教材語料詞義分佈量化考察*

肖航
教育部語言文字應用研究所　北京
exiaohang@hotmail.com

摘　要：本文通過對已標注了多義詞義項信息的義項標注教材語料庫進行詞義分佈量化統計，分析了多義實詞不同詞義的重現情況及其分佈規律。多義詞普遍高頻，但各詞義分佈普遍具有不均衡的特點。統計資料顯示，只出現一個高頻義項的多義詞占到 30%～40%的比例；同時具有兩個或以上的高頻義項的多義詞只占 10%～20%的比例。文章指出掌握並利用多義詞詞義分佈特點及規律對詞義消歧研究和語文教學都有重要意義。

關鍵字：多義詞、詞義分佈、義項標注、語料庫、量化分析

Quantitative Analysis of Word Sense Distribution in Chinese Teaching Materials

Hang Xiao
Institute of Applied Linguistics of Ministry of Education, Beijing
exiaohang@hotmail.com

Abstract: This paper conducts a quantitative analysis of word sense distribution of polysemous words in the sense tagged Corpus of Chinese teaching materials. The paper investigates the reappearance rate of polysemous words in sense tagged corpus and analyses its distributive pattern. Finally, the paper points out that sense distribution study is important to word senses disambiguation research and is helpful to improve the quality of vocabulary teaching.

Key words: Polysemous words; word sense distribution; sense tagging; corpus; quantitative analysis

1　前言

在自然語言中，一詞多義（polysemy）是非常普遍的現象。瞭解多義詞的詞義分佈情況對自然語言處理和語文教學都有重要意義。考察多義詞詞義分佈需要有達到一定規模的、經過詞義標注的語料庫的支持。由於詞義標注語料庫構造困難，多義詞詞義分佈已有研究成果較少。本文通過對經過基於詞典的義項標注的語文教材語料中多義詞的詞義分佈進行統計，考察了語文教材語料中的多義詞詞義複現情況及其義項分佈規律，並根據統計資料對多義詞的不同詞義分佈特點進行了分類分析。本文還探討了如何利用詞義分佈不均衡的特點降低詞義自動消歧（Word Sense

* 本文部分摘自硕士学位论文《基于词典的语料库词义标注》（新加坡国立大学中文系，2009）。本研究得到新加坡教育部学术基金资助（AcRF – Tier 1, WBS No. R -102-000-046-112）。

Disambiguation，WSD）的複雜度，同時分析了掌握教材語料詞義分佈規律對教材編寫、詞彙教學與學習的積極作用。

2 語料庫

本文的研究基於"義項標注教材語料庫"[1]。該語料庫收入了多個出版社的中小學語文教材。本文使用了其中的初中和小學語文教材部分語料，總字數約為 200 萬字。語文教材中的課文多以文學作品為主，詞義表現較為豐富，適合進行詞義分佈量化研究。

義項標注教材語料庫已經經過以《現代漢語詞典》（第五版）為詞義體系的詞義標注。單義詞與多義詞的區分均根據《現代漢語詞典》（第五版）。以詞典中劃分為多個義項的詞作為多義詞。

本文重點考察了語料庫中多義實詞，並限於名詞、動詞和形容詞，的詞義分佈特點，下文所說的多義詞皆指多義的名詞、動詞和形容詞。表 1 列出了研究所用的語料庫中多義詞占總詞數的比例情況。

表 1：研究所用語料庫的詞語組成情況

語料中的詞語組成情況				
總詞數	52798 個		總詞頻	1143120 次
※實詞（名動形）	37603 個	占總詞數約 71%		占總頻次的 63%
※多義實詞	5514 個	占實詞數的 15%		占實詞頻次的 43%
※多義名詞	2758 個			占多義實詞頻次的 28%
※多義動詞	2473 個			占多義實詞頻次的 59%
※多義形容詞	945 個			占多義實詞頻次的 13%

說明：由於存在兼類，多義名詞、動詞、形容詞數量之和大於多義實詞數量。

3 詞義標注方式

語料庫詞義標注的內容為根據語境為多義詞選擇合適的詞典義項，採用设计计算机程序进行預標注後再加以人工校對的方式進行。

詞義標注的人工校對由六個以漢語為母語的、語言學專業的研究生完成。校對方式是由校對者根據《現代漢語詞典》的釋義，通過多義詞的語料上下文判斷其合適的詞典義項。校對者通過上下文判斷詞義的方式如表 2 所示。

