情报学报 ISSN 1000-0135
第31卷第3期297-302，2012年3月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 31 No. 3, 297-302 March 2012

doi:10.3772/j.issn.1000-0135.2012.03.009

洛特卡现象在汉语词汇句法功能分布复杂度中的呈现[1)]

王东波 朱丹浩 苏新宁
（南京大学信息管理系，南京 210093）

摘要 本文基于大规模清华树库，从中统计了汉语词汇在句法结构中充当的句法成分，获取了汉语词汇的句法功能分布，并给出了汉语词汇句法功能分布复杂度的定义。在对汉语词汇按照汉语词汇句法功能分布复杂度的高低排序后，本文发现两者之间呈现洛特卡现象。本文的这一发现一方面揭示了汉语词汇在句法结构中的分布规律，对于汉语的研究具有重要的促进作用；另一方面对于中文信息处理中的词性标注、自动消歧和句法分析等研究具有重要的影响。

关键词 洛特卡现象 汉语词汇 句法功能分布复杂度 清华树库

Lotka's Phenomenon in Distributional Complexity of the Chinese Words with Syntactic Function

Wang Dongbo, Zhu Danhao and Su Xinning
(*Department of Information Management, Nanjing University, Nanjing* 210093)

Abstract The Chinese word syntactic constituents in the syntactic structure are calculated based on large-scale Tsinghua Treebank, and the Chinese word syntactic function distribution is gained in the paper. The definition of Chinese word syntactic function distribution complexity is given in this paper. The Lotka's phenomenon presents between Chinese word syntactic function distribution complexity and Chinese word amount after the Chinese word is sorted according to the Chinese word syntactic function distribution complexity. On the one hand, the discovery in the paper reveals Chinese word distribution law which will promote the Chinese researches in the syntactic structure, on the other the discovery will influence the researches of part-of-speech tagging, automatic disambiguation and syntactic analysis in the Chinese information processing.

Keywords Lotka's phenomenon, Chinese word, syntactic function distribution complexity, tsinghua treebank

1 引 言

在20世纪20年代，美国学者A. J. 洛特卡[1]揭示了科学工作者人数与其所著论文之间的规律，即写有 x 篇论文的作者频率与 x 的平方呈反比，这一规律被称为洛特卡定律。在基于大规模数据考察汉语词汇句法功能分布复杂度与其对应的词汇数量之间的关系过程中，本文发现两者之间的关系与论文和论文作者呈现的关系一致，对于词汇在句法结构中的这种分布现象，本文称之为洛特卡现象。本文的这一发现说明了汉语词汇在分布上不仅符合齐普

收稿日期：2011年5月3日

作者简介：**王东波**，男，1981年生，博士研究生，主要研究方向：自然语言处理与文本挖掘、语料库、知识库。E-mail: wangdongbo0102@gmail.com。**朱丹浩**，男，1986年生，硕士研究生，主要研究方向：数据挖掘，中文信息处理。**苏新宁**（通讯作者），男，1955年生，教授，博士生导师，主要研究方向：信息处理与检索、知识管理、引文分析等。

1) 项目支持：本文系教育部人文社会科学重点研究基地重大项目"基于智能信息处理的知识挖掘技术及应用研究"（项目批准号：08JJD870225）和南京大学研究生科研创新基金资助项目"基于网络的英汉/汉英平行语料对自动获取"（项目编号：2010CW02）的研究成果之一。

夫定律[2]这一静态分布规律，而且在词汇与词汇或短语结构的组合，这一动态分布上也是有规律可循的。同时，这一现象的揭示一方面进一步深化了人们对于汉语的认识，更加有助于研究者从语言学的角度把握汉语句法功能分布的规律；另一方面对于中文信息处理的词性标注、自动消歧和句法分析都会有重要的促进作用，如通过词汇的汉语句法功能分布复杂度就可以了解汉语到底有多少词汇在句法上是有歧义的，并且也可以知道词汇的歧义程度。

关于汉语词汇句法功能的研究，先前的学者在相关的研究中有一些涉及。从按照词的语法功能对汉语的词进行分类这一设想出发，朱德熙[3]提出了获取词汇的各种语法功能的构思；陈小荷[4]类比北京大学的《现代汉语语法信息词典》[5]，构建详细描述词汇语法功能的语法知识词典，由于当时没有大规模的语料库，该理念一直没有得到有效地验证；徐艳华[6]按照汉语语法功能分布的理念，借助人民日报大规模语料库，在计算机辅助的前提下，通过人工分析了名词、动词、形容词和高频副词的语法功能，构建了3514个汉语词汇的简单语法功能知识库。

