现代汉语词汇历时检索系统的建设与应用*

荀恩东 饶高琦 谢佳莉 黄志娥
北京语言大学 汉语国际教育技术研发中心 北京 100083
Email: {edxun,raogaoqi,xiejiali,hze}@blcu.edu.cn

摘要：词汇是语言系统中最具活力的子系统。在语言演化的过程中，词汇的历时变化是语言学、历史学、社会学等多学科所关注的信息。我们收集了时间跨度约为六十年的同质新闻语料。基于自然语言处理技术我们开发了现代汉语词汇历时检索系统。基于该平台可以利用频率、累积和与累积频率等方法从微观和宏观的角度上对词汇的在语义、语用等方面进行研究。

关键词：历时信息；词汇演化；历史计算；语料库

Diachronic Retrieval for Modern Chinese Word: System Construction and its Application

Xun Endong, Rao Gaoqi, Xie Jiali, Huang Zhi-e
International R&D Center for Chinese Education, Beijing Language and Culture University, Beijing 100083

Abstract: Vocabulary is the most active and time sensitive sub system of a language. During the evolution of a language, diachronic changes in vocabulary are focused by linguist, historian and sociologist etc. We collected large scale of corpora with a large time span, and formed Diachronic Retrieval for Modern Chinese Word with natural language processing technology. The relative features based on this system like frequency, cumulative sum, cumulative frequency etc. could be used to give a research on vocabulary semantic, pragmatic etc.

Key words: diachronic information; vocabulary change; historical computing; corpus

1. 引言

词是语言中有意义，能独立运用的最小单位，也是最能够体现语言生活变迁的语言单位。每一个词都有在其所属语言社团中独特的发展过程。从微观上说，一个词语包括其使用情况的历时信息，可以反应特定时间乃至特定领域在不同时期所受到关注的情况。从宏观上讲，整个词汇的丰富程度是语言生活情况的重要体现，从一个侧面反应了社会变迁和人民生活的变化。每个时间断面上的词汇都带有以往的语言历史，是共时和历时的混合产物[1]。

计量语言学关注今天的词汇始于哪个历史时期，还关注现在词汇的使用状况是如何形成的。语言的历时信息同样为计量史学所关注。而利用计量史学方法进行的观念史研究，则更注重特定词语的历时使用变化。金观涛、刘清风[2]使用晚清至民国有影响力的报刊杂志一亿两千万字作为数据源，通过表达同样观念的不同词在不同时期使用频率和上下文特征的研究，观察并分析了一百个中国现代政治术语的形成和发展。在史学界引起很大反响，但是其史料库规模和选材偏执也引起了争议[3]。刘长征运用 1981-2009 共 29 年的《深圳特区报》进行了新词语监测和词语生命力的研究[4]。涵盖面更广的语料库如 LIVAC 则收集泛华语地区的新闻语料四亿字，在共时性和历时性上都有突出贡献[5]。在囊括两岸三地新闻语料的基础上，持续更新，在此基础上发布港台京沪双周、全年名人榜，热词榜等信息，并对两岸三地的词汇使用异同做出了定量的分析。LIVAC 新闻语料库建设始于 1995 年，历时仅 17 年。对于语言现象的变迁，这样的跨度还略显不足。谷歌公司 2010 年上线的服务 Google Books N-gram Viewer，利用其数字化的 520 万册图书制作了可实现五元文法的词汇历时查询[6]。

* 国家自然科学基金 60973062 和 61170162 支持了本文与相关工作的开展。

覆盖了 1800 年-2000 年间两个世纪的语料。但其汉语图书量较少，未对语料进一步分类，且有效的查询跨度少于 200 年。此外，图书对于现实语言现象的变迁存在一定的滞后。

