汉语常用名词的自动提取研究
—兼论《汉语水平词汇与汉字等级大纲》的词语更新问题

王治敏
北京语言大学
E-mail:wangzm000@gmail.com

摘　要：本文利用大规模语料，提取汉语名词在语料中的统计特征，探索出一种可以自动发现汉语常用词语的有效方法，从而建立了汉语常用名词统计词表。通过对比分析《汉语水平词汇与汉字等级大纲》与汉语常用名词统计词表的异同，提取和发现词汇大纲未收入的常用名词，同时也可以把不再常用或很少使用的历史词汇自动过滤掉，使词汇大纲具有更新代谢功能，满足对外汉语教学的需要。

关键词：统计特征，教材编写；统计词表

Research on Automated Abstract of Common Chinese Nouns — The Discussions of "Syllabus of Graded Words and Characters for Chinese Proficiency" Update

Wang　Zhimin
Beijing Language and Culture University, Beijing,100083
E-mail: wangzm000@gmail.com

Abstract: This paper extracts the statistical features of Chinese nouns in large-scale corpus. A new effective way is proposed to find common Chinese nouns automatically, thus a statistical nouns database is established. By comparison between "Syllabus of Graded Words and Characters for Chinese Proficiency " and the statistical nouns database, it can automatically find and extract information of new common nouns which Syllabus did not contained and the historical vocabulary is removed from Syllabus. Therefore Syllabus may have a self-renewal potential and satisfies the needs of Chinese teaching as a second foreign language.

Key word: Statistical feature, textbook compilation; Statistical database

1　前言

《汉语水平词汇与汉字等级大纲》（简称 HSK 词汇大纲）是一个规范性的汉语水平大纲，是我国对外汉语教学总体设计，教材编写，课堂教学和成绩测试的重要依据；也是我国汉语水平考试的主要依据。词汇大纲作为规范性的词表，在对外汉语教学领域发挥了重要的作用。

但是由于社会的发展，词语也发生了很大变化，词汇大纲中的很多常用词语在今天看来已经成了历史词汇，有些已经弃置不用或者很少使用，如："倒爷、缎子、的确良、走狗、资本家"。同时也有很多词走入了我们的社会生活并成为常用词语。如："手机、邮件、短信、视频、上传、传送、光盘、证券、基因、总裁、社区、出版社、平台"，这些词语还没有被收入到词汇大纲中。昔日的词汇大纲已经不能满足对外汉语教学日新月异的发展需求，使得很多学者在制定教学大纲或者教材编写时不得不采用人工或者专家经验的办法解决收词问题。如何使词汇大纲保持一种自我更新能力，使其真正成为一个名副其实的标准是亟待解决的问题。

专家们对于大纲的修订提出各种设想和建议。赵金铭（2003）提出在大型语料库进行精细的

词频和义频统计之后重新进行词语筛选和分级。李红印（2005）提出把大于词的短语、结构、成语和习用语归入新增的“语汇大纲”，与已有的“汉字等级大纲”“词汇等级大纲”相照应。姜德梧（2004）从词汇的发展变化、收词标准、词性标注、同形词和一词多义的处理、轻声和儿化等多个方面提出了解决这些问题的原则和方法。刘长征（2008）提出了利用语言监测的相关结果，实现对外汉语教学用词表定期更新的设想。但并未见到针对大纲中名词的量化研究及实验。

《汉语水平词汇与汉字等级大纲》词条总数 8822 个,包括甲、乙、丙、丁四个等级，其中甲级词汇 1033 条，乙级词汇 2018 条，丙级词汇 2202 条，丁级词汇 3569 条。经考察我们发现，词汇大纲中名词共计 3456 条，在大纲中所占比例最多，占全部标注词条的 47%。名词的收取将是解决词表更新的关键。

