 都採用關係數據庫二維表的數據結構，詞作爲各個數據庫文件 (file) 的每個記錄 (record) 的登錄項 (entry)。文件中設立了很多字段 (field) 描述詞的特徵屬性。CCD 是與英語 WordNet 兼容的中文詞彙知識庫，用同義詞的集合 (synset) 描述“概念 (concept)”的定義，而且進一步描述概念和概念之間的各種關係。儘管 GKB，CSD 和 CCD 都涉及了詞的聚合信息和組合信息，不過，從總體上看，GKB 和 CSD 更偏重於詞的組合信息，而 CCD 則更多地反映詞的聚合信息。GKB 是單語的，只包含漢語詞彙的知識，而 CSD 和 CCD 都是雙語的，包含了漢語和英語的詞的對譯知識。

GKB，CSD 和 CCD 這類詞庫，相當於語文詞典，關於語義知識的表達都是顯性的，但也是靜態的，存在“不確定性”（即一詞多義或歧義）。這種不確定性在文本中出現時要依靠其他手段排除。在真實的文本語料中，詞的每次出現都有一定的語境，其詞義、句法功能、語義角色都是確定的，但表現方式卻是隱性的，語料加工就是使隱含的信息顯性化，加工越深，顯性化的信息就越多。在完成詞語切分和詞性標注的基本標注語料庫（俞士汶等 2003）中，詞語及其詞性顯性化了，在詞義標注語料庫 CSC 中對每個詞都標注它在當前語境中的義項，詞義也就顯性化了。

語料加工可以依靠專家的手工作業，也可以依靠機器自動進行，現在廣泛採用的技術手段還是人和機器的相互配合。

4. 現代漢語語法信息詞典 GKB

《現代漢語語法信息詞典》是 ICL/PKU 建設的語言知識庫大廈的第一塊基石。其成功得益於北大中文系朱德熙、陸儉明、郭銳教授早期的合作與指導。GKB 以詞的句法信息翔實著稱於學界，其實，GKB 也有豐富的詞彙語義知識。GKB 至少在以下幾方面包含了詞彙語義知識。

一、區分了詞的粗粒度義項

GKB 的各個數據庫文件的主關鍵項（數據庫中每個記錄的唯一標識，即 ID）都是：“詞語”＋“詞類”＋“同形”。對於“同形”這個字段的含意需要做些解釋。如果數據庫中一個記錄中的“詞語”與其他記錄中的“詞語”至少有一個漢字不同，那麼“詞語”就可以作爲該記錄的主關鍵項。不過，漢語中有漢字相同而詞類不同的詞，就要用“詞語”＋“詞類”作爲主關鍵項。如“地道 (di4dao4)”是名詞 n，“地道 (di4dao5)”是形容詞 a，“鎖 (suo3)”只有一個讀音，也分屬兩類詞（“門上的鎖”的“鎖”是名詞 n，“鎖好房門”的“鎖”是動詞 v），像這樣的情況，詞類的不同也蘊含了詞義的不同。進而，屬於同一詞類的漢字相同的詞仍可能是不同的詞或者需要區分爲不同的“義項”。在影響廣泛的《現代漢語詞典》（中國社會科學院語言研究所詞典編輯室 2005）中，採用不同的編排方式區分“詞項”和“義項”。在數據庫中，僅用“詞語”＋“詞類”就不足以區分了，需要另加一個字段“同形”：對於讀音不同的情況，如“挨”作爲動詞，也有兩種讀音：“ai1”和“ai2”，“同形”分別填“A”和“B”；還有讀音一樣的，如動詞“抄”也有兩個不同的詞項，“同形”也還是分別填“A”和“B”。對於“義項”不同的情況，如動詞“保管”要區分爲兩個不同的

“義項”（分別爲“保存”和“擔保”），這時同形字段分別填“1”、“2”。簡單來講，“A、B”大致相當於語言學上同形層面的義項區分 (homograph)，“1、2”類同於義項層面 (sense) 的區分。

一個詞的義項區分很不容易把握，在研製 GKB 的前期，除了參考《現代漢語詞典》外，GKB 區分義項遵循一個原則：區分了義項的，其語法功能也一定有所區別。像“保管”之所以區分爲兩個不同的“義項”，是因爲這兩個義項不同的“保管”的語法功能也有顯著區別：作爲“保存”義的，只能帶體詞性賓語，而作爲“擔保”義的卻可以帶謂詞性賓語。對於那些即使意思不同但語法功能沒有明顯區別的（如“沖”，可區分爲“沖膠捲的沖”，“沖咖啡的沖”和“沖盤子的沖”），GKB 就未作區分，在這個意義上講，GKB 對詞的義項的區分只是粗粒度的。

GKB 總庫包含 8 萬記錄，其中約有 1,100 組詞語區分了同形信息，佔記錄總數不到 3%。〈表 1〉是 GKB 總庫的記錄的抽樣示例（“詞語”＋“詞類”＋“同形”是主關鍵項）。

二、在體賓動詞分庫中，指明了及物動詞的體詞性賓語可能擔任的語義角色（語義格）以及各種語義格的格標記。如對於“告訴”，描述了它的賓語可能是受事（“告訴大家一個好消息”中的“一個好消息”），可能是與事（“告訴大家一個好消息”中的“大家”），而且受事的格標是介詞“把”（“把這個好消息告訴大家”中的“這個好消息”是受事，“把”是格標）。動詞“坐”的賓語，還可以是“施事”（如：“前排坐嘉賓”中的“嘉賓”就是動詞“坐”的施事）。

