基于词汇聚类方法的现代汉语分期与分期体系构建[1]

饶高琦，李宇明

（北京语言大学 北京市 100083）

摘要：当前对现代汉语史的研究多借用政治-社会史的分期方式将现代汉语分为新文化运动到 1949 年、1950 年到 1966 年、1967 年到 1976 年和 1977 年至今四个时期，并在这一基础上开展了许多研究。语言尤其是书面语虽然与社会政治生活有密切联系，但语言系统有其自身的演化规律。从语言数据出发对语言进行分期是更加合适的选择。

本文将语言的分期问题视作历时语料的分期问题，进而成为历时文本的聚类问题。本文工作基于历时报刊语料库遴选出的时间敏感程度较好的词汇。使用机器学习领域中广泛使用的 K 均值和期望最大算法进行聚类，以该部分词汇频率为特征对 70 年跨度（1945-2015 年）的历时报刊语料进行聚类，并在不同的聚类数量下绘制了具有层次性的词汇使用分期树。构建了战后现代汉语的词汇层次分期模型，揭示了改革开放的开始作为二战后词汇使用变迁最重要分水岭的地位。

关键词：现代汉语、分期、词汇、历时演变、聚类

Lexicon Clustering based Period Dividing of Modern Chinese

Rao Gaoqi, Li Yuming

(Beijing Language and Culture University, Beijing 100083, China)

Abstract: State-of-art research tend to divide modern Chinese into 4 periods according to the political history: new culture movement to 1949, 1950-1966, 1967-1976, and 1977 till now. Though written language is quite influenced by the social and political movements, language evolve by its own pattern. Periods should be divided based on language data.

In this paper, we regards the period dividing as a text classification problem. Based on the time sensitive words and its frequency as features, K-means and EM algorithm are carried out to cluster the corpus of 70 years of "*People's Daily*". Hierarchical dividing system is formed and revealed the beginning of Reform and Open Policy as divide crest of written language use in the past century.

Key words: Modern Chinese, period dividing, lexicon, diachronic evaluation, clustering

1. 引言

现代汉语研究的基础问题是现代汉语的起源与变迁。如此则无法不涉及现代汉语的历时分期问题。以往对现代汉语史的研究多直接借用政治史的分期方式将现代汉语分为新文化运

[1] 本文研究受以下项目资助：北京市语言资源高精尖创新中心项目（TYR17001J）、北京语言大学校级项目（中央高校基本科研业务费专项基金）（17PT05）、教育部人文社科重点研究基地重大项目（16JJD740004）

动到 1949 年、1950 年到 1966 年、1967 年到 1976 年和 1977 年至今四个时期，并在这一基础上开展了许多研究[1-4]。虽然语言生活，尤其是本文使用的报刊历时语料，在内容上与政治生活有密切联系，但语言系统有其自身的演变规律。从语言数据出发对语言进行分期是更加合适的选择。传统的分期方法缺乏定量分析，往往根据相对孤立的语法现象和语法点进行分析，无法根据广泛的语言使用情况来获得叫合理的分期依据。

本文将语言的分期落实在语料的分期中。语料的分期则可以视作不同时间文本的自然分组任务，即聚类问题。本文基于历时的词汇分层工作的结果[5,6]，使用机器学习方法对历时语料库中的文本进行自动聚类。以期从词语使用的角度，进行定量的历时语料时期划分。

2. 基础工作

2.1 历时语料库

本文使用的历时报刊语料为 BCC 历时检索系统[2][7,8]中 1946 年到 2015 年的《人民日报》语料 [3]，时间跨度 70 年，规模 12 亿字。使用 GPWS 通用分词系统[4][9]并辅之以小规模人工修正对历时语料库进行分词，词种数约 220 万。

2.2 时间敏感词

饶高琦（2016）基于 70 年跨度的历时报刊语料库，使用了包括 TF・IDF、互信息、联合熵、变异系数、词项随机采样、修正频率、累积频率等 9 种统计方法计算了词汇在历史中的使用稳定性，并通过对稳定性、覆盖度和时间区分性能的考察，确定了以月为划分文本的时间颗粒度，TF・IDF 为主的计量方法，并获得了规模为 3013 词的历时稳态词候选词集。其中词语的时间敏感性极差，包括功能词和基本名词等，构成语言生活的底层，即基干层[8]。