因此，本文尝试利用大规模语料的统计结果，赋予词汇大纲中名词的多种统计特征，为名词常用词语的自动提取提供可靠的依据，使词汇大纲名词的收取更具有科学性。

2 语料的选择

定量研究对语料的选取有着很高的要求，不同的语料在定量研究中会显示出不同的功用。语料的规模、影响力、时间跨度等因素往往都是考虑的因素。

2005 年国家语言监测与研究中心平面媒体、有声媒体、网络媒体三个分中心在《中国语言生活状况报告》（以下简称《语言生活报告》）中发布了针对中国内地报纸、广播电视和网络的用字用词调查结果。三种媒体的语料共计 892 034 个文本文件，909 429 700 字符次，其中汉字出现 732 143 010 字次。该调查基于超大规模语料，考虑了平面媒体、有声媒体、教材媒体等多方面的因素，而且发布了年度流行语的监测。上述监测结果无疑是教学词表更新的有益参考。

但是由于时间跨度的局限性，上述调查还无法判断词语的持续性。比如几年前流行的“呼机、非典、倒扁”等流行语随着时间的流逝，渐渐推出了历史舞台。

张普(2003)指出，年度流行语是具有迅速流行、广泛传播的词语。并根据流行语的传播曲线，上升到一定高度后保持一段时间。与 2008 年大事相关的流行词语如“金融危机、地震、矿难、破产”等词语的词频会大大增加，但是随着历史事件的消失，这些词语出现的次数也会随之减少，流行语是否能成为社会生活日常用语还需要时间的考验。

流行词语在年度调查中的词频非常高，具有较高的关注度，但这对于对外汉语教学而言并不是最重要的，对外汉语教学需要的是能够在社会中产生广泛影响，能够保持稳定且常用的词语。因此，本文在选择语料时并不是像《语言生活报告》那样选择一个年度的报纸、广播、电视等多种数据，而是扩大时间跨度，选择了 1999-2003 年五年的《人民日报》。选择《人民日报》也考虑到语言的规范、发行量、影响力等多方面的因素。另外，选择六年前的语料，也是为了让词语有一个沉淀的过程。任何一种新词语经过六年的沉淀可能都会有一个去留的结果。五年《人民日报》的用字用词统计如表 1 所示。

表 1　《人民日报》的用字统计

《人民日报》年份	字数	字符数
1999 年	24,855,170	26,122,608
2000 年	27,195,578	28,487,138
2001 年	27,002,454	28,290,764
2002 年	25,615,279	28,250,598
2003 年	31,355,909	31,404,077
5 年合计	136,024,390	142,555,185

本次调查使用的分词软件是北京大学计算语言学研究所的分词标注系统①。经过切分标注

①本文使用的分词软件是 2004 年的旧版。

得到的切分单位经过整理，最后得到《人民日报》每个季度的词语切分单位。请见表2。

表2 《人民日报》的用词统计

语料来源	4季度	3季度	2季度	1季度	年词条
人民日报	词条总数	词条总数	词条总数	词条总数	平均数
2003年	80212	78122	78453	79351	79035
2002年	64335	69394	68751	67256	67434
2001年	69045	70879	69476	66728	69032
2000年	69385	72454	72781	66399	70255
1999年	65110	70881	78681	62765	69359

在这些切分单位中抽取了其中的全部名词，最后得到5年20个季度的全部名词切分单位。见表3：

表3 《人民日报》的名词统计

语料来源 人民日报	4季度 名词数	3季度 名词数	2季度 名词数	1季度 名词数	名词词条 年平均数
2003年	51361	49119	49199	50081	49940
2002年	38881	42254	41858	40515	40877
2001年	41819	43341	42299	40295	41939
2000年	41820	44328	44773	39432	42588
1999年	39033	43009	49397	37060	42125

为了考察名词切分单位的准确性，本文对北京大学计算语言学研究所的分词软件作了小规模测试，所选语料为2003年的测试文本，文件大小为83KB，经分词标注后文本包含11730个切分片断（包含1837个标点），其中名词有2906条，切分标注测试结果见表4：

表4 分词软件名词测试

测试指标	统计结果
切分标注出来的正确名词个数	2749
切分标注出来的名词个数	2906
人工测试语料中名词个数	2826
准确率	94.60%
召回率	97.28%