三、從名詞庫中分化出來的時間詞、處所詞乃至時間詞庫中的“時態”字段以及語素庫中的“姓氏”、“人名”、“地名”、“水名”等字段都給機器提示了語義信息。

〈表 1〉GKB 總庫的記錄的抽樣示例

詞語	詞類	同形	拼音	釋義	-----
挨	v	A	ai1	觸，碰，靠近	
挨	v	B	ai2	遭受，忍受	
安裝	v		an1zhuang1		
保管	v	1	bao3guan3	保存	
保管	v	2	bao3guan3	擔保	
抄	v	A	chao1	照原稿寫	
抄	v	B	chao1	走近道	
地道	a		di4dao5		
地道	n		di4dao4		

鎖	n		suo3	門上的鎖
鎖	v		suo3	請鎖好門
儀表	n	A	yi2biao3	儀器設備
儀表	n	B	yi2biao3	外貌風度

四、至於各個庫中的“釋義”（原稱“義項”，這個字段將更名）、“備註”字段都含有的語義知識，也可利用它們爲機器學習（如：詞義自動消歧）提供參考信息。

無需諱言，GKB 的重心是在句法信息上，對語義信息的描述屬於試驗性質。有關詞彙的更多的語義信息由 ICL/PKU 的其他詞彙知識庫提供。

5. 現代漢語語義詞典 CSD

構建一部現代漢語語義詞典首先需要回答的問題是：如何在詞典中有效地描述詞語的詞彙語義知識？在詞典中對詞語意義的描述要有利於真實本文中或者言談交際時詞義歧義的消解，要有利於詞義知識的推理。現有的詞語意義表徵主要有以下四種方式。一、意義訓釋 (meaning definition)，用自然語言的語句、詞語來定義一個詞語的不同意義；二、語義原語 (semantic primitive)，將一個複雜的意義進行分解 (decomposition)，分解成爲易於理解的、不可再分的若干簡單意義，組織這些簡單意義以描述原來的複雜意義；三、語義關係 (semantic relation)，表達不同詞語意義之間的各種各樣的關係，例如同義（近義）關係、上下位關係、部分整體關係、反義關係等；四、框架網絡 (FrameNet)，詞的意義描述必須與某個語義框架相聯繫，框架網絡的分析單元不是一個個的詞，而是一個個的框架，每個分析單元都力圖涵蓋該框架所引發的所有詞元 (lexical unit)。上述四種詞義表徵方式都沒能充分地描述詞語分布的上下文環境，而人和計算機都是利用語境來理解詞語意義的。Véronis 做了一個有趣的實驗：把 600 個詞形分配給 6 名語言學專業的學生，由他們依據傳統辭書中的釋義對真實語料進行詞義標注。結果發現，不同標注者之間的一致性非常低，對於某些詞語，不一致性甚至和隨機標注一樣糟糕 (Véronis 2003)。由此認識到，應該建立一種適合於計算機（和人）進行詞義理解的詞語多義表徵形式，這種形式應該能夠容納豐富而又不冗餘的上下文信息。

《現代漢語語義詞典》(CSD) 繼承了《現代漢語語法信息詞典》(GKB) 的數據模式。當代若干計算語言學語法理論（詞彙功能語法 LFG，中心語驅動的短語結構語法 HPSG，等等）都以採用複雜特徵集 (Complex Feature Set) 的詞彙知識

表示和基於合一 (unification) 的分析算法爲特徵。CSD 和 GKB 就是在這些語法理論的啓示下，採用在大致分類的基礎上，以“屬性–屬性值”的形式詳細描述詞語的句法、語義知識。爲了採用成熟的關係數據庫技術而又便於語言學家直接參與詞典的構建，又將“屬性-屬性值”的描述形式轉換爲數據庫二維表的字段與值。CSD 依據詞義理解的需要設定多個不同的特徵屬性，依據屬性值的不同即可辨別出不同的義項，而且 CSD 描述的語義知識和 GKB 描述的句法知識採用統一的描述形式，便於實現“句法–語義接口”(syntax/semantic interface)。CSD 完全繼承 GKB 的“詞語”、“詞類”、“同形”這三個字段的信息，但“詞語”＋“詞類”＋“同形”不足以作爲 CSD 的主關鍵項，在 CSD 中，增加了“義項”字段，“詞語”＋“詞類”＋“同形”＋“義項”才是 CSD 的主關鍵項，“同形”和“義項”兩個屬性字段共同構成一個詞語的意義編碼。“義項”是在“同形”的基礎上所做的更爲細緻的意義區分。“同形”可看作是粗粒度的詞義區分 (coarse-grained sense)，“義項”是細粒度的詞義區分 (fine-grained sense)。CSD 對於動詞，還設置了“語義類、子類框架、配價數、主體、客體”等多個屬性字段。王惠等 (2003) 還沒有關於“子類框架”的介紹，這個字段是在開發詞義標注語料庫的過程中新設置的，故在此作簡單說明：子類框架 (subcategory frame) 是指在動詞性短語內動詞所能組合的姊妹節點，本質上就是動詞對賓語的語法選擇。