饶高琦（2017）在基干层之外发现月颗粒度下 TF·IDF 降序 10000 到 50000 词间的词汇在 70 年跨度，以年为颗粒度的条件下对时间变化较为敏感，且频次较大，如：合作社、非典、拨乱反正等。它们与瞬时出现迅速退出使用的命名实体很大不同。这部分词汇称为时间敏感层[9]。许多时间敏感的社会语言现象多由这一层中的词语构成。流行语和年度词亦多出自此层。

2.3 聚类算法

本文选择 K 均值算法和期望最大化算法对历时语料库中的文本进行聚类，并使用机器学习平台 Weka[5][10]实现。

K 均值（K-means）算法是一种十分常用的聚类机器学习算法，也是一种基于距离的迭代聚类算法[11]。本文中的 K 均值算法采用欧式空间距离。其优点是可以确保一个类中每个实例到中心的距离平方和最小。但聚类数量 K 需要人工指定，且只能获得局部距离平方和

---

[2] http://bcc.blcu.edu.cn/hc

[3] 由于种种原因，本文实验过程中没有获得 2003 年到 2008 年的《人民日报》语料，该部分由实验室积累的相应年份的《贵州日报》替补。

[4] 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/

[5] http://www.cs.waikato.ac.nz/ml/weka/

最小值。通常通过对不同 K 值进行多次实验来寻找较优的聚类数量，对特定 K 值进行多次实验则可以一定程度上克服无法获得全局最优聚类的缺陷。根据经验 K 均值方法中的聚类数 K<<N，N 为样本数量，在文中就是历时语料的年数。

期望最大算法（Expectation-Maximum Algorithm）是一种基于统计的聚类方法，其基础是建立一个限混合（finite mixtures）统计模型[12]。期望最大算法在给定一个（随机）初始值后不断进行不断迭代，进行重新估计直到收敛。该算法的优势在于其无需事先制定聚类数量（分布的数量），但同样不能保证收敛于全局极大值。为了有机会获得全局巨大值，需要对同一组数据进行多次试验。

3. 对历时语料进行聚类

本文在历时语料库中提取各年度的词表，使用基干层词汇和时间敏感层词汇两个时间敏感性几乎相反的词集进行处理，以获得进行聚类实验的特征集，处理方法如下。

处理 A：在第 i 年词表 Lex_i 中保留出现的时敏层词汇 S_{sens}，即 $F_{Ai} = Lex_i \cup S_{\text{sens}}$。$F_{Ai}$ 第为第 i 年的特征集。

处理 B：在第 i 年词表 Lex_i 中去除基干词集 S_{Base}，即 $F_{Bi} = Lex_i \cup \overline{S_{Base}}$。$F_{Bi}$ 为第 i 年的特征集。

将两种方法处理后获得特征集 F_{Ai} 和 F_{Bi} 中的词当做聚类特征，其在当年的频率当做特征值，分别使用 K 均值和期望最大化两种方法进行聚类。语料的时间颗粒度为年。

由前文已知 K 均值和期望最大化算法的缺陷，每种实验设定均对相同数据进行 5 次实验以获得稳定的聚类结果。表 1 是 K 均值选取不同聚类数时的聚类结果。

表 1.　K 均值在不同聚类数时的实验

K 值		处理 A	处理 B
K=2	类 1： 类 2：	1946-1979; 1980-2015	1946-1980, 2008-2015; 1981-2007;
K=3	类 1： 类 2： 类 3：	1946-1979; 1980-2000; 2001-2015;	1946-1980, 2004-2015; 1981-2002; 2003;
K=4	类 1： 类 2： 类 3： 类 4：	1946-1979; 1980-1997; 1998-2007; 2008-2015	1946-1980; 1981-2002; 2003; 2004-2015;
K=5	类 1：	1946-1977;	1946-1979;

	类 2：	1978-1988;	1980-1991;
	类 3：	1989-2001;	1992-2002;
	类 4：	2002-2007;	2003;
	类 5：	2008-2015;	2004-2015;
K=6	类 1：	1946-1964;	1946-1955,1958-1960,1963-1972;
	类 2：	1965-1978;	1956-1957,1961-1962,1973-1981;
	类 3：	1979-1989;	1982-1991;
	类 4：	1990-2001;	1992-2002;
	类 5：	2002-2007;	2003;
	类 6：	2008-2015;	2004-2015;
K=7	类 1：	1946,1950-1952,1973-1978;	1946-1955, 1958, 1963-1972;
	类 2：	1947-1948,1965-1972;	1956-1957, 1973-1981;
	类 3：	1949,1953-1964;	1959-1962;
	类 4：	1979-1989;	1982-1990;
	类 5：	1990-2001;	1991-2002;
	类 6：	2002-2007;	2003;
	类 7：	2008-2015;	2004-2015;

