文章编号：1003-0077(2017)05-0008-06

词语序差的分布特点与文本间词汇异同

刘 锐[1,4]，孙碧泽[2]，龙云飞[3]，王 珊[4]

（1. 厦门大学 中文系，福建 厦门 361005；
2. 南京大学 中文系，江苏 南京 210023；
3. 香港理工大学 电子计算学系，香港；
4. 香港教育大学 中国语言学系，香港）

摘 要：该文在已有关于“频级”“频序”研究的基础上，结合两种不同类型的语料，采用词汇计量分析方法，考察词语的“序差”所具有的分布特点。该研究发现，对于两种文本的共有词集，词的序差呈对称分布，且集中分布于中位数附近，存在离群值序差。这一特点在序差图上表现为“中段平直，双尾翘曲”的“双尾分布”形态。根据词语序差的分布规律，可以将文本共有词划分为“中段”“下尾”“上尾”三个层次。“中段”词语反映两个文本的共性特征，“下尾”及“上尾”词语反映两个文本的差异性特征，这些特征具有反映文本的主题内容和文体风格的语言学意义。

关键词：序差；双尾分布；主题内容；文体风格

中图分类号：TP391 **文献标识码**：A

Lexical Frequency Rank Difference Distributions Between Texts

LIU Rui[1,4], SUN Bize[2], LONG Yunfei[3], WANG Shan[4]

(1. Department of Chinese Language and Literature, Xiamen University, Xiamen, Fujian 361005, China;
2. Department of Chinese Language and Literature, Nanjing University, Nanjing, Jiangsu 210023, China;
3. Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China;
4. Department of Chinese Language Studies, The Education University of Hong Kong, Hong Kong, China)

Abstract: Based on previous studies on frequency and frequency rank of words, this paper focuses on the analysis of the frequency rank difference (FRD) from the perspective of lexical quantitative analysis. This paper reveals that for the common words between texts, the FRDs are distributed symmetrically and gathered around the median. This characteristic assumes a "two-tailed distribution", which is flat in the middle and curving in both ends. Three lexical levels, i.e. middle, downward end and upward end, are summarized based on the FRD distributions. The middle lexicon reflects the common characteristics of the two texts, while the lexicon that belongs to both ends reflects their own distinctive features. These features are of linguistic significance in reflecting the thematic content and stylistic features of the texts.

Key words: frequency rank difference; two-tailed distribution; thematic content; stylistic features of the texts

1 引言

在词汇计量分析中，对词语的频率信息关注最多，如高频词、低频词、独有词、共有词。基于词汇的文本特征分析也以词语的频率信息为基础，较为常用的做法是在 TF-IDF 的基础上，结合词语分布比例、词语的类分布、词语位置因子、本体语义关联等因素进行文本特征词分析和提取[1-4]。

词语除了频率信息以外，还有“位序”的信息。美国学者 Zipf 发现人类语言的真实文本中，词出现的频数与其频数秩（位序号）之间具有反比例关

收稿日期：2017-03-03 **定稿日期**：2017-05-16

基金项目：香港教育大学(Internal Research Grant; Project No.: 15214, Activity Code: R3733, Reference Number: RG 92/2015-2016)

系[5-6]。进而研究者对英语中从音素到语句等不同结构层次的频数—序号关系进行了统计研究[7]，汉语的相关统计规律也得到了实证[8]。此后在本体研究[9]、词典编纂、词表研制[10-12]、中文信息处理[13]、语言监测[14]、风格分析[15]、词语历时稳态分析[16]等研究中对字词的位序信息均有讨论和运用。值得注意的是，文献[15]进一步发展了对词语位序信息的使用，从共有词的“序差”入手，以“同中求异”的思路来提取文本的区别特征。但是该研究把词语的位序信息作为一个统计量来使用，对其内在规律却少有讨论。

