

基於大規模語料的成語調查與教學成語詞表的提取

王治敏

楊爾弘

北京語言大學

北京語言大學

wangzm000@gmail.com

yerhong@126.com

摘要：本文利用大規模語料，以季度作為統計節點，考察了連續五年的新聞語料和口語語料成語的分佈特點，提取了漢語教學常用成語統計詞表，該詞表考慮了成語的分佈領域、分佈時間、頻繁程度、穩定性等多方面的因素，為成語教學及漢語辭彙大綱提供了有益參考。同時也對三個系列對外漢語教材中出現的成語和《等級劃分》、常用成語統計詞表進行了對比分析，考察發現目前對外漢語教材中的成語中的絕大多數不在《等級劃分》的收取範圍內，也有相當一部分不是真實語料中的流行的常用成語，教材成語的選擇有必要借鑒大規模語料的統計資料，把統計詞表中出現頻繁出現，穩定程度高的成語吸收到教材中去。

關鍵詞：分佈特徵；統計時段；統計詞表；重現率

Idioms' Survey Based on Large-scale Corpus and the Extraction of Idioms Wordlist for Teaching

Wang Zhimin

Yang Yerhong

Beijing Language and Culture University Beijing Language and Culture University

wangzm000@gmail.com

yerhong@126.com

Abstract: In this paper, we explore the distribution of idioms in news corpus and spoken text within 5 years based on large-scale corpus which is calculated quarterly. Furthermore, we extract the Common Idioms Wordlist in Chinese teaching. This list considered many factors of idioms: the distribution fields and time, frequency, stability and so on. It can be used as a reference for idioms' teaching and the syllabus of Chinese words. At the same time, we compare and analyze the idioms appeared in three text book for the Graded Chinese Syllables (GCS) and Common Idioms Wordlist. It proves that most of the textbook idioms are not included in GCS and many textbook idioms are not popular in real language material. Therefore, it is necessary that we can choose the high frequency and stable idioms as the textbook idioms.

Keywords: distribution characteristics, Statistical period, Statistics wordlist, repetition rate

1. 前言

成語形式固定，言簡意賅，是漢民族文化的精華所在。其不僅是對外漢語教學的重要組成部分，也是中高級階段留學生的教學重點和難點。專家們對於如何進行成語教學進行了深入探討。例如：張亞茹（2006）指出學生學習成語的難點在於不能全面深刻地理解成語的意義與用法，不能準確恰當地使用成語，應加強語用、語境等方面的教學。夏俐萍（2010）提出多種教學方法進行成語教學。雖然不同年級、不同層次的教材都會有成語出現，但是哪個年級應該教授哪些成語還沒有清晰統一的認識。

劉長征（2007）根據前人對辭書的統計，指出目前辭書中的成語可達一兩萬條，

其根據國家語言監測中心 2005 年度的語料統計得到 8637 條成語。曾小兵（2010）針對國家語言監測中心 2006-2008 年度的語料，統計得到三年的共有成語 7437 條。上述資料對於漢語教學提供了很好的借鑒。但是這些成語都是基於年度統計，無法看出其穩定程度，因此用於教學還需要進一步的篩選。目前《漢語水準辭彙與漢字等級大綱》（辭彙大綱）所收成語只有一百多條，遠遠不能滿足需求。

因此本文將在大規模語料的基礎上，建立一個可以反應歷時變化的成語詞表，然後再結合目前的教材成語，提供一個直接用於教材編寫的成語教學備選。

2. 語料的選擇

成語不僅廣泛存於書面語中，在日常口語中也常常涉及。因此，我們在選擇語料方面，綜合考慮了書面語和口語兩個方面的因素，同時還考慮成語在歷時語料中的持續時間。這就是說，一個成語的常用程度不僅和它出現的頻次密切相關，同時也和語料的領域和時間有關。所以我們選擇了 2005-2009 國家語言監測中心的《人民日報》作為書面語文本，同時也選擇 2005-2009 國家有聲媒體分中心的廣播電視文本作為口語文本。

為了能夠觀察到成語在五年時間內更加細緻的變化，我們以季度作為中斷點，分別考察兩種語料 5 年 20 個季度的成語的發展變化。首先利用中科院自動化研究所的分詞軟體對 5 年內的語料切分標注，提取出全部成語及其在各個季度的頻次。這樣就為每一個成語繪製出反應 5 年歷時變化的曲線。我們可以清晰看到每一個成語的發展變化。例如：

