基于ＤＣＣ的流行语动态跟踪与辅助发现研究[1]

张普

北京语言大学　应用语言学研究所　北京 100083

E-mail:zhangpu@blcu.edu.cn

摘　要：本文介绍了基于ＤＣＣ（Dynamic Circulating Corpus 动态流通语料库）的流行语动态跟踪发布研究的基本情况。着重介绍了流行语的界定与特点，流行语的动态曲线特点和意义，计算机辅助发现的可能等。最后还指出今后的研究目标与方向。

关键词：动态流通语料库、流行语、动态流通曲线、词汇曲线类型

Study on the Dynamic Tracing and Computer-aided Finding of Popular Words and Phrases Based on DCC

ZHANG pu

Beijing Language and Culture University

E-mail:zhangpu@blcu.edu.cn

Abstract: This paper introduced the current situation on the tracing and issuing of the popular words and phrases based on the Dynamic Circulating Corpus (DCC). It laid emphasis on the delimitation and features of the popular words and phrases, characteristics and meaning of the dynamic curve drawn according to certain values of popular words and phrases in a continuous period of time, and the possibility of finding them aided by computer. At last it pointed out the research goal and direction for the future.

Keywords: Dynamic Circulating Corpus (DCC), popular words and phrases, dynamic circulating curve, typology on lexical curve

一、 引言

2002 年底，北京语言大学、中国新闻技术工作者联合会、中国中文信息学会三家机构共同主办了中国报纸流行语的跟踪研究、评选发布活动，经过对 15 家中国主流报纸的约

[1]本文得到国家９７３重点基础研究发展规划项目“面向大规模真实文本的汉语计算理论、方法和工具（项目批准号：Ｇ１９９８０３０５０７－２）的子项目资助。同时得到国家语言文字应用“十五” 科研项目“报纸流行语跟踪研究”（项目号：ＹＢ１０５－６３Ｅ）的资助。

5 亿语料的动态统计分析，12 月 25 日零点，30 个侯选词语在华夏大地教育网等 10 多家授权网站（包括已授权的 15 家大报网站）上进行网上投票评选活动。2003 年 1 月 6 日在北京语言大学会议中心召开新闻发布会，由国内语言学界和语言信息处理界的著名专家向全社会发布了 2002 年中国主流报纸的十大流行语：**十六大、世界杯、短信、降息、三个代表、反恐、数字影像、姚明、车市和 CDMA。**

会后，国内有数十家报纸和 2 2 4 0 余网页陆续刊登转载了十大流行语的评选工作和评选结果。同时，中国评选十大流行语的新动态还引起国外报刊的重视，，2003 年 1 月 26 日《参考消息》转发了俄罗斯《消息报》1 月 16 日亚历山大·丘多杰耶夫的文稿，他指出“近日，《人民日报》刊登了国家 15 家大报 2002 年最流行的词语，说明经济改革 20 多年来，中国人，特别是城里人的思想发生了深刻变化。”

对流行语的动态跟踪研究与发布工作是以北京语言大学多年来的动态流通语料库（ＤＣＣ）的研究成果为基础的。我们在 1998 年正式提出建立动态流通语料库的设想，发表了关于对语言进行动态观察分析的一系列研究论文，2000 年该课题正式立项，建立了报纸动态流通语料库，采集国内十种（2002 年扩大到１５种）流通度最高的报纸语料。正是这一系列的先期工作，才保证了流行语的动态跟踪和十大流行语的顺利发布。

本文将主要介绍基于ＤＣＣ（动态流通语料库）的有关流行语动态跟踪研究与发布工作，特别是介绍流行语的界定、分类、动态流通曲线特征，计算机辅助发现的可能等，提出了新的词汇曲线类型研究课题。