从“序差”入手进行“同中求异”的分析利用了词语位序分布的什么性质？“序差”信息能够反映哪些词语在文本间分布的规律？分析技术和操作程序上有无进一步改进和规范化的空间？本文在已有关于“频级位序的差比”[17]、“序差”[15]研究的基础上，对两个文本共有词的序差进行整体性的分析，考察其分布上的规律和特点，并分析该分布所反映的文本间词汇使用异同，进而讨论其在文本词汇特征分析中的作用。

2　序差的分布特点

2.1　“频级”与“频序”

为了保证实验数据具有可比性，我们选择长度接近、时期相同的文本作为实验材料。文献[18]指出，影响文本词汇分布有两大因素，一是文本的主题内容，二是文本的文体风格。因此，将“主题内容”和“文体风格”作为一组控制变量，经过网络检索筛选，确定了 T_A 和 T_B 两种文本，内容主题都是“厦门”，T_A 是纪录片解说词文本，T_B 是散文文本。语料基本情况见表 1①。

表 1　T_A、T_B 语料基本情况

文本	字数	词语数	词种数	共有词数
A	11 535	7 143	2 382	507
B	9 580	6 735	2 305	

文献[14]指出，“频级”是在由调查对象形成的列表中根据频次的多少所划分的级别，相同频次或某一频次段的调查对象可划为一个频序。在已有研究中，“频级”既指按照“某一频次段”的划分，也指按照“相同频次”而进行的划分。前者是根据研究需要而进行的主观划分[9,11,13,19]；后者是由频次统计而自然形成的[5-6,8,10,12,15-16]。本文讨论的是后者，称为“频序”(frequency order)。故频序指调查对象按照频次由高到低而形成的自然数序列。频次最高的对象，其频序为 1；频次相同的对象，其频序相同。分别求出两个文本中词语的频序，如表 2 所示。

表 2　T_A、T_B 的频序

词语(T_A)	词频	频序	词语(T_B)	词频	频序
的	291	1	的	503	1
厦门	199	2	厦门	116	2
在	78	3	在	116	2
是	53	4	我	101	3
了	46	5	是	100	4
……	……	……	……	……	……
人民日报	1	39	欢	1	35
人人	1	39	话	1	35
拳拳之心	1	39	怀旧	1	35

2.2　序差

序差(frequency order difference，FOD)是指两个自然语言文本或文本集合的共有词的频序之差。根据定义，某词的序差就等于该词在两个文本中的频序之差。例如，“的”在 T_A 和 T_B 中的频序都是 1，则其序差为 0；“我”在 T_A 里的频序是 37，在 T_B 里的频序是 3，则其序差为 34；“海峡”在 T_A 里的频序是 8，在 T_B 里的频序是 35，则其序差为 −27。

如果序差零散地排列，将无助于发现其数据特征，因此要对数据进行处理和分析。序差是一组有正有负的数字，可以进行升序或者降序排列，得到序差序列。将 T_A 和 T_B 共有的 507 个词按序差升序排列得到序差序列，如表 3 所示。

关于“序差”有以下三点需要说明：

(1) 词语序差的大小反映该词在文本间的地位差别。文献[15]指出“序差的大小反映了该词在不

① 解说词文本“《风从大海来》——献给厦门经济特区建设30周年三集电视专题片解说词”，来源：http://www.xm.gov.cn/xmyw/201112/t20111225_448528.htm，日期 2011-12-25。散文文本由三篇合成，分别为：《悠闲的厦门》：http://www.tianya.cn/publicforum/Content/no16/1/70934.shtml，日期 2006-2-4；《难忘厦门风姿》：http://lpssyy.blog.163.com/blog/static/387398200941822843373/，日期 2009-6-29；《怀念厦门》：http://blog.sina.com.cn/s/blog_5045f7f40100d9bf.html?tj=1，日期 2009-4-2。所用分词软件为 ICTCLAS2016 分词系统。