表 2 是期望最大化算法自动获得聚类数时的聚类结果。

表 2. 期望最大化算法实验

	处理 A	处理 B
期望最大化	类 1：1946-1979; 类 2：1980-2015;	类 1：1946-1979; 类 2：1980-2000; 类 3：2001-2015;

语言的演变是连续而渐进的，则依据语言特征的时间演变进行聚类，其结果在时间轴上也应当是连续的，即一个聚类应由年份连续的语料组成。

通过对表 1 的观察，发现处理 A 仅在聚类数量增加到 7 的时候开始出现类和类之间互相穿插的现象，如类 1 和类 2 将彼此切割成了 3 段和 2 段。处理 B 在许多实验的聚类数量下容易出现类和类之间互相穿插的现象。如 K=2 和 3 时类 1 被类 2 截为两段；K=6 时类 1 和类 2 则互相穿插多次，而且 K>2 的聚类中始终都存在一个孤立点（2003 年）。这与语言演变具有渐变性的认知有较大的冲突。因而处理 B 在使用 K 均值算法进行的聚类中并不是一种好的选择。处理 A 在聚类数量增加到 7 时，聚类质量也开始变差。下一节将选择聚类数

了 2 到 7 的实验进行分析。

我们尝试对处理 B 的较差效果进行解释。处理 B 是将每年词表去除出现在当年的基干层词后的结果。每年语料的特征数量过于庞大，词汇繁杂。其中既有时间敏感性略差的介于基干层与时间敏感层之间的词汇，也有大量超低频的，出现时间极短的词汇（大多是命名实体）。这些特征对聚类过程形成了一定干扰。

与 K 均值不同，在期望最大化算法的结果中，两种处理方式都有较好的表现。期望最大化算法在处理 B 时的聚类结果和 K 均值算法中聚类数量 K=3 时的结果一致。期望最大化算法中使用处理 A 时的聚类结果和 K 均值算法中聚类数量 K=2 时的结果一致。这也在一定程度上使我们可以更确信的在后文中使用 K=2 和 3 时的 K 均值的聚类结果。

4. 聚类结果分析

语言变化的速率是不均匀的。当变化较快，在一个特定时间单位（如年）内无法刻画变化过程的时候，该时间单位就形成一个**边界**。而较为缓慢的变化可以在几个时间单位内被观察到，这就形成了若干时间单位构成的一个**过渡**（地带），更加缓慢的变化以至于在很长一段时间内保持稳定，那么就形成了前文中所描述的一个聚类，其现实意义就是一个**时期**。本节通过这三种方式对聚类结果进行分析，从而获得语言使用的时期信息。

K 均值在不同聚类数量下的聚类结果给我们提供了一扇观察历时语料分期，尤其是历时词汇使用分期的窗口。K 均值在处理 A 的特征集 F_{Ai} 的实验中存在一些较为稳定的聚类边界。如 1979-1980 年的边界在聚类数为 2、3 和 4 时均无变化，在聚类数为 5、6 时变化为 1977-1978 年边界和 1978-1979 年边界。聚类边界的移动或变化是算法受聚类数量影响的结果。但其移动幅度小则说明类的聚类比较稳定。聚类的小幅度移动也符合对语言使用演变是渐变的假设，刚性的边界在语言使用的变化中可能不多见。因此本节将聚类的边界模糊处理，即将较少的若干个样本（即若干年的语料）视作两个聚类的过渡，如 1979-1980 年边界可以扩大为 1977-1980 年过渡。

聚类数为 3 时的 2000-2001 年边界在聚类数量为 4 到 6 时可以扩大为 1997-2002 年过渡。聚类数为 5 时的 1988-1989 年边界在聚类数量为 6 时可以扩大为 1988-1990 年过渡。聚类结果中也存在不变化或移动的边界，如 2007-2008 年边界在聚类为 4 时出现后，在后来的实验中保持稳定，并未变化。而 1964-1965 边界出现的很晚（K=6 时才出现），也未移动或变化。

