现代汉语词类使用情况统计

邢红兵

提 要 本文依据"现代汉语研究语料库系统"[注]中带有词性标记和频度的词典对现代汉语词类的使用情况进行了统计分析。文章主要统计了语料库中各类词的数量及词次(词在语料库中的使用次数之和);并按照词次分等级,分析各类词在各频度等级的分布情况;文中还分析了各频级词在语料库中的覆盖情况。

1. 关于语料库中的切分单位

本文的统计是依据"现代汉语研究语料库系统"(以下简称"语料库")中的电子词典来进行的,该词典包括三方面的内容:切分单位、对切分单位的标记和频度。切分单位包括以下几类:各类词、短语、词缀、非汉字、段落标记、标点及其他。"语料库"中被切分的不同形的切分单位共48972个,如果将兼类的词看成不同的切分单位(比如"把"可以是介词、量词、概数词、动词、名词,我们就把"把"看成五个单位),那么,切分单位共有55506个,各类切分单位的数量如下:

各类词	短语	词缀	段落标记	标点	其他
50137	4607	281	1	34	446

其中各类词50137个,占总切分单位的90.33%。本文只对"语料库"中50137个词进行统计分析。

2. 各类词的数量及词次

2.1 实词和虚词

我们知道,实词和虚词的区别在于能否充当句子成分。但它们在"语料库"中的数量和使用情况有没有显著的差别,这是我们首先要分析的。"语料库"标注的词类有:名词、动词、形容词、数词、量词、介词、副词、连词、助词、语气词、象声词和叹词,共13类,我们参照王还(1995)将名词、动词、形容词、数词、量词、代词归为实词;介词、连词、副词、助词、语气词、象声词、叹词、语气词归为虚词。下表是"语料库"中实词和虚词的数量和使用情况。

	数量	占总词数的比例	总词次	比例	平均词次
实词	47133	94.01%	936156	76.29%	19.86
虚词	3004	5.99%	291008	23.71%	96.86

从上面的数据可以看出:实词的数量是虚词的近16倍,但从平均词次来看,虚词却比实词高。在数量上虚词虽然只占5.99%,但总词次已占23.71%,就是说平均每4个词中就有一个虚词;语料库的总词次为1227164,总词数为50137,词的总平均词次为24.48次,虚词的平均词次为96.86,是总平均词次的近4倍。

2.2 各类词的词次及数量

"语料库"中使用的标记共112个,其中词类标记85个,这些词类标记采用层级体系,最多有三个层级,如"npm"中"n"是最高层,表示名词,"p"是第二层,表示名词中的专有名词,"m"是第三层,表示名词中的专名中的人名。最高层有13个大类,即上述的六类实词和七类虚词(孙宏林等,1997),它们分别是:名词(n)、动词(v)、形容词(a)、数词(m)、量词(q)、代词(r)、介词(p)、连词(c)、副词(d)、助词(u)、语气词(y)、象声词(o)、叹词(e)。

2.2.1 各类词的总词次

我们首先统计了各类词在语料库中的总词次,同时分析了各类词的总词次占全部词总词次的比例。

© 1995-2004 Tsinghua Tongfang Optical Disc Co., Ltd. All rights reserved.

词性	名词	动词	助词	副词	代词	形容词	数词
总词次	357842	304171	95618	94434	89878	76552	59309
比例(%)	29.16	24.79	7.79	7.70	7.32	6.24	4.83

词性	介词	量词	连词	语气词	叹词	象声词
总词次	49679	48404	27004	20931	2444	898
比例(%)	4.05	3.94	2.20	1.71	0.20	0.07

从上表我们可以看出各类词在语料库中的动态分布情况。名词的分布最广，占 29.16%，其次是动词，占 24.79%，这些动词包括系动词、助动词、动词作谓语、动词作修饰语、动词作补语、动词作中心语等，名词和动词合起来占一半以上，就是说语料库中的词有一半以上是名词和动词。形容词的分布相对于名词、动词来说比较低，只占 6.24%，排在第六位。

2.2.2 各类词的数量

上面我们分析了各类词在语料库中的动态分布，下面我们统计一下语料库中各类词的数量。

词性	名词	动词	形容词	副词	数词	量词	连词
总词次	26754	12623	5158	1937	1708	591	302
比例(%)	53.36	25.18	10.29	3.86	3.41	1.18	0.60

词性	代词	介词	象声词	助词	语气词	叹词
总词次	299	224	223	137	93	88
比例(%)	0.60	0.45	0.45	0.27	0.19	0.18