〈表 2〉就是 CSD 對動詞“開”的不同義項的描述。

以“屬性–屬性值”形式表徵詞語意義時，所有特徵屬性構成的集合才形成一個計算機可以利用的關於詞義的定義，而“釋義”中用自然語言對意義的描述僅僅是一種參照（目前供人參照，將來或許機器也可以參照）。另一方面，借助基於詞義的特徵屬性描述，又可以定義義項之間的區別性特徵 (Distinguishing Features, DF)，計算機依據這些區別性特徵可以進行詞義辨識，從而自動完成詞義消歧的任務（或許人也可以利用，特別是在對外漢語教學的某些場合）。下面給出 DF 的形式化定義。

DF 定義：設詞 W 可區分爲 n 個義項 S_1，S_2，……S_n $(n > 1)$，義項 S_i 有複雜特徵集

$$S_i \begin{bmatrix} f_1 = v_{1i} \\ f_2 = v_{2i} \\ \cdots\cdots \\ f_m = v_{mi} \end{bmatrix} (m \geq 1)$$，另一個義項 S_j 也存在相同的特徵屬性項，並且 S_j $(f_k = v_{ki})$，

若 $v_{ki} \neq v_{kj}$，則稱 f_k 是 S_i 和 S_j 的區別性特徵，且 $f_k = v_{ki}$ 是 S_i 對 S_j 的區別性特徵值，對應的 $f_k = v_{kj}$ 是 S_j 對 S_i 的區別性特徵值。

例如在〈表 2〉中，“語義類”、“子類框架”、“配價數”、“主體”都構成“開 1-1”和“開 1-2”的區別性特徵，“語義類”、“客體”是“開 1-3”和“開 1-4”的區別性特徵。（“開 1-1”指〈表 2〉中“同形”和“義項”字段的值都是“1”的那個“開”，其他類推。）

吳雲芳、俞士汶 (2006) 探討了面向漢語語言信息處理的詞語義項區分應該遵守的原則和方法。面對大規模真實文本，詞語義項區分應具有可操作性，即應具有完備性和離散性。所謂“完備性”是指，根據詞語義項劃分，操作者（計算機或者人）可以對規模足夠大的語料中的每一個目標詞標注出義項，即不遺漏語料中出現的每一個義項。所謂“離散性”是指，根據詞語義項的劃分，給定充足的上下文環境，操作者（計算機或者人）可以順利地對語料中的每一個目標詞標注出一個確定的義項。信息處理用詞語義項的區分應該是“實證型”的，即語言學者或是計算機可以依據詞語所在的上下文尋找出可清晰辨析詞義的、可進行形式化描述的證據，這就是詞語的句法行爲 (syntactic behavior)，而與詞義相關的世界知識在詞義區分中則顯得不是那麼重要或暫時用不上。目前，《現代漢語語義詞典》(CSD) 總庫包含 6 萬餘條記錄。在細粒度詞義區分層面上，目前已對 CSD 中的 170 個動詞、796 個名詞進行了仔細的詞義辨析和描述。所完成的工作量是相當大的。

〈表 2〉《現代漢語語義詞典》示例

詞語	同形	義項	釋義	語義類	子類框架	配價數	主體	客體	WORD	特殊句法位置
開	1	1	打開	其他行爲	[NP]	2	人		open	
開	1	2	展開，分開	變化	[~]	1	～人		open out	
開	1	3	支付；開銷	領屬轉移	[NP]	2	人	錢財	discharge	
開	1	4	發動或操縱	其他行爲	[NP]	2	人	交通工具｜武器	drive	
開	1	5	開辦	創造	[NP]	2	人	建築物	open	
開	1	6	舉行	社會活動	[NP]	2	人	事件	hold	
開	1	7	寫出	創造	[NP]	2	人	票據	write	
開	1	8	開創（抽象事物）	創造	[NP]	2	人	抽象物	open	
開	1	9	打開（電器）使運作	其他行爲	[NP]	2	人	電器	turn on	

開	1	10	開闢	其他行爲	[NP]	2	人	地理	open up	
開	1	11	開通	社會活動	[NP]	2	人	符號	open up	
開	1	12	使開闊	其他行爲	[NP]	2	人	心理特徵	open up	
開	2		開始	其他行爲	[NP]	2	人	事件	begin	
開	3		用在動詞或形容詞後作補語							A＋～｜V＋～｜V＋得＋～｜V＋不＋～

6. 中文概念詞典 CCD

第 3 節已概要介紹，“現代漢語語義詞典”(CSD) 和“中文概念詞典”(CCD) 都是詞彙語義知識庫，且都是雙語的，包含了漢語和英語的詞的對譯知識。除了結構和知識表達形式有所不同，CSD 和 CCD 在應用層面上也是互補的。

CSD 適用於漢外機器翻譯。請看下面兩個例句：

(1) 她的儀表很精密。
(2) 她的儀表很端莊。

如果要譯成英語，自然要進行詞義消歧，因爲兩個例句中的“儀表”是不同的詞。這兩個句子的句法形式完全一樣，對“儀表”的詞義消歧沒有作用，只能依據其後做謂語的形容詞“精密”和“端莊”的語義選擇特性。在 CSD 的形容詞庫中，也設立了表徵形容詞語義選擇特性的字段，如“主體”。根據“精密”和“端莊”的“主體”字段中信息的不同，可以判斷例 (1) 中的“儀表”是指電錶等人工物，例 (2) 中的“儀表”是指人的外貌風度。