如果将上述边界和过渡都整理到一张树状图（图 1）中，就可以较清晰的看到 1946 年到 2015 年历时语料由词汇使用来划分的年代分期情况。

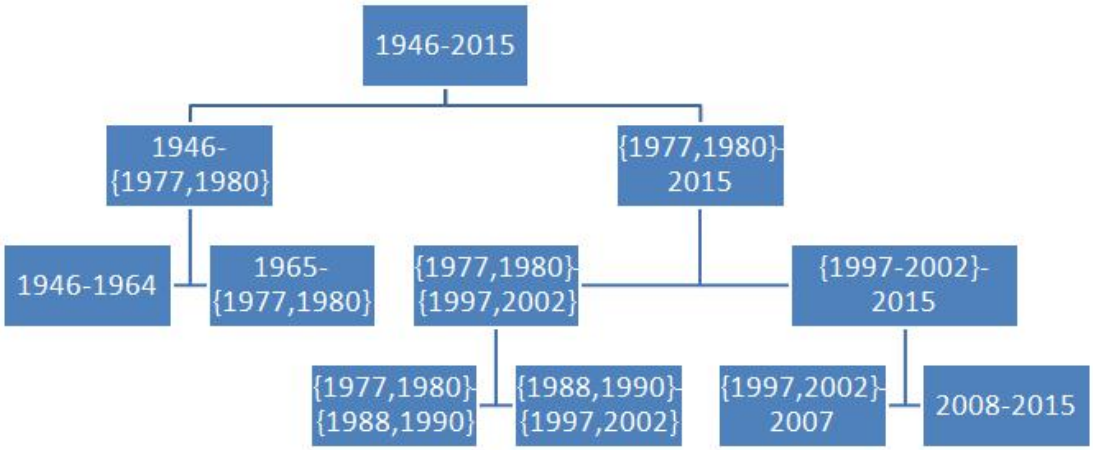

图 1. 基于 K 均值方法的历时语料词汇使用分期

图中{m,n}表示年份 m 到年份 n 形成的过渡，如{1988,1990}为前文中描述的 1988-1990 年过渡。我们将边界、过渡和聚类的数量同时映射到一张树形图上，可以得到图 2，以反映分期和聚类数量之间更直观的关系。