同文本中的地位差别”。例如,“东南”的序差为2,反映该词在 T_A 和 T_B 的地位差别不大;而“海峡”的序差为-27,说明其地位差别比“东南”大。

(2) 序差的大小指的是序差的绝对值,其正负反映的只是频序相减的顺序。例如,“城市”序差为3,“环境”序差为-3,序差的大小(绝对值)一样,但“环境”的序差-3表示该词在 T_A 里的频序要高于在 T_B 里的频序,“城市”则相反。

表3 T_{A-B} 序差序列表

词语	T_A 频序	T_B 频序	T_B 修正频序	序差	修正序差
海峡	8	35	39.00	-27	-31.00
台湾	6	29	32.29	-23	-26.29
中国	9	31	34.53	-22	-25.53
全国	21	35	39.00	-14	-18.00
等	21	34	37.88	-13	-16.88
……	……	……	……	……	……
如	39	18	20.00	21	19.00
看	38	16	17.76	22	20.24
鼓浪屿	34	8	8.82	26	25.18
去	37	11	12.18	26	24.82
我	37	3	3.24	34	33.76

(3) 修正序差。从表2可以看到 T_A 和 T_B 的频序总数是不相等的,分别是39和35。两个文本的频序在大小值上是不对等的,这会对其反映词语地位差别造成系统性影响,因此需要进行两端对齐的修正操作。两端对齐是指以较大频序数为基准,将频序数少的文本的频序按比例放大。在这里就是以 T_A(频序数39)的频序为基准,将 T_B(频序数35)的频序进行放大修正。修正公式如式(1)所示。

$$f_B^* = \frac{(s_A-1)}{(s_B-1)} \times (f_B-1)+1 \qquad (1)$$

s_A 是 T_A 的频序总数,s_B 是 T_B 的频序总数,f_B 是一个词语在 T_B 中的频序,f_B^* 为修正序差。再根据修正后的频序求差值,就得到修正序差①。下文讨论中的序差如非特别说明,均指修正后的序差。

2.3 序差的“双尾分布”特点

序差序列把词语的“地位差别”集中并有序地表现出来。文献[15]认为,序差序列把不同文本之间的差异有序地排列出来,何者是有价值、价值最大的,区别特征一目了然。本文认为序差序列对序差起了组织整理的作用,但由于序差数量众多,并没达到“一目了然”的效果。因此,本文用图表方法对序差数据进行描述和分析。

在表3的基础上,按照序差(D)升序的排列顺序,给每个词从1开始顺次标号(r),则一个词的位置在坐标系中为(r, D),将全部507个词按此方法表示在坐标系中,得到序差的散点分布图(图1)。从修正前后来看,序差分布的趋势基本相同,散点图整体向 x 轴平移,散点的分布更加平滑。

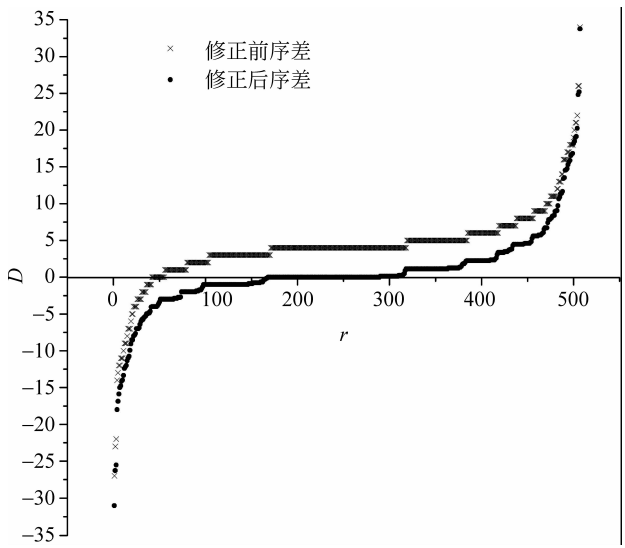

图1 T_{A-B} 序差双尾图

词语序差分布散点图在形态上很有特点,呈“中段平直,双尾翘曲”状。众所周知,齐普夫图反映出词频和词的序号之间呈“长尾分布”(long-tailed distribution)。根据序差序列的图形分布特点,类比称之为“双尾分布”(two-tailed distribution)。序差的双尾分布反映出词的序差和排序号之间的关系。