本文在前人研究的基础上，基于大规模清华树库，统计了汉语词汇的句法功能分布，并且界定了汉语词汇句法功能分布复杂度，同时基于汉语词汇句法功能分布复杂度的数据，进行了相关的分析，发现并验证了洛特卡现象在汉语词汇句法功能分布复杂度的呈现。

2 数据源简介和相关概念界定

2.1 数据源简介

在16种句法结构的基础上，清华大学计算机系智能技术与系统国家重点实验室构建了一个100万字的汉语树库。该库对每个句子都进行了全面的句法分析，并标出了每一个词汇和短语的句法功能和句法结构。清华汉语树库的详细信息见表1和表2。

表1 清华汉语树库的基本统计数据①

文体	文件数	句子数	词项数	汉字数	平均词长（词/句）
文学	139	16 335	340 208	415 040	20.83
新闻	154	6 877	173 942	246 757	25.29
学术	15	5 589	158 780	240 289	28.41
应用	195	3 169	66 586	97 924	21.01
合计	503	31 970	739 516	1 000 010	23.13

表2 清华汉语汉语树库的句子长度分布数据①

	简单句子			复杂句子		
文体	句子数	词项数	平均长度	句子数	词项数	平均长度
文学	9 692	102 895	10.62	6 643	237 313	35.72
新闻	3 025	34 023	11.25	3 852	139 919	36.32
学术	2 021	24 204	11.98	3 568	134 576	37.72
应用	1 870	16 946	9.06	1 299	49 640	38.22
合计	16 608	178 068	10.72	15 362	561 458	35.90

本文关于词汇句法功能分布复杂度洛特卡现象的研究都是基于清华汉语树库统计的数据进行的。清华汉语树库的具体标注例子如下：

[zj-XX [dj-ZW 报名者/n [vp-XX [vp-ZZ 须/d [vp-ZZ 正楷/n [vp-PO 填写/v [np-LH 姓名/n 、/、[np-DZ 通信/vN 地址/n] 和/c 邮编/n]]]] [dlc-BC (/([vp-ZZ [vp-ZZ 不/dN 要/vM] [vp-ZZ 另/rD [vp-PO 写/v 信/n]]])/)]]] 。/。]

2.2 汉语词汇句法功能分布复杂度概念界定

根据具体研究的需要，本文给出了汉语词汇句法功能分布复杂度概念的定义。

定义：在语料集合 U 中，词汇 A 在且仅在 n 个不同类别的句法结构中出现，则 A 在 U 中的句法复杂度为 n。

如：在清华树库中，“打开”这个词在且仅在附加结构（ZW）、连谓结构（LW）、述宾结构（PO）、状中结构（ZZ）和主谓结构（ZW）中出现过19次、2次、26次、6次和1次，那么“打开”的句法功能分布复杂度为5，“必然”这个词在且仅在定中结构（DZ）、状中结构和联合结构（LH）中出现过7次、3次和1次，那么“必然”的句法复杂度为3，根据定义本文认为“打开”在句法功能分布上比“必然”复杂。

3 相关数据的获取

3.1 基于清华树库的句法结构调整

清华树库的所有句子中共有16种句法结构，这些结构共充当了16种词性类的语法成份。本文在清华汉语树库的基础上，按照汉语语法功能分布的

① 该数据为清华大学计算机系智能技术与系统国家重点实验室的技术资料，名称为“汉语树库构建——标注规范”。在此表示感谢！

表 3　调整后的句法结构表

编号	结构名称	符号表示	编号	结构名称	符号表示
1	主谓结构	ZW	14	兼语结构	JY
2	带的定中结构	DZ < de >	15	比况结构	BK
3	无的定中结构	DZ	16	“的”字结构	DE
4	数量结构	SL	17	一般附加结构	AD
5	带“地”状中结构	ZZ < de >	18	枚举结构	MJ
6	无“地”状中结构	ZZ	19	时间短语	SJ
7	简单述补结构	SB	20	“所”字结构	SZ
8	带“得”述补结构	SB < de >	21	语气结构	YQ
9	带“不”述补结构	SB < bu >	22	框式结构	KS
10	述宾结构	PO	23	顺序结构	SX
11	方位结构	FW	24	联合结构	LH
12	介宾结构	JB	25	复句结构	FH
13	连谓结构	LW	26	分句结构	FJ