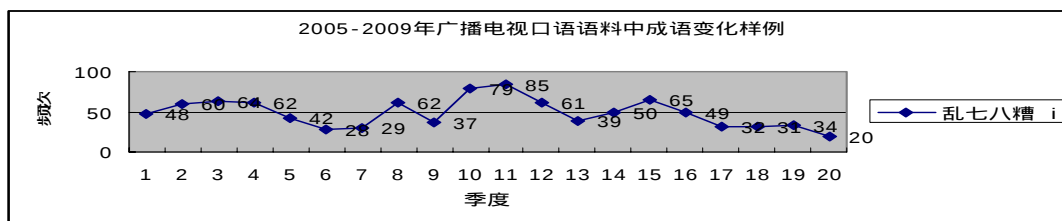


圖 1 “亂七八糟”在廣播電視口語語料中的頻次分佈

“亂七八糟”出現在漢語教材中的常用成語，它在 5 年廣播電視語料中的季度節點中連續出現，可以看出其比較常用。

3 常用成語的提取

我們對 2005-2009 年中《人民日報》和廣播電視語料展開調查，最後得到 5 年內連續出現的常用成語詞表。結果如下：

表 1 《人民日報》和廣播電視語料中成語的分佈

語料	人民日報（2005-2009）	廣播電視（2005-2009）

年份	2005	2006	2007	2008	2009	2005	2006	2007	2008	2009
一 季 度	2887	2733	2706	2684	2501	2369	3409	3699	3478	3670
二 季 度	2907	2759	2833	2551	2770	2775	3544	3814	3485	3775
三 季 度	2900	2833	2955	2645	3205	3059	3364	3639	3536	3837
四 季 度	2709	2676	2564	2590	2903	2944	3408	3556	3493	3556
季度共有數	1688	1562	1588	1461	1641	1536	2540	2853	2558	2894
年度共有數	851					1316				
兩種語料交集	612									
兩種語料並集	1292									

從季度節點看，出現在季度中的成語超過千條，但是考慮季度節點中連續出現的因素，數量就會大大降低。《人民日報》連續出現的成語有 851 條，廣播電視語料中連續出現的成語有 1316 條，兩種語料的成語總數為 1292 條，而能夠滿足在兩種語料中連續出現的成語只有 612。曾小兵（2010）通過三年的國家語言監測語料的考察得出一個量化的指標，即“800 條左右的成語頻次已經覆蓋了全部成語的使用頻次的 90%”，這說明，對外漢語教學只考慮 800 條左右的成語就基本滿足教學的需要。這 1292 條可以作為對外漢語教學的成語備選。

王治敏（2009）曾利用頻次資訊和時段資訊設計一個提取常用名詞的模型，成功提取出《漢語水準與漢字等級大綱》中過時的辭彙和最近新出現的常用名詞。

$$U = \frac{\bar{f}}{stdev(f)} \quad (1)$$

U 為成用度，式（1）中， $\bar{f}$ 表示詞語出現的平均頻次，其計算公式如式（2）所示。 $stdev(f)$ 表示詞語出現頻次的標準差，其計算公式如式（3）所示。

$$\bar{f} = \frac{f_1 + f_2 + \cdots + f_n}{n} = \frac{\sum f}{n} \quad (2)$$

$$stdev(f) = \sqrt{\frac{\sum (f - \bar{f})^2}{n-1}} \quad (3)$$

式（2）、式（3）中， n 為詞語統計頻次 f 的個數。

該模型對於漢語常用成語提取同樣具有借鑒意義，我們以廣播電視語料為例。

表 2 廣播電視語料中成語的常用度

2005-2009 口語—成語			
詞語	stdev	aver	U=aver/stdev
解放思想	156.657008847522	143.55	.916333083697012
千方百計	105.179345580975	124.2	1.18084020502278
全力以赴	98.5800636405172	124.85	1.26648325624214
前所未有	92.8102194917417	170.5	1.83708217622706
堅定不移	91.2762723534262	129.1	1.41438729552976

從平均頻次(aver)來看,“解放思想”的使用平均次數最高。但是從標準差(stdev)來看,“解放思想”最不穩定,一個詞語的平均使用次數最高,說明其常用,但是其標準差最大,說明其最不穩定,其常用程度也會受到限制,因此,我們利用上面的模型,就可以看到“解放思想”的常用度(U)相對於其他詞語有所下降。