二、 关于流行和流行语

我们查阅了《现代汉语词典》、《辞海》、《辞源》和《汉语大词典》，四部工具书都没有收录“流行语”或相关条目，但是都收录了“流行”一词。“流行”的定义大致如下：

《现代汉语词典》【流行】：传播很广；盛行。一例为：流行性感冒，一例为：这首歌在我们家乡很流行。

《辞海》【流行】的一个义项为：迅速传播或盛行一时。

《辞源》【流行】：传布、盛行。

《汉语大词典》【流行】的三个义项之一为：广泛传布、盛行。

根据以上情况，我们认为：公认的“流行”就是“迅速传播，盛行”，“流行语”就是**“在某一时期，某一地域或者某一人群中迅速传播、盛行的语汇。”**例如：“２０００年中国青年流行语”，就是在２０００年，在中国，在青年中迅速传播、盛行的语汇。语汇包括词、语和词语总称。

总之，流行语的界定应涵盖如下几大要素：扩散性、时效性、地域性、密集性。

“流行”的可以是褒义或中性的事物，如：孟子公孙丑上：“德之流行，速于置邮而传命。”又如“流行歌曲” “流行款式”“流行词语”等。但并不一定仅仅是褒义的可以流行，贬义的事物也可以流行，例如：左传僖公十三年：“天灾流行，国家代有。”又如：“瘟疫流行”、“妖教流行”、“流行性感冒”等。

三、关于“十大流行语”的发布

“十大流行语”是在流行语中通过一定方式选择出的社会认知度最高的十条流行语。例如：中国十大流行语、青年十大流行语、美国十大流行语、２０００年十大流行语等等。

所谓通过一定方式，通常是采用社会调查的方式，社会调查的选票可以在网下按常规抽样进行，也可以在因特网上随机点击投票。而事先提出的流行语侯选词语表，则是由一定范围的专家、学者、记者等根据语感提出决定的。

国内外对流行语的研究与“十大流行语”的发布工作都已经有多年的历史。许多国家都有流行语的研究机构、发布机制，有的国家每年定期发布。例如美国方言协会参与的 2001 年度美国流行词语的评选，“零地带”（世贸废墟）、“９１１”、“后９１１时代”等都入选侯选词语。日本每年由“自由国民社”主办当年“流行语大奖”，同时还要举行隆重的颁奖仪式，甚至是首相，也会介入流行语的评选和颁奖活动。２００１年的“日本流行语”中，“没有圣域的改革”、“米百俵”（为了美好的明天要忍耐今天的艰苦生活之意）、“不要恐惧、不要怯懦、不要被束缚”等都出自日本首相小泉纯一郎，他成为２００１年创造流行语最多的人。此外随着韩国文化在亚洲各国的传播，韩剧、韩国流行歌曲、韩国电影、韩国服饰、发型为“哈韩族”所钟爱和追逐。来自韩国的流行语随着“韩风”、“韩流”一起渗入中国。有人列出２０个来自韩国流行语，：阿里郎、流氓兔、红魔、金喜善、安在旭、《我的野蛮女友》、《蓝色生死恋》、宇田公司、汉城音乐厅、BabyVox（有“韩国妖姬”之称的美少女歌舞团体）等等均是这两年大家熟悉的。

国内对于流行语的观察、研究与分析，也日渐受到社会各界人士的关注，１０多年来有不少媒体参与十大流行语的评选与发布工作，例如：１９９３年《大学生》杂志最早公布“十大流行语”，《青年研究》、《中国青年报》、《中央电视台》、《北京电视台》等也有评选和发布；一些网站和公司也参与其中，例如：搜狐网、新华网、华夏大地教育网、北京零点公司等；甚至还有个人调查发布的十大流行语，例如：郑欣公布了１９９８年的“十大流行语”。各种部门和机构发布的流行语涉及到了各种类别，应有尽有。从流行的人群看，有“大众流行语”、“青年流行语”、“中学生流行语”等；就流行的地域看，有“中国流行语”、“港台流行语”、“都市流行语”等；就流行的行业和领域看，有“高校流行语”、“军营流行语”、“旅游流行语”、“娱乐流行语”、“零售行业流行语”等；就流行的时间而言，有“初春十大流行语”、“二十世纪９０年代流行语”、“改革开放以来２０年流行语”等；还有就流行语刊载的媒体进行分类的，如：“网络流行语”、“报纸流行语”、“手机短信流行语”等。