漢外翻譯一方面提出了詞義消歧的需求，另一方面也可視爲檢驗詞義消歧的手段。不過，詞義消歧的應用不限於機器翻譯，譯詞選擇也不是檢驗詞義消歧的充分手段。詞義消歧在（跨語言）信息檢索或信息提取中也會廣泛用到。像“病毒”這個詞，對應的英文也是一個 virus，但在當前社會生活中，“病毒”指稱的兩個概念“生命體”或“惡意代碼”都是常用的。對於（跨語言）信息檢索或信息提取研究，區分這兩個概念是十分必要的。ICL/PKU 開發的第三個大型詞彙知識庫“中文概念詞典”(CCD) 可以爲這類消歧提供必要的詞彙語義知識。

CCD 呈現給用戶的基本數據結構是若干棵樹 (tree)，樹上的每個結點 (node)

都是一個概念 (concept)，而概念則用同義詞集 (synset) 來表示。當一個詞的不同義項分屬不同的概念時，該詞就在不同的 synset 中出現。樹結構自然反映了對概念進行分類的上下位關係 (hypernymy-hyponymy)。此外，CCD 還進一步描述了反義關係 (antonymy)、整體–部分關係 (holonymy-meronymy)、蘊涵關係 (entailment)、致使關係 (cause) 等。所有這些信息附加在樹結構之上，實際上構成更複雜的網結構。這種網結構體現了概念之間的關係和約束。CCD 另給出了由詞（按義項分開）到概念的索引，索引文件實際上也是關係數據庫的二維表，以每個詞（按義項分開）爲登錄項，每個記錄也設立了若干字段，描述登錄項的所在概念 (synset)、詞性 (part-of-speech)、語義範疇 (category) 等屬性。

由於 CCD 基於 WordNet 開發並考慮了兼容和複用的方便，事實上已是漢英雙語概念詞典。但 CCD 不是 WordNet 的簡單的中文化，它充分考慮了漢語語言的實際特點，融入了中國語言文化的元素。比如，經過分析比較，在漢語概念與英語概念的對應上，提出了一系列要求遵守的準則，包括成詞及詞性原則、準確性原則、完備性原則等多項準則。CCD 的開發綜合利用了多項漢語語言資源，並充分發揮機器和人工的不同優勢。

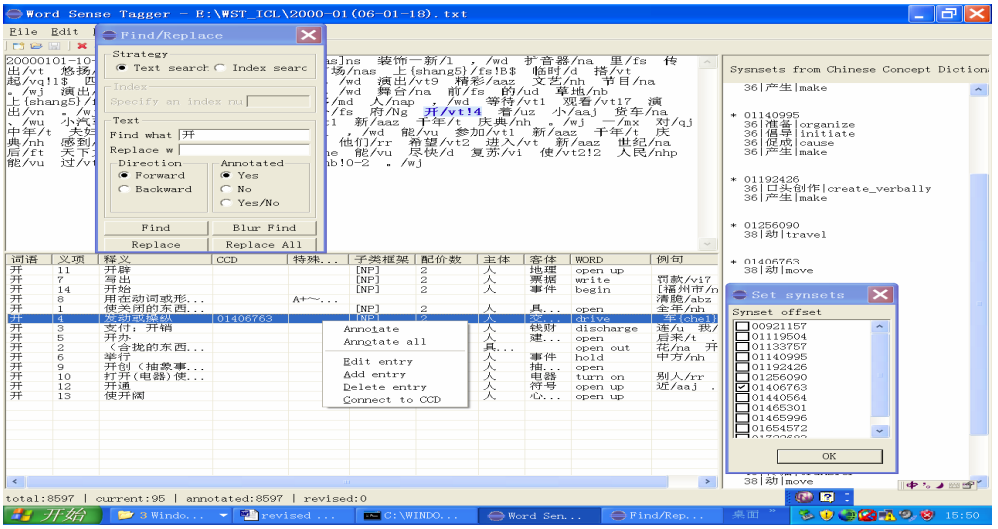
目前，CCD 對 WordNet 1.6 的覆蓋面在 94% 以上。ICL/PKU 正在把 CCD 發展爲多語概念詞典 (Multi-lingual Concept Dictionary, MCD)，除漢語和英語外，再加上日語和韓語。再進一步，要發展爲多語多學科的概念詞典 (Multi-lingual Multi-discipline Concept Dictionary, MMCD)。MCD 彙集的知識主要是常識，而 MMCD 則包括多個學科中的專業知識，目前正在增加醫藥學的知識。

7. 現代漢語詞義標注語料庫 CSC

詞義消歧 (Word Sense Disambiguation, WSD) 一直是語言信息處理中的熱點和難題。長期以來，詞義消歧研究未能取得突破性進展，主要原因之一是缺乏高質量的規模足夠大的詞義標注語料，缺少了機器和人學習詞義知識的真實文本語境。無論是詞義消歧研究還是詞彙語義研究，都期待著一個大規模、高質量的現代漢語詞義標注語料庫的誕生。ICL/PKU 在大規模現代漢語基本標注語料庫（俞士汶等 2003）和 GKB，CSD 這些寶貴的語言數據資源的基礎上，致力於開發一個大規模、高質量的現代漢語詞義標注語料庫 (Chinese Semantic Corpus, CSC)，期望它成長爲用於現代漢語詞義消歧研究的訓練和測試的基準語料，爲現代漢語詞彙語義研究添一把柴火。