漢語教學所使用的成語不僅體現在頻次的多少,還體現在其穩定程度和使用語體方面,因此我們的實驗模型充分考慮使用頻次、統計時點、穩定程度方面的因素,同時還要滿足一個最常用成語必須滿足在《人民日報》和廣播電視兩種語料同時出現且常用度排名靠前的成語。

表3 《人民日報》和廣播電視語料中排名前五位的成語統計

廣播電視語料常用度排名前五位的成語			《人民日報》常用度排名前五位的成語		
詞語	U(口)=aver/stdev	U(書)	word	U(書)	U(口)=aver/stdev
所作所為	3.07574613196902	1.57390738800586	任重道遠	7.79805231314112	1.71802478404538
如履薄冰	2.97543726835511	2.15394927030854	豐富多彩	6.09211699525128	1.82392840291405
自然而然	2.92951666341976	2.26134532487002	名副其實	5.36032620794818	2.0522214418795
可乘之機	2.89734606286377	2.26704546834264	引人注目	5.17795773113992	2.08544799656777
喜怒哀樂	2.87250850679569	2.54441153069761	迫在眉睫	4.64546466618951	2.02398085971721

上述成語是在兩種語料中常用度排名靠前的成語,雖然這些成語在兩種語料中均有出現,但是他們在各自語料中的常用度完全不同,“所作所為”在廣播電視口語體中的常用度要高於《人民日報》書面語體。而“任重道遠”在《人民日報》書面語體的常用度要遠遠高於廣播電視口語體。常用度相差越大,說明其語體使用的差異性。通過這種差異可以很好地描寫一個成語的語體風格,是很好的參考資料。

這裏我們有兩種考慮。一方面我們可以把在兩種語料中常用度數值相差大的詞語全部提取出來,作為不同語體的常用成語備選。另一方面我們還要考察只在一種語料中出現使用度較高的成語,這種成語在語體方面具有很強的特性,也可以作為教學的成語備選。

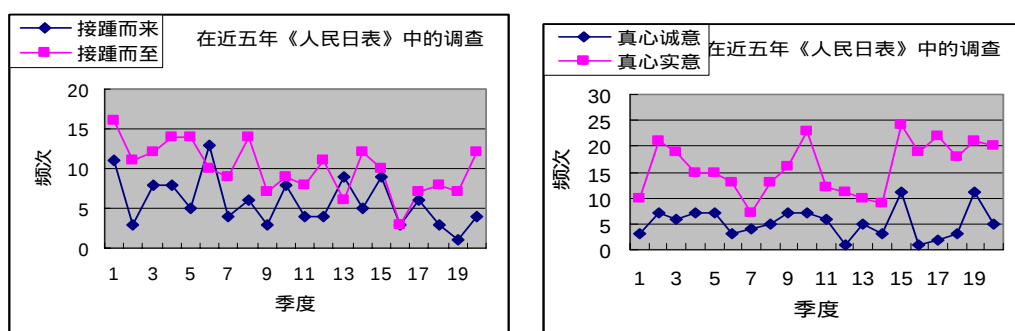
同時，我們還考察在兩種語料中出現，平均頻次和使用度相差不大的成語。因為這類成語非常穩定，不因為語體的原因而改變。這類成語在我們的詞表中比重最高。是我們考察的重點。

對於結構相近的同義成語，也可以根據其統計資訊擇優收錄。例如：

接踵而來 真心誠意 引人矚目 經久不息
接踵而至 真心實意 引人注目 經久不衰

他們從結構上看完全一致，只有一字之差。它們在文本中是否有差別？哪個更常用一些？如果有一些資料支援，教材編寫就會有更多的參考依據。下列兩組成語在文本中存在明顯差別。例如：

表 4 近義成語舉例



在《人民日報》中“接踵而至”相對於“接踵而來”更常用一些，“真心實意”口語化意味相對更濃一些，比“真心誠意”更加常用。關於這一點我們在廣播電視口語語料中也得到認證。例如：

表 5 “真心實意”在廣播電視語料中的統計

詞語	詞性	aver	sum	stdev	aver/stdev
真心實意	i	10.95	219	5.74433359242416	1.90622634006515

“真心實意”在廣播電視語料中的 20 個季度節點中全部出現，平均頻次為 10.95，而“真心誠意”並不出現。因此，常用成語統計詞表可以為漢語教學的同義詞辨析提供重要資料。