理论，对原先清华树库的句法结构进行了适当的调整。重新调整后的句法结构见表 3。汉语词汇的句法功能分布复杂度数据主要是基于调整后的清华树库统计的。

3.2　基于清华树库的句法功能分布统计

本文定义了一个多叉树来存放清华树库中的每一个句子，用来存储句法树的短语结构或词汇。树的节点分两种：分支节点和叶子节点。分支节点中存储了句法结构信息，每一个分支节点至少有一个或一个以上的子节点；叶子节点中有两个数据段，一个数据段存储标点或者词汇，另一个数据段存储词性，一般标点的词性被标记为标点本身。

生成树结构的算法流程图如图 1 所示：

格式化句子：去除句子编号，并压缩掉句子前后的空格、制表符等符号；

发现节点：遇见“[”表示出现了句法节点，例如“[pp-JB”表示遇到的句法节点是介宾结构。遇到“/”表示遇到了叶子节点，例如“财政/n”表示词汇是财政，词性是名词；

建立叶子节点：新建一个叶子节点，分别存如词汇和词性。并将该叶子节点粘贴到父节点上；

建立分支节点：新建一个分支节点，初始化其子节点队列，并存入句法信息。将该节点粘贴到父节点上通过上述流程，获取汉语词汇在树库句法结构中的句法功能分布，并存入数据库中。

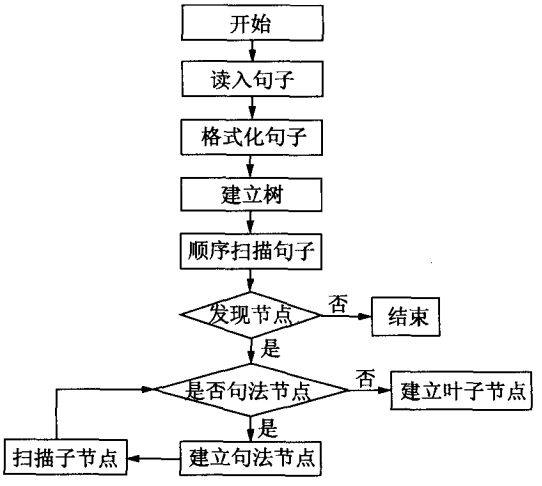

图 1　生成树结构的算法流程图

图 2 是获取的一部分样例。

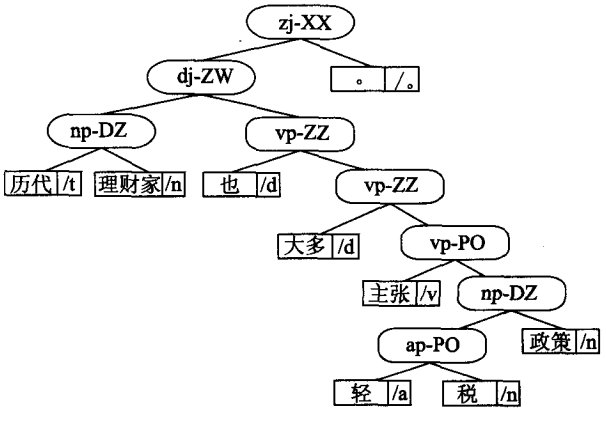

图 2　树结构样例图

按照以上流程，对句子“［zj-XX［dj-ZW［np-DZ 历代/t 理财家/n］［vp-ZZ 也/d［vp-ZZ 大多/d［vp-PO 主张/v［np-DZ［ap-PO 轻/a 税/n］政策/n］］］］］］。/。］”进行处理，得到结果如图2所示：

树检索使用了根节点深度优先的方式遍历了整个树结构，每当遍历到叶子节点时，提取词汇，例如“历代”。其父节点，也就是存储句法结构信息的节点也将被记录下来。统计算法如图3所示。

最后将统计的数据写入数据库，获取到汉语词汇的句法功能分布表。图4是从句法功能分布表中截取的部分字段和部分数据的样例。

3.3 汉语词汇句法功能分布复杂度的获取

基于汉语词汇的句法功能分布表，建立“词汇-句法结构”表T，以词汇为行，句法结构为列。

$$T = \begin{cases} 0 & \text{第 } i \text{ 个词汇在第 } j \text{ 个句法结构中未出} \\ 1 & \text{第 } i \text{ 个词汇在第 } j \text{ 个句法结构中出现} \end{cases}$$

