

Research on Chinese Words and Characters Based on Statistics

邢红兵 著

基 于 统 计 的

汉语字词研究

良	ab	1	4
良辰	ng0	1	1
良好	ag0	129	138
良机	ng0	5	5
良久	az	1	9
良久	nt	7	9
良师益友	in	1	1

警	b	敬	言	5	廿	勺	口	女	言
景	b	日	京	4	日	京			
境	a	土	竟	3	土	立	日	儿	
敬	a	苟	女	4	廿	勺	口	女	
镜	a	钅	竟	3	钅	立	日	儿	
竞	b	音	儿	3	立	口	儿		
拒	a	扌	巨	2	扌	巨			



语 文 出 版 社

<http://www.ywcbs.com>

基于统计的汉语字词研究

邢红兵 著



语文出版社

图书在版编目(CIP)数据

基于统计的汉语字词研究/邢红兵著, —北京: 语文出版社, 2005. 6. 6

ISBN 7-80184-499-8/H. 49

I. 基... II. 邢... III. ①汉字—研究②汉语—词
汇—研究 IV. H1

中国版本图书馆 CIP 数据核字(2005)第 053053 号

基于统计的汉语字词研究

邢红兵 著

*

语文出版社出版

100010 北京朝阳门南小街 51 号

E-mail: ywp@ywbs. com

新华书店经销 北京市昌平百善印刷厂印刷

*

850 毫米×1168 毫米 32 开 7.625 印张 212 千字

2005 年 6 月第 1 版 2005 年 6 月第 1 次印刷

印数: 1—1000 定价: 25.00 元

本书如有缺页、倒页、脱页, 请寄本社发行部调换。

目 录

现代汉语字词基础部件统计分析	(1)
附录 1 汉字基础部件的使用次数、构字数、构词数表	(11)
附录 2 “现代汉语研究语料库系统”简介	(27)
现代汉语词类使用情况统计	(31)
附录 3 部分出现 10 次以上词的使用情况表	(40)
从分词的不一致性看汉语分词的难点	(50)
附录 4 出现次数在 10 次以上的有分歧词语表	(59)
词性标注中难归类词语分析	(65)
汉语词语重叠结构统计分析	(73)
附录 5 重叠结构表	(84)
现代汉语插入语研究	(103)
附录 6 语料库中出现的插入语及其使用次数	(114)
信息技术领域术语用字分析	(115)
附录 7 术语用字在术语中的出现频率表	(124)

信息技术领域汉英术语的特征及其分布规律	(139)
《(汉语水平)汉字等级大纲》汉字部件统计分析	(148)
附录 8 基础部件成字、构字情况及其所构字表	(160)
《(汉语水平)汉字等级大纲》中形声字声旁表音特点分析	(185)
附录 9 “等级汉字”中可用作声旁的汉字及其表音能力表 ..	(194)
《(汉语水平)词汇等级大纲》语素数据库建设	(201)
留学生形声字声旁规则性效应调查	(210)
留学生偏误合成词的统计分析	(219)
附录 10 偏误合成词分类表	(235)
后 记	(239)

现代汉语字词基础部件统计分析*

提要 本研究依据《信息处理用 GB13000.1 字符集汉字部件规范》所规定的汉字基础部件,对“现代汉语研究语料库系统”中使用的全部汉字和词进行基础部件拆分,并对拆分出的汉字基础部件进行统计和分析。文章分三部分进行分析:一部分是对“语料库”中使用的汉字进行拆分,对拆分出的结果进行静态(不考虑该字在“语料库”中的使用频度)和动态(考虑该字在“语料库”中的使用频度)的分析;另一部分是以词为单位,对“语料库”中使用的全部的词直接进行基础部件拆分,在此基础上,对词和部件的关系作一些统计和分析。第三部分是在上述的统计和分析的基础上,对基础部件构字、构词的情况进行分析。