表 2：詞義標注人工校對表

保證/v　bǎozhèng				
①擔保；擔保做到：我們～提前完成任務。				
②確保既定的要求和標準，不打折扣：～產品品質｜～科研時間。				
句編號	左窗口	詞	義項	右窗口

[1] "义项标注教材语料库"由新加坡国立大学、北京大学、商务印书馆三方合作开发完成，项目负责人是新加坡国立大学王惠博士。本文使用了该语料库中的部分内容，在此向语料库研发课题组致谢！

#1	信貸資金的目的出發，	保證	2	重點，區別緩急，堅持
#2	保證貨源，保證品質，	保證	1	及時向外商供貨，有意
#3	分佈的維管束稱葉脈，	保證	1	葉內的物質輸導。
#4	全世界作出莊嚴承諾，	保證	1	香港繼續她的生活方式
#5	均高出岷江的洪水位，	保證	2	內江與外江相互分隔。
#6	業務，為了保證貨源，	保證	2	品質，保證及時向外商
#7	辦法、規定外，並	保證	1	恪守下列條款：
#8	叔叔來之前，我們得	保證	2	那東西完好無損。
#9	使洪水順暢排出，從而	保證	2	進入灌區的水量不致為患
#10	卵石甩向飛沙堰，從而	保證	2	寶瓶口和下游灌渠不致淤塞

表 2 中，首先給出的是多義詞的詞性、中文拼音和多義詞在《現代漢語詞典》（第五版）中的全部義項信息。其後的数据表格中給出了語料庫中需要標注義項的包含目標詞的所有語句。為提高詞義辨識的準確率和一致性，一個多義詞的所有語句都在一起顯示，由同一個人標注，以利於對比區分。語句採用關鍵字居中（KWIC）方式顯示，按目標詞的位置分為左右兩個部分，即上文和下文，分別是表格中的左視窗和右視窗列。可以選擇按上文和下文排序，將具有類似上文和下文的語句排到一起。校對者需要根據上下文在義項列填入合適的義項編號。

4 詞義分佈統計

基於詞義標注語料庫，本文對其中的多義實詞的義項頻率分佈進行了統計。

（一）多義詞的比例

本文首先對單義詞和多義詞的比例分佈進行了統計。僅從詞形來看，研究所用的語料中出現了 37603 個實詞，其中單義詞占大部分（85%），多義詞只占少數。

但是如果考慮詞頻（出現次數）因素，單義詞比例大幅降低至 43%，多義詞的比例則上升到 57%。由此可見多義詞普遍具有高頻的性質。

（二）具有一個最常用高頻義項的多義詞

根據對多義詞在真實語料下的詞義情況統計，絕大部份多義詞的義項頻率分佈是很不均衡的。具體表現為只有個別義項高頻，其他義項低頻。若以出現多個義項並且有一個義項的出現頻率超過 80%作為衡量標準，語料中約有 31%多義詞符合只出現一個高頻義項這個特點；若以 70%作為衡量標準，則有 41%的多義詞滿足條件。表 3 列出了語料庫中多義詞最常用義所占的比例。

表 3：語料庫多義詞最常用義平均比例

常用義標注	詞數	總詞頻	最常用義比例
多義實詞（名、動、形）	2525	177181	79%
※兩個義項	1771	71116	88%

※三個義項	456	37605	81%
※四個義項	125	15098	74%
※五個義項及以上	112	53362	67%

具有一個最常用高頻義項的多義詞又可分為兩種情況：

1）語料中只出現一個義項的多義詞

語料庫中多義詞詞義分佈不均衡的一種極端情況是一個詞雖然有多個詞典義項但除一個高頻義項外其他義項不出現。具有這種分佈特徵的詞約占語料多義詞的 35%。考慮語料庫規模帶來的影響，低頻的多義詞有些義項未出現與其總詞頻低有直接關係，而高頻的多義詞則真實反映了其所具有的詞義分佈不均衡的特點。表 4 列出了部分語料庫中只出現一個義項的多義詞。