可见，进行语言历时信息研究，尤其是词语历时信息的研究，需要大规模、长时间跨度的语料。我们收集了时间跨度五十七年的某省日报语料，为汉语词汇的历时信息提供了良好的基础。在第二部分中，我们将介绍历时新闻语料的构成。对于特定词语的微观研究，频次、频率和频序是计算语言学中的使用的经典表征形式。在对宏观语言现象的历时研究中，采用前 N%频率累积和（TNFA）与总词表前 N%累积频率历时分布（TNFD）两种可计算指标对词汇使用丰富程度和高频词汇来源的历时分布进行表征。这些可计算特征将在第三部分中进行讨论。基于这几项表征，搭建了现代汉语词汇历时检索系统 Diachronic Retrieval for Modern Chinese Word)。在线上开放数据为广大研究者所用。第四部分将介绍该系统的设计和原理。最后一部分简要列举了几项基于该系统的应用，并展望了未来的研究方向。

2. 历时语料的构成

我们收集了自 1949 年 11 月创刊至 2007 年间的某省日报，全部语料 7 亿字。该语料时间跨度大，覆盖了共和国自成立以来的绝大部分历史，记录了期间的语言生活与社会生活的巨大变迁，对于各个学科的追踪研究具有格外高的研究价值。以年为单位，对语料进行整理。经过分词并去除标点符号、拉丁字母与低频命名实体等，共有 328000 个词形。各时间段语料规模如表 1 和图 1 所示。可见，随着时间的推移，语料规模逐渐扩大，在 1996 年前后达到最高峰，接近 1970 年最低点的两倍。这是报刊信息量加大，社会传媒发展的结果。

时间段	平均词形数	平均词次数	平均字数
1950-1959	99.1K	7.7M	12.5M
1960-1969	71.6K	6.2M	10.1M
1970-1979	69.1K	5.5M	9.2M
1980-1989	112.6K	6.2M	10.1M
1990-1999	143.1K	9.4M	15.3M
2000-2007	136.3K	9.8M	16.1M

表 1. 各时间段语料规模统计表

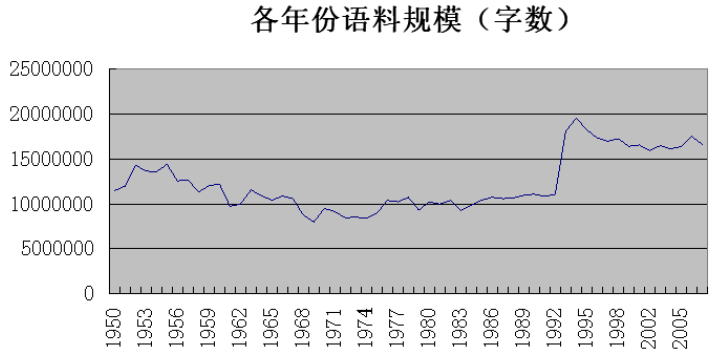

图 1. 各年份语料规模（字数）

3. 词汇历时信息的表征方式

基于词语历时信息的研究，可以分为微观的对特定词语历史信息的分析、跟踪和宏观的对整个语言基于词语信息的历时研究。对于前者，频次、频率和频序是较为经典的表征方式。后者又分为基于词的历时语言丰富程度的度量与高频词历时分布的研究。基于词的历时语言丰富程度的度量，我们借用类似香农熵的思想，使用前 N%频率累积和（TNFA）。高频词历时分布则用总词表前 N%累积频率历时分布来加以描述。

3.1 微观词语历时信息的表征形式

词语出现的频次是语料中最能直接表征其使用情况的特征。由于不同时间段的语料规模不一，使用词语出现的频率作为衡量改该词使用情况的标准显然更为为科学。频率的定义如下：

$$q(word)=\frac{freq(word)}{Count}$$

其中 q(word)为词语 word 的频率，freq(word)是它在当年语料中的出现的频次，Count 整个语料的全部词次数。

另一种表征词语使用状况的方式是特定词语在当年词表中的排名，如果该词表是按照的频率降序排列的话，这种排名被称作频序[8]。相对于频率，这项指标更能反映出一个特定词语在当年相对于其他词语的使用情况，显示出其在整个语言生活中所占的地位。