双尾图的分布形态特点反映出词的序差不是无规律的。为了进一步发掘序差的数据特点,我们使用箱式图及相关参数来描述和分析。箱式图(box-plot)也称箱须图(box-whisker plot),采用一组数据中的最小观测值(Lower bound)、第一四分位数(Q_1)、中位数(Q_2)、第三四分位数(Q_3)、最大值观测值(Upper bound)和中间四分位数极差(interquartile range, IQR)来反映数据分布的中心位置和散布范围,可以对数据的离散分布程度、对称性、异常值等进行观察和分析。

使用OriginPro 9.1计算并绘制507个词的序

① 文献[20]中未进行修正处理,而是从序差集中的角度进行了讨论。本文通过修正处理,使得序差的计算更加严密合理,也使序差的分布图能更加直观地反映序差分布特点,简化相关讨论。

差箱式图(图 2)，左边的点状图为序差数据的箱式分布，右边为序差的箱式图，相关参数如图 2 所示。根据箱式图可以发现：

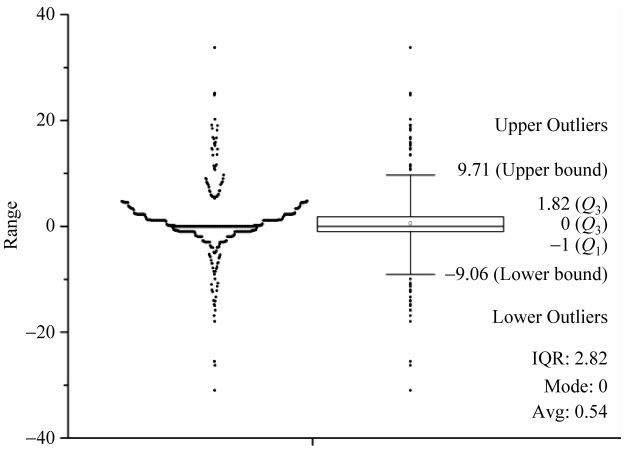

图 2　T_{A-B}序差箱式图

(1) 序差呈对称分布。从数值上来看，Q_1(−1)和 Q_3(1.82)的绝对值差为 0.82，Upper bound(9.71)和 Lower bound(−9.06)的绝对值差为 0.65，差距非常小；从箱式图上可以看出，箱子的上下边(Q_3 和 Q_1 位置)和上下触须(whisker)基本呈对称分布；

(2) 序差集中分布于中位数附近。50%的序差都分布在−1(Q_1)到 1.82(Q_3)之间，箱子的长度仅为 2.82(IQR)，箱子显得非常扁平，这说明序差分布集中。结合序差的众数(Mode)为 0，平均数(Avg)为 0.54，可以看出序差集中分布于中位数附近，稍稍向上偏移；

(3) 序差中存在离群值(Outlier)。在箱式图分析中根据某个数据与观测值的关系来认定其是否游离于数据的整体特性之外，并单独汇出。在图 2 中我们将观测值的系数(coef)设置为三个 IQR，来确定观测值(Upper bound 和 Lower bound)，那么大于 Upper bound(9.71)和小于 Lower bound(−9.06)的序差就属于离群值①，见图 2 所示 Upper Outlier 和 Lower Outlier 部分。

综上所述，词语序差的分布特点可以概括为：对于文本的共有词集，词的序差呈对称分布，且集中分布于中位数附近存在离群值序差。这一特点在序差图上表现为“中段平直，双尾翘曲”的“双尾分布”形态。

