中 文 信 息 学 报
第9卷 第4期 JOURNAL OF CHINESE INFIOMATION PROCESSING Vol.9 No.4

汉语输入编码中简码字、词的合理选配

韩布新　　任雪松

(中国科学院心理研究所)　　(北京信息工程学院)

【摘要】本文分析了几种常用汉字编码方案的简码字表，发现有很多不一致之处。考虑到简码字的合理选取数量、记忆量和键位安排等因素，提出汉字的使用频度／构词能力级比率要较单纯使用频度指标更为合理。根据这一指标，在"汉字属性信息数据库"基础上，找出了 78 个简码字和 120 个简码双字词，并进行了相应的键位安排以便于实际应用。

一、各汉字编码方案的简码字表比较

为了提高汉字输入的速度，减小记忆量，各种编码方案都安排了简码字和简码词。简码字就是把频度最高的 26×N 个汉字分配给小键盘上的每个键位，输入时只需击一次字母键再加一空格键或选字数字键这些高频字就可上屏。简码词就是可以根据词中简码拼写输入的汉语多字词。输入时只需要击单字的声母或部首键再加一空格键或选字数字键即可。简码字、词方法确可简化输入，缩短码长，方便使用。

影响键盘录入速度的因素有键位安排、码长、界面转换、编码规律性、重码率等因素[1]。理论上所有编码方案选择简码字的原则都是高频字优先，但一涉及键位安排则产生了很多问题，因为不论是音托还是形托，都必须照顾独体基根字的键位安排，同时还受汉字集合的使用频度、音节和形态分布特征的制约，故各种方案的简码字选取结果差别较大，实际上削弱了对字频因素的考虑。表 1 列举了几种常见编码方案的首选简码字[2]。

从表 1 中可以看出如下几方面问题①各方案的首选简码字大多不一致，形码方案尤其严重；②简码字的频度差别较大，缺乏严格的选配依据，失去了简码字的意义。此问题音码、形码方案中均存在，以苍颉码、自然码比较突出。如"欧"字、只是因"0"键空着，聊以凑数而已；③简码字的键位安排比较混乱，盲目扩大选字范围。可以说，各方案中简码字、词的选取和键位设置情况差异太大，以致几乎无规律可循。

这种情况从一个侧面反映出各编码方案的设计思想、软件功能水平差别较大，用户在一种技术方案中学到的知识，不适用于其它方案。这很不利于计算机中文信息处理技术的

①本文是在陈一凡教授指导下完成的，并承蒙李公宜、何厚存两位教授允许使用"汉字属性信息数据库"，谨致以衷心的感谢!

普及和提高。本文就简码字、简码词选取依据和键位配置等问题进行了探讨。

表 1　几种常用汉字编码方案的简码字比较

字母键	双拼	自然	普及	声数	郑码	五笔	仓颉
A	这	按				工	日
B	不	从	不	不	一	了	弓
C	从	的	从	次	地	以	女
D	的	二	的	的	现	在	木
E	而	发		分	的	有	水
F	分	个	地	个	世	地	火
G	个	和	和	和	要	一	土
H	和	出		产	在	上	竹
I	是	及	就	级	成	不	戈
J	级	可	看	可	上	是	十
K	可	了	了	了	中	中	大
L	了	没	没	们	是	国	中
M	们	你	你	年	用	同	
N	年	欧			我	民	一
O	欧	片	都	平	他	为	人
P	批	七	去		中	这	心
Q	起	人	人	起	所	我	手
R	人	三	是	人	月	的	口
S	所	他	他	是	多	要	尸
T	他	是	有	他	说	和	甘
U	产	着		这	度	产	山
V	者	我	我	时	为	发	月
W	我	小	小	我	没	人	田
X	学	一	要	学	这	经	金
Y	一	在	在	有	对	主	卜
Z	在			在	了		