3.2 基于词语信息的宏观语言现象表征

3.2.1 基于词语信息的历时语言丰富程度度量

词形数的增减从一个方面反映了语言生活的丰富程度。而更具有说明力的指标是香农熵。香农熵的公式如下[9]：

$$Entropy=-\sum_{i=1,w_i\in W}^{i=n}p(w_i)\log p(w_i)$$

其中 W 为语料中的全体词汇，设共 n 个词，w_i 为第 i 个词。$p(w_i)$ 为第 i 个词在语料库中出现的概率。熵值的增高表明所有词间使用频率的差异较小，系统趋于平均和混乱。熵值的降低则表明词语使用的频率并不那么平均。图 2 为各年词的熵值变化。与图 1 类似，在 1970 年前后落到谷底，而随着改革开放的开始而逐渐回升。香农熵的计算中带有词语使用的概率信息，较词形数变化，可以更全面的反应语言生活的丰富度。

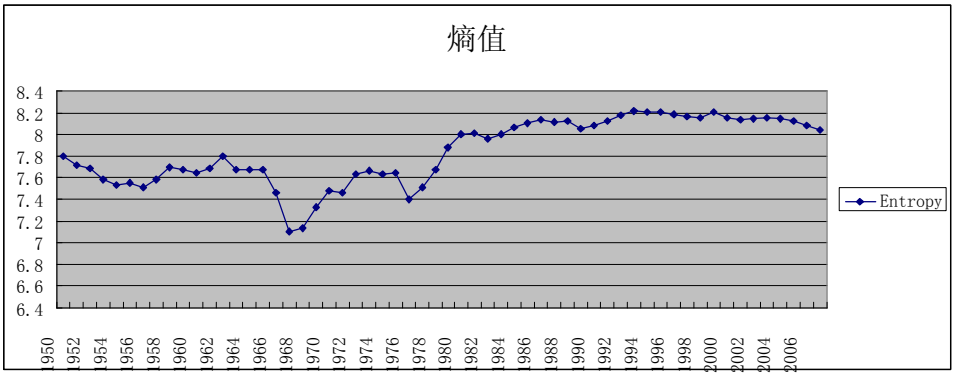

图 2. 各年语料的词熵变化

香农熵的计算是基于当年全部词汇进行。我们提出一种更加直观而灵活表现语言丰富程度的方式——年内前 N%累积和。其定义如下：每年词表中的词目，按频率降序排列，累积频率（也被称作覆盖率）达到 N% 时的词数 Y（如，公式 1）。

$$\sum_{1}^{Y}q(w_i)=N$$

公式 1　Y 代表年内 topN 累积和，即达到累积频率时词的个数；q(w)为词表中词 w 的频率，词表按频率大小从大到小排练；N 为待选定的累积频率。

显然，当达到指定累积频率所需的词越多（即频率累积的越慢），表明词汇使用的越分

散，丰富程度越高。反之亦然。图 3 为 1950-2007 年的年内前 30%累积和。与图 2 类似，只是更为明显。词汇使用的丰富程度改革开放前总体低于改革开放后，文革十年是一个明显低谷。这符合我们的生活直觉与传统语言学对语言发展的认识 [10]。