表 4：只出現一個義項的多義詞示例

多義詞	詞類	語料庫義項數[*1]	詞典義項數[*2]	高頻義項	語料頻次[*3]
發生	動詞	1	2	①_273	273
唱	動詞	1	2	①_273	273
臉	名詞	1	4	①_267	267
先生	名詞	1	3	②_267	267
風	名詞	1	2	①_253	253
朋友	名詞	1	2	①_252	252

說明：*1 語料庫義項數是指該詞類的多義詞在語料庫中出現的義項總數

*2 詞典義項數是指該詞類的多義詞在《現代漢語詞典》中的義項（不含語素義）總數

*3 語料庫頻次是指該詞類的多義詞在語料庫中出現的次數

*①_273 表示義項①出現了 273 次。後表同此。

只出現一個義項的高頻多義詞中的一部分可能是“偽多義詞”，除一個義項外，其他義項在現代漢語中都不再使用。區分出這種類型的多義詞對減少多義詞數量、降低詞義消歧複雜度有明顯作用。

2）出現多個義項但只有一個高頻義項的多義詞

多義詞不同義項在頻率分佈上不均衡，個別義項高頻，其他義項低頻是普遍現象。表 5 列出了部分有多個義項但只有一個高頻義項的多義詞及其詞義分佈情況。這部分詞只有一個義項高頻常用，其他義項均不常用。

表 5：出現多義項但只有一個高頻義項的多義詞示例

多義詞	詞類	語料庫義項數	詞典義項數	義項頻率分佈	語料頻次
世界	名詞	3	5	①_460 ③_23 ⑤_4	487
老	形容詞	2	8	①_448 ⑤_14	462

在	形容詞	2	5	②_441 ③_1	442
笑	動詞	2	2	①_440 ②_2	442
學	動詞	2	2	①_368 ②_37	405

對詞義消歧研究而言，這部分多義詞容易使用預標注最常用義的方式提高詞義識別的準確率。

（三）同時具有多個高頻義項的多義詞

1）有兩個以上高頻義項的多義詞

多義詞不同義項在頻率分佈上不均衡的另外一種情況是多個義項中有兩個或以上的義項同時高頻常用，其他義項不常用。若以有兩個義項的出現頻率都在40%以上作為衡量標準，語料中有約 9%的多義詞同時出現兩個高頻義項；若以30%作為衡量條件，則比例提高到約 20%。這類詞的高頻義項間的區分是提高詞義消歧正確率的難點之一。表 6 列出了部分有兩個以上高頻義項的多義詞及其詞義分佈情況。

表 6：有兩個以上高頻義項的多義詞示例

多義詞	詞類	語料庫 義項數	詞典 義項數	義項頻率分佈	語料 頻次
天	名詞	4	6	①_174 ⑩_24 ③_154 ⑦_32	384
爬	動詞	3	3	①_166 ②_195 ③_19	380
落	動詞	5	6	①_244 ⑩_9 ②_92 ⑥_8 ⑨_5	358
問題	名詞	4	4	①_102 ②_219 ③_3 ④_16	340
指	動詞	4	4	②_2 ③_191 ⑤_2 ⑥_84	279

2）所有出現的義項都高頻的多義詞

多義詞的義項分佈主要呈不平衡狀態，少數義項高頻，部分義項低頻、甚至不出現。其例外就是義項頻率分佈差異很小的，或者說所有義項都高頻的多義詞。如果以沒有任何一個義項的出現頻率低於 30%作為衡量標準，語料中義項分佈平均的多義詞約占 15%；若以 40%作為衡量標準，則比例降低到約 8%。表 7 列出了部分所有義項均高頻的多義詞及其詞義分佈情況。

表 7：所有義項都高頻的多義詞示例

多義詞	詞類	語料庫 義項數	詞典 義項數	義項頻率分佈	語料 頻次
長	動詞	2	3	①_310 ②_138	448
送	動詞	3	3	①_204 ②_112 ③_106	422
怕	動詞	2	3	①_183 ③_140	323
準備	動詞	2	2	①_179 ②_89	268
表示	動詞	2	2	①_81 ②_182	263