根据给出的汉语词汇句法功能分布复杂度定义，假定表中有 n 个词汇，m 个句法结构，对于第 i 行求和得到 $a_i = \sum_j^n$，就是第 i 个词汇的句法功能分布复杂度。

将句法功能分布复杂度相同的词汇并入同一类，并对每一类中词汇的数目进行统计，得到表示句法功能分布复杂度和其词汇数目对应关系的汉语词汇句法功能分布复杂度表，并将表按照词汇数目的大小降序排序。本文在进行词汇句法功能分布复杂度数据分析和一元线性回归分析的过程中，去除了其中词汇数目在15以下的数据。具体的表见第5部分。

5 基于数据的相关分析

5.1 汉语词汇句法功能分布复杂度分析

为了更真实地验证词汇句法功能分布复杂度在

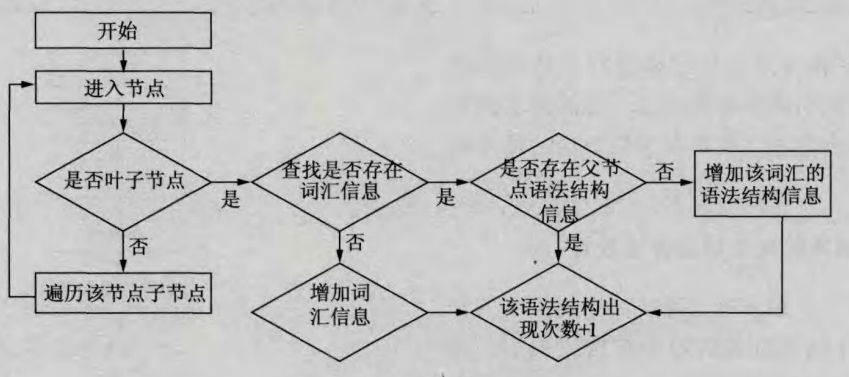

图3 统计算法图

元素	原词性标记	主语	谓语	定语	定中中心语	带“的”定	2DZ<de>	带“的”定	数词	量词	状语
新	aD a	0	0	29	0	30	0	0	0	0	6
世纪	n	0	0	0	60	0	0	0	0	0	0
能	vM	0	0	0	0	0	0	0	0	0	24
最	dD	0	0	0	0	0	0	0	0	0	59
他	rN	34	0	3	0	17	0	0	0	0	6
所	u	0	0	0	0	0	0	0	0	0	0
不同	a	0	10	16	0	7	0	1	0	0	6
理论	n	0	0	25	16	1	0	10	0	0	0
知识	n	4	0	8	17	2	0	10	0	0	6
人类	n	8	0	30	1	11	0	1	0	0	0
地	u n	1	0	0	0	1	0	0	0	0	6
神	a n	12	0	0	9	4	0	7	0	0	0
形成	vN v	0	4	0	0	0	0	6	0	0	6
观点	n	0	0	0	31	0	0	15	0	0	0
还	d	0	0	0	0	0	0	0	0	0	48
人们	n	16	0	11	0	9	0	2	0	0	0
哲学家	n	4	0	0	32	0	0	5	0	0	0
关系	vN n	0	0	0	23	0	0	21	0	0	0
自然科学	n	9	0	5	7	14	0	2	0	0	6
思维	vN n	1	0	5	28	1	0	1	0	0	0
产生	vN v	0	8	0	0	0	0	5	0	0	6
则	c	0	0	0	0	0	0	0	0	0	0
自己	rN	3	0	6	1	26	0	0	0	0	6
同	b a v c d p	0	2	2	0	0	0	0	0	0	2
他们	rN	27	0	1	0	11	0	0	0	0	6
实践	vN v	2	0	6	14	5	0	2	0	0	0
形式	n	2	0	2	27	0	0	8	0	0	6
大	a	0	0	25	0	0	0	0	0	2	0
等	u	0	0	0	0	0	0	0	0	0	6
一切	rN	9	0	27	1	0	0	3	0	0	0
自然	a n	1	0	19	1	4	0	1	0	0	6

图4 汉语词汇的句法功能分布样例

汉语中的分布情况以及揭示其中呈现的洛特卡现象，本文在清华树库的基础上按照全部树库、2/3 树库和 1/3 树库的规模上进行了三组实验。在获取的三组汉语词汇句法功能分布的数据基础上，使用 3.3 中的方法，本文得到了三组汉语词汇句法功能分布复杂度的数据，具体数据见表 4。