一些人认为：中国的“十大流行语”的发布工作已经失之过滥，任何媒体、机构、个人都可以进行一场自己感兴趣的评选与发布，而且动则冠以“十大”也令人觉得俗套。更有的流行语的推介已经走向媚俗或庸俗，诸如：“怪怪流行语”、“现代两性流行语”、“男女关系流行语”等等。也有的人认为：中国的流行语评选和发布，与国外相比，还远远不够，不够细、不够多、不够频繁。流行语是社会的反光镜、时代的透视机，透过流行语，我们可以观察社会的发展、也可以了解大众的心态。人们的价值观念、社会思潮、取舍好恶、心路历程都可以在流行语中展露无疑。

我们认为：目前中国的流行语的评选与发布，既存在过滥的现象，也存在远不够的方面，甚至有些东西是不是可纳入流行语范畴都需要探讨和推敲。流行语和新词语的界限，流

行语和切口、黑话的界限，流行语和行业语的界限都需要仔细厘定。

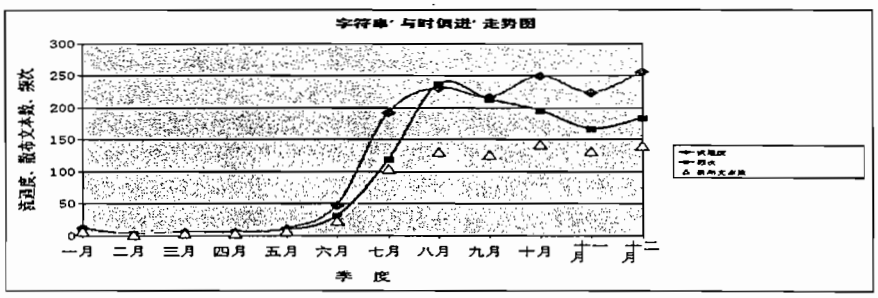
我们希望：通过系统化的跟踪研究，将流行语的评选与发布品牌化，即：科学、动态、权威、深入、全面。

四、基于ＤＣＣ（动态流通语料库）的流行语研究

为了使流行语的发布具有科学、动态和权威的品牌，我们基于动态流通语料库（ＤＣＣ）来进行２００２年的流行语候选词语表的筛选工作，而不是像一般的流行语候选词语表那样，由一个人、几个人或一些人来凭印象决定。

我们的流行语的侯选词语表是有科学的定量分析依据的。我们建立了庞大的报纸动态流通语料库，动态流通语料库的两大特点是动态性和流通性。流通度，是动态流通语料库的新属性，是在前人频度、使用度、通用度基础上的一个新推进[2]。基于动态流通语料库统计出的词语的频度、使用度和流通度，都不是一个共时的数字点，而是一条由若干个数字点构成的变化曲线（又称词语历时变化“走势图”）。

不同的词语有不同的变化和自己的“走势”，下图是词语“与时俱进”在２００１年的走势，黑色、灰色和白色分别代表频次、流通度和散布文本数（下同）：



我们建立的２００２年的中国主流报纸动态流通语料库，是在２０００多种报纸中筛选了１５种流通度高的报纸作为入选媒体，这１５种报纸是（按汉语拼音音序排列）：

北京青年报　北京日报　北京晚报　法制日报　光明日报　环球时报
经济日报　今晚报　人民日报　深圳特区报　文汇报　新民晚报
羊城晚报　扬子晚报　中国青年报

五、流行语的界定与特征

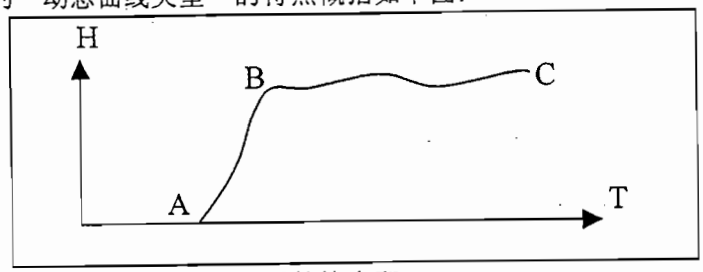
我们对１５种主流报纸约５亿字的语料进行了动态加工统计，提取的所有新词语都有一条“历时变化曲线”（又称词语历时变化“走势图”）。根据这些走势图我们可以对流行语的特点进行研究，依据这些量化数据，我们可以比较科学地判断该新词语是否已经“迅速传播”和“盛行”。再进一步利用计算机辅助我们进行流行语的筛选。