说明：• 表示该字频度／构词能力比未达一级简码标准，详见后文。

二、合理选择简码字、词的理论思考

1. 简码字、词输入的几个基本关系

一般而言，简码字或简码词的输入码式可用式（1）、（2）表示，其中 K_s 的作用为重码选定兼隔断上屏。

$$E_{sc} = K_1 K_s \tag{1}$$

$$E_{sw} = K_1 K_2 K_s \tag{2}$$

由于字词频度是简码字、词的主要选择依据，所以式（3）、（4）分别表示文本中简码字和简码词的累计使用频度。其中几个符号的含义分别为：f_{ci}（字使用频度）、f_{wi}（词使用频度）、N_c（简码字总数）和 N_w（简码词总数）。

$$F_c = \sum_{i=1}^{N_c} f_{ci} \tag{3}$$

$$F_w = \sum_{i=1}^{N_w} f_{wi} \tag{4}$$

简码字、词的键位安排是否合理关系到其使用难易程度，尤其是记忆问题。如果硬性安排，则不但使用时大脑负担重，额外增加时间，而且记忆量增加。记忆量如式（5）、(6) 所示。设 M_c 表示简码字的记忆量，M_w 表示简码词的记忆量，则有：

$$M_c = K_c N_c \tag{5}$$

$$M_w = K_w N_w \tag{6}$$

K_c、K_w 为由简码字、词键位安排确定的常数。由式（5）和（6）可知，要真正发挥简码字、词的效能，用户必须记住哪些字、词可以用简码输入以及它们的键位分配。在提示行上选定上屏字、词的时间 t_s 表示为式（7）（即海曼公式），其中 A_s、B_s 分别为与录入者和特征信息元键位分布有关的常数，n 为提示行供选择的字或词数。

$$t_s = A_s + B_s l_n n \tag{7}$$

式（7）表明：提示行的字、词数越多，则单字或词的输入时间也会大大增加。由式 (5) ～ (7) 得出结论：从易记和减少选键时间这两个方面考虑，简码字、词的数量不可太多。根据式（2）～（7），得出 N_c、n_c 和 N_w、n_w 的确定原则：F_c 和 F_w 要大；M_c、M_w 和 t_s 要小。

2. 字的构词能力

简码字的设计目的是为了尽快地输入单字，在可以词输入的时候，则不需要字输入。因此确定简码字时还必须考虑到字的构词能力，以使高构词能力的单字可以词方式输入。构词能力是指一个单字在汉语中可以组词的次数。字的构词能力越高，则作为字方式输入的可能性就越小。

3. 简码字的选定指标

连续文本输入的主导技术是以词为主导的（国标词库）。刘源等的研究表明[3]，连续文本以双字以上词输入的字数比例为 62%（双字词占总字数 53.3% 以上，三字词占 4.46%），余下的 38% 左右需用字方式输入，因此补充技术则是简码字输入方法。如何根据字频和构词能力确定简码字呢？设 S 为简码字选定指数，则有：S∝F（F 为字的使用频度）和 S∝1／E（E 为字的组词能力级），综合起来以式（10）表示就是：

$$S = k\frac{F}{E} \tag{10}$$

在 E＝1，…，6 级划分情况下[4]，令 K＝1(级／%)，则有式（11）。其中 S 为无量纲参数，实际上就是汉字使用频度／构词能力级之比，简称“频构比”。

$$S = \frac{F}{E} \tag{11}$$

44. 简码字的合理数量

根据心理学研究，短时记忆容量为 7±2 个组块单位。在文本输入时，每个汉字都可作为一个记忆组块，其短时记忆容量约为 4 左右[5]。考虑到汉字集合的实际分布情况，每一键位以安排 3 个简码字为宜。综合考虑用户记忆量、指示行选字时间和简码字在输入过程中的实际使用频度等因素后可确定简码字总数为 78 个。

三、简码字的选定与键位分布方案

1. 简码字的甄选

根据上海交通大学计算机系以 GB2312-80 中 6763 个汉字的属性特征建立的"汉字属性信息数据库"[6]的数据，得出了表 2 中的几项统计结果，作为选定 78 个简码字的依据。