詞義標注語料庫的詞義知識主要來源於 CSD，當然也參照《現代漢語詞

典》，並根據語料的實際狀況對詞義區分作了調整，其成果還反饋於 CSD。語料選自人民日報基本標注語料庫，即已經完成詞語切分和詞性標注的加工了的語料（俞士汶等 2003）。對應於 GKB 和 CSD 中“同形”、“義項”的區分，詞義標注也區分爲粗粒度的同形標注和細粒度的義項標注。本文著重討論細粒度的義項標注。詞義標注語料庫的建設其實是 CSD 不斷改進和語料標注不斷進展的互動增長的過程。CSD 中的詞義辨析和描寫是進行詞義標注的前提條件，在某種意義上可以說是詞義標注語料庫成敗的關鍵。Véronis (2003) 曾經指出，如果不建立一部充分描述了詞語分布信息的語義詞典，詞義消歧研究將不可能取得突破性進展。面向大規模文本信息處理，詞義的區分和特徵屬性描寫要在充分觀察語料的基礎上進行。開發 CSC 的工作就是一方面根據語料中詞語的實際使用狀況對詞義區分進行確認、調整和修改，一方面依據詞典中的意義區分對語料中出現的多義詞語賦予一個確定的義項。〈圖 1〉是用 JAVA 語言設計的詞義標注軟件的界面。



〈圖 1〉詞義標注軟件界面

界面上半部是待標語料，下半部是 CSD 中的詞義信息。通過觀察語料上下文，標注者單擊義項條目就可將相應的意義（“同形–義項”）編碼標注到語料中。標注者也可以根據語料中詞語的上下文對詞典中的詞義區分進行調整，如增加義項、刪除義項或者重新編輯一個義項的相關信息。

中文概念詞典 CCD 是 ICL/PKU 的另一個重要的詞彙語義知識庫（見本文第

6 節）。它描述了漢語概念之間豐富的語義關係。現代漢語語義詞典 CSD 採用“屬性-屬性值”的形式描述了詞語的組合分布信息，對詞語之間聚合關係的描述相對粗糙。只有當一部語義詞典既描述了組合分布信息又描述了聚合關係信息時，才是完備的。因此，建立 CSD 和 CCD 之間的鏈接 (linking) 顯得格外重要。當這種鏈接建立起之後，利用 WordNet 進行詞義消歧的眾多理論和方法就可方便地移植到現代漢語詞義標注語料庫 CSC 的建設上。〈圖 1〉的詞義標注界面提供了 CSD 和 CCD 之間便捷的鏈接操作。界面右端顯示的是 CCD 中相應詞語的義項區分情況，顯示了不同義項所在的同義詞集 (synset) 及其逐級上位。Set synsets 窗口列出了同義詞集在 CCD 中的偏移量，標注者選中相應的偏移量即可建立起當前詞語的義項和 CCD 同義詞集的映射。詞義標注語料庫 CSC，現代漢語語義詞典 CSD 和中文概念詞典 CCD 構成一個相互支持的整體。

目前，在 2000 年《人民日報》基本標注語料庫上，已經對 642 萬字的語料，標注了 76,519 個詞語的“同形-義項”編碼。限於作者的知識範圍，它應該是當今規模最大的現代漢語詞義標注語料庫，作爲詞義消歧研究的訓練語料和測試語料使用，基本上可以滿足要求。

已在 ICL/PKU 網站 (http://icl.pku.edu.cn) 上公布了 10 天《人民日報》的詞義標注語料，歡迎下載，期望得到批評和反饋。

在這個語料庫基礎上，ICL/PKU 已經開展了最大熵模型 (ME)、貝葉斯分類器 (Bayes)、支持向量機 (SVM) 等詞義消歧算法研究，期望取得良好效果。

8. 語言數據資源建設的理念與心得

借此機會，將 ICL/PKU 多年來開發語言數據資源的理念與心得加以整理，期望於人於己皆有裨益。

一、語言數據的規模要足夠大。規模並非僅指詞庫中詞的多少，更重要的指標是知識庫對詞的詞法、句法、語義知識描述的詳細程度。已經取得階段性成果和相當好的效益的《現代漢語語法信息詞典》GKB 收詞達 8 萬，現代漢語語義詞典 CSD 約 6 萬，中文概念詞典 CCD 包含中文概念（同義詞集，synset）約 10 萬，基本標注語料庫已有 6,000 萬字。詞義標注語料庫 CSC 自 2004 年剛開始建設，就規劃第一階段的目標要達到數百萬字。質量是語言數據資源的生命線。GKB 自 1986 年始，CSD 自 1993 年始，CCD 自 2000 年始，都經歷了長期的校正、使用、提高過程，力求精益求精，絕不懈怠，常以“行百里者半九十”自勉。不過，開發者又要有清醒的工程意識，在一定的投入和限期內，對規模和質