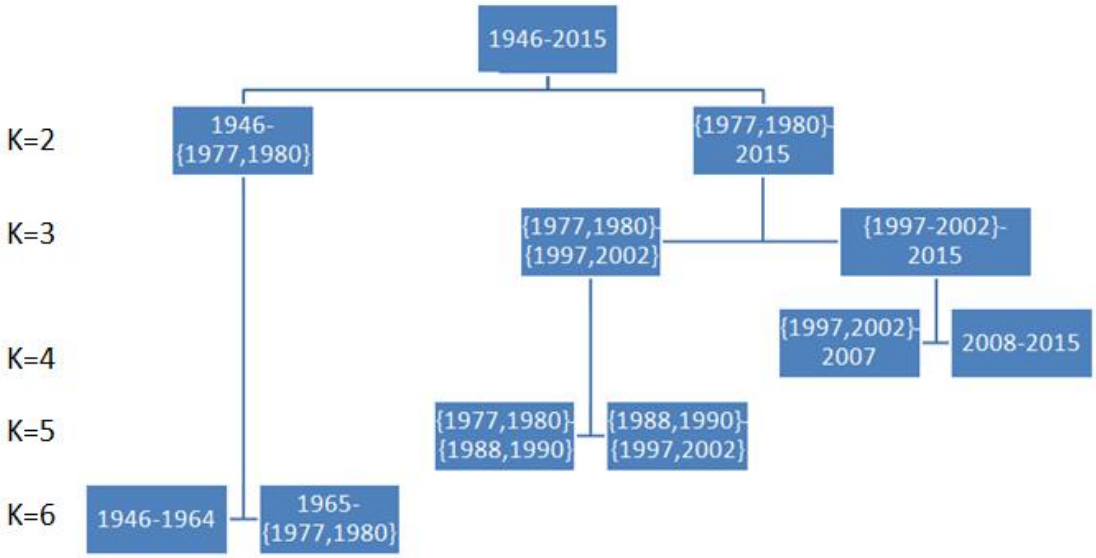

图 2. 基于 K 均值方法的历时语料词汇使用分期及其聚类数量

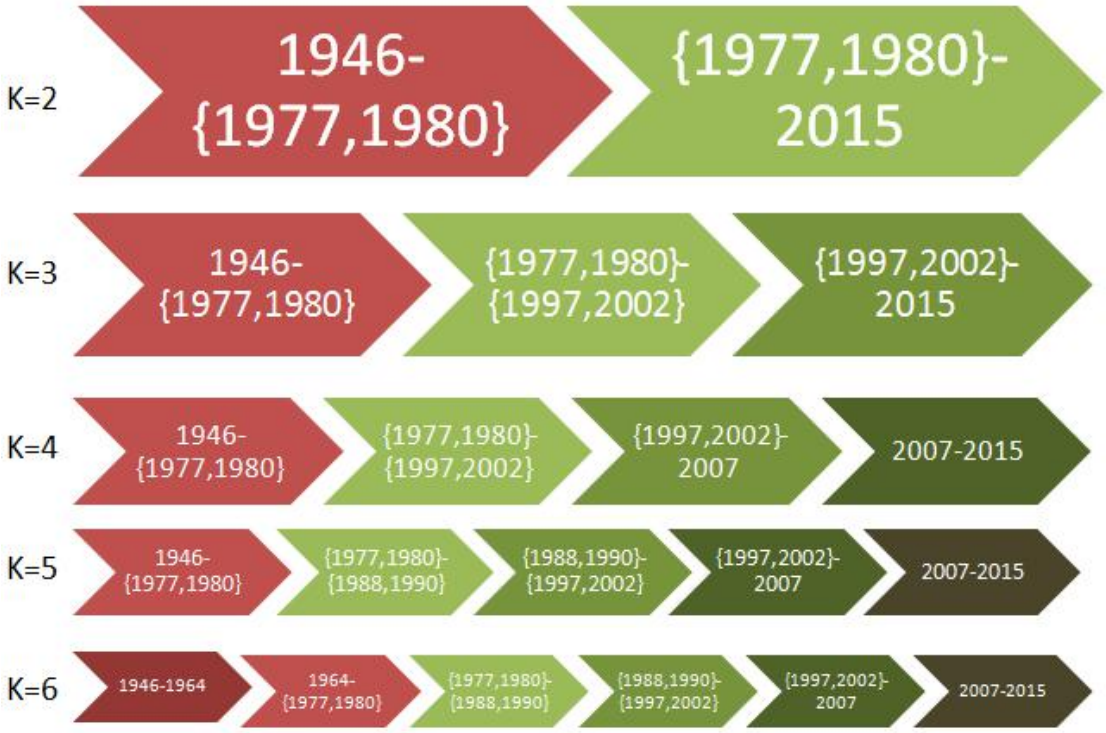

图 3. 历时语料词汇使用分期及其聚类数量

图 3 则将不同聚类数目中的分期结果绘制在时间轴上。颜色之间的差异用来表示不同的聚类分裂的早晚和亲疏关系。

如前文所述，这些“边界”或“过渡”中，1977-1980 年过渡和 1997-2002 年过渡也在期望最大化算法中出现，将语料分为两个或三个相对稳定聚类。而它们出现时的实验设定（K=2 和 3）也表明，如果只将历时语料划分为两段，那么应该选择 1977-1980 年进行切分。如果划分为三段，应该再将八十年代至今划分为八十年代到两千年和新世纪以来两个阶段。

在这一划分结果和对过去 70 年语言生活变迁的直观感受基本相符。1977-1980 年过渡是改革开放政策开始并逐渐起步的阶段，语言使用的情况随着国人思想的变化焕然一新。可以说改革开放的开始是过去 70 年词汇使用变迁最重要的分水岭。

1997-2002 年过渡则是在改革开放渐入佳境，共和国综合国力高速上升的阶段。语言生活和媒体的发展步入新的阶段。但是这一过渡在已有的研究中很少被注意到。刁晏斌（2006a、b）和 Gaoqi Rao（2015）的研究中，都借用政治史将共和国建立后的现代汉语白话以文革运动为界分为三段。王建华（2006）意识到“跨世纪稳定发展期”的存在，但是将 1990 年至今的时段划为此段。涉及到新政权建立前的语料，则简单的以 1949 年为边界分为两段。并没有注意到在词汇使用的层面，新政权建立在语言上所产生的冲击不如改革开放，甚至不如进入新世纪的影响大。