表 2　频度／构词级比最高的前 78 个汉字

首选				次选				末选			
字序	字	频度(‰)	频度/构词级	字序	字	频度(‰)	频度/构词级	字序	字	频度(‰)	频度/构词级
1	的	38.342	12.781	27	就	3.383	0.846	53	学	3.421	0.570
2	在	9.441	2.360	28	以	4.204	0.841	54	下	3.410	0.568
3	也	2.348	2.348	29	地	4.888	0.815	55	说	2.720	0.544
4	这	6.392	2.131	30	用	4.730	0.788	56	都	1.627	0.542
5	一	12.512	2.085	31	部	3.099	0.775	57	年	3.225	0.537
6	了	8.187	2.047	32	工	4.551	0.758	58	种	2.675	0.535
7	是	9.782	1.957	33	时	4.504	0.751	59	义	3.187	0.531
8	们	5.223	1.741	34	动	4.453	0.742	60	与	1.575	0.525
9	和	7.469	1.494	35	国	4.449	0.741	61	成	3.133	0.522
10	我	4.114	1.371	36	对	3.627	0.725	62	现	2.064	0.516
11	不	8.046	1.341	37	于	3.546	0.709	63	而	2.533	0.507
12	它	1.328	1.328	38	十	2.022	0.674	64	可	3.036	0.506
13	他	3.960	1.320	39	那	1.334	0.667	65	出	3.000	0.500
14	个	4.970	1.242	40	会	3.861	0.643	66	方	2.918	0.486
15	又	1.193	1.193	41	作	3.856	0.643	67	等	1.931	0.483
16	有	6.927	1.155	42	多	2.510	0.627	68	加	2.404	0.481
17	大	6.857	1.143	43	来	3.751	0.625	69	同	2.869	0.478
18	很	1.076	1.076	44	第	1.870	0.623	70	把	1.874	0.469
19	到	4.064	1.016	45	民	3.085	0.617	71	组	1.402	0.467
20	主	6.095	1.016	46	分	3.679	0.613	72	各	1.365	0.455
21	中	5.957	0.993	47	生	3.664	0.611	73	或	1.359	0.453
22	人	5.836	0.973	48	但	1.197	0.599	74	量	2.222	0.444
23	你	0.926	0.926	49	较	1.195	0.598	75	过	2.646	0.441
24	上	5.496	0.916	50	能	2.926	0.585	76	最	1.321	0.440
25	要	4.499	0.900	51	进	2.874	0.575	77	只	1.262	0.421
26	为	5.260	0.877	52	些	1.713	0.571	78	克	1.194	0.398
累计频度(‰)		156.494				243.662				284.795	

说明：(1) 单字频度和构词能力级数据均以"汉字属性信息数据库"为准，并参考《现代汉语频度词典》[7]。
(3) 简码字按频度／构词级之比降序排列选取。

总的来看，汉字集合中频度和构词能力高度相关，相关系数高达 0.49，$p<0.001$。说

明两者一致性较高，同时也说明在选择简码字时构词能力是一个不可忽却的指标。两者也有不一致的地方，如表 2 中“地”、“也”、“下”三个字的频度相差不多，但根据构词能力排序后，则可将它们区别开来，分别属于首选、次选、末选。使用构词能力这一指标，可将简码字中能以简码词方式输入的字去除，有效地提高了效率。

现在回过头来以频度／构词级这一新指标来评价表 1 中各编码方案的简码字，可发现有很多字是不符合“录取”标准的，如表 1 中打上“＊”号者。

2. 简码字的键位安排

以上述“频度／构词级”指标为依据选择前 78 个字为简码字。采用音托为主、辅以形托的方法，在每个键位上安排 3 个字。前 26 个为一级简码字，如表 3 所列。这些高频单字的构词级均不高，故能将字输入方式与词输入方式区别开来，使单字输入也达到高效、快速的目的。

表 3　本方案选定的简码字及其键位安排

键位	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
首选	音托 形托 记忆	大	不 他	的	一	上	个	和 这	就	很	了	们	你	要	到	中	人	是	它	有	为	我	又	也	在
次选	音托 形托 记忆	年	部	成	地	而	分	工	会	义	较	可	来	民	那	只 都	种	下	时	同	于	多	组	些	以	主
末选	音托 形托 记忆	生	把	出	动	加	方	国	或 过	进	克	量	第	能	各 等	但	最	十 现	与 说 对					学	用	作

说明　表头栏中“音、形、记”三字表示其下的汉字分别以“音托、形托、记忆”方式定位于对应键。
78 个简码字以音托为主（52 个），辅以形托（15 个），少数字需要记忆（11 个）。