从上表可以看出，名词、动词和形容词这三类词，是汉语词汇的主要构成，这三类词共 44535 个，占总词数的 88.83%，其中名词已经占总词数的 53%以上，其他十类词的数量只占 11.17%。

2.2.3 各类词的平均词次

上面我们统计了各类词的数量以及它们在语料库中的动态分布情况，根据这两项数据我们可以得出各类词的平均词次。

助词	代词	语气词	介词	连词	量词	副词	数词	叹词	动词	形容词	名词	象声词
697.9	300.6	225.1	221.8	89.4	81.9	48.8	34.7	27.8	24.1	14.8	13.4	4.03

从上表可以看出，虚词的平均词次都很高，比如助词、语气词、介词、连词等。因为虚词常常只表示一种语法关系，这种关系只是类的关系，而这种类的关系又很普遍，比如表示结构关系的助词“的”，凡是具有领属、修饰等关系时都用它，所以它的词次很高。再比如时态助词“过”，凡是表示经历态的动词都要用这个词，而动词在语料库中的分布又很多，所以它的词次也比较高，因为它是辅助动词的。代词的平均词次也很高，这是因为代词是一个特殊类的实词，它代替具体事物或动作，而且在不同的环境中它代替不同的对象。动词、形容词和名词则排在倒数 2～4 位，我们就拿名词为例做一分析，名词中词次大于 5000 的 1 个；词次大于 1000 的 28 个；词次大于 500 的 76 个；词次大于 100 的 553 个；词次小于 100 的 26201 个；词次小于 10 的 22735 个；词次等于 1 的 11439 个，甚至还有一部分名词在语料库中没有出现，因而它们的平均词次就很低了。

3. 各类词频度分布情况

语料库中使用的 50137 个词中，各个词的出现次数从 65669 次到 1 次不等，而且，很多词都是兼类或同形，例如“一”兼数词和副词；动态助词“了”和补语“了(liǎo)”是同形。我们这里所说的频度是指一个词的词次占语料库总词次的比例，这个词不包括它兼类或它的同形，比如“的”在“语料库”中的词次是 65669 次，频度为 5.35%，位居第一，但是“的”兼助词和语气词，它们的词次分别是 64047 次和 1536 次，频度分别为 5.22%和 0.13%。因此我们仅仅统计出“的”是高频词还不准确，因为语气词“的”的使用频度并不高。我们这里分析的是各类词的准确频度(不包括兼类或同形)。

3.1 各类词的频级分布

我们按照词次划分八个等级，一级：大于 10000 次；二级：大于 5000 次；三级：大于 1000 次；四级：大于 500 次；五级：大于 100 次；六级：大于 50 次；七级：大于 10 次；八级：小(等)于 10 次。具体数据见下表：

	一级	二级	三级	四级	五级	六级	七级	八级	总数
名词	0	1	28	80	590	1162	4249	22505	26754
动词	1	4	36	86	459	878	2862	9761	12623

© 1995-2004 Tsinghua Tongfang Optical Disc Co., Ltd. All rights reserved.

形容词	0	0	6	10	139	267	968	4190	5158
副词	1	3	14	26	126	210	552	1385	1937
数词	1	1	10	21	60	101	219	1489	1708
量词	1	2	10	18	71	109	212	379	591
连词	0	1	3	7	40	65	111	191	302
代词	2	4	17	23	66	80	125	174	299
介词	1	1	12	20	48	56	88	136	224
象声词	0	0	0	0	1	3	11	212	223
助词	2	2	8	9	17	19	43	94	137
语气词	0	1	8	9	14	17	24	69	93
叹词	0	0	0	0	6	16	23	65	88
合计	9	20	152	309	1637	2983	9487	40650	50137

从上面的统计可以看出，词的频度差别很大，词次最高可达 64100 次，频度为 5.22%，词次最低的为 1，各类词中词次为 1 的词共有 19854 个，占总词数 50137 的 39.6%，其中名词、动词、形容词分别为 11439、4318、1917 个。词次高于 1000 次的 152 个高频词的排列顺序、词性及频度参见附录。

3.2 各类词中的高频词

各类词中的每个词的频度差别也很大，比如介词中频度最高的是“在”，它的词次为 12395 次，排在第二的“对”的词次仅为 3551 次，词次为 1 的介词就有 60 个，占介词总数的 26.8%。下表是各类词中频度排在前 10 位的词，词后面的数字是它们的词次。