3　基于序差的文本间词汇差异分析

词语的序差分布特点有哪些语言学上的意义？对于分析文本特征又有哪些作用呢？下面从词语序差的“双尾分布”特点入手，提取出不同层次的词语来分析其类聚特点，并尝试解释其语言学意义，从而揭示序差分布所反映的文本间词汇异同。

3.1　共有词层次的划分

词语的序差代表的是词语在文本间中的“地位差别”。序差的“双尾分布”特点显示，文本词汇的使用具有层次性，可以凭借前面的分析结果客观地划分出词语的层次。

结合序差图可以发现：“双尾分布”可以分为三段——中段、上尾、下尾。中段词语就是在 T_A 和 T_B 中的地位差异不大的词语。越往两边的“尾巴”，词语的序差越大，也就代表词语在 T_A 和 T_B 中的地位差异越大。更具体地说，下尾(也就是序差值为负)是在 T_A 中频序高、地位高的词语，因而反映了 T_A 的文本特点；上尾(也就是序差值为正)是在 T_B 中频序高、地位高的词语，因而反映的是 T_B 的文本特点。

共有词的层次可以依据箱式图来进行划分。中段词语的序差位于 Upper bound 和 Lower bound 之间；下尾词语的序差为小于 Lower bound 的离群值；上尾词语的序差为大于 Upper bound 的离群值。根据这个方法计算得出中段词语的序差范围是[−9.06, 9.71]，包含词语 465 个，下尾词语的序差范围是[Min value, −9.06)，包含词语 18 个，上尾词语的序差范围是(9.71, Max value]，包含词语 24 个。

影响文本词汇分布有两大因素，一是文本的主题内容，二是文本的文体风格[18]。主题内容是文本构建的概念意义。不同于逐字逐句理解文本的具体意义，概念意义可以说是文本具体意义的抽象，可以通过对词汇的分析达到对文本概念意义的概括和表征。主题内容对文本词汇分布的影响是显性的。文体风格从语篇角度来讲，它是文本表义倾向性模式的概括；从语言交际的角度来讲，是说话者对语言形式的有意识选择。不同的交际功能会作用于语言的使用，从而使得文本在词汇方面具有选择性。文体风格对词汇分布的影响相对隐性一些。文本的主题内容和文体风格是我们分析共有词的不同层次反映文本特征时采用的两个主要维度。

① 在文献[20]中，根据试验总结出利用序差的平均数和 1.5 个标准方差的和来确定分界，划分序差层次的方法。在本文中用箱式图分析取代了经验做法。虽然使用的方法不同，但得出的结果却非常接近，从而相互印证了方法的正确性。

3.2 中段词语与文本间的词汇共性

465个中段词语是T_A、T_B两个文本中序差比较小的一群词，也就是说它们在两个文本中的频序接近，地位接近，是两个文本的共性体现。但是中段包含了大量的低频序词和少量的高频序词，比如“的”在两个文本中的频序都是1，所以序差（修正前）为0，而“居民”在两个文本中的序差都是35，序差（修正前）也为0。可见，单看序差会掩盖两者的差异，有必要分为高频中段词和低频中段词来讨论。

高频中段词是指频序在两个文本中都在前50%的中段词，共计12个：的、厦门、在、是、了（助词）、年、城市、有、到、之、与、了（语气词）。从常用度来看，这些词大多是常用词。考察它们在《现代汉语频率词典》中的频序，“的”(1)、“在”(7)、“是”(3)、“了”(2)，“年”(41)、“有”(8)、“到”(24)、“与”(182)、“之”(289)、“城市”(557)的频序均在前3 000之内，属于常用词的范围。这里的“的、了、与、之、是、有、到”是大多数文本中都存在的常用助词、介词和动词，反映的是两个文本与整个词汇系统之间的共性连接，在区分文本特点上的意义不大。而“城市”和“厦门”作为常用度稍低的词语，且作为名词指称了相关的概念，直接体现了两个文本在主题内容上的共同点。