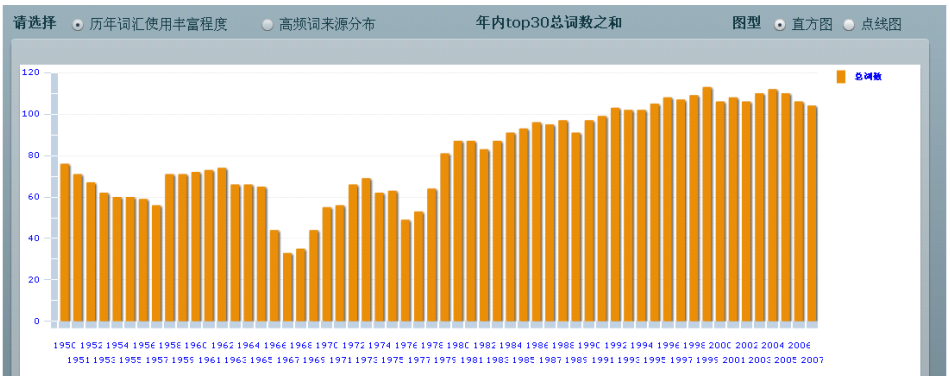

图 3. 年内前 30%累计和

3.2.2 基于词语分布的高频词历时分布描述

我们使用总词表前 N%累积频率的历时分布来描述高频词的来源，定义如下：使用全部语料形成的总词表，按照频率降序排列，当累积频率达到 N%时，该范围内的词语（公式 2-1,2-2）在各年中出现频率之和（如公式 2-3）。以前 50%为例，总词表中按频率降序，当频率累积到达 50%时，共有 t 个词。这 t 个词在 1959 年中，出现频率之和，即为 1959 年对总词汇的贡献情况。这一指标表征了高频词的历时性分布与构成。

$$\sum_{rank=1}^{rank=t} q(w_i) = N \text{.....(1)}$$

$$S = (w_1, w_2, w_3 \ldots w_t) \text{...(2)}$$

$$Y = \sum_{w_i \subset S, i=0}^{t} p(w_i) \text{.....(3)}$$

公式 2 （1）：N 为待选定的累积频率；$q(w_i)$为全部语料形成的总词表中词 w_i 的频率，词表按频率降序排列；（2）：S 是从总词表中按照频率从大到小取词，其累积频率达到 N 时所取出词组成的集合。（3）：$p(w_i)$为 w_i 在某一年（横坐标所指示的年份）中出现的频率，将公式 2 上所取出的集合 S 里所有的词累加得到的频率和即为当年语言对总高频词汇的贡献和 Y。

图 4 是总词表前 50%累积频率的历时分布直方图。从变化幅度上可以看出该项指标对词汇历时分布的敏感性。同时，也可以看出改革开放后的词语使用对总词汇中使累积频率达 50%的词汇有更重要的贡献，即改革开放后的词语使用对今天的影响更大。

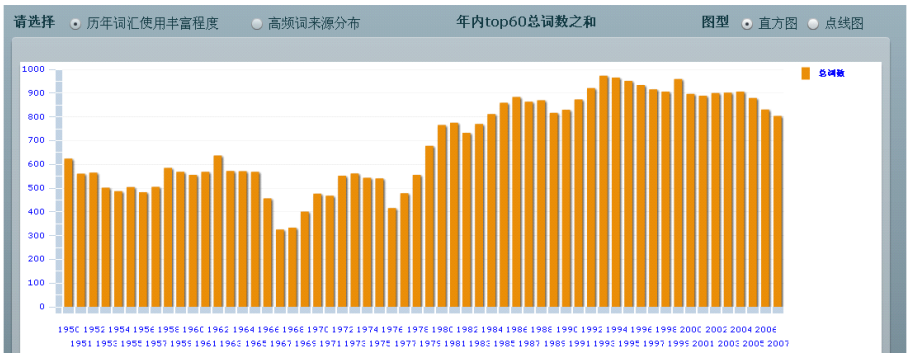

图 4. 总词表前 50%累计频率历时分布

4. 现代汉语词汇历时检索系统的设计与实现

基于上一部分所讨论的几种表征词语历时使用状况的要素，我们设计了现代汉语历时检索系统，提供在线词语查询和语言丰富度计算。我们将所收集语料，按照来源时间，以年为单位分割。使用北京语言大学研发的 GPWS（通用自动分词系统）对其进行分词和命名实体识别[11]。经过此步骤后即可抽取出各年的词表与总词表。通过全文检索系统对全部语料建立了倒排索引，并在索引中加入时间标记。基于此，计算所有词在各年和全部时间段的频次、频率、频序与累积频率（覆盖率），形成支撑服务的后台数据。系统设计流程图如图 5 所示。