助词： 的 64100 了 13818 着 4902 地 3671 得 1741 等 1585 过 1402 之 1185 所 817 来 401

动词： 是 17795 有 8556 说 6053 要 5369 到 3824 来 3784 去 3049 看 3010 发展 2927 能 2841

数词： 一 16473 两 2831 十 2451 三 2348 第 2183 二 2000 几 1872 万 1704 多 1438 四 1183

代词： 我 16252 这 11162 你 9553 他 8506 我们 3721 她 3536 那 3396 什么 2419 各 1927 自己 1865

介词： 在 12395 对 3551 把 3283 给 2350 为 2333 从 2284 到 2139 以 1546 向 1416 与 1399

副词： 不 12133 就 7236 也 6137 还 4216 都 4134 又 2812 一 1943 很 1769 最 1444 没 1424

量词： 个 10498 年 5179 种 2246 次 1691 月 1682 位 1340 天 1314 条 1098 些 1087 元 1081

连词： 和 9385 而 1868 但 1341 与 992 并 678 可 617 或 589 因为 473 如果 471 可是 465

语气词： 了 7327 啦 2108 呀 1740 吧 1739 的 1536 呢 1504 吗 1489 啊 1027 哪 859 嘛 391

名词： 人 6763 上 4927 中 3756 里 3002 中国 2910 经济 2872 企业 2446 国 2186 工作 2001 国家 1849

形容词： 大 3557 好 3269 新 2081 多 1844 小 1416 高 1129 重要 913 全 900 老 743 副 618

叹词： 哎 408 啊 343 嗯 169 噢 155 哎呀 149 唉 148 哦 98 嘿 97 哟 86 嗬 82

象声词： 哈 104 哈哈 78 嘿 60 锵 24 喷 21 咚 19 叭 18 嘿嘿 15 轰 15 嗬 14

从上面的表中我们可以看出：(1)助词“的”的词次为 64100 次，远远高于其他助词，也远远高出其他类词。(2)系动词“是”、数词“一”、介词“在”、否定副词“不”、量词“个”、连词“和”、语气词“了”等都在各自的类中遥遥领先，这些词可看成为特高频词。

4. 各频级词的分布情况分析

我们要考察的各频级词的分布情况包括两个方面，一是该频级词的数量和它们占总词数的比例，这个数据是静态的，不包括频度；另一是该频级词总词次和它们占语料库总词次的比例。下表是各级词在语料库中的分布情况表，其中一级：10000 次以上；二级：5000－9999 次；三级：1000－4999 次；四级：500－999 次；五级：100－499 次；六级：50－99 次；七级：11－49 次；八级：10 次以下(包括 10 次)。

	一级	二级	三级	四级	五级	六级	七级	八级
词数	9	11	132	157	1328	1346	6504	40650
比例(%)	0.0180	0.0219	0.2633	0.3131	2.6487	2.6846	12.9725	81.0778
累计比例(%)	0.0180	0.0399	0.3032	0.6163	3.2650	5.9469	18.9194	100
总词次	174626	80064	247861	107618	274428	93904	144000	104663
比例(%)	14.2300	6.5243	20.1979	8.7697	22.3628	7.6521	11.7344	8.5289
累计比例(%)	14.2300	20.7543	40.9522	49.7219	72.0846	79.7367	91.4711	100

从上面的表我们可以看出，前 9 个高频词虽然数量只占 0.018%，但在语料库中已经覆盖了 14.23%；前 20 个高频词数量只占 0.0219%，但覆盖率高达 20.75%；词次在 50 次(包括 50 次)以上的词共 2983 个，数量

© 1995-2004 Tsinghua Tongfang Optical Disc Co., Ltd. All rights reserved.

上只占 5.95%，而覆盖率却高达 79.74%；10 次以上的 9487 个词数量上只占 18.92%，覆盖率已达到 91.47%，而词次在 10 次以下的 40650 个词，数量上占 81.08%，覆盖率却只有 8.53%。

注："现代汉语研究语料库系统"是由北京语言文化大学语言信息处理研究所开发研制而成，该语料库系统包括 2000 万字的精选的粗语料、200 万字的经过分词和词性标注的精语料。