量都要有適當的把握，爲研究集體進入良性循環的軌道做全盤的、周密的考慮。

二、語言知識的選取，從詞彙著手，以句法知識爲基礎，逐步向語義深入。知識庫的類型，也從詞典向語料庫發展。這些決策既符合基礎研究的內在發展規律，也適應 NLP 應用技術發展和實用系統開發的需要。儘管《現代漢語語法信息詞典》GKB 遵循詞組本位語法體系，現代漢語語義詞典 CSD 以擴充的配價語法爲理論基礎，但知識庫所彙集的詞彙知識一定是漢語中客觀存在的，反映漢語使用的實際。1980 年代一大批基於複雜特徵集和合一算法的新型語法理論發展起來了，GKB 和 CSD 順應技術發展的大趨勢，無論框架設計、表述形式乃至特徵屬性項目的選取都適應了這些先進的計算語言學理論的發展。北大的多級加工語料庫的建設也是適應了自 1990 年代初重新興起的語料庫語言學的發展。需要特別強調的是，在這些詞彙知識庫中，語言知識及其表述形式皆獨立於特定的信息處理系統和實現算法，爲實現信息共享提供了便利。特別是在大致分類的基礎上，以“屬性–屬性值”的形式描述詞語的知識，適應了詞彙知識庫動態發展的需求。

三、充分發揮軟件工具和專家的各自優勢，協同攻關。爲提高開發效率，必須探索科學的語言數據資源構建方法，開發合用的語言知識庫構建工具。ICL/PKU 在建設語言知識庫的同時也積累了語言數據資源建設和管理的多種工具軟件（如《現代漢語語法信息詞典》的演化管理平台、中文概念詞典的可視化開發軟件、服務於 CSC 開發的詞義標注和輔助校對軟件等）。在開發語料庫加工（基本加工以及詞義標注）軟件時，基於規則的方法和基於統計的方法並舉；反過來，大規模深加工高質量的語料庫的出現又促進了計算語言學的發展。由於語言的複雜性，無論軟件工具多麼好，都不可能代替語言專家的努力和奉獻。ICL/PKU 充分重視專家的指導和關鍵決策的作用。長期的知識庫建設決定於人才培養，在 ICL/PKU，實現了人才與科研成果的同步增長。

四、對知識產權等其他相關問題也給予充分重視。

9. 綜合型語言知識庫的建設方略

9.1 綜合型語言知識庫建設的目標

儘管 ICL/PKU 多年積累的語言數據資源數量龐大，種類相對齊全，在邏輯上相互支撐，但在物理上卻基本是獨立存在的，尚未形成一個協同工作的整體。計畫把這些語言數據資源集成起來，形成一個綜合型的語言知識庫（俞士汶等

2004）。綜合型語言知識庫的建設目標是：

一、支持各成分數據資源之間便捷的準確的交叉參照，方便用戶（包括人和機器）從結構各不相同的多種語言資源獲取豐富的語言知識。

二、提供統一的應用程序接口 (API) 和風格一致的友好的用戶界面 (UI)。

三、提供數據挖掘工具，發展機器學習機制，支持知識發現，充分展現綜合型語言知識庫的價值和作用。

四、提供知識傳播和信息服務的機制，既實現知識共享，對知識產權又有妥善處理。

9.2 多個詞彙知識庫集成的指導思想——以詞義爲主軸

要把現有的語言數據資源無縫地整合到一起，建成綜合型語言知識庫，就必須填補其構成成分之間的“縫隙 (gap)”。《現代漢語語法信息詞典》GKB、《現代漢語語義詞典》CSD 與“大規模基本標注語料庫”之間就有縫隙。正在研製的詞義標注語料庫將逐步填平這些縫隙。粗粒度的詞義標注語料庫以“詞語”＋“詞類”＋“同形”爲軸連接了標注語料庫和 GKB；細粒度的詞義標注語料庫以“詞語”＋“詞類”＋“同形”＋“義項”爲軸連接了標注語料庫和 CSD。這就是「以詞義爲主軸」把詞典知識庫與標注語料庫連接起來的基本構思（俞士汶等 2006）。

進一步還可以把 CCD 集成進來。這就要利用 CCD 的由詞（按義項分開）到概念的索引文件。於是，詞義標注語料庫 STD，GKB，CSD 和 CCD 都集成到一起了，綜合型語言知識庫的大廈便基本落成了。接著，還可以進一步集成以不同單位對齊的多語言知識庫、專業術語庫等等。

由於採用「以詞義爲主軸」的基本構思，綜合型語言知識庫對北大計算語言所以外的其他語言知識庫也是開放的（集成方法類似於 CCD）。綜合型語言知識庫將永保青春與活力。

9.3 異構知識庫集成技術的實現

儘管有了「以詞義爲主軸」構造綜合型語言知識庫的基本構思，但由於《現代漢語語法信息詞典》GKB 是二維表格式的數據庫文件，是一張“平面”，而標注語料庫是文本文件，是線性結構，要把不同結構的知識庫有機地無縫隙地集成到一起，仍需要找到一個實現方案。經過長期的思考，觀察各個知識庫管理軟件

分別實現的理想中的綜合型語言知識庫的部分功能，並受到本所張化瑞老師提出的把標注語料庫的文本文件轉化爲數據庫文件的方案（爲了進行詞語頻度、匀度計算等研究的需要）的啓發，俞士汶終於想出了一個集成 GKB 和標注語料庫的方案。