我们在上文曾经提到：公认的“流行”就是“迅速传播，盛行”，“流行语”就是“在某

[2]参见张普，《关于语感与流通度的思考》，载《语言教学与研究》，1999年第2期。

一时期，某一地域或者某一人群中迅速传播、盛行的语汇。”

我们把流行语的“动态曲线类型”的特点概括如下图：

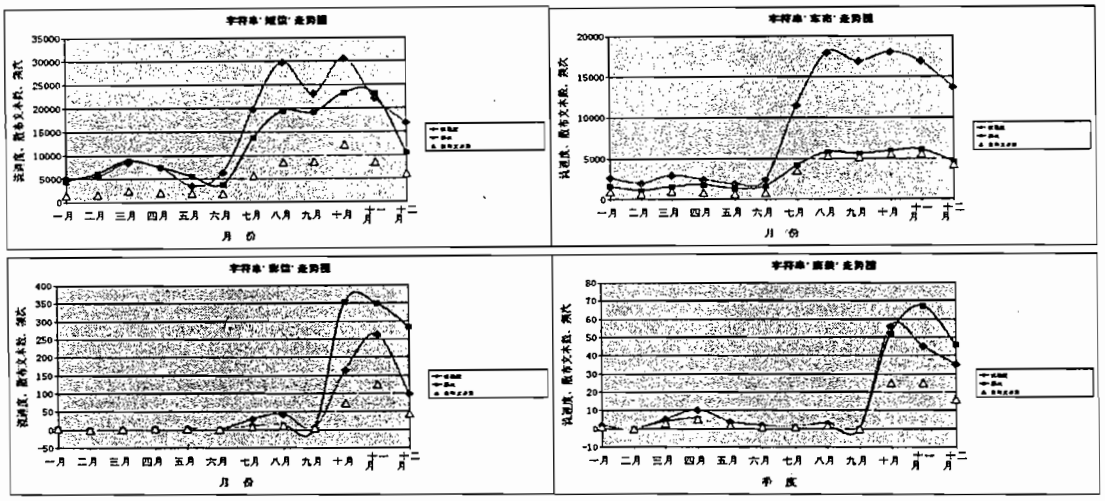


通过上图我们可以分析出流行语的特点即：

- 曲线几乎从“0”开始（图中的A点）；
- 上升迅速（图中A到B的时间），具有一定的“斜度”（A到B的时间越短，B的绝对高度越高，斜度越大）；
- 上升有一定绝对“高度”（图中B点）；
- 上升到一定高度后保持一定时间（图中B到C点）；

根据这样的形式化的特点，我们可以探讨计算机辅助发现流行语的方法，进而提取流行语的侯选词表，编纂流行语词典等。

下面就是2002年和2001年的一些与流行语的走势图相似的词语动态走势图，其中2002年“短信”和“彩信”的图形走势均与流行语的走势图相合，但是前者绝对高度较高，比较成熟，后者绝对高度较低，说明成熟度还不够。“唐装”也是2001年的走势，那时的绝对高度也是比较低的。