参考文献：

[1]朱德熙：《语法讲义》，商务印书馆 1997 年版。

[2]胡明扬主编《词类问题考察》，北京语言学院出版社 1996 年版。

[3]王还主编《对外汉语教学语法大纲》，北京语言学院出版社 1995 年版。

[4]黄伯荣、廖序东主编《现代汉语》，高等教育出版社 1991 年版。

[5]陈原：《现代汉语定量分析》，上海教育出版社 1989 年版。

[6]孙宏林、黄建平、孙德金、李德钧、邢红兵：《"现代汉语研究语料库系统"概述》，胡明扬主编《第五届国际汉语教学讨论会论文选》，北京大学出版社 1997 年版。

附录：词次高于 1000 次的前 152 个高频词词性及频度表

表中的竖栏依次为：序号，词形，词性代码，频度(‰)。

1 的 u 52.2343	32 她 r 2.8814	63 一 d 1.5833	93 吗 y 1.2134	123 之 u 0.9656
2 是 v 14.5009	33 那 r 2.7674	64 各 r 1.5703	94 最 d 1.1767	124 四 m 0.9640
3 一 m 13.4236	34 把 p 2.6753	65 走 v 1.5654	95 多 m 1.1718	125 得 v 0.9616
4 我 r 13.2435	35 好 a 2.6639	66 会 v 1.5304	96 家 n 1.1710	126 可以 v 0.9445
5 了 u 11.2601	36 去 v 2.4846	67 几 m 1.5255	97 没 d 1.1604	127 世界 n 0.9412
6 在 p 10.1005	37 看 v 2.4528	68 而 c 1.5222	98 小 a 1.1539	128 公司 n 0.9249
7 不 d 9.8870	38 里 n 2.4463	69 自己 r 1.5198	99 向 p 1.1539	129 高 a 0.9200
8 这 r 9.0958	39 发展 v 2.3852	70 他们 r 1.5189	100 过 u 1.1425	130 同志 n 0.9062
9 个 q 8.5547	40 中国 n 2.3713	71 国家 n 1.5067	101 与 p 1.1400	131 条 q 0.8947
10 你 r 7.7846	41 经济 n 2.3404	72 多 a 1.5027	102 做 v 1.1164	132 打 v 0.8907
11 和 c 7.6477	42 能 v 2.3151	73 社会 n 1.4912	103 起 v 1.1164	133 每 r 0.8858
12 有 v 6.9722	43 两 m 2.3069	74 市场 n 1.4481	104 更 d 1.1148	134 些 q 0.8858
13 他 r 6.9314	44 又 d 2.2915	75 很 d 1.4415	105 人民 n 1.1123	135 管理 v 0.8825
14 了 y 5.9707	45 出 v 2.1179	76 得 u 1.4187	106 怎么 r 1.1009	136 成 v 0.8825
15 就 d 5.8965	46 十 m 1.9973	77 呀 y 1.4179	107 已 d 1.0977	137 前 n 0.8809
16 人 n 5.5111	47 企业 n 1.9932	78 吧 y 1.4171	108 但 c 1.0928	138 元 q 0.8809
17 也 d 5.0010	48 什么 r 1.9712	79 万 m 1.3886	109 位 q 1.0919	139 听 v 0.8711
18 说 v 4.9325	49 给 p 1.9150	80 次 q 1.3780	110 技术 n 1.0854	140 文化 n 0.8597
19 要 v 4.3751	50 三 m 1.9134	81 月 q 1.3706	111 被 p 1.0781	141 开 v 0.8564
20 年 q 4.2203	51 为 p 1.9011	82 您 r 1.3275	112 天 q 1.0708	142 这样 r 0.8491
21 上 n 4.0149	52 从 p 1.8612	83 问题 n 1.3185	113 使 v 1.0618	143 才 d 0.8450
22 着 u 3.9946	53 种 q 1.8302	84 建设 v 1.3112	114 时 n 1.0545	144 于 p 0.8401
23 还 d 3.4356	54 国 n 1.7813	85 为 v 1.2997	115 下 v 1.0292	145 谁 r 0.8377
24 都 d 3.3687	55 第 m 1.7789	86 等 u 1.2916	116 它 r 0.9999	146 啊 y 0.8369
25 到 v 3.1161	56 到 p 1.7430	87 以 p 1.2598	117 没有 v 0.9950	147 叫 v 0.8304
26 来 v 3.0835	57 啦 y 1.7178	88 后 n 1.2574	118 事 n 0.9925	148 市 n 0.8287
27 中 n 3.0607	58 新 a 1.6958	89 的 y 1.2517	119 吃 v 0.9762	149 政府 n 0.8239
28 我们 r 3.0322	59 工作 n 1.6306	90 让 v 1.2394	120 钱 n 0.9762	150 找 v 0.8173
29 地 u 2.9915	60 二 m 1.6298	91 呢 y 1.2256	121 下 n 0.9738	151 知道 v 0.8157
30 大 a 2.8986	61 想 v 1.5955	92 改革 v 1.2166	122 再 d 0.9705	152 住 v 0.8149
31 对 p 2.8937	62 上 v 1.5882			

（责任编辑　蔡根林）

© 1995-2004 Tsinghua Tongfang Optical Disc Co., Ltd. All rights reserved.