4·常用成語詞表及教材成語的對比分析

常用成語詞表的收詞完全來自大規模語料的統計，這些廣泛出現在大規模語料中的成語

在教學中是否使用，筆者為此專門對三個系列教材課文中出現的成語進行了統計。三個系列教材分別為北京語言大學初中高級漢語本科系列教材《北語系列》，博

雅漢語初中高級教材《博雅系列》和北大初中高級教程《北大系列》。三個系列教材包含了漢語初級、中級、高級三個階段的教學內容。三個系列教材不同階段的成語分佈如下：

表 6 系列教材不同階段的成語分佈

	北語系列		博雅系列		北大系列	
教材	教材名稱 及本數	成語 個數	教材名稱 及本數	成語 個數	教材名稱 及本數	成語 個數
初級	漢語教程(5)	31	博雅初級(2)	1	北大初級 教程(4)	2
中級	橋樑(2)	86	博雅中級(4)	128	北大中級 教程(2)	62
高級	三年級現代漢語 高級教程《高漢》 (2)	233	博雅高級(2)	183	漢語高級 教程(2)	265
共有 成語	橋樑∩高漢	11	博雅中級∩ 博雅高級	7	北大中級∩ 北大高級	4

從三個系列教材成語數量來看，成語分佈極不均勻。初級階段分佈數量很少，其中博雅系列和北大系列的初級階段只出現 1 到 2 例成語，北語系列出現 31 例成語，這說明博雅和北大系列教材編寫者一般都把成語的教學放在中高級階段。只有北語系列在初級階段涉及到成語教學。

中級階段，三個系列教材成語數量急劇上升，特別是博雅系列和北大系列教材中分別出現 128 例和 62 例成語，兩套教材的成語大量出現都是從中級階段開始，基礎相同，成語數量卻相差如此之大，這說明成語教學長期以來被人忽視，專家在編寫教材時缺乏相應的依據。

高級階段，北語系列和北大系列的成語數量出現 3-4 倍增長，達到 200 多條，博雅系列成語增長速度放慢，低於 200 條。三個系列成語數量參差不齊，相差懸殊，中高級兩個層次中，同一系列教材內部出現的共有成語非常少，重現率極低。

除此之外，我們對三個系列教材的中高級教程進行了橫向比較。結果如下：

表 7 系列教材中的共有成語

教材系列	中級	總數	占比例	高級	總數	比例
北語∩博雅∩北大	0	276	0%	6	681	0.88%
北語∩博雅	11	148	7.43%	20	416	4.81%
北語∩北大	4	214	1.87%	37	548	6.75%
博雅∩北大	7	190	3.68%	16	448	3.57%

中級階段，三個系列教材的共有成語為 0。高級階段三個系列教材的共有成語只

有 6 例，占這個層次成語總數的 0.88%，它們是“百無聊賴、氣喘吁吁、吞吞吐吐、微不足道、無可奈何、興高采烈”。這三個系列教材中的獨有成語已經達到 99%。我們進一步縮小範圍，對任意兩個系列的中高級出現的成語進行了調查，發現教材中共有成語所占比例非常少，均在 10% 以下。這說明對外漢語系列教材內部成語收取缺少客觀依據，教材和教材之間，以及教材內部不同層級之間缺乏有效的銜接。

為了能驗證教材共有成語在真實文本中的分佈，我們以成語統計詞表為參照，考察三個系列教材中在大規模語料中的分佈。

表 8 教材共有成語在成語統計詞表中的比例

教材系列	中級	出 現 在 詞 表 中 的 成 語	比 例	高級	出 現 在 詞 表 中 的 成 語	比 例
北語∩博雅∩北大	0	0	0	6	5	83.33%
北語∩博雅	11	10	90.91%	20	15	75.00%
北語∩北大	4	4	100.00%	37	24	64.86%
博雅∩北大	7	5	71.43%	16	11	68.75%

調查發現高級階段三個系列教材中出現的 6 例成語中，5 例出現在成語統計詞表中。兩個系列的共有成語絕大比例都出現在常用成語統計詞表中。但是我們應該看到，三個系列教材的共有成語非常有限，絕大多數成語都是教材的獨有成語。2010 年國家漢辦和教育部社科研制了《漢語國際教育用音節漢字辭彙等級劃分》簡稱《等級劃分》，這是目前最新的漢語教學辭彙參考資料，共收錄了 11093 個詞語。其中四字成語為 423 條。