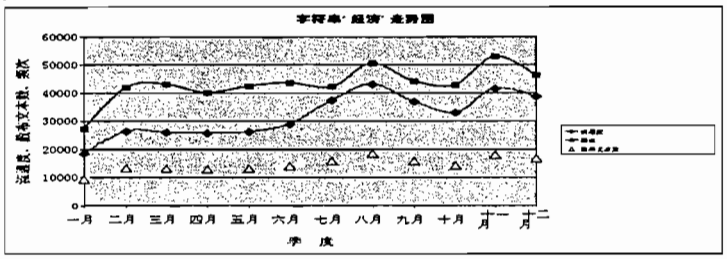


六、词汇曲线类型学

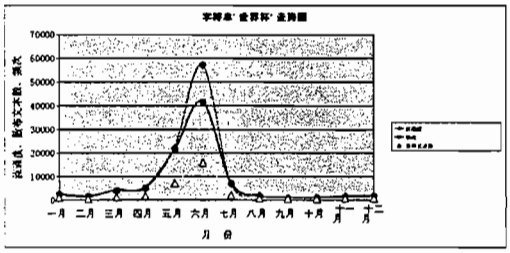
依据词语的曲线特点可以构成不同类型的曲线特征，我们可以依据不同的曲线类型，对词汇进行分类研究，这种研究并不仅仅限于流行语。根据这些走势图我们可以研究词语的特点，进行计算和提取、聚类与组合，这将是我们在计算语言学特别是计算词汇学方面的又一个重要创新。对于这些词语的“走势图”的类型和分类进行研究，将有可能形成计算语言

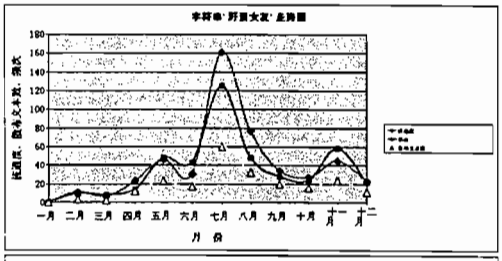
学领域的一门新学科。例如也许会形成一门新的“词汇曲线类型学”，当然这是后话。

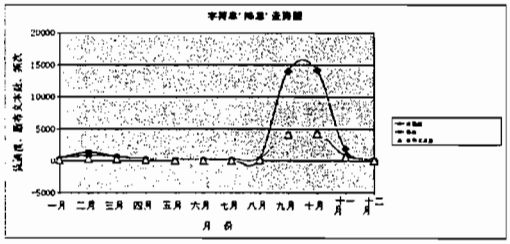
例如：根据流行语的曲线特点，“经济” 显然不符合流行语的变化曲线特点，其曲线特点长期居高不变并持平，所以它是一般词汇。我们还可望由计算机辅助把基本词汇提取出来，形成一个基本词汇表（迄今为止现代汉语的基本词汇集还是一个未知数）。下面是词语“经济”在２００１年的走势图。

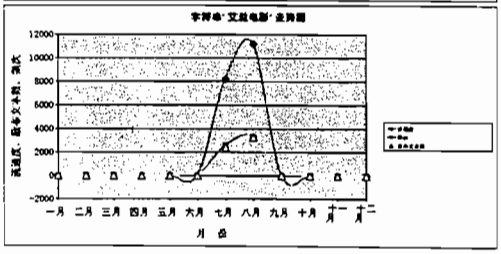


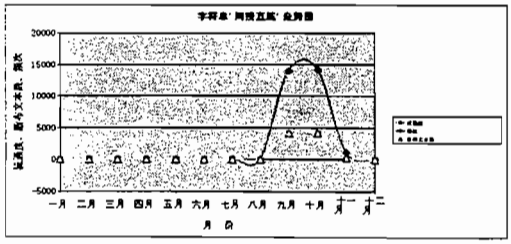
又例如：标准的流行语走势图已如上述，但是今天进入流行语的还有其他一些类型。流行词语的“流行过程”将依据曲线类型决定，正如在流行病学研究的“流行过程”分类中，也有流行趋势是属于“散发性”或“多发性”一样。一些现在常常被人们视为流行的词语，它们的曲线类型没有上述的第四个特点，即流行的势头很快过去，曲线下迭，没有一个持续的时间，只是一个一度流行的词语。这些词语常常与事件或人物相关。例如下述的“世界杯”、“野蛮女友”、“艾滋电影”、“户籍改革”、“间接直航”、“降息”、“刘晓庆”等词语的曲线：