聚类的分裂率先从图 2 第二层的右侧（也就是图 3 的第一层的绿色部分）开始。当二分类时的类 2 已经分裂为四个类的时候，二分类时的类 1 才开始分裂。这从一个侧面展现了改

革开放前后词汇使用更新速度的差异。改革开放以前词汇使用总的来说变化缓慢，该聚类较之改革开放后的更为稳固，分裂的晚。

5. 两层三段分期体系

综上所述，本文将 1946 年到 2015 年共 70 年的历时语料的时期划分任务分为两个层次。第一层次分为两段，1946 年到 1977-1980 过渡为一段（E1），之后为一段（E2），并以 1980 年为实际操作时的边界。第二层次分为三段，即第一层次中的第二段进一步分为 1980 年到 1997-2002 年过渡为一段（E2.1），之后为一段(E2.2)，并以 2000 年为实际操作时的边界。为行文简便，后文中也使用 E1、E2、E2.1 和 E2.2 指代该分期体系中的时期。该体系示意如图 4 所示。

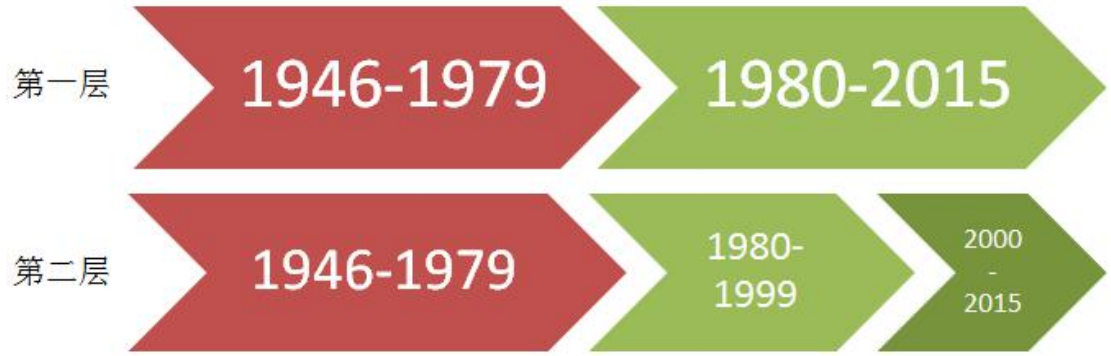

图 4. 两层三段分期体系示意图

对应 Gaoqi Rao（2015）在传统分期方法下进行的词语使用统计，本文在新的分期体系下对用词情况进行了统计，如表 3、4 和 5 所示。

表 3 在第一层次分两段时各段的用词情况

时间段		**E1:1946-1979**	**E2:1980-2015**	**全部时期**
总词种数		989060	1455099	2118768
均词种数		102751	145620	124798
总词次数		319749805	415512150	735261955
均词次数		9689388	11871776	10503742
各年间共用词	词种数	12396		
	比例%	1.25	0.85	0.59
	词次数	279021752	342777889	619562690
	比例%	87.26	82.50	87.41
时期间共用词	词种数	312384		
	比例%	31.58	21.47	14.74
	词次数	317952769	408460755	69973807
	比例%	99.44	98.30	98.76
时期独用词	词种数	676676	1142715	1819391
	比例%	68.42	78.53	85.87

	词次数	1797036	7051395	8848431
	比例%	0.56	1.70	1.20

表 4 在第二层次分三段时各段的用词情况

时间段		**E1:1946-1979**	**E2:1980-2015**		**全部时期**
			E2.1:1980-1999	**E2.2:2000-2015**	
总词种数		989060	1071206	663249	2118768
均词种数		102751	165850	120333	124798
总词次数		319749805	247719547	167792603	735261955
均词次数		9689388	13037871	11186174	10503742
各年间共用词	词种数	12396			
	比例%	1.25	0.12	0.19	0.59
	词次数	279021752	205666217	137111672	619562690
	比例	87.26	83.02	81.71	87.41
时期间共用词	词种数	167508			
	比例%	16.93	15.64	25.26	7.91
	词次数	316424457	243417870	164310734	677722638
	比例%	98.96	98.26	97.92	98.44
时期独用词	词种数	676676	670460	360409	1707545
	比例%	68.42	62.59	54.34	80.59
	词次数	1797037	1379665	1042603	4219305
	比例%	0.56	0.56	0.62	0.57