低频中段词是除开高频部分的中段词，共453个，数量相对较多。单独的一个低频词不足以反映文本的特点，但是大量的低频词聚集在一起则会使文本内容特征得到某种程度的浮现（emergence）。如低频中段词里的名词“海、岛、风、城、机场、海域、闽南、旅游、地方、客轮、岸、沙滩、梦、花园、炮、故事、电话、白鹭、钢琴、书、码头、音乐、涛声、游人、海滩、日光、时间、蓝色、小巷”可以勾勒出主题对象“厦门”的环境特征，而“海防、林语堂、建筑、郑成功、集美、街巷、传统、本岛、漳州、北京、时光、腾飞、屈辱”则对厦门的历史、地理图景进行了呈现。中段词随着共有词数量的增加，其“异质性”程度也会增加，需要用更具有概括度的方法对词语聚类进行描写和分析，比如借助语义分类体系[21]。无论是高频还是低频，中段词都可以反映出文本在主题内容上的共性。

3.3 双尾词语与文本间的词汇差异

双尾部分为序差中的离群值，对应到文本的词汇特征上就是能反映文本差异的词。越是“尾端”的词，序差越大，也就说明该词在两个文本中的地位越不对等，就越能体现文本的差异性特点。

下尾词语代表了解说词文本T_A的特点，包括词语“海峡、台湾、中国、全国、等、这、大陆、大、以、经济特区、交流、从、为、个、金门、最、大海、国家”。上尾词语代表了散文文本T_B的特点，包括词语“我、鼓浪屿、去、看、那、如、小、不、上、中、很、这个、着、她、人、也、下、自己、得、过、就、地、走、听”。

可以发现，解说词T_A的特征词中名词多，如“海峡、台湾、中国、全国、大陆、经济特区、交流、金门、大海、国家”等，这些词语反映出解说词T_A的主题内容偏向政治、经济、社会等方面，文体风格上更倾向于叙述说明。相比较而言，散文T_B中更多的是代词“我、那、这个、她、自己”，动词如“走、听、去、看、如”，以及方位词“上、中、下”，而名词则很少，仅“鼓浪屿”和“人”。这些词虽属于不同的词类，但都说明散文T_B在主题内容上更注重个人的体验，文体风格上更倾向于记叙描写。

综合上面的分析，通过对词语序差分析，能够科学地划分文本间词汇使用的不同层次，中段词语可以反映文本主题内容的共性，而上尾和下尾词语反映文本在主题内容和文体风格特征上的差异。

序差在风格分析、文本相似度计算以及语言的统计特性方面具有应用价值和启示：第一，本文建立了一套描述和分析序差的程序，能将文本间词语地位差异加以量化，并在分析其分布规律的基础上划分层级，也就是依据序差给词语对文本特征的反映能力赋予了权重，具有用于文本特征提取的价值；第二，序差反映文本间“共性中的差异性”，对文本的分析更加微观细致，可以满足颗粒度更小的文本风格分析、相似度分析；第三，本文的分析显示，序差的分布具有形态上的规律性，其中可能存在的、具有普遍意义的语言统计规律值得进一步探讨。

4 总结

本文在已有关于“频级”“频序”研究的基础上，着重考察词语“序差”的分布特点。通过对解说词和散文文本中共有词集的序差的分析，本文发现：共有词的序差呈对称分布，且集中分布于中位数附近，存在离群值序差。这一特点在序差图上表现为“中段平直，双尾翘曲”的“双尾分布”形态。根据词语序差的分布规律，可以将文本共有词划分为“中段”“下尾”“上尾”三个层次。中段词语反映文本的共性特征，下尾和上尾词语反映两个文本的差异性特征，这