经人工统计，名词的准确率和召回率约为94.60%和97.28%。这里的错误主要表现在未登录词的切分错误，例如“非典/n”标注成“非/d 典/Ng”；国家名或者地名的标注错误。例如“西班/nr 牙/n、河/n 北/f 省/v ”；词性标注错误。例如“拉面/n 铺/v”，其中未登录词的错误比例最高，而我们在准确率和召回率的统计时同时考虑了切分和标注两方面的因素。如果排除未登录词和国家名的切分标注问题，切分标注的准确率会大大提升。因此本文使用的分词标注程序完全可以满足常用词提取的需求。

3 名词统计词表的设计

对于任何一个词语，判断其常用或不常用常常根据直觉的判断，但是这种直觉判断往往带有主观的个人因素，不同的专业背景可能有不同的结果，因此制定一个词语收取的客观标准非常重要。学者们在这方面进行了较多的尝试，最常用的莫过于根据词语的频次，也就是词语在文本中出现的次数。人们做了很多以频次为基础的计量研究，比如2005、2006年的语言生活报告给出了词语的词种数。张普（2003）根据词语的流通度来发现新词，但这些都是针对一年的时间点

上给出的词语计量方法。

词语在时间的某一个点上的频次只是一个点的纪录，如果在时间的两个点上就是一条线，在时间的 n 个点上就是一个变化曲线。如果把词放在一个以时间为横坐标、以词语频次为纵坐标的二维空间里，任何一个词语都会在过去持续的时间内留下一个频次的变化曲线。可以根据这个变化曲线看到词语的变化轨迹。因此本文选取五年的《人民日报》，经过分词标注后按照季度分成 20 个文件。将五年《人民日报》的词的集合 S 按照季度划分为 20 个子集。S1,S2……S20。设词 W 在这些子集中的频次分别为 F1,F2……Fn，然后提取词语在 20 个季度词频统计数据。建立了一个可以反映词语变化曲线的《人民日报》统计词表。这个词表记录了 1999-2003《人民日报》5 年间 20 个季度的时间节点的频次。

名词统计词表必须满足词语在 20 季度的数据库中均有出现，通过这样的筛选，在任意一个季度不出现，都会被过滤掉。例如："万元户"在 20 个季度的频次非常低，在 99 年第三季度和 02 年的第二季度的频次为 0，在其他季度从 1 到 8 次不等，该词就会自动排除在统计词表之外。

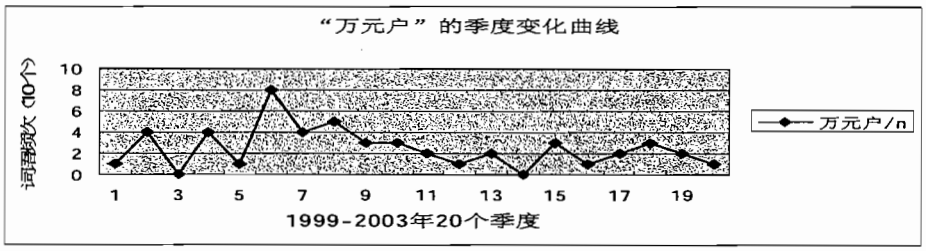

图 1　"万元户"的词语变化曲线

词表中也有在 20 个季度都出现，但是频次非常低的词语。例如："寻呼机"，这个词语现在已经不用，但是语料中还有少量纪录。不过这样的词语相对于频次高的词语，它的变化曲线也几乎为零。例如：

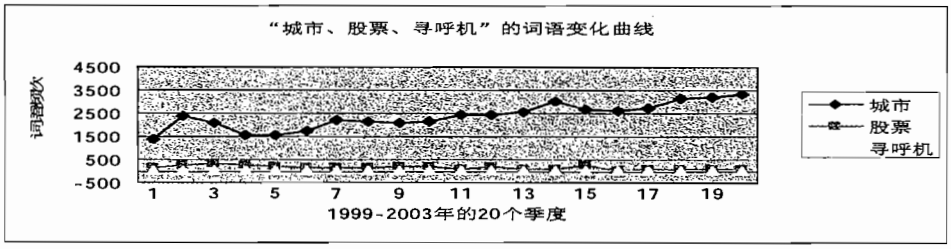

图 2　"城市、股票、寻呼机"的词语变化曲线

通过这样的筛选办法，可以把很多不常用的词语过滤掉，笔者对 HSK 词汇大纲的名词进行了筛选，发现 HSK 名词在统计词表不出现的词语约 179 条。在四个级别的分布如下：