如图 6 所示，用户在下拉框选择历年或全时高频词的覆盖率（如前 20%，前 30%等等），可通过高频词历时分布统计从宏观上观察语言使用状况。在检索框中输入待查询词语，检索词语历时信息（历年频次、频率、频序）以直方图和折线图的形式可视化显示。在直方图或折线图上点击某特定年份，便可获得当年待查询词的使用实例。以查询词为中心，上下文窗口为 20 个字，显示检索结果实例，方便研究者在统计数据之外能更详实直观的了解某时间点上的语言现象。

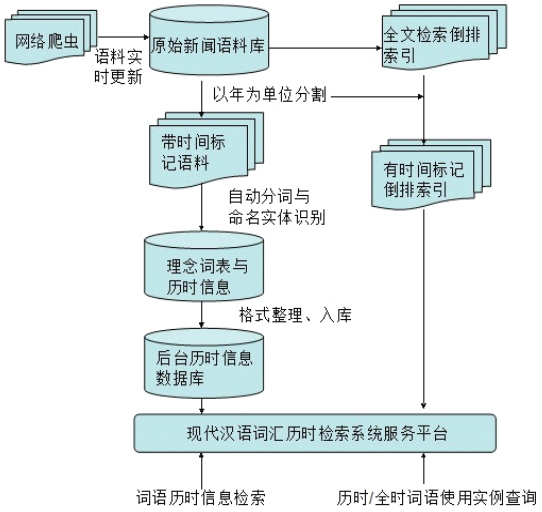

图 5. 系统设计流程图

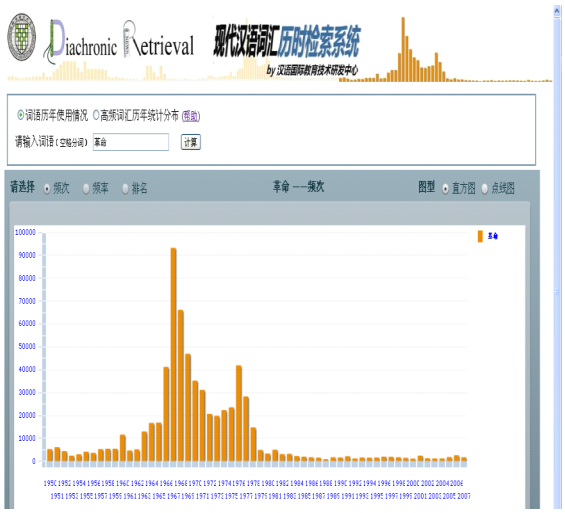

图 6.用户界面

5. 系统应用与未来工作

现代汉语历时检索系统自 2012 年 5 月初上线以来，展现出了较高的实用性与可用性。期间进行了一次语料扩充（延伸为 1951 年至 2012 年）和两次用户界面改版。用户的高频查询主要是新词和公共领域相关概念两方面。由于报刊新闻语料的特点，本系统主要功能体现是后者。对于新词，如宅女、忽悠等随着经济文化事业产生的词，不如网络语料反应快，但可以通过实时的新语料抓取来得到部分满足。公共领域相关概念有环保、减肥、听证会等。单个词语使用的变化，从一个侧面揭示了一类社会问题、社会现象发生发展以及受关注的过程。而这类词总数的增多和使用频率的增加，表明了公共空间作为社会发展标志，从无到有、从小到大的过程，是符合生活直觉和社会发展规律的[12]。

2002 年，教育部发布了《第一批异形词整理表》[13]，对 338 个异形词对进行了整理和规范。异形词的整理工作需要照顾到语言事实并充分考虑文化传承，在大时间跨度上的统计分析是十分重要的。以“身份-身分”为例。“身份”为推荐词形。从图 7 中可以看出，两者长期稳定共存（两者都一直使用，无间断），但是“身份”在 1961 年及其后均占据了绝对优势。该异形词对的选择都得到了“大数据实证”上的支持。对于未涵盖的词对，以“交待-交代”为例，从图 8 中可以看出在七十年代以后两者频率降低并逐渐趋同。

