流通度—字词使用情况测定的新方法

郑泽之[1,2]　王强军[1,3]　张普[1]

（1. 北京语言大学网络教育学院 2. 太原师范学院计算机系 3. 河北大学人文学院）

E-mail: zezhi@blcu.edu.cn, wangqj@blcu.edu.cn, zhangpu@blcu.edu.cn

摘要：流通度理论是由张普教授提出的，是对字词的使用情况进行多层次信息分析加工的一种动态的方法。我们使用流通度的方法对北京语言文化大学网络教育学院 DCC 博士研究室的大规模动态流通语料库的汉字使用情况进行了统计，在此基础上给出了统计结果的分析。并由此阐明流通度理论及其加工方法是对语言文字使用情况进行研究的一种行之有效的、直观快捷的新方法。

关键词：频度 使用度 通用度 流通度 动态流通语料库

Circulation, a New Mensuration in Determining the usage factor of words

ZHENG Ze-zhi[1,2], WANG Qiang-jun[1,3], ZHANG Pu[1]

1. Network Education College, Beijing Language and Culture University
2. Department of Computer Science and Technology, Taiyuan Teachers' College
3. College of Humanities and Social Sciences, Hebei University

E-mail: zezhi@blcu.edu.cn, wangqj@blcu.edu.cn, zhangpu@blcu.edu.cn

Abstract: Circulation and the arithmetic is a new and dynamic approach used in processing natural language. It was proposed for the first time by Pro. Zhang_pu, in China and abroad, and is firstly used in statistic Chinese characters. Based on the statistic result, which is procured from the large scale *Dynamic Circulation Corpus* (DCC) of Network Education College, Beijing Language and Culture University, an analysis on Chinese characters in circulating was given. All of those show that Circulation arithmetic is an effective and visual method in processing language factors.

Keywords: Frequency, Usage, Currency-usage, Circulation, Large Scale Dynamic Circulation Corpus

1. 引言

处于信息时代的今天，信息交流的程度直接影响到科技进步，语言是信息的载体，一个民族的用字和用词情况对语言信息处理、情报检索、机器翻译、人工智能、文字改革、语音识别、汉字录入，以及对字词的语言学、语言教学、词书编纂、认知科学研究都有重要的影

响。要搞清楚一个民族用字或用词的规律就需要对其现实生活中的用字或用词的情况进行调查统计，但采用不同的统计方法对统计结果有重大影响，统计结果能否真实反映实际使用情况也是语言学家和应用系统开发者都关心的问题。为此语言学家、统计专家进行了不懈的努力，研究使用了不同的方法并进行了改进。本文作者将就汉字使用情况的通用度和流通度两种统计方法进行对比分析。

2．统计方法的简单回顾

因频度的简单易行，一开始人们自然想到的是使用频度（所谓的频度是指一个汉字或词在所处语料库中出现的次数）去测定字词的使用情况。随测定语料范围的增大人们逐渐发现频度在测定字词使用情况中的缺陷，当字词在语料中分布不均匀时频度不能正常反映字词的使用情况，而发生偏移。以后各国的统计语言学家提出了各种算法如“使用度”、“扩散率”。在汉字词的测量中人们也使用了使用度、通用度，在这里我们使用的是流通度。

使用度、通用度就是在频度这个“基”数上增加调整因子修正后的频度，只不过前一个是线性压缩了的频度，后一个是对频度的非线性处理，虽然都考虑到了字词的散布情况，但两种方法均是静态统计。流通度以其本身的层次性、动态性，不仅考虑到了字词的散布情况，而且也考虑了字词的载体——媒体的的流通情况，既可以完成静态的统计，也可以进行动态统计。