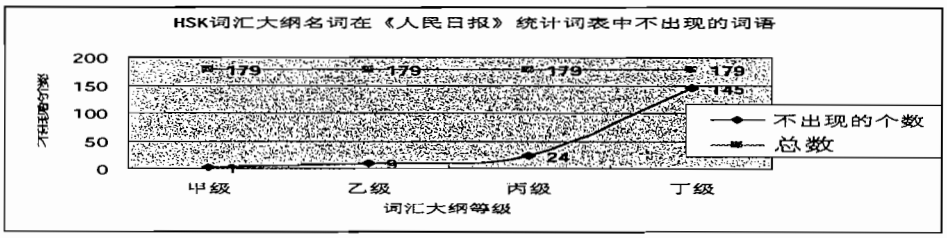

图 3　HSK 词语在五年《人民日报》统计词表中不出现的词语

从上面的统计可以看出，丁级词汇 145 条，比例最高，占全部不出现词语的 81%。丙级词汇 24 条，占全部不出现词语的 13%，乙级词汇 9 条，占全部词语的 5%，甲级词汇最少，只有一例"汽水"。"汽水"原来是人们生活中常用词，但由于生活水平的提高，出现了种类繁多的饮

品，人们常常统称为“饮料”，而“汽水”已经基本消失。“饮料”在5年《人民日报》20个季度中的平均出现次数为47次，因此可以考虑用“饮料”替换掉“汽水”。《人民日报》名词统计词表过滤掉的HSK词汇如表5所示：

表5　《人民日报》过滤掉的HSK词汇

词语	等级	个数
汽水	甲级.	1
板、墨水儿、画报、火柴、礼拜天、衬衣、鼓、礼拜日、老大妈	乙级.	9
管子、电炉、冰棍儿、国营、电铃、底片、钩子、炉子、笼子、猿人、毛泽东思想、染料、马克思主义、害虫、猎人、资本家、棍子、胶卷、农具、窟窿、边疆、嗽叭、寡妇、工钱	丙级.	24
靴子、叛徒、走狗、齿轮、山冈、妖怪、刑场、老太婆、算术、炮火、烧饼、槐树、禾苗、元件、蝇子、桑树、烟卷儿、斧子、粉末、贫民、译员、俘虏、仆人、筛子、山岭、锯、晌午、气流、便道、灯泡、稻子、倒爷、现钱、导体、罪状、云彩、蛾子、鞭子、闺女、蝗虫、钳子、校徽、年头儿、蜘蛛、方程、大炮、腮、抹布、函授、蝉、木匠、出品、分母、汞、来年、汉奸、穗、代数、国有、重型、混纺、总务、葡萄糖、螺丝钉、半截、镰刀、车床、交点、统战、小鬼、莲子、甲板、稿纸、跳远、脸盆、两口子、桅杆、唯心论、唯物论、缎子、半径、整数、头子、葵花、雹子、灾荒、定理、梗、马力、祆、马戏、半边天、唯心主义、框、高产、炊事员、镁、尼龙、腊月、老天爷、霉、文言、跳高、工事、大气压、梧桐、恩人、焦炭、国际主义、胶片、蜂、国库券、田间、牲口、手巾、珠子、牢房、前线、的确良、侦探、匪徒、背面、瑞雪、眼下、舵、巫婆、照会、骡子、流寇、渡船、殿、旅店、再生产、政变、半路、弹、公债、箩筐、伙计、倍数、屁、抗战、糠、大雁、电钮	丁级	145

上述词语绝大多数都是不常用的词汇。有的甚至已经基本不用。例如“火柴、冰棍儿、校徽、尼龙、的确良”等词语所指的事物已经在人们生活中基本消失，应该考虑剔除。

一些用于农业生产的常用词语。例如：“害虫、农具、斧子、筛子、锯、鞭子、钳子、镰刀、牲口、骡子、箩筐、糠”等词语虽然还偶尔使用，但是已经不再常用，也应该考虑从词表中剔除出去。