5 多義詞詞義分佈考察的意義

（一）詞義消歧研究

由於詞語多義現象的普遍存在，如何識別文本中多義詞的詞義成為自然語言處理的一個重要課題。詞義消歧研究就是為解決這個問題而產生，其任務一般定義為根據語境為文本中的多義詞選擇合適的詞義。Weaver（1955）指出實現翻譯的前提是知道詞語在當前語境下的詞義，因此機器翻譯系統必須具有詞義識別能力。

通過對語料多義詞詞義分佈的統計分析，可以看到義項頻率分佈不均衡是一個普遍的特點。這個特點對詞義消歧有十分重要的意義，例如掌握多義詞的高頻義項就可以通過預標注其最常用義來提高詞義消歧準確率的下限值（Base Line），從而提高詞義消歧的準確率。此外，對詞義消歧系統而言，詞義的使用頻率應當成為詞義劃分時需要考慮的一個重要因素，可以減少"偽多義詞"的產生，降低詞義消歧複雜度。

正如 Wilks（1997）所指出的，考慮到詞義的具體語言使用，詞義消歧並沒有根據對詞典詞義區分方式進行分析所看到的那麼複雜；充分利用多義詞的義項頻率分佈特點，設計一個實用的、高準確率的詞義消歧系統是可以實現的。

（二）教材編寫與詞彙教學

教材編寫者普遍認為，語言學習必須經過多次重複，教材的詞彙安排必須要考慮常用詞的重現率。然而，常用詞大多是多義詞，考慮到詞頻的同時如何考慮其不同詞義的複現情況仍未受到充分重視或尚未有好的解決方案。語文教材中還普遍存在對多義詞某一義項進行生詞釋義，而該詞其他重要義項出現時不再作為生詞釋義的情況，不利於學生正確掌握詞義。另外，教材中常用詞只出現一個或幾個義項，出現不常用義項而常用義項缺失，也容易給學習者帶來困擾。在語文教材設計中，準確瞭解多義詞詞義的分佈情況及出現順序，在考慮詞彙複現次數的同時，考慮詞義複現的次數，有助於教材編者編寫出更準確地反映多義詞的詞義特點的教材。

在詞彙教學中，準確掌握多義詞詞義的分佈情況及出現順序，特別是基本義和引申義的出現次數和順序、從低年級到高年級的教材中的詞義分佈差異、未出現的詞義等，有助於教師把握教材中詞義的呈現順序和規律，從而更為高效地、有針對性地引導學生學習掌握多義詞詞義。此外，多義詞詞義分佈的量化資料也可以成為詞典按詞義常用度安排義項的重要依據。

6 小結

義項標注教材語料庫中多義詞的詞義分佈統計資料顯示，只出現一個高頻義項的多義詞占到 30%～40%的比例，同時具有兩個或以上的高頻義項的多義詞只占 10%～20%的比例，證明了詞義分佈的不均衡性。詞義分佈情況的量化描寫有助於我們掌握不同詞的不同詞義的使用情況，從而在自然語言處理或是語文教學過程中制定合適的策略。但本文研究所用語料僅限於中小學語文教材，並且只分

析了名詞、動詞、形容詞三種詞類多義詞的詞義重現及分佈情況，統計資料尚不能全面反映多義詞詞義的分佈情況。

參考文獻

[1] Weaver, W. (1955). Translation. In Locke W. N. and Booth A. D. (eds.), Machine Translation of Languages: Fourteen Essay, pp.143-172.

[2] Wilks, Y. (1997). Senses and texts. Computers and the Humanities, 31(2), pp.77-90.

[3] 符淮青，詞義的分析和描寫，語文出版社，1996。

[4] 劉頌浩，現象和解釋:詞彙重現率及其他，暨南大學華文學院學報， 2006（1）

[5] 王惠，詞義·詞長·詞頻——《現代漢語詞典》（第 5 版）多義詞計量分析，中國語文，2009（2）。

[6] 肖航，基於詞典的語料庫詞義標注，新加坡國立大學碩士學位論文，2009。

[7] 中國社會科學院語言研究所詞典編輯室編，《現代漢語詞典》（第 5 版），商務印書館，2005。