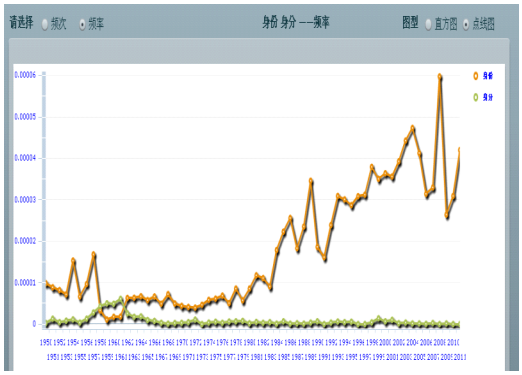

图 7. 身份-身分频率变化图

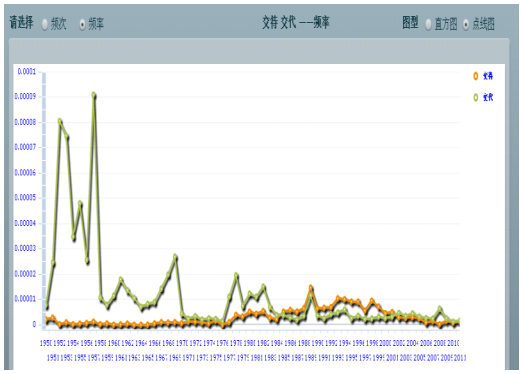

图 8. 交待-交代频率变化图

就同一字/词而言，其使用和语义在漫长的时间流转中也会发生巨大的变化。以“炒”为例，1950 年检出的 45 次使用中，全部为“把食物放在锅里加热并随时翻动使熟”，然而在 1996 年检出的 245 次中仅有 101 次为此义，其余为表示“频繁买卖”，或者是南方方言中表示解雇的“炒鱿鱼”，以及表示“扩大影响”。一个有趣的现象是南方方言中表示解雇的“炒鱿鱼”。在 80 年代初进入新闻出版语言的时候共检出两次，均是在双引号中引用；在 1993 年 17 次检出中有 11 次在双引号中；而到了 2004、2005 年各有一次检出，均不在双引号中。期间所伴随的事件便是 1999 年开始修订的《现代汉语词典》最终收录了“炒鱿鱼”。

词语的历时信息体现了词语在语言社团中的使用，对语言社团中重大事件的发生有着很好的表现作用。词语取代现象还可以微观的体现出语言生活的许多变迁。以南朝鲜-韩国两词的频率查询为例。如图 9 所示，南朝鲜在 1960 前后出现使用高峰，恰好对应了冷战进入高潮，武装对峙白热化。韩国和南朝鲜的使用频率在 1992 年出现交叉。1992 年之前，几乎不使用韩国这一称谓，之后则迅速停用了南朝鲜这一称谓。这一节点所标示的历史事件即中韩于 1992 年建立外交关系。图 10 为科学技术-科技的频率图，直观地显示出了“科技”取代“科学技术”的过程。