取出了词汇数目为 15 以上的数据之后，本文在三个规模不同的树库基础上，分别获取了 44 739 个、34 305 个和 20 579 个词汇的语法功能分布复杂度，并且对应的语法功能分布复杂度类别分别为 13 个、12 个和 11 个。为了更加直观地说明词汇句法功能分布复杂度与词汇总量之间的关系，本文在表中增加了对词汇数目取自然对数后的数值，同时，这一组数据也被用于了后面的一元线性回归分析中。

从表 4 中可以看出，三个不同规模的树库同时呈现出随着词汇句法功能分布复杂度的增加，词汇数目整体上基本呈现出指数级下降的趋势，这一趋势在对数值中呈现的更加明显。

从表 4 中可以看出，汉语词汇的句法功能分布复杂度为 1，即词汇在一种句法结构出现的词汇在三个不同规模的树库中分别占了整个词汇的 59.71%、60.26% 和 62.64%，这一数据有力地说明了绝大部分汉语词汇在组成句法结构上表意的单一性和明确性。但 40.29%、39.74% 和 37.36% 的词汇充当多种句法结构的成分，这也从一定程度上说明了汉语句法结构歧义问题非常严重这一事实。在三组不同规模的树库上，汉语词汇的句法功能分布复杂度为 2 的分别占了整个词汇的 17.88%、18.12% 和 17.91%，分别为整个词汇充当两种或两种以上句法结构成分的 44.37%、45.61% 和 47.93%，这两组数据说明解决了这部分词汇的歧义问题对整个汉语歧义问题的解决具有重要的意义。

基于表 4，在汉语词汇句法功能分布复杂度和词汇数目的基础上，计算出三个树库中汉语词汇的平均句法功能分布复杂度，分别为 1.97、1.92 和 1.84，在这三个值的基础上，可以得出整个汉语词汇的平均句法功能分布复杂度为 1.91。根据齐普夫定律，结合本文所使用树库中的词汇规模，可以推测出在其他规模的树库中，汉语词汇的平均句法功能分布复杂度应和本文中的相仿，这说明汉语词汇的平均句法功能分布复杂度是一个固定的值。

5.2 汉语词汇句法功能分布复杂度的洛特卡现象

从表 4 可以看出，尽管汉语词汇的句法功能分布复杂度和词汇总量之间的关系与洛特卡定律的含义是没有任何关系的，但其呈现形式确实是一致的，

表 4 汉语词汇句法功能分布复杂度表

编号	全部树库			2/3 树库			1/3 树库		
1	汉语词汇句法复功能分布复杂度	对应句法功能分布复杂度的词汇数目	对词汇数目取对数（ln）后的值	汉语词汇句法复功能分布复杂度	对应句法功能分布复杂度的词汇数目	对词汇数目取对数（ln）后的值	汉语词汇句法复功能分布复杂度	对应句法功能分布复杂度的词汇数目	对词汇数目取对数（ln）后的值
2	1	26 713	10.19 291	1	20 673	9.936 584	1	12 890	9.464 207
3	2	7 998	8.98 6947	2	6 217	8.735 043	2	3 685	8.212 026
4	3	3 987	8.290 794	3	3 003	8.007 367	3	1 645	7.405 496
5	4	2 291	7.736 744	4	1 774	7.480 992	4	940	6.84 588
6	5	1 444	7.275 172	5	1 029	6.936 343	5	555	6.318 968
7	6	942	6.848 005	6	686	6.530 878	6	367	5.905 362
8	7	567	6.340 359	7	391	5.968 708	7	222	5.402 677
9	8	351	5.860 786	8	232	5.446 737	8	139	4.934 474
10	9	199	5.293 305	9	155	5.043 425	9	76	4.330 733
11	10	121	4.795 791	10	81	4.394 449	10	35	3.555 348
12	11	68	4.219 508	11	42	3.73 767	11	25	3.218 876
13	12	36	3.583 519	12	22	3.091 042			
14	13	22	3.091 042	13					

在对词汇总量取对数后，这种呈现形式更加明显和直观，对于汉语词汇的这种分布现象本文称之为汉语词汇句法功能分布复杂度的洛特卡现象。以汉语词汇的句法功能分布复杂度和词汇数目取对数后的值做成的散点图具有非常明显的线性趋势，具体见图5。