这样安排考虑了如下几个方面问题：①记忆量：汉字短时记忆容量约为 4 个左右，从易学、易用角度考虑，每键之字数安排（即提示行的项目数）以不多于 4 个为宜；②汉字集合的实际允许程度：汉字数量多，但在形、音等特征量上分布很不均匀。加上频／构比限制后，为尽可能地使键位安排有规律可循，简码字的数量以每键安排 3 个为宜；③选字上屏时间：提示行字数过多则难以简便快速地输入；④78 序号以后的字频度不高，不到 1‰，再增加对快速输入的意义不大，反而增加记忆量，减小简码字的作用。

每键简码字不超过 3 个，既使一般用户只需记住 26 个一级简码字和简码词若干，不必费多大劲即可快速高效地输入；又为专职录入员提供更快、更高效率条件，他们可以记住次选、末选简码字。所以首选简码字必须记住，其累计使用频度可达 156.49‰；次选、末选的累计频度总和才为 128.30‰，能记住当然更好。去除简码字的累计频度中与构词能力相关的词输入可能性，需全码输入的单字在全部文本中所占比例约为 38–32%。

四、简码双字词的选定

众所周知，双字词是汉语文章中出现次数最多的词语，所以提高双字词输入速度可以大幅度地提高整篇文章的输入速度。双字词简码提取每个字的第一特征，从而减轻用户输入汉字的输入键数。显然，若将每字的声母做为简码统一各种编码方案的双字词输入方法，则可以大大地提高双字词的输入效率。因为，简码双字词输入的码长为 1.5 键／字，比全码输入缩短 0.5～1.0 键／字，即减少了 33%～66%的击键数量。

但是，在标准词库中常用的双字词多达三万余条。若对这样大量双字词全部采用简码输入，则必将产生难以处理的大量重复简码。为此，需要限制简码双字词的数量。兹以下列几方面作为选定简码双字词的主要原则：

1、简码双字词中使用频度最高者不超过 0.66%，即“我们”的使用频度。再如：“中国”的使用频度为 0.30%；“北京”的使用频度为 0.08%。双字词的使用频度可以分为六个等级，即所谓的“频级”。在一万字的连续文本中，双字词所占的字数约为 5300，即双字词的使用次数为 2650。其中仅出现一、二次的词做为简码双字词意义不大。为此，我们选用使用频度大于 0.10%的 120 个常用双字词作为简码双字词，例见表 4（详见附表 1）。

表 4　前 20 个高频简码双字词示例

序号	词语	频度(‰)	频级	LJPD(‰)	序号	词语	频度(‰)	频级	LJPD(‰)
1	我们	0.66	6	0.66	11	就是	0.33	6	4.36
2	可以	0.51	6	1.17	12	问题	0.31	6	4.67
3	他们	0.43	6	1.60	13	国家	0.30	6	4.97
4	进行	0.37	6	1.97	14	中国	0.30	6	5.27
5	没有	0.36	6	2.33	15	革命	0.27	6	5.54
6	工作	0.35	6	2.68	16	这样	0.27	6	5.81
7	人民	0.34	6	3.02	17	不能	0.26	6	6.07
8	生产	0.34	6	3.36	18	自己	0.26	6	6.33
9	这个	0.34	6	3.70	19	所以	0.24	6	6.57
10	发展	0.33	6	4.03	20	由于	0.24	6	6.81

说明：缩写词LJPD为“累计频度”的拼音缩写。

2、选择简码双字词表要考虑用户的记忆能力，即用户需记住哪些双字词才能实现“盲打”。简码双字词数量应使累计使用频度与人的记忆量之间的矛盾取得较好的平衡。

3、重码双字词不应超过三个，这样既有利于用户的记忆，又可以减少重码双字词的选择时间，以实现“盲打”。

4、为提高简码双字词的键盘输入过程实际使用频度，也可以高频的双字代词、介词、连词、副词、中国各省市自治区及直辖市和主要国家国名构成简码双字词表，（见附表 2）。因为这是利用用户已掌握的中文背景信息来确定简码双字词，所以可以将对简码双字词的记忆量减少到最低。由于此法更便于用户记忆，故有利于提高简码双字词的实际使用率。

五、讨论

1. 关于单音输入和部件频度的考虑

在字输入水平上，简码字的选择可以按频度／构词级之比，但如果考虑到简码字当码元，则应结合部件频度考虑优先情况。问题在于简码字的选取目的决定了哪一个优先的原则，这又取决于编码方式以音还是以形为主。前者当以字为首要考虑，后者则首要考虑部件码元的组字次数，以充分提高键盘的利用效率[8]。