关键词 汉字部件 基础部件 部件拆分

一 前言

《信息处理用 GB13000.1 字符集汉字部件规范》(以下简称“规范”)已经颁布实施,该规范的实施是我国中文信息处理发展中的一件大事。本研究以该规范规定的基础部件为依据,对“现代汉语研究语

* 本文载于黄昌宁主编的《1998 中文信息处理国际会议论文集》,清华大学出版社 1998 年 9 月出版。收入本书时略有修改并增加了附录。

料库系统”(参见附录2,以下简称“语料库”)中所使用的全部汉字和全部词进行基础部件拆分,并对拆分结果进行分析。本文做这样的统计研究是基于以下几个方面的考虑:(1)为规范提供动态的相关统计数据;(2)为汉字信息处理,特别是汉字的编码提供相关的统计数据;(3)为汉字教学提供一些相关统计数据;(4)为汉字的研究提供可查询的相关资料。

去除一些标点、符号、英文字母、非汉字符号等,“语料库”所使用汉字共4868个,这些汉字共使用1857660次,平均每字使用次数为381.61次。我们对这4868个汉字进行部件拆分,得到“语料库”用字基础部件序列表。此表包括字形、该字的使用次数、使用频率、部件数、基础部件序列等信息。

我们知道,能够独立使用的最小的语言单位是词。一个汉字可以是一个词,但很多时候一个汉字只代表语素,是构词的单位;有的汉字甚至只表示音节,例如:“猩猩”中的“猩”、“蝴蝶”中的“蝴”等。从这个角度考虑,我们还将对“语料库”中的词包含的汉字进行部件分析。“语料库”共有47006个不同词形(其中已经排除用阿拉伯数字表示的数词,英文词以及一些特殊符号等),共使用1236897次,平均每个词使用11.12次。我们对这四万七千多词按照每个词中的汉字的部件进行拆分,得到“语料库”用词汉字基础部件序列表。此表包括词形、词长、词的使用次数、词的汉字部件序列等信息。

在此基础上,对拆分出的基础部件进行统计,得出“语料库”基础部件表,此表包括基础部件、该部件能否单用、是否是单笔、该部件构字数、该部件构词词数、该部件使用次数等信息。

二 汉字包含的基础部件数量

我们首先分析一下汉字的平均部件数:库中共使用汉字4868个,

拆分的基础部件总数量为 15216 个,平均每字的部件数为 3.1257 个;加上各字的使用次数,库中共使用汉字 1857660 字次,基础部件总使用次数为 4324716 次,平均每字的部件数为 2.3280 个。表 1 是我们对部件数不同的汉字的字数及在“语料库”中的出现次数进行统计的结果:

表 1 不同部件数汉字的数量及其使用情况表

	字数	百分比	字次	百分比	平均使用次数
单部件字	219	4.50	401145	21.59	1831.71
两部件字	1195	24.55	680833	36.65	569.73
三部件字	1867	38.35	588364	31.67	315.14
四部件字	1070	21.98	144276	7.77	134.84
五部件字	400	8.22	38878	2.09	97.20
六部件字	94	1.93	3776	0.20	40.17
七部件字	20	0.41	379	0.02	18.95
八部件字	2	0.04	8	0.0004	4.00
九部件字	1	0.02	1	0.0001	1.00

从上述统计数字可以看出,在这 4868 个汉字中,三部件字最多,两部件字和四部件字的数量也不少,共 4132 个,占 84.88%;单部件字只占 4.5%,四部件以下字共占 90.34%。在实际使用中,两部件字占多数,四部件字却占 8% 不到;单部件字、两部件字、三部件字占“语料库”总字次的 90%;单部件字虽然只有 219 个,而实际使用中却占五分之一强,使用率最高,平均每字在“语料库”中使用 1831.71 次,比平均次数 381.61 多好几倍。五部件以上的字共 517 个,只占 10.62%,使用 43042 次,占总字次的 0.22%。六部件以下字使用次数最高的前二十个字分别是:

单部件:口一人日土白又大八木十寸女小匕立王也心不
 两部件:有这国他和来要们地出家生么着时作以多对好

三部件:的是在你会到发经学行没同后都实部动展起定
 四部件:说得能高给资提党通路做您统院调增深常感啊
 五部件:就题解领商候德续影搞脸演献街读检喝险额器
 六部件:验歌嘴疑傻衡龄赢滴敲灌嚷攀摘歇蔬耀囊懈翠
 七部件的字有 20 个:颤疆窿凝颧颙攘嚷衢瀚瀛鸢骥骧鹞翳蹶蹴

鰐

八部件的字只有“壑”和“鰐”两个;九部件的字只有“懿”一个。

三 词包含的基础部件数量

词包含的基础部件数量是指对构成一个词的汉字进行部件拆分所得到的基础部件数量,例如“白”拆分为“白”,它的基础部件数量为 1;“他们”拆分为“亻也亻门”,基础部件数量为 4。“语料库”中使用的词共有 47006 个,我们对这些词包括的汉字进行部件拆分,得到的部件总数是 277448 个,平均每个词的部件数是 5.9024 个。加上频度,“语料库”中使用词的总次数为 1236897 次,基础部件总使用次数为 4324716 次,平均每词的部件数为 3.4964 个。

下面我们先分析一下不同部件数词的分布情况。“语料库”中的词,包含的部件数最少只有一个,最多有 24 个。具体分布情况见下表:

表 2 不同部件数词的数量及其使用情况表

部件数	词数	百分比	累计	使用次数	百分比	平均使用次数
1	203	0.4319	0.4319	179981	14.5510	886.61
2	1685	3.5846	4.0165	243789	19.7097	44.68
3	4506	9.5860	13.6025	296449	23.9672	65.79
4	7589	16.1447	29.7472	176334	14.2562	23.24

现代汉语字词基础部件统计分析

5	9124	19.4103	49.1575	150522	12.1693	16.50
6	8374	7.8147	66.9722	101134	8.1764	12.08
7	5581	11.8730	78.8452	47894	3.8721	8.58
8	3727	7.9288	86.7740	21049	1.7018	5.65
9	2302	4.8972	91.6712	9868	0.7978	4.29
10	1609	3.4230	95.0942	4747	0.3838	2.95
11	946	2.0125	97.1067	2280	0.1843	2.41
12	617	1.3126	98.4193	1278	0.1033	2.07
13	295	0.6276	99.0469	685	0.0554	2.32
14	223	0.4744	99.5213	517	0.0418	2.32
15	84	0.1787	99.7000	117	0.0095	1.39
16	73	0.1553	99.8553	120	0.0097	1.64
17	21	0.0447	99.9000	46	0.0037	2.19
18	30	0.0638	99.9638	62	0.0050	2.07
19	3	0.0064	99.9702	3	0.0002	1.00
20	4	0.0085	99.9787	6	0.0005	1.50
21	2	0.0043	99.9830	2	0.0002	1.00
22	7	0.0149	99.9979	12	0.0010	1.71
24	1	0.0021	100.0000	2	0.0001	2.00

从上表我们可以看出:在这四万七千多个词中,五部件词占 19% 多,四部件、六部件、七部件词的比例也很高,九部件以下词已占 90% 多,十三部件以下词已占 99% 以上,十四部件以上的词仅占 1% 不到。而在实际使用中,三部件词占近 24%,五部件以下词占近 85%,九部件以下词已占 99.2%。下面我们分析一下音节不同的词所包含的部件情况。