另外，一些和战争相关的词语。例如：叛徒、走狗、刑场、统战、大炮、汉奸、头子、匪徒、抗战、工事、牢房、前线、弹、小鬼等，对留学生的汉语教学作用不大，也可以征求专家意见后考虑有选择剔除。

除此之外，还有一些词语已经改变了说法。比如“国库券”现在经常使用“国债”，“旅店”经常使用“酒店、宾馆”，“牢房”经常使用“监狱”。而“国债、酒店、宾馆、监狱”在5年《人民日报》20季度中均有出现，季度平均次数如图4所示。这些词语可以考虑替换。

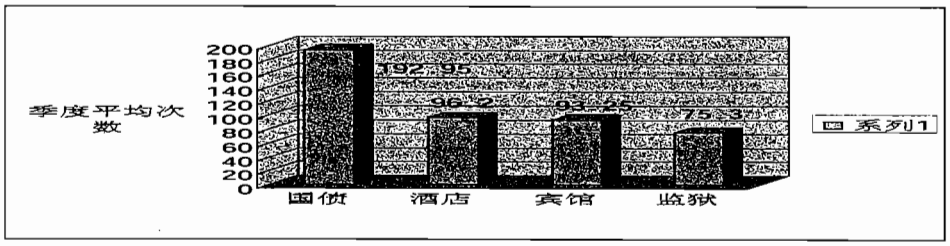

图4　词语分布4例

4 汉语常用名词的自动提取方法

本文提出用一个衡量词语稳定程度的参数U来作为选取HSK词表的依据，那么U应该能反映词语在语料中出现的平均频次及频次的波动性等因素。因此本文采用公式（1）计算该参数。

$$U = \frac{\overline{f}}{stdev(f)} \tag{1}$$

式（1）中，$\overline{f}$ 表示词语出现的平均频次，其计算公式如式（2）所示；$stdev(f)$ 表示词语出现频次的标准差，其计算公式如式（3）所示。

$$\overline{f} = \frac{f_1 + f_2 + \mathrm{L} + f_n}{n} = \frac{\sum f}{n} \tag{2}$$

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

式（2）、式（3）中，n 为词语统计频次 f 的个数。

从公式（1）可以看出，参数 U 与词语在语料库中出现的平均频次成正比，而与词语出现频次的标准差成反比。词语的季度平均值反映了使用该词语的频繁程度，一个词语使用的越频繁，其在语料中的季度平均值越高。

标准差 $stdev(f)$ 反映了该词语出现频次的波动程度，一个词语在 20 个季度中的分布越不稳定，其标准偏差的值越大，U 的值就越小，比如和年度突发事件的词语标准偏差很大，参数 U 就会把这些词语排除在外。

按照评价参数 U，可以提取和发现 HSK 词汇大纲没有的备选词语。词汇稳定度排名靠前的 15 个词语如表 6 所示：

表 6　HSK 备选词语

词语	季度平均条数	评价参数	词语	季度平均条数	评价参数
全国	6039.2	7.43343747137599	高等院校	48.15	5.98306139769554
领导	2462.15	7.33679330961664	重点	1227.05	5.91200472365504
作用	1360.45	7.10508412861899	网络	1008.35	5.62631508996312
贡献	616.35	6.86296338238374	董事长	161.55	5.61301089221299
关系	1441.6	6.79169060363905	力度	804.65	5.5832961676006
乡镇	716.25	6.5464285446849	总裁	134.35	5.54036424398084
总经理	201.4	6.46537164806482	热点	218	5.48310680894976
难点	114.15	6.09857146025414			

姜德梧(2004)认为，“词频是选词的主要依据，但不是唯一依据，有时要进行人工干预。因此这些备选词语将来可以考虑人工干预的方式有选择加入 HSK 词汇大纲。

同时，统计词表也有其他的功用，如利用评价参数 U 也可提取和发现重大历史事件。因为在历史事件发生之后，往往与之相关的词汇会急剧增加，而且它们在季度的频次分布会很不均匀，请见表 7。