給出同形標注語料庫（粗粒度詞義標注）中兩個句子的實例如下：

例句 1：

19980321-05-003-003/m　　二/m　是/v　此類/r　編著/vn　內容/n　抄/v!A　自/p　別人/r　的/u　多/a　，/w　多/a　到/v　被/p　人/n　告/v　到/v　了/u　法庭/n　；/w

例句 2：

19980804-09-006-007/m　　炮兵/n　學院/n　原來/d　圍牆/n　殘缺/v　，/w　周邊/n　群眾/n　進城/v　，/w　習慣/v　抄/v!B　近道/n　。/w

每句前面的一串數字標誌該句在語料庫中的位置，例句 1 是《人民日報》1998 年 3 月 21 日第 5 版第 3 篇文章第 3 段（同一段落中的句子是連續的，中間沒加特殊的位置標誌，假定例句 1 是該段的第 1 句）。原文句子切分成了詞語，用“/”和兩個空格隔開，在“/”後再標注詞性（如：m 是數詞，v 是動詞，r 是代詞，n 是名詞，等等）。注意，動詞“抄”不僅標注了詞性 v，而且還有“!A”和“!B”的區分。這就是標注了第 4 節中〈表 1〉中的“同形”信息，“抄/v!A”代表“抄寫的抄”，“抄/v!B”代表“抄近道的抄”，即同形標注語料庫完成了粗粒度詞義標注。

由例句 1 生成一個初始的只包含兩個字段的數據庫文件：

〈表 3〉由文本文件生成的數據庫文件示例

切分單位	長（字節）
19980321-05-003-003/m	21
二/m	4
是/v	4
此類/r	6
內容/n	6
抄/v!A	6
自/p	4
別人/r	6

各個記錄的第 1 列（“切分單位”字段）實際上就是同形標注語料庫中一個段落的文本內容的轉置（行變成列）。第 2 列（“長度”字段）就是第 1 列的值的字符數目（以字節 byte 爲單位），第 2 列的值是第 1 列所含信息的顯性化。

爲了同 GKB 連接，把“切分單位”拆分成“詞語”、“詞類”、“同形”3 個字段，爲了顯性地給每個詞語精確定位，由每段前的那一串數字派生出“年”、“月”、“日”、“版”、“篇”、“段”、“句”、“位”。如此得到的結構化文本數據庫文件 TDB 如〈表 4〉所示。

在〈表 4〉所示的數據庫中，每一個記錄都是唯一的，因而“頻度”字段的值都是 1。如果要統計某個範圍內，例如，1998 年 3 月份詞的頻次，只要略去“日”、“版”、“篇”、“段”、“句”、“位”這些字段，計算“年”=1998 且“月”=3 且“詞語”、“詞類”、“同形”3 個字段相同的記錄的數目，就可以得到互不相同的“詞語”、“詞類”、“同形”在 1998 年 3 月份的頻次。由此可知，將線性的文本文件改造成平面的數據庫文件，對於各種統計是十分方便的。

〈表 4〉每個詞語在文本中的位置信息顯性化了的數據庫文件 TDB

切分單位	長	詞語	詞類	同形	年	月	日	版	篇	段	句	位	頻次
19980321-05-003-003/m	21												
二/m	4	二	m		1998	3	21	5	3	3	1	1	1
是/v	4	是	v		1998	3	21	5	3	3	1	2	1
此類/r	6	此類	r		1998	3	21	5	3	3	1	3	1
內容/n	6	內容	n		1998	3	21	5	3	3	1	4	1
抄/v!A	6	抄	v	A	1998	3	21	5	3	3	1	5	1
自/p	4	自	p		1998	3	21	5	3	3	1	6	1
別人/r	6	別人	r		1998	3	21	5	3	3	1	7	1

〈表 4〉所示的數據庫文件 TDB 和《現代漢語語法信息詞典》GKB 有相同的 3 個字段：“詞語”、“詞類”和“同形”。因此，這兩個數據庫文件根據“這 3 個字段的值相等”條件聯結 (join)，便可以把這兩個數據庫文件的絕大多數記錄結合到一起，形成一個新的一體化的數據庫文件，它就是“綜合型語言知識庫”的主體部分。由此更可以認識到“以‘詞語’＋‘詞類’＋‘同形’爲主軸”的含義，它恰似一根軸，把兩個平面聯結在一起。

9.4 基於綜合型語言知識庫的語言知識發現

基於綜合型語言知識庫，可以進行數據挖掘，從而得到更深層的詞彙知識，例如：

一、詞頻、帶詞性的詞頻、粗粒度詞義頻度、細粒度詞義頻度。

二、詞的分布均勻度（關於時間或關於領域）（朱學鋒等 2004）。

三、動詞向名詞漂移的動態過程的考察（俞士汶等 2005）。

四、詞的各種屬性值的概率表示。

五、詞的組合規律的定量考察。例如，已經抽取出“數詞”＋“量詞”＋“名詞性短語”的組合實例，通過計算，將得到名詞（包括作爲名詞性短語中心語的名詞）和量詞搭配的計量數據，一直可以細到與同一個名詞搭配的各小類量詞的分布概率乃至同一小類中的不同量詞的分布概率。其他句法結構的計量研究也在進行中。