表3 不同音节词的数量及其使用情况表

	词数	百分比	词次	百分比	平均 使用次数	平均 部件数
单音节	3199	6.8055	682251	55.1583	213.27	2.99
双音节	29906	63.6217	504951	40.8240	16.88	5.16
三音节	8026	17.0744	34438	2.7842	4.29	7.10
四音节	5454	11.6028	14370	1.1618	2.64	9.45
五音节	286	0.6084	660	0.0534	2.31	11.44
六音节	82	0.1744	133	0.0108	1.62	13.21
七音节	53	0.1128	94	0.0076	1.77	15.60

“语料库”中双音节词占优势,比例为64%,单音节词只占6.8%,单音节词双音节词合在一起占70%多,四音节以下词占99.1%,五音节以上词仅占0.9%;但是,在实际使用中,单音节词的使用比例最高,占55%多,平均每词使用213次多;双音节词占近41%,平均每个词使用近16次,单音节词和双音节词加起来的使用次数占词的总使用次数的95.98%,三音节以上词的使用次数仅占词的总使用次数的4.02%。

我们对四音节以下词的部件数进行统计,统计结果见表4:

表4 四音节以下词的部件数表

部件数	单音节词		双音节词		三音节词		四音节词	
	词数	比例	词数	比例	词数	比例	词数	比例
1	203	6.3457						
2	878	27.4461	805	2.6918				
3	1214	37.9494	3175	10.6166	118	1.4702		
4	625	19.5374	6482	21.6746	433	5.3950	49	0.8984
5	225	7.0334	7733	25.8577	1041	12.9703	122	2.2369

现代汉语字词基础部件统计分析

6	45	1.4067	6394	21.3803	1590	19.8106	344	6.3073
7	8	0.2501	3283	10.9777	1677	20.8946	603	11.0561
8	1	0.0313	1430	4.7816	1415	17.6302	860	15.7682
9			437	1.4612	928	11.5624	890	16.3183
10			119	0.3979	492	6.1301	958	17.5651
11			38	0.1271	218	2.7162	625	11.4595
12			9	0.0301	77	0.9594	481	8.8192
13			1	0.0033	21	0.2616	233	4.2721
14					7	0.0872	178	3.2637
15					4	0.0498	40	0.7334
16					2	0.0249	41	0.7517
17					1	0.0125	7	0.1283
18					2	0.0249	16	0.2934
19							1	0.0183
20							1	0.0183
22							4	0.0733
24							1	0.0183

单音节词中,部件最少的是一个部件,最多的是八个部件,平均每个词的部件数是 2.99 个,其中三部件字占多数,其次是两部件字和四部件字,这三类共占单音节词总数的 85%;双音节词中,部件最少的是两个部件,最多的是十三个部件,平均每个词的部件数为 5.16 个,五部件词最多,三、四、五、六、七部件词共占 90.5%;三音节词数量最多的是七部件词,平均每词的部件数为 7.01 个;四音节词数量最多的是十部件词,平均每个词的部件数为 9.45 个。

四 基础部件分析

4.1 成字部件和非成字部件

成字部件是指可以独立成字的部件,例如“口”可以成字,也可以构成“啊、知”等汉字;不能独立成字的部件叫做非成字部件,例如“乚”,只能构成“吃、每、复”等字,不能单独成字。

在“语料库”使用的 4868 个汉字中,成字部件共有 214 个。这些成字部件从构字数量来看,有两种情况:一种是尽管它是基础部件,但它只构成它自己,不具备构成其他字的能力;一种是除了构成它自己以外,还可以构成其他字。在这 4868 个汉字范围内只构成自己的成字部件有 15 个,见下表。

表 5 不构成其他字的成字部件

部件	使用次数	部件	使用次数	部件	使用次数
个	11882	承	564	凹	15
年	9337	飞	549	凸	11
事	4448	竹	87	毋	6
已	2373	丫	34	卅	3
书	1861	秉	22	乚	2

既单独使用,又可以构成别的字的基础部件有 199 个。下面列出的“语料库”中使用次数、构字数在前十位的这类成字部件:

使用次数前 10 位的部件是:口一人日土白又大八木

构字数前 10 位的部件是:口一日木人八土十大又

4.2 单笔部件和复笔部件

只包含一个笔画的部件叫做单笔部件,包含两个或者两个以上笔画的部件叫做复笔部件。在拆分出的 440 个基础部件中,单笔部件共 13 个,占总部件的 2.95%,其中“一”、“乙”为成字部件,可以单用,单

用次数分别为 27021 次和 30 次;复笔部件 427 个,占总部件的 97.05%。单笔部件使用及构字情况见下表。

表 6 单笔部件表

单笔部件	一	丶	丿	丨	㇀	㇁	7
使用次数	175500	115270	44457	19730	9990	3684	3404
构字数	434	135	137	16	13	21	15
单笔部件	乙	乚	乚	㇃	㇄	㇅	
使用次数	3197	2921	1122	736	477	369	
构字数	12	23	12	3	5	4	

4.3 部件的频度统计

“规范”规定了 GB 13000.1 字符集中的基础部件 393 组,包括主形部件和变体部件共 560 个。“语料库”中用到的汉字基础部件 440 个,没有出现的汉字基础部件 120 个,这些基础部件可以概括为以下几类:(1)G 列(该字符集中祖国大陆提出的汉字)以外汉字专用基础部件,这类部件共有 55 个,例如:“冂”、“冊”、“𠂇”、“卅”等;(2)繁体汉字专用基础部件,例如:“貝”、“車”、“東”、“為”等;(3)少数不常用字专用部件,例如:“𠂇”、“𠂈”等。

我们统计了“语料库”中使用的 440 个汉字基础部件在“语料库”中的使用次数、构字数、在 20902 汉字中的构字数以及它们在“语料库”中的构词数。从使用频度来看,频度最高的基础部件是“口”,为 5.7%。使用频度排列在前十位的部件的累计频度为 26.9%,前二十位的部件的累计频度为 38.9%,使用频度排列在前 33 位的基础部件的累计频度已达 50.5%,前 100 个基础部件的累计频度为 79.3%,前 160 个基础部件的累计频度为 90%,前 310 个部件的累计频度为 99% 以上,其余 129 个部件的累计频度不到 1%。使用次数、构字数和构词数排在前 20 位的基础部件如下:

使用次数:口一、亻人日土勺白又大八ム讠木才十彳亍

构字数:口一日彳亍亻木人艹八才土十一又大ㄣ又女贝ソノ

构词数:口一日亻人木八彳大土十又ㄣ才一艹ムノ、讠

从统计数据可以看出,“口”、“一”这两个基础部件构字、构词及使用都排在第一、二位;构字数排在前二十位的基础部件中成字部件就有12个,而从使用频度和构词数来看,排在前二十位的基础部件中成字部件就有10个,常用偏旁“彳”、“亻”、“艹”、“才”等构字、构词能力都很强,而且在“语料库”中的使用比例也很高。

参考文献

《汉字部件规范》研制组(1998)关于制定《汉字部件规范》的若干问题,《计算机世界》1998年4月27日。

陈原(1989)《现代汉语定量分析》,上海:上海教育出版社。

罗振声、袁毓林(1996)《计算机时代的汉语和汉字研究》,北京:清华大学出版社。

苏培成(1994)《现代汉字学纲要》,北京:北京大学出版社。

王均(1995)《语文现代化论丛》,济南:山东教育出版社。

信息处理用 GB13000.1 字符集汉字部件规范(1998) 北京:语文出版社。

邢红兵、崔永华、张普(1997)GB13000.1 字符集汉字部件拆分与规范,《第二届术语学、标准化与技术传播国际学术会议论文集》,北京:中国大百科全书出版社。

邢红兵、崔永华、张普(1998)对部件称说的建议,《计算机世界》1998年4月27日。

张普(1992)《汉语信息处理研究》,北京:北京语言学院出版社。

现代汉语字词基础部件统计分析

附录1 汉字基础部件的使用次数、构字数、构词数表

表中的数据按照顺序是:部件、使用次数、使用频率(按照千分之多少来计算的)、构字数、构字的比例(百分比)、构字次数、构词数、构词比例(百分比)。表中部件是按照使用次数的降序排列的。