2．1 使用度

在大量的实践过程中人们发现，仅以频度去测量字词的使用情况不能完全反映字词的实际使用情况，１９６４年A. Juilland（尤兰德）等在统计西班牙语词频时曾利用了一个计算使用度的计算公式如下：

$$S_k=\sqrt{\sum_{i=1}^{n}(N_{ki}-N_k)^2/n} \quad ， \quad D_k=1-S_k/N_k\times(n-1)^{\frac{1}{2}} \quad ，$$

$$U_k=D_k\times F_k, \quad (0\leqslant D_k\leqslant 1)$$

其中 S_k 为 k 号词的标准分布偏差，N_k 表示 k 号词在各类中的平均词次，N_{ki} 表示 k 号词在 I 类例的词次。N 为分类数。D_k 为 k 号词的散布系数。U_k 为 k 号词的使用度，F_k 为 k 号词的词次。

随着时间的推移，字词的使用情况会发生变化，字词在语料中出现的类数、出现的文本数都有可能影响到统计结果能否反映字词在生活中真实情况。常宝儒教授曾经改进过这个计算公式，但受时代所限他那时的公式依然没有考虑到也不可能考虑到语言的动态发展过程中时间对字词使用情况的影响。

2．2 通用度

所谓词语的“通用度”，是指词语在语言应用的各个领域里常用性的综合指标。通用度已经兼顾到词语的分布率和频率两个方面，并且把两者有机地结合起来[1]。通用度的基本计算公式为：

$$T = (\sqrt{n_1} + \sqrt{n_2} \cdots\cdots + \sqrt{n_k})^2 / k$$

其中，T为某词的通用度.k为抽样统计的全部语料的分组数，而且每组的语料数量大致相等，$n_1 n_2 \cdots n_k$为该词在各组中分别出现的次数。

特别应该指明的是通用度与使用度虽然都是考虑了语料的散布问题，但是通用度比使用度更进一步，尹斌庸、方世增在他们的文章中指出："通用度概念中所说的'领域'，既可以指'空间'，也可以指'时间'，它既可以指一个词在共时的语言应用中各领域里的通用程度，也可以指一个词在历时的各个时期里的语言应用中的通用程度。"实际上就是考虑了词语在时间轴的一种散布。

事实上，通用度是一种频度的非线性叠加。这种方法可以从类散布空间上和时间轴方向两个方向上对字、词的使用情况进行统计。比使用度多了一个时间维度，而少了篇数这个因子。从这个时间维而言，通用度使字、词使用情况的统计又往真实语感上靠了一步。而且这个公式有效的调节了由于字、词在类或时间轴上的频度散布不均衡给统计结果是否反映字词的真实使用情况带来的负面影响。

但众所周知，字词依赖于语言得以流通，今天的语言又是越来越主要依赖于媒体的传播而得以传播的，所以研究字或词使用情况不能不涉及到媒体流通情况。若要模拟字或词的真实使用情况，媒体的流通情况是无法忽视的。所以我们提出了流通度的统计，目的是想一步步"把对语言成分的一般性的统计分析推向对语感的推测性统计分析和验证，从而探索使电脑可以逐步获得语感并随时增强和调整语感的路径。"[2]

2．3 流通度

"流通度"（circulation）是一种语言现象在社会传播中的流行通用程度。语言的流通度与社会传媒的流通度密切相关。流通度不仅是判定新词、新义、新用法的重要条件，也是判定方言词语、术语、文言词语、外来词语是否进入普通话、是否进入通用领域、是否合乎规范的极为有效的量化操作标准。语料的流通度的选择，首先是社会传媒的流通度的选择，就显得十分重要[4]。

流通度把字、词的使用情况从字词本身的特性（频度）、拓展到了篇章级（散布篇数、散布类数），进而推广到了媒体层（媒体流通系数）。加工方法上，由静态上升到动态。

流通度的动态特性体现在它的加工方法的动态上，它既可以以某时间段为一共时点进行加工，也可以把某时间段进行合理的切分，以切分的时间段为轴动态加工，然后给出字或词的动态曲线走势，依据这个动态曲线，研究者可以很直观地观察字或词使用的变动情况。而且因其参数的动态性，流通度更能反映字词的实时使用情况。

仅考虑字词的流通情况，流通度有三个部分构成：媒体流通属性、文本流通属性（散布系数）、字词流通属性。但流通度决不是仅用于字词的流通研究。[4]