10. 結語與致謝

面向信息處理的詞彙語義研究仍待深入，詞彙知識庫仍有廣闊的發展空間。綜合型語言知識庫大家庭還將增添新成員：廣義虛詞知識庫，成語知識庫，縮略語知識庫、隱喻知識庫。

筆者衷心感謝 CLSW-7 的組織者台灣交通大學（新竹）劉美君教授和鄭錦全院士、黃居仁教授。正是有了他們的激勵，筆者才下決心，擠時間回顧了 ICL/PKU 開展漢語詞彙語義研究、開發詞彙知識庫的歷程，同時環視學科全局的發展，初步整理了指導這項研究和開發工作的思想脈絡，總算在由經驗向理論提升的道路上邁出了一小步。

感謝陸儉明教授對本文初稿提出了許多寶貴的意見。本文所述關於漢語詞彙知識庫各項成果的取得實賴 ICL/PKU 內各位同仁以及 ICL/PKU 以外眾多師友的指教和奉獻。在以往介紹單項成果的論著中，均曾具名致謝，恕不在此贅述。

引用文獻

Véronis, Jean. 2003. Sense tagging: does it make sense? *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*, ed. by Andrew Wilson, Paul Rayson & Tony McEnery, 273-290. Frankfurt & New York: Peter Lang.

于江生, 劉揚, 俞士汶. 2003.〈中文概念詞典規格說明〉,《漢語語言與計算學報》13.2:177-194。

中國社會科學院語言研究所詞典編輯室編. 2005.《現代漢語詞典》(第五版)。北京：商務印書館。

王惠, 詹衛東, 俞士汶. 2003.〈現代漢語語義詞典規格說明書〉,《漢語語言與計算學報》13.2:159-176。

朱學鋒, 張化瑞, 段慧明, 俞士汶. 2004.〈《漢語高頻詞語法信息詞典》的研製〉,《語言文字應用》2004.3:98-104。

吳雲芳, 俞士汶. 2006.〈信息處理用詞語義項區分的原則和方法〉,《語言文字應用》, 2006.2:126-133。

俞士汶, 朱學鋒, 王惠, 張化瑞, 張芸芸. 2003.《現代漢語語法信息詞典詳解》(第二版)。北京：清華大學出版社。

俞士汶, 段慧明, 朱學鋒, 孫斌, 常寶寶. 2003.〈北大語料庫加工規範：切分・詞性標注・注音〉,《漢語語言與計算學報》13.2:121-158。

俞士汶, 段慧明, 朱學鋒, 張化瑞. 2004.〈綜合型語言知識庫的建設與利用〉,《中文信息學報》18.5:1-10。

俞士汶, 段慧明, 朱學鋒. 2005.〈詞語兼類暨動詞向名詞漂移現象的計量分析〉,收錄於孫茂松、陳群秀主編《自然語言理解與大規模內容計算》, 70-76。北京：清華大學出版社。

俞士汶, 朱學鋒, 段慧明, 張化瑞. 2006.〈以詞義爲主軸的綜合型語言知識庫〉,收錄於蘇新春、王惠主編《第六屆漢語詞彙語義學研討會論文集》, 156-170。新加坡：COLIPS Publications。

張志毅, 張慶雲. 2001.《詞彙語義學》。北京：商務印書館。

符淮青. 2004.《現代漢語詞彙》。北京：北京大學出版社。

陳群秀. 1995.〈現代漢語述語動詞機器詞典的研究(二探)〉,收錄於陳力爲、袁琦主編《計算語言學進展與應用》, 207-212。北京：清華大學出版社。

黃居仁, 洪嘉馡. 2006.〈感官動詞的近義辨析〉,收錄於蘇新春、王惠主編《第六屆漢語詞彙語義學研討會論文集》, 1-10。新加坡：COLIPS Publications。

蘇新春. 2001.《漢語詞彙計量研究》。廈門：廈門大學出版社。
蘇寶榮. 2000.《詞義研究與辭書釋義》。北京：商務印書館。

[Received 31 December 2006; revised 4 September 2007; accepted 1 November 2007]

俞士汶
北京大學計算語言學研究所
中國 100871 北京市
yusw@pku.edu.cn

Research on Chinese Lexical Semantic and the Construction of Lexical Knowledge Base

Shiwen Yu, Xuefeng Zhu, Huiming Duan, Yunfang Wu, and Yang Liu
Peking University

Information technology, especially search engine technology popular on the Internet, offers promising opportunities for future research into Chinese Lexical Semantics. This paper demonstrates the significance and the urgent need for lexical semantics in natural language processing. Chinese lexical resources developed by the Institute of Computational Linguistics of Peking University (ICL/PKU) is introduced in this paper, which contains mainly four components: (1) Grammatical Knowledge Base of Contemporary Chinese (GKB), (2) Chinese Semantic Dictionary (CSD), (3) Chinese Concept Dictionary (CCD), and (4) Word-Sense Tagging Corpus of Contemporary Chinese (CSC). Building a comprehensive language knowledge base by combining all existing language resources is proposed in the last section; main ideas, technological measures, progress to date are outlined.

Key words: Chinese Lexical Semantics, Lexical Knowledge Base, Grammatical Knowledge Base of Contemporary Chinese, Chinese Semantic Dictionary, Chinese Concept Dictionary, Word-Sense Tagging Corpus of Contemporary Chinese, Comprehensive Language Knowledge Base