2. 简码字特点的分析

前 78 个高频简码字中有 20 个独体字，合理安排在常用键位，同时兼做码元，比较理想。78 个简码字的音位分布特征是：字母 a、i、o、u 和 v 无字，e、p、r 少字，解决办法是尽量利用汉字的发音与字母相近者音托在该键上，而 a 键三个字都要记忆。

3. 关于字频的一些问题

"汉字属性信息数据库"中个别字的情况显然与现时使用情况不符合，反映了某些特定字、词使用频度的时代特征，比如"阶、级、斗、争、彪"字等等，应该去除。另外"改革、开放、公司、股票"等词的使用频度增加，影响字频度，在研究和使用中也会有所反应。如能重新统计反映当前汉语使用特征的确切数据，当可尽量减少这一类影响。

在"汉字属性信息数据库"中有 2901 个字的构词能力数据缺如，其中频度较高（百万字语料中出现次数＞100）的独体字有"勿、日、尹、戊、皿"等。这与该库所引用的数据来源有关[9]。这种情况对本研究也有一定影响。

六、结语

本文讨论了当前几种汉字编码方案中简码字、词的混乱状况，并结合频度和构词能力两项指标，探讨了简码字、词合理选择和键位安排。初步研究表明，以频度／构词能力比率为指标选取简码字、词并指导简码字的键位安排，比单纯以频度为依据更加合理。本文据此初步选定了 78 个简码字和 120 个简码双字词，欢迎中文信息处理系统设计者使用，并提宝贵意见。

参考文献

[1] 何秀全，影响键盘输入速度的原因分析，中文信息，1992，(1)：63

[2] 陈朝、汉字操作系统及文字处理系统使用指南，北京，清华大学出版社，1992.12

[3] 刘源等，汉语处理的基础工程—"现代汉语词频统计"75 攻关项目，中文信息学报，1986，3

[4]《汉字信息字典》编委会编著，汉字信息字典，北京，科学出版社，1988，12，第 9 页。

[5]张武田，彭瑞祥，汉语字词的短时记忆容量，心理学报，1986，(2)：133—139。

[6]上海交大计算机科学系编制。"汉字属性信息数据库"使用手册，上海科技文献出版社，1988.10

[7]北京语言学院语言教学研究所编，现代汉语频度词典，北京语言学院出版社，1986.6

[8]韩布新，汉字部件信息数据库的建立——部件和部件组合频度的统计分析，心理学报，1994，26（2）：147—152.

[9]付兴岭，陈章焕主编，常用构词能力词典，中国人民大学出版社，1982.

The Equitable Arrangement of brevity Coding Characters and Words for Chinese Input

Han Buxin Institute of Psychology, Chinese Academy of Sciences

Ren Xuesong Beijing Information Technology Institute

Abstract

The brevity coding characters lists of several coding systems in common use were quite different. From the respects of number, memory span and arrangement of key position, a new parameter was proposed instead of usage prequency of character, that is the ratio of usage frequency of a character and its grade of forming words. 78 brevity coding characters and 120 brevity coding words combining with double characters were selected according this parameter.