教材成語在《等級劃分》和常用成語統計詞表中分佈如何？我們做了進一步考察，結果如下：

表 9 教材成語在《等級劃分》和常用成語統計詞表中的分佈

級 別	教材系列	教材名稱及本數	教 材 成 語	教 材 成 語 在 統 計 詞 表 中 的 個 數	在 本 部 教 材 中 占 比 例	教 材 成 語 在《等 級 劃 分》 中 的 個 數	在 本 部 教 材 中 占 比 例
中 級	北語	橋樑（2）	86	59	68.60%	36	41.86%
	博雅	博雅中級（4）	128	71	55.47%	40	31.25%
	北大	北大中級（2）	62	40	64.52%	14	22.58%
高 級	北語	三年級高漢（2）	233	135	57.94%	46	19.74%
	博雅	博雅高級（2）	183	74	40.44%	24	13.11%
	北大	漢語高級教程（2）	256	125	47.17%	52	20.31%

上表的統計可以看出，教材中出現的成語出現在《等級劃分》數量很少，占本部教材的比例很低，這意味著 3 個系列教材的成語遠遠超出了《等級劃分》收詞範圍，而教材成語出現在成語統計詞表的數量遠遠高於《等級劃分》的數量，這說明，教材成語雖然超出《等級劃分》的範圍，但是它們連續 5 年的大規模語料中出現，其仍然可以看做是常用成語，可以作為成語教學的一部分。

北語系列教材中的成語在大規模文本中出現的比例高於其他兩個系列教材。其中北語系列的中級《橋樑》中出現的成語在統計詞表中比例最高，占本部教材的 68.60%。但是也應該看到，近一半的教材成語並不是我們日常生活中常出現的成語，教材成語的出現缺乏系統性和科學性。特別是高級階段的成語，她們在常用成語統計詞表和《等級劃分》中的比例普遍偏低。

由於教材之間缺乏聯繫，同一層次的留學生因為漢語教材不同，成語的教學內容幾乎毫不相關，這樣就會帶來成語教學的混亂。同時教材中的很多成語並非日常生活中常用成語，這勢必會導致教學和語言實際應用的脫節。因此，教材成語的選擇有必要吸收大規模語料的統計資料，對教材中的成語逐一排查，把統計詞表中出現頻繁出現，穩定程度高的成語吸收到教材中。

5. 結語

本文參照大規模語料，考察了連續 5 年來《人民日報》和廣播電視語料中成語的分佈，提取了漢語教學常用成語統計詞表，同時也對三個系列對外漢語教材中出現的成語和《等級劃分》、常用成語統計詞表進行了對比分析，考察發現目前對外漢語教材中的成語中絕大多數不在《等級劃分》的收取範圍內，也有相當一部分不是真實語料中的流行的常用成語，而具有時間連續性的常用成語統計詞表包括了成語在規模語料中的分佈領域、分佈時間、頻繁程度、穩定性等多種資訊，在一定程度上為成語教學提供了客觀資料。

參考文獻：

- [1] 曾小兵, 邱麗娜, 主流平面媒體中成語的使用情況及特徵分析, 語言教學與研究, 2010, 6。
- [2] 劉長征, 秦鵬, 基於中國主流報紙動態流通語料庫(DCC)的成語使用情況調查, 《語言文字應用》2007 年第 03 期。
- [3] 王治敏, 2009, 《漢語常用名詞的自動提取研究—兼論“漢語水準辭彙與漢字等級大綱”的詞語更新問題》, 全國第十屆計算語言學學術會議論文集(CNCCL) [C], p52-58。
- [4] 張亞茹, 試論高級階段的成語教學, 語言文字應用, 2006 年 2 月第 1 期。
- [5] 石慧敏, 論中高級階段韓國留學生的成語教學, 雲南師範大學學報, (對外漢語教學與研究版), 2007 年第 4 期。
- [6] 沈莉娜, 近, 年來對外漢語教學中的成語教學綜述, 語文學刊(高教版) 2007 年第 7 期
- [7] 勞培萱, 對外漢語教材的成語研究, 語言教學研究, 2009, 5。
- [8] 洪 波, 對外漢語成語教學探論[J]. 中山大學學報, 2003, (2)。
- [9] 國家漢辦, 教育部社科司, 漢語國際教育用音節漢字辭彙等級劃分, 北京語言大學出版社, 2010。