文本流通度和媒体流通度之间是一种继承与被继承的关系。媒体流通度的一切特征，在文本流通度中都应该有所体现。而文本流通度在继承了媒体流通度的全部特征之后，又会生成自己特有的属性。

媒体流通属性包含发行地域、发行内容、发行数量、发行周期、受众对象、阅读率等几个因素。只要能够科学合理地分析这些因素，就可以制定出确定流通度的标准。

我们提出书面媒体流通度的最简单最基础的计算公式：

$$Ct=Vc \cdot Dc \cdot Ac \cdot Fc \cdot \cdots$$

即：流通度＝流通量·流通密度·流通空间·流通率·…

（具体参数估计见参考文献 5）

字、词流通度=字词频率·文本流通度系数·媒体流通系数

这种流通度的计算公式使得只在某一类（这里为某种报纸）中高频出现，只在某一地区高频出现等，可以通过流通度参数对其出现频度与以缩小或增大。可以使得影响大的报纸对语感的冲击更明显，大家可以观察到当前正流行的字、词或新词一般首先出现在文化发达地区比如北京、上海、广州，然后在全国蔓延开去，我们目前的流通度计算公式中，对发达地区的或全国性报纸的加权系数正体现了这一原则。所以对于新词新义热点字词选出起了推动作用，我们的统计结果也表明了这种必然。

我们认为："……只有在流通度高的真实文本的基础上计算使用度才是真实的使用度"，"……语言的生命力就在于这种稳定中的变化，这些变化的端倪就隐藏在大规模的真实文本（无论它们是经典的还是非经典的文本）之中……"[2]。所以我们选择北京语言大学网络教育学院 DCC 博士研究室的动态流通语料库作统计对象进行了一次大规模流通度统计实验。

3．统计结果

我们抽取北京语言大学网络教育学院 DCC 博士研究室的动态流通语料库 2002 年的 2 亿 3 千多万的语料进行了流通度的统计，同时为了有个比较，我们还选用了时间轴分类的通用度算法进行了通用度的统计。

频度、流通度和通用度的统计情况见表 1：

表 1

字数	频度（frq）	累加比率 (%)	通用度（tyl）	累加比率 (%)	流通度（ltd）	累加比率 (%)
10	25897128	11.12556	13370860.94	11.182542	318253.2136	23.46366
100	89629507	38.50537	46176669.36	38.61924	871451.1139	64.24895
500	179349243	77.04951	92273482.86	77.1717	1286264.74	94.83165
1000	212595960	91.3325	109317583.1	91.42631	1346221.797	99.25207
1500	224067628	96.2608	115172237.1	96.32277	1354412.468	99.85594
2000	228748006	98.27152	117559656.8	98.31946	1355899.483	99.96557
2500	230885048	99.18605	118641374.6	99.22413	1356251.756	99.99154
3500	232331443	99.81098	119365370.7	99.82964	1356356.37	99.99926
7376	232771417	100.00000	119569068.3	100.00000	1356366.44	100.00000

从上表中可以看出流通度收敛速度较快。

我们对统计结果分别以频度、通用度、流通度降序排序，并把三种排序结果进行了比较，查看各种表中字序的变化情况，比较的对象为：流通度——频度、通用度——频度。

结果见表 2：

表 2

流通度		通用度	
总上移数	3307	总上移数	4617
上移（>5）	3100	上移（>5）	3512
上移（>10）	2917	上移（>10）	2897
上移（>50）	1785	上移（>50）	1173
上移（>100）	1075	上移（>100）	489
上移（>200）	544	上移（>200）	65
上移（>500）	125	上移（>500）	0
上移（>1000）	10	上移（>1000）	0
总下降数	4013	总下降数	2444
下降（>10）	3664	下降（>10）	1632
下降（>50）	2542	下降（>50）	911
下降（>100）	1457	下降（>100）	515
下降（>200）	429	下降（>200）	270
下降（>500）	8	下降（>500）	24
下降（>1000）	0	下降（>1000）	1
总移位数	7320	总移位数	7061