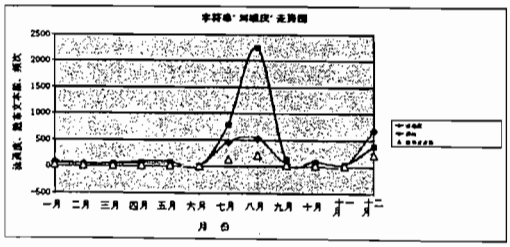






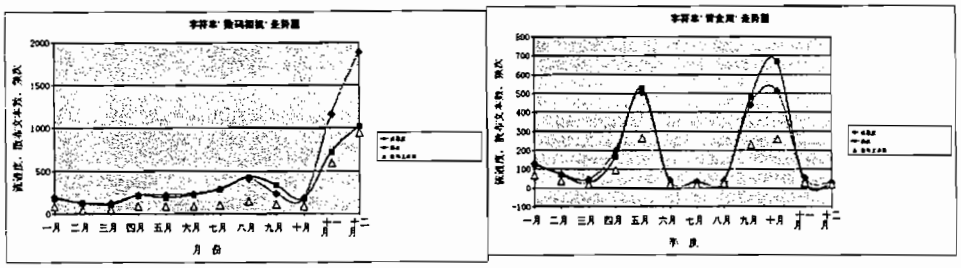






“黄金周”类也是特殊的流行类型，一年一度或两度、三度流行，其趋势与“散发性”或“多发性”的流行病的走势更加类似。这一类的词语不多见，类似的还有“春运”、“考研”、“换季打折”、“扫黄打黑”、“严打”等，但是能否成为流行语是值得研究的。词

语“数码相机”也是可能的流行语，只是其突然上升正值年末，还为见其持续一个时间段的第四个特点，如果继续滚动追踪，与其次年年初的动态曲线相接，可能就是典型的新的流行词语。做十大流行语的计算机辅助提取，年头和年尾的曲线，需要与相邻的年头衔接来考虑曲线的完整性。另外，一些需要更粗或更细的时间颗粒度来观察的词语，还需要考虑一次动态加工的结果多次粗化或细化的问题。下面是“黄金周”、“数码相机”的动态走势图：



七、结束语

总之，基于ＤＣＣ语料库的“报纸十大流行语”的科学研究与发布，仅仅是一个开始，流行语的动态、持续跟踪研究与发布，将作为一项经常性的工作更加广泛而深入地进行下去。

随着各种条件的具备，今后将陆续解决媒体的扩大问题、流通度的精度问题、词语表的垃圾问题、流行语的界定与分类问题、流行语的走势图问题、十大流行语发布的时间和领域的细化与粗化问题、流行语的定义辅助提取问题、流行语的词典编纂问题等一系列的问题。同时也将尽快推出“报纸动态词语表”，加速推进ＤＣＣ动态流通语料库的上网服务。

参 考 文 献

[１] 陈原主编：《现代汉语定量分析》，上海教育出版社１９８９年。

[２] 于根元：《二十世纪的中国语言应用研究》，书海出版社，１９９６年。

[３] 黄昌宁：《关于大规模真实文本的谈话》，载《语言文字应用》１９９３年第２期。

[４] 张普：《关于大规模真实文本语料库的几点理论思考》，载《语言文字应用》，1999 年第 1 期。

[５] 张普：《关于语感与流通度的思考》，载《语言教学与研究》，1999 年第 2 期。

[６] 张普：《信息处理用动态语言知识更新的总体思考》，载《语言文字应用》，2000 年第 2 期。

[７] 隋岩、张普：《1997 中文报纸媒体流通度分析》，载《计算语言学文集》清华大学出版社，1999。

[８] 张普、石定果：《论历时中包含有共时与共时中包含有历时》，２００２年首届社会语言学国际研讨会。

[９] 隋岩、张普：《基于动态流通语料库的〈动态词典〉编纂》，载《中国辞书论集 2000》中国大百科全书出版社，2001 年 10 月。

[１０] 隋岩、张普：《基于“动态流通语料库”的词语评估和新词语发现》，２００２年中国辞书学会年会。