表 5 各时期用词覆盖度

覆盖率	达到 80%		达到 90%		达到 99%		达到 100%
	词种数	比例	词种数	比例	词种数	比例	词种数
E1: 1946-1979	2105	0.21%	6081	0.61%	71214	7.20%	989060
E2: 1980-2015	2973	0.20%	8692	0.60%	108758	7.47%	1455099
E2.1: 1980-1999	3020	0.28%	9083	0.85%	109477	10.22%	1071206
E2.2: 2000-2015	2602	0.39%	7207	1.09%	80283	12.10%	663249
全部语料	3007	0.14%	8784	0.41%	111463	5.26%	2118768

表 3 容易发现 E2 在词语使用的丰富程度大大超越 E1，这在总词种数和年均词种数上都得到体现。在表 4 中比较年均词种数可以发现 E2.1 处于最高峰。表 5 统计了各时期达到特定词语覆盖度（按词频降序获得的词汇累积频率）所需的词数。该组数据也刻画了诸时期的词语使用丰富程度，比例越高说明词汇使用越丰富。表 5 数据体现出 E2 高于 E1，E2.2 高于 E2.1 高于 E1 的趋势。表明虽然高频段词种数在 E2.1 较多，但内部分布于 E2.2 更为平均，词汇分布在高频段更为多样。总体而言，改革开放后（E2）的词汇丰富程度有了明显提高，并且呈现出先增长（E2.1）后回调（E2.2）的态势，词汇使用的多样性持续增长。

表 3、4 和 5 所示数据与传统分期中用词简况[3]所展现的数据趋势差异并不悬殊。首先是因为双层分期体系依据时敏层词的使用状况进行的分期而非全体词汇的频率，表 3 和表 4

中着重分析的共用词恰是基干层的重要部分；其次是因为时期的划分本身具有一定的模糊性；这也恰恰表明仅仅通过对词汇系统做整体的频次统计，难以获得时期划分的线索。

6. 小节

不同于过去借用政治史对现代汉语白话文进行分期的方法，本文工作使用统计聚类方法，以具有时间敏感性的词汇的使用频率为特征对 70 年跨度的报刊语料进行了聚类，寻找到了较稳定的聚类，并在不同的聚类数下绘制了具有层次性的词汇使用分期树。本文以 1980 年和 2000 年为实际操作边界，构建了两层三段分期体系。从纯粹的语言学数据出发进行语言分期，打破了现代汉语白话文分期借鉴政治史分期的局面，揭示了改革开放的开始作为过去 70 年间词汇使用变迁最重要分水岭，世纪之交为第二重要的地位，并显示了语言使用相对于社会变革的短暂滞后效应。

参考文献

1. 刁晏斌. 现代汉语史概论[M] 北京：北师大出版社.2006a.
2. 刁晏斌. 现代汉语史[M] 北京：人民出版社.2006b.
3. Gaoqi Rao, Endong Xun. Words and Characters in Official Newspapers since the Foundation of PRC: Guizhou Daily and People's Daily as Examples[C], International Journal of Knowledge and Language Processing (IJKLP), 2015, 6(2):23-33.
4. 王建华，周明强，刘福根. 信息时代报刊语言跟踪研究[M]，杭州：浙江大学出版社，2006.
5. 饶高琦,李宇明. 基于 70 年报刊语料的现代汉语历时稳态词抽取与考察[J]. 中文信息学报,2016,(06):49-58.
6. 饶高琦,李宇明. 基于词频逆文档频统计的词汇时间分布层次[C]. 第十八届汉语词汇语义学研讨会，乐山，2017.
7. 荀恩东，饶高琦，肖晓悦，臧娇娇. 大数据背景下 BCC 语料库的研制[J]. 语料库语言学，2016, 3(1):93-118
8. 荀恩东，饶高琦，谢佳莉，黄志娥. 现代汉语词汇历时检索系统与应用研究[J]，中文信息学报，2015(3): 169-176.
9. 罗智勇. 现代汉语通用分词系统的技术与实现[D]. 北京工业大学, 2002.
10. Ian H. Witten, Eibe Frank, Mark A. Hall. Data Mining: Practical Machine Learning Tools and Techniques (3rd Edition). Press. Morgan Kaufmann, 2011.
11. Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 46 (3): 175-185.
12. Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B. 39 (1): 1-38.