附　录　　附表1. 简码双字词（根据频度选定）

序号	词语	频度(‰)	频级	累计频度(%)	序号	词语	频度(‰)	频级	累计频度(%)
1	我们	0.66	6	0.66	61	它们	0.15	6	14.49
2	可以	0.51	6	1.17	62	影响	0.15	6	14.64
3	他们	0.43	6	1.60	63	产生	0.14	6	14.78
4	进行	0.37	6	1.97	64	而且	0.14	6	14.92
5	没有	0.36	6	2.33	65	技术	0.14	6	15.06
6	工作	0.35	6	2.68	66	例如	0.14	6	15.20
7	人民	0.34	6	3.02	67	生活	0.14	6	15.34
8	生产	0.34	6	3.36	68	时候	0.14	6	15.48
9	这个	0.34	6	3.70	69	许多	0.14	6	15.62
10	发展	0.33	6	4.03	70	以及	0.14	6	15.76
11	就是	0.33	6	4.36	71	增加	0.14	6	15.90
12	问题	0.31	6	4.67	72	政府	0.14	6	16.04
13	国家	0.30	6	4.97	73	重要	0.14	6	16.18
14	中国	0.30	6	5.27	74	比较	0.13	6	16.31
15	革命	0.27	6	5.54	75	表示	0.13	6	16.44
16	这样	0.27	6	5.81	76	采用	0.13	6	16.57
17	不能	0.26	6	6.07	77	活动	0.13	6	16.70
18	自己	0.26	6	6.33	78	具有	0.13	6	16.83
19	所以	0.24	6	6.57	79	开始	0.13	6	16.96
20	由于	0.24	6	6.81	80	利用	0.13	6	17.09
21	这些	0.24	6	7.05	81	其他	0.13	6	17.22
22	因此	0.23	6	7.28	82	使用	0.13	6	17.35
23	情况	0.22	6	7.50	83	变化	0.12	5	17.47
24	如果	0.22	6	7.72	84	代表	0.12	5	17.59
25	什么	0.22	6	7.94	85	地方	0.12	5	17.71
26	一般	0.22	6	8.16	86	地区	0.12	5	17.83
27	作用	0.22	6	8.38	87	规定	0.12	5	17.95
28	必须	0.21	6	8.59	88	经过	0.12	5	18.07
29	方法	0.21	6	8.80	89	实际	0.12	5	18.19
30	社会	0.21	6	9.01	90	世界	0.12	5	18.31
31	因为	0.21	6	9.22	91	温度	0.12	5	18.43
32	主要	0.21	6	9.43	92	一切	0.12	5	18.55
33	当前	0.20	6	9.63	93	知道	0.12	5	18.67
34	经济	0.20	6	9.83	94	出来	0.11	5	18.78
35	要求	0.20	6	10.03	95	存在	0.11	5	18.89
36	不是	0.19	6	10.22	96	得到	0.11	5	19.00
37	但是	0.19	6	10.41	97	工业	0.11	5	19.11
38	起来	0.19	6	10.60	98	美国	0.11	5	19.22
39	条件	0.19	6	10.79	99	农民	0.11	5	19.33
40	为了	0.19	6	10.98	100	认为	0.11	5	19.44
41	研究	0.19	6	11.17	101	完全	0.11	5	19.55
42	一定	0.19	6	11.36	102	物质	0.11	5	19.66
43	根据	0.18	6	11.54	103	只有	0.11	5	19.77
44	需要	0.18	6	11.72	104	注意	0.11	5	19.88
45	发生	0.17	6	11.89	105	作为	0.11	5	19.99
46	过程	0.17	6	12.06	106	并且	0.10	5	20.09
47	同时	0.17	6	12.23	107	成为	0.10	5	20.19
48	我国	0.17	6	12.40	108	出现	0.10	5	20.29
49	一些	0.17	6	12.57	109	工人	0.10	5	20.39
50	已经	0.17	6	12.74	110	基础	0.10	5	20.49
51	方面	0.16	6	12.90	111	决定	0.10	5	20.59
52	提高	0.16	6	13.06	112	企业	0.10	5	20.69
53	通过	0.16	6	13.22	113	全国	0.10	5	20.79
54	现在	0.16	6	13.38	114	人们	0.10	5	20.89
55	以后	0.16	6	13.54	115	速度	0.10	5	20.99
56	第一	0.15	6	13.69	116	提出	0.10	5	21.09
57	对于	0.15	6	13.84	117	现象	0.10	5	21.19
58	时间	0.15	6	13.99	118	形成	0.10	5	21.29
59	它们	0.15	6	14.14	119	应该	0.10	5	21.39
60	影响	0.15	6	14.29	120	有的	0.10	5	21.49