口	247647	57.2658	891	18.3032	1009	12938	27.5241
一	175500	40.5826	434	8.9154	466	8557	18.2041
丶	115270	26.6550	135	2.7732	139	2679	5.6993
亻	110975	25.6618	295	6.0600	301	5170	10.9986
人	101168	23.3941	283	5.8135	316	4924	10.4753
日	97265	22.4915	335	6.8817	349	5246	11.1603
土	93582	21.6399	233	4.7864	257	4418	9.3988
勹	79159	18.3047	102	2.0953	102	1325	2.8188
白	76533	17.6975	63	1.2942	63	1165	2.4784
又	66473	15.3712	187	3.8414	214	3707	7.8862
大	58783	13.5930	206	4.2317	207	4456	9.4796
八	56908	13.1594	247	5.0740	247	4574	9.7307
厶	56897	13.1568	126	2.5883	130	2776	5.9056
乚	54664	12.6405	108	2.2186	108	2388	5.0802
木	52018	12.0286	293	6.0189	325	4674	9.9434
扌	49161	11.3680	246	5.0534	246	3213	6.8353
十	49129	11.3606	220	4.5193	236	3973	8.4521
艹	48441	11.2015	117	2.4035	117	1586	3.3740
辶	46310	10.7087	295	6.0600	295	4466	9.5009
亼	45168	10.4446	210	4.3139	210	3035	6.4566
㇀	44568	10.3059	190	3.9030	190	3307	7.0353
丿	44194	10.2194	132	2.7116	132	2680	5.7014
寸	43067	9.9588	84	1.7256	84	1941	4.1293
女	42876	9.9146	151	3.1019	151	2081	4.4271

基于统计的汉语字词研究

讠	40780	9.4300	113	2.3213	113	1921	4.0867
ナ	40445	9.3525	62	1.2736	62	1386	2.9486
目	39960	9.2403	108	2.2186	109	2096	4.4590
小	37796	8.7399	69	1.4174	69	1720	3.6591
ㄨ	36861	8.5237	140	2.8759	140	2114	4.4973
匕	34183	7.9045	103	2.1159	106	1677	3.5676
立	33380	7.7188	99	2.0337	103	2073	4.4101
リ	32425	7.4979	86	1.7666	86	1765	3.7548
β	31964	7.3913	106	2.1775	106	1827	3.8867
王	31232	7.2221	99	2.0337	103	1828	3.8889
也	30906	7.1467	11	0.2260	11	446	0.9488
心	30305	7.0077	95	1.9515	97	2125	4.5207
纟	29828	6.8974	101	2.0748	101	1914	4.0718
不	28861	6.6738	18	0.3698	18	1050	2.2338
禾	28718	6.6407	84	1.7256	84	1570	3.3400
力	28422	6.5723	71	1.4585	73	1876	3.9910
贝	28387	6.5642	148	3.0403	154	2118	4.5058
儿	28043	6.4847	42	0.8628	46	2096	4.4590
子	27521	6.3639	58	1.1915	62	2194	4.6675
一	27119	6.2710	121	2.4856	121	2153	4.5803
口	27037	6.2520	59	1.2120	59	1253	2.6656
目	25210	5.8296	87	1.7872	92	1298	2.7613
夕	24211	5.5985	54	1.1093	61	1105	2.3508
了	23662	5.4716	12	0.2465	12	146	0.3106
工	22996	5.3176	50	1.0271	50	1425	3.0315
文	22992	5.3167	24	0.4930	24	709	1.5083
攴	22492	5.2010	74	1.5201	74	1474	3.1358
二	22283	5.1527	43	0.8833	43	1249	2.6571