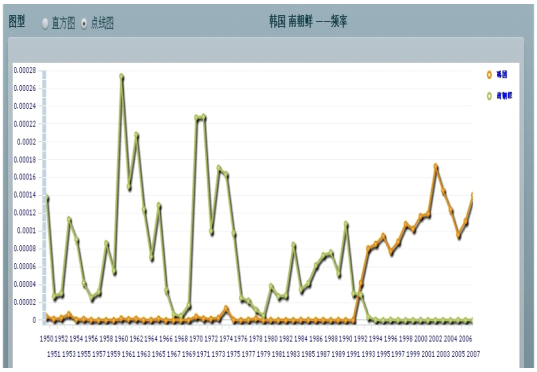

图 9. 南朝鲜-韩国频率图

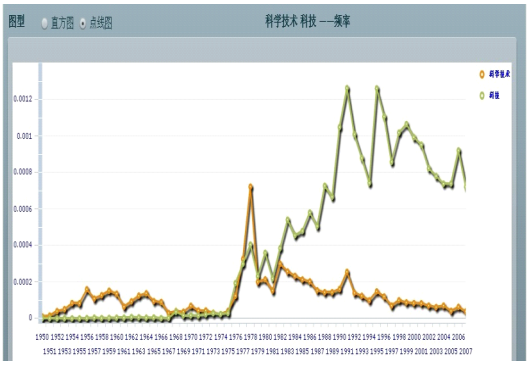

图 10. 科学技术-科技频率图

缩略语随着原短语使用的增长，自身使用也增长，基于人类交际的最省力原则，最终取代本词。基于社交网络、微博和 tweeter 的公共事件预测研究方兴未艾[14-16]，与本系统探测事件发生和语言趋势的原理本质上类似，都是利用了群体智慧。历时的语料数据，尤其是词信息数据在何等程度上有助于语言使用情况的预测，乃至热点的追踪和挖掘，将是十分值得深入研究的问题。

许多词在不同时代有迥异的语义，其使用情况亦大为不同。我们通过历时语言实例的查询能够对其进行一定区分。在词语的研究方面上，现在的词语历时检索系统是面向词语使用情况的历时变化，等于说是基于一元语法（Unigram）的统计研究，怎样合理地注入更多上下文信息，利用报纸语料中版面、板块这一天然分类信息，提供分领域的查询和对比，提供更可靠的自动化分析也是未来的研究方向。

此外，基于统计的自动分词技术并不考虑语言的历时特性。前文示例中词语浅层特征在不同时间段上有着明显的差异，这是否可以对统计自动分词提供一定反馈？从资源建设上来讲，单一媒体作为语料来源，必然有其偏执，如何平衡的融合其他不同时间跨度上的语料；如何基于语料特点，寻找具有应用价值的衡量指标，这些都是在这套系统的研发过程中产生的新的学术问题，并期待系统的使用者和开发者共同进行更深入的研究与探索。

参 考 文 献

[1]葛本仪. 词汇的动态研究与词汇规范[A]. 载《词汇学理论与应用》苏新春，苏宝荣主编. 北京：商务印书馆. 2004.

[2]金观涛，刘庆峰. 观念史研究[M] 北京：法律出版社.2009.

[3]张仲民. “局部真实”的观念史研究.《东方早报》2010 年 5 月 23 日 B05 版.

[4]刘长征. 基于动态流通语料库的新词语监测研究[M] 北京：世界图书出版社.2011.

[5]邹嘉彦，邝蔼儿，陆斌，蔡永富. 汉语共时语料库与追踪语料库[J]. 中文信息学报. 2011 年 11 月 25 卷第六期 P38-45.

[6]Jean-Baptiste Michel, Yuan Kui Shen，Aviva Presser Aiden etl. Quantitative Analysis of Culture Using Millions of Digitized Books. Science 331, 176(2011); DOI: 10.1126/science.1199644.

[7]李宇明. 权威方言在汉语规范中的地位[J]. 清华大学学报. 2004 年第五期 P24-29

[8]教育部语言文字信息管理司. 中国语言生活状况报告. 北京：商务印书馆.2009 P525-534

[9]克劳德·艾尔伍德·香农: 《通信的数学理论》 (A mathematical theory of communication) 贝尔系统技术月刊 l,27 卷,379-423

[10]叶蜚声，徐通锵. 语言学刚要（修订版）[M] 北京：北京大学出版社.2010. P264.

[11]宋柔，罗智勇.现代汉语通用分词系统(GPWS v3.5）http://democlip.blcu.edu.cn:8081/gpws/

[12]尤尔根-哈贝马斯. 公共领域的结构转型[M] 上海：学林出版社.1999. P62-74.

[13]《第一批异形词整理表》，中华人民共和国教育部.2002

[14]Shen Yu，Subhash Kak. A Survey of Prediction Using Social Media[C]. ArXive-prints. March, 2012.

[15]路荣，张旸，杨青. 社交网络中新闻趋势的预测分析[J]. 中文信息学报. 2012 年第五期 P85-90.

[16]洪宇，张宇，刘挺，李生. 话题检测与跟踪的评测及研究综述[J]. 中文信息学报. 2007 年第六期 P71-87.