些特征具有反映文本的主题内容和文体风格的语言学意义。

与前人的研究相比，本文的贡献在于引入结合散点图和箱式图的分析方法，改进了基于序差的文本词汇特征分析程序，更直观地刻画了词语的序差分布形态，对序差数据的分布特点进行了讨论和概括，并结合具体文本对序差的语言学意义进行了初步探讨。但本文对这一问题的讨论仍然具有深入的空间，后续研究我们将以本文提出的分析程序，对更多类型的文本进行考察，进而发掘和测定更广泛层面上序差分布的统计学规律；另一方面，探索序差分布规律应用于词汇计量、文本风格分析、文本分类的方法和途径，例如用序差指标来选取文本词汇特征，用于文本相似度计算、文本聚类等。

参考文献

[1] 鲁松，李晓黎，白硕. 文本中词语权重计算方法的改进[J]. 中文信息学报，2000，14(6)：8-13.

[2] 廖浩，李志蜀，王秋野. 基于词语关联的文本特征词提取方法[J]. 计算机应用，2007，27(12)：3009-3012.

[3] 熊忠阳，黎刚，陈小莉. 文本分类中词语权重计算方法的改进与应用[J]. 计算机工程与应用，2008，44(5)：187-189.

[4] 徐建民，王金花，马伟瑜. 利用本体关联度改进的 TF-IDF 特征词提取方法[J]. 情报科学，2011，29(23402)：279-283.

[5] G K Zipf, The Psycho-biology of language: An Introduction to dynamic philology [M]. London: George Routledge & Sons Ltd., 1936.

[6] G K Zipf, Human behavior and the principle of least effort: An introduction to human ecology [M]. New York & London: Hafner Publishing Company, 1965.

[7] G A Mitier, E B Newman, Tests of a statistical explanation of the rank-frequency relation for words in written English [J]. American Journal of Psychology, 1958 (71): 209-218.

[8] 关毅，王晓龙，张凯. 现代汉语计算语言模型中语言单位的频度-频级关系[J]. 中文信息学报，1999，13(02)：9-16.

[9] 邢红兵. 现代汉语词类使用情况统计[J]. 浙江师范大学学报(社会科学版)，1999(03)：27-30.

[10] 上海交通大学汉字编码组，上海汉语拼音文字研究组. 汉字信息字典[M]. 北京：科学出版社，1988.

[11] 安华林，曲维光.《现代汉语词典》释义性词语的统计与分级[J]. 语言文字应用，2004(01)：105-111.

[12] 苏新春. 计量方法在词汇研究中的作用及频序统计法[J]. 长江学术，2007(02)：118-124.

[13] 韩布新，任雪松. 汉语输入编码中简码字、词的合理选配[J]. 中文信息学报，1995，9(04)：41-50.

[14] 侯敏. 语言资源建设与语言生活监测相关术语简介[J]. 术语标准化与信息技术，2010(02)：30-33.

[15] 陈海波. 序差在文本区别特征研究中的应用[J]. 长江学术，2010(04)：112-117.

[16] 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，2016，20(06)：49-58.

[17] 苏新春. 词汇计量及实现[M]. 北京：商务印书馆，2010.

[18] G., Dee. Vocabulary input through extensive reading: A comparison of words found in Children's narrative and expository reading materials [J]. Applied Linguistics, 2004, 23(1): 1-37.

[19] 杨继本. 认知心理学在《汉字教学字典》研编中的应用[J]. 心理科学，1995(01)：43-47.

[20] 刘锐. 词语的“序差”与文本词汇特征研究[D]. 厦门大学硕士学位论文，2016.

[21] 苏新春.《现代汉语语义分类词典》(TMC)研制中若干问题的思考[J]. 中文信息学报，2008，22(05)：12-21.

刘锐(1990—)，硕士，研究助理，主要研究领域为词汇计量、语料库语言学。

E-mail：liuruioscar@hotmail.com

龙云飞(1991—)，博士研究生，主要研究领域为计算语言学。

E-mail：csylong@comp.polyu.edu.hk

孙碧泽(1990—)，硕士，主要研究领域为现代汉语语法。

E-mail：sunbize_erlangshen@foxmail.com