表 7 人民日报统计词表中稳定度（U）最低的部分词语

词语	肺炎	暴行	军品	志愿军	疫情	种族主义	疫病	人防	传染病	病例
Freq99-1	9	11	3	8	25	8	8	2	14	5
Freq99-2	13	572	2	15	12	9	5	5	22	21
Freq99-3	6	27	4	15	3	7	5	10	14	21
Freq99-4	9	10	6	12	6	10	5	3	14	7
Freq00-1	4	36	9	3	10	17	12	3	23	11

Freq00-2	15	27	4	13	15	8	4	36	19	15
Freq00-3	3	19	4	15	6	6	8	61	15	14
Freq00-4	5	27	2	349	17	13	10	300	21	16
Freq01-1	1	14	7	4	32	19	31	32	44	20
Freq01-2	9	22	1	28	26	7	33	4	12	7
Freq01-3	3	28	6	12	28	243	12	7	28	13
Freq01-4	7	14	4	20	22	21	17	2	33	26
Freq02-1	2	9	14	7	14	14	7	16	22	7
Freq02-2	7	13	4	12	33	10	10	30	35	17
Freq02-3	8	11	11	6	9	10	10	5	23	7
Freq02-4	3	4	196	11	9	7	12	17	22	7
Freq03-1	385	3	7	4	1115	3	78	8	296	342
Freq03-2	179	11	6	9	692	5	68	7	215	131
Freq03-3	90	3	13	13	428	3	26	3	159	126
Freq03-4	3185	3	8	20	2942	1	451	11	1117	803
稳定度 U	0.277 992	0.3460 41	0.364 841	0.380 88	0.393 65	0.400 913	0.411 276	0.427 273	0.429 977	0.431 013

按照 U 的降序排列，本文发现排序前 10 位的词语都是和 2003 年发生的“非典肺炎”相关，比如“肺炎、疫情、传染病、病例”。如果以 2003 一年的数据做统计数据，无法发现这些有价值的数据。因此，扩大语料的时间跨度对于提取和发现词语的特征大有帮助。

5 结语

本文利用大规模语料，建立《人民日报》名词语料统计词表，通过对比分析《汉语水平词汇与汉字等级大纲》和语料词表的名词统计特征，提取和发现词汇大纲没有收入的新词语，同时也可把不再常用或很少使用的历史词汇过滤掉，使词汇大纲具有更新代谢功能，满足对外汉语教学的需要。

参 考 文 献

[1] 赵金铭，张 博，程 娟，关于修订《(汉语水平) 词汇等级大纲》的若干意见[J] ，世界汉语教学，2003 ，(3)。

[2] 李红印，《汉语水平词汇与汉字等级大纲》收“语”分析[J]，语言文字应用，2005，4。

[3] 姜德梧，关于《汉语水平词汇与汉字等级大纲》的思考[J]，世界汉语教学，2004，(1)。

[4] 刘长征，对外汉语教学用词表的多元化与动态更新[J]，语言文字应用，2008，2。

[5] 张普，关于大规模真实文本语料库的几点理论思考[J]，语言文字应用，1999，1。

[6] 苏新春， 对外汉语词汇大纲与两种教材词汇状况的对比研究[J] ， 语言文字应用，2006 , (2) 。

[7] 国家汉语水平考试委员会办公室考试中心，汉语水平词汇与汉字等级大纲(修订本) [S] ，经济科学出版社，2001。

[8] 刘英林，宋绍周，论汉语教学字词的统计与分级(代序)[A]，汉语水平词汇与汉字等级大纲(修订本)[S]，经济科学出版社，2001。

[9] 国家语言资源监测与研究中心，中国语言生活状况报告 2005 下编[R]，商务印书馆, 2006。

[10] 国家语言资源监测与研究中心，中国语言生活状况报告 2006 下编[R]，商务印书馆, 2007。

[11] 北京语言学院教学研究所，现代汉语频率词典，北京语言学院出版社[S]，1986。

[12] 张普，基于 DCC 的流行语动态跟踪与辅助发现研究[C]，语言计算与基于内容的文本处理，清华大学出版社，2003。