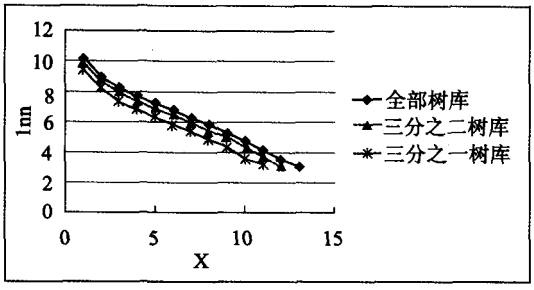

图5 汉语词汇句法功能分布复杂度与词汇数目的散点图

图5中呈现的是表4中三个不同规模树库中汉语词汇句法功能分布复杂度与词汇数目对数(ln)值之间关系的散点图，X 代表词汇的句法功能分布复杂度，ln$n$ 表示汉语词汇数目取对数后的数值。从图5中可以看出，相关点之间基本呈现一条直线态势，与洛特卡定律的呈现图基本上是一致的。下面对汉语词汇句法分布复杂度和汉语词汇数目取对数后的数值之间的关系进行一元线性回归分析。在使用SPSS[7]对三个不同规模的清华汉语树库中的数据处理后，得到一元线性回归的相关值，具体见表5。

表5 汉语词汇句法功能分布复杂度和词汇总量关系的一元线性回归结果

	全部树库	2/3 树库	1/3 树库
句子个数	31 138	20 758	10 379
词汇数目	44 739	34 305	20 579
coefficient	-0.548	-0.569	-0.580
constant	10.181	9.975	9.444
R Square	0.991	0.989	0.983
F sig	0.000	0.000	0.000
T sig	0.000	0.000	0.000

从表5中可以看出，Square 分别为0.991、0.989和9.444，这说明词汇句法功能分布复杂度与词汇总量共现比率非常高。而基于表5得到的三种不同树库下的 $\ln n = -0.548x + 10.181$，$\ln n = -0.569x + 9.975$ 和 $\ln n = -0.580x + 9.444$ 三个一元线性回归方程，从 F sign 和 T sign 的显著水平上看，都是非常显著的。这也充分说明了汉语句法功能分布复杂度和词汇总量之间的这种洛特卡现象是非常显著的。

结合图5和表5可以看出，虽然词汇的句法分布洛特卡现象在不同大小的树库中都有呈现，但又有所不同。在规模比较小的树库中，汉语词汇的洛特卡现象斜率相对大一些。但从总体的态势看，这种斜率又是基本一致的。

6 结 语

本文基于清华汉语树库，统计了汉语词汇句法功能的分布，并给出了汉语词汇句法功能分布复杂度的定义。在汉语词汇句法功能分布复杂度数据表的基础上，本文详细地分析了汉语词汇的句法功能分布复杂度情况，计算得出了汉语词汇的平均句法功能分布复杂度值。通过与文献计量学中的洛特卡定律外在形式相比较，本文把句法功能分布复杂度和词汇总量之间的这种关系称为词汇句法功能分布复杂度的洛特卡现象，对于这种关系，进行了一元线性回归分析，分析数据显示这种关系非常显著。本研究下一步打算选取与汉语一样都存在分词问题的日语和与汉语差异较大的英语，通过同样的方法进行相关的实验，验证这种洛特卡现象是否在这两种语言上也有呈现。

参 考 文 献

[1] Lotka A J. The frequency distribution of scientific productivity [J]. Journal of the Washington Academy of Sciences, 1926(12): 317-324.

[2] Zipf G K. Human Behaviour and the Principle of Least-Effort, Addison-Wesley[M]. Cambridge MA, 1949.

[3] 朱德熙. 语法讲义[M]. 北京：商务印书馆，1999：109-126.

[4] 陈小荷. 从自动句法分析角度看汉语词类问题[J]. 语言教学与研究，1999(03)：63-72.

[5] 俞士汶，朱学锋，王惠，等.《现代汉语语法信息词典》的新进展[J]. 中文信息学报，2000(01)：59-65.

[6] 徐艳华. 现代汉语实词语法功能考察及词类体系重构[D]. 南京：南京师范大学，2006：15-37.

[7] Berry M J A, Linoff G S. Data Mining Tech-niques: For Marketing, Sales, and Customer Relationship Management [M]. Hoboken: Wiley Computer Publishing, 2004.

(责任编辑 马 兰)