表 2 中的“上移”和“下降”意思是字的频度排序位序，按流通度（通用度）排序其序号是往前移了还是往后移了。

从上表我们可以看出流通度的调整范围比通用度大，流通度上调的范围小，下调范围大；通用度上调范围大，下调范围小。实际上这种结果与它们采用的计算公式有关系，我们从上面的计算公式可以观察到流通度根据语言、社会发展的现实状态制定参数，比如经济、文化发达地区对语言发展影响较大，这种现象在媒体系数制定过程中有所体现，所以对于一些语言的新特点体现较好，下面的例子将予以说明；而通用度是一种类或时间刻度上的一种分布调整，它把所有的字词一视同仁地进行了一次基于散布的调整，对语言发展中的新现象体现力度显得不如流通度了。

虽然流通度调整的范围和幅度较大，但分别频度和流通度为序选取它们的前 2500 个，进行差集的运算，我们发现有 38 个不同，取 3500 个时也只有 61 个不同。

就以上的统计结果我们做了一个集合求差运算得出通用度与频度的前 2500 常用字集中只有 6 个字不同。同时我们对流通度也做了相同的比较，结果有 39 个不同。

从流通度的上移表（3307）中我们观察到这样一种现象，关系到热门事件、焦点人物、现实社会正盛行的现象等用的字词在流通度的上移表中频频出现。从中拿一部分大家可以观察一下（见表 3），我们对这些字做了一个粗分类：

- 流行词语：野“蛮”女友，“网吧”，“新锐”，“酷”，“爽”，“帅”‘……
- 现代一些人心态：“寂寞”，“闷”，无“聊”，“厌”，“侃”，“尴尬”，“嬉”，“醉”，“悠”，

“悦”，“疯”，“怨”，“遗”憾”，“孤独”，“忧”……

- 政治热点：弹“劾”，反“贪”，“贬”官，“攀”比，“滥”用职权，“祸”国秧民，“卫”闽，朱“镕”基……
- 揭露社会丑恶现象：“裸”，“傍”大款，“大亨”，“瘾”君子，“妓”女，“陪”……
- 反映现代人生活：“宵”夜，“宠”物，“吻”，“诱惑”，“淘汰”，“夕”阳红，“恋人”……
- 与科技有关：“网”络，“光碟”，……
- 商业用语：“渠”道，“滞”销，“盈亏”，“劣”制产品……

等等。可以看出这个流通度的计算公式对热点词语做的一种调整。

表 3

贬	汕	次	名	三	间	两	期	济	网	闻	解	数	式	共	源	料	原	举	推
单	难	直	准	布	声	反	见	具	真	星	需	优	效	贵	晚	益	光	显	热
识	昨	严	低	率	批	负	港	介	友	边	双	市	富	访	短	券	希	载	善
觉	算	域	欢	独	绝	减	征	抓	迎	夜	映	苦	您	频	键	诚	吧	脱	贫
累	纺	恶	坏	贡	艰	衡	圈	刷	纠	漫	欲	暗	渠	辛	阻	浦	忧	盈	肃
颇	劣	妹	奉	烦	尝	悲	涵	夕	恋	牢	媳	陪	霸	绪	乃	仲	汨	挤	踏
镕	穷	漂	耗	敲	巩	叹	吹	轰	惨	淘	裕	憾	闪	锐	舆	孤	贪	惑	肩
诱	衰	德	汰	叁	扣	酷	埋	怨	猜	帜	疲	魂	愤	赋	昂	弘	姿	侨	胸
畜	汪	碎	狮	拘	盼	朴	呀	框	闯	稀	疏	歧	凉	卫	挫	疯	卓	悦	滥
裸	傍	亨	叛	瘾	擒	墓	靓	舌	碱	氨	擂	蛮	韶	寞	裹	殴	氮	搅	赂
溢	遏	蕴	爽	苯	蕊	彦	嗓	赃	贼	背	寂	磷	醇	丙	吻	宵	厌	颂	…