附表2. 加入常用专有名词和虚、实词后的简码双字词表

序号	词语	频度(‰)	频级	累计频度(‰)	序号	词语	频度(‰)	频级	累计频度(‰)
1	我们	0.66	6	0.66	61	以前	0.05	4	8.51
2	他们	0.43	6	1.09	62	以为	0.05	4	8.56
3	没有	0.36	6	1.45	63	英国	0.05	4	8.61
4	这个	0.34	6	1.79	64	有时	0.05	4	8.66
5	中国	0.30	6	2.09	65	正在	0.05	4	8.71
6	这样	0.27	6	2.36	66	只要	0.05	4	8.76
7	自己	0.26	6	2.62	67	非常	0.04	3	8.80
8	所以	0.24	6	2.86	68	如下	0.04	3	8.84
9	这些	0.24	6	3.10	69	仍然	0.04	3	8.88
10	由于	0.24	6	3.34	70	如何	0.04	3	8.92
11	因此	0.23	6	3.57	71	如此	0.04	3	8.96
12	什么	0.22	6	3.79	72	是否	0.04	3	9.00
13	如果	0.22	6	4.01	73	那个	0.04	3	9.04
14	因为	0.21	6	4.22	74	通常	0.04	3	9.08
15	但是	0.19	6	4.41	75	此外	0.04	3	9.12
16	为了	0.19	6	4.60	76	往往	0.04	3	9.16
17	一定	0.19	6	4.79	77	法国	0.04	3	9.20
18	根据	0.18	6	4.97	78	相当	0.04	3	9.24
19	同时	0.17	6	5.14	79	每个	0.04	3	9.28
20	以后	0.16	6	5.30	80	一面	0.04	3	9.32
21	对于	0.15	6	5.45	81	一直	0.04	3	9.36
22	它们	0.15	6	5.60	82	于是	0.04	3	9.40
23	以及	0.14	6	5.74	83	以下	0.04	3	9.44
24	而且	0.14	6	5.88	84	除了	0.04	3	9.48
25	只有	0.11	5	5.99	85	那样	0.04	3	9.52
26	美国	0.11	5	6.10	86	各个	0.04	3	9.56
27	有的	0.10	5	6.20	87	总是	0.04	3	9.60
28	并且	0.10	5	6.30	88	德国	0.03	0	9.63
29	还是	0.09	5	6.39	89	河南	0.02	0	9.65
30	这里	0.09	5	6.48	90	广东	0.02	0	9.67
31	关于	0.09	5	6.57	91	天津	0.02	0	9.69
32	以上	0.09	5	6.66	92	台湾	0.02	0	9.71
33	当时	0.09	5	6.75	93	四川	0.02	0	9.73
34	然后	0.08	4	6.83	94	朝鲜	0.02	0	9.75
35	日本	0.08	4	6.91	95	香港	0.02	0	9.77
36	或者	0.08	4	6.99	96	山东	0.02	0	9.79
37	北京	0.08	4	7.07	97	湖北	0.02	0	9.81
38	虽然	0.08	4	7.15	98	安徽	0.01	0	9.82
39	那么	0.08	4	7.23	99	江苏	0.01	0	9.83
40	最后	0.07	4	7.30	100	澳门	0.01	0	9.84
41	不过	0.07	4	7.37	101	辽宁	0.01	0	9.85
42	你们	0.07	4	7.44	102	甘肃	0.01	0	9.86
43	可是	0.07	4	7.51	103	陕西	0.01	0	9.87
44	不仅	0.07	4	7.58	104	西藏	0.01	0	9.88
45	按照	0.06	4	7.64	105	贵州	0.01	0	9.89
46	上海	0.06	4	7.70	106	河北	0.01	0	9.90
47	甚至	0.06	4	7.76	107	湖南	0.01	0	9.91
48	首先	0.06	4	7.82	108	江西	0.01	0	9.92
49	随着	0.06	4	7.88	109	浙江	0.01	0	9.93
50	只是	0.06	4	7.94	110	福建	0.01	0	9.94
51	不要	0.06	4	8.00	111	云南	0.01	0	9.95
52	怎么	0.06	4	8.06	112	广西	0.01	0	9.96
53	当然	0.05	4	8.11	113	山西	0.01	0	9.97
54	并不	0.05	4	8.16	114	新疆	0.01	0	9.98
55	上述	0.05	4	8.21	115	海南	0.00	0	9.98
56	那些	0.05	4	8.26	116	吉林	0.00	0	9.98
57	同样	0.05	4	8.31	117	韩国	0.00	0	9.98
58	从而	0.05	4	8.36	118	内蒙	0.00	0	9.98
59	经常	0.05	4	8.41	119	宁夏	0.00	0	9.98
60	以来	0.05	4	8.46	120	青海	0.00	0	9.98