再看一些流通度排序调整到后面的字（表 4）：

表 4

妈	矿	奶	爸	犬	猴	枣	铅	醛	鸽	豹	卵	鲁
蚊	堰	羚	窟	佼	蒜	蝗	酪	舅	鲸	娥	鲨	喳
栓	礁	秤	翡	豚	椰	罹	粟	芹	胰	疳	鹊	蜥
砚	藻	鹭	鸵	箔	悻	蛤	龋	蟋	釉	蜂	氢	螨
筏	蛾	刘	鲁	…								

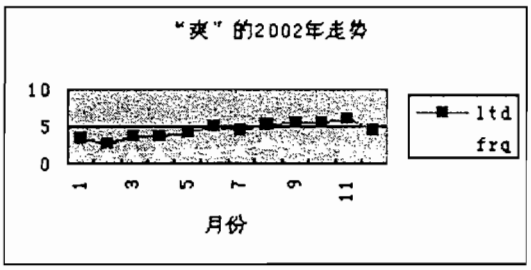
可以一眼看出的分类有：

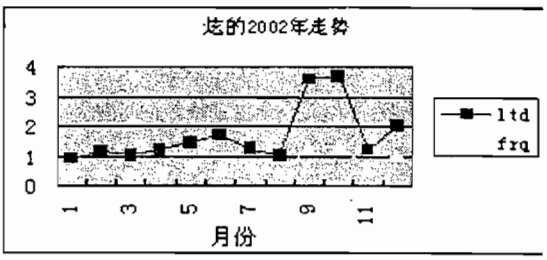
- 具体事物名：犬，猴，枣，铅，醛，鸽，豹，卵，蚊，羚，窟，佼，蒜，蝗，酪，舅，鲸，娥，鲨……
- 各种称呼：爸、妈、婶……
- 一些脏器：胰、肝……
- 姓氏：刘、翟、鲁……

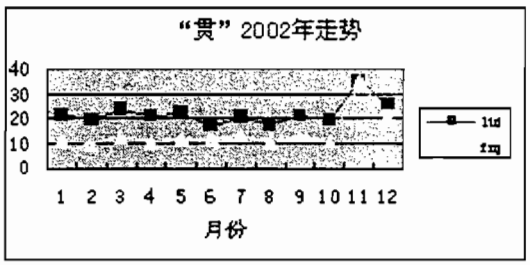
等等。

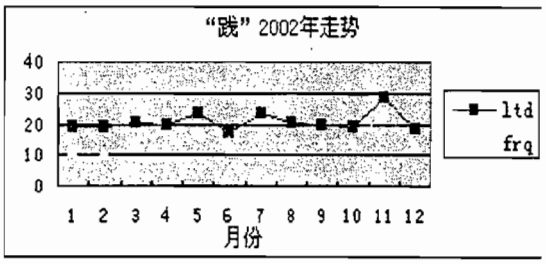
此外，流通度的加工结果可以是动态的，因为，流通度可以动态的展示字或词的时间轴走势，直观上即可以一览字词的使用情况。静态的流通度加入媒体流通系数后进一步模拟人

的语感，对频度的调整可从下图得到展示。









4. 结语

流通度的计算公式不是只能是这种方式，关于流通度的计算公式还在进一步的探索研究中，这次的统计结果仅是一次流通度在汉字统计中的检验，已经呈现出来的结果基本达到我们的预期目标。

汉字的使用在一年的每个时间段的分布变化情况较慢，词语的使用情况在一年中的变化与汉字的情况肯定有不一样之处，流通度对词语的使用情况的统计正在试验中。

生活实际中的语言材料是动态的，不断变化的，而流通度的计算，目的就是以客观量化的手段去揭示这些发展变化的规律，用可视化的方式展示字词的流通变化趋势，对大规模语料进行实时处理，甚至对语言未来的发展趋势进行预测。这对语言应用研究和语言规范化工作都会有极大的帮助。

参考文献：

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

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

[3] 张普 关于第三代大规模真实文本语料库的几点理论思考，《自然科学基金重点项目结题报告》（内部），清华大学，1998 年。

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

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

[6] 陈原主编 《现代汉语定量分析》，上海教育出版社，1989 年。

[7] 尹斌庸、方世增 词频统计的新概念和新方法，《语言文字应用》1994 年第二期