The Most Frequent English Words
5000+5000

英语常用词

5000+5000

雷秀云　徐兆昌　编
周国勤　陈庆昌
黄人杰　审　校

上海交通大学出版社

英语语料库最新统计结果

英语常用词

5000＋5000

雷秀云　徐兆昌
周国勤　陈庆昌　编

黄人杰　审校

上海交通大学出版社

图书在版编目(CIP)数据

英语常用词 5000+5000/雷秀云等编. —上海：上海交通大学出版社，1999.4

ISBN 7-313-02217-4

Ⅰ.英… Ⅱ.雷… Ⅲ.英语-词汇-手册 Ⅳ.H313

中国版本图书馆 CIP 数据核字(1999)第 12056 号

英语常用词 5000+5000

雷秀云 徐兆昌
周国勤 陈庆昌 编

上海交通大学出版社出版发行
上海市番禺路 877 号 邮政编码 200030
电话 64281208 传真 64683798
全国新华书店经销
立信会计常熟市印刷联营厂·印刷
开本：700×960(mm)1/32 印张：19 字数：627 千字
版次：1999 年 5 月 第 1 版
印次：1999 年 5 月 第 1 次
ISBN 7-313-02217-4/H·429
定价：24.00 元

前　言

英语常用词汇的确定是英语教学研究的一个重要方面。1982年上海交通大学对100万词的英语语料进行过统计，并根据统计结果编制了《新编科技英语分级词汇》，由上海交通大学出版社出版，受到读者的欢迎。

最近，JDEST语料库在华中理工大学、大连理工大学、哈尔滨工业大学、南通医学院、江汉石油学院和上海交通大学等校共同努力下，已扩大到350万词左右。其中专业的种类也大大增加，已成为初具规模的包括文、理、医、工各专业的学术英语语料库。在对新语料库统计结果的深入分析、研究基础上，以原书为模版，我们编制了《英语常用词5000＋5000》。本书收词10000个，分为10级，每级1000词左右。与原书相比，本书收词数增加了一倍，统计结果也由于语料的增加而更具科学性、稳定性和代表性。

为便于读者使用，词表分为两部分。0～4级为第一部分，5～9级为第二部分。这是因为前5000词出现频率较高，是最常用词，掌握了这些词，可比较顺利地阅读英语文章。后一部分是次常用词，读者可以酌情学习。书中每个词都注有国际音标和简明释义，为读者学习提供方便。另外还有一篇文章介绍语料库的组成、统计的方法和对统计结果的分析等，供读者参考。书后附有统计数据。

本书由雷秀云、徐兆昌、周国勤、陈庆昌合作完成，分工如下：

数据整理、词表编制及注释	雷秀云
计算机编程	徐兆昌、陈庆昌
统计	周国勤

全书由上海交通大学黄人杰教授审校。

《英语常用词 5000＋5000》可供大学生、科研人员、工程技术人员以及其他读者学习英语时使用。对于英语教师和研究人员进行教学、编写教材和研究英语词汇，本书亦可提供可靠的科学依据。

由于学识、经验及时间的限制，书中难免有疏误之处，请专家及广大读者不吝指教。

编　者

1999 年 4 月

使用说明

1. 单词词条用黑正体按字母顺序排列。同形异义词不分列。词组用斜体表示，列在其关键词词条内。

2. 每个单词均注有国际音标。凡发音差异涉及词义或词性不同时，用罗马数字Ⅰ、Ⅱ分列。用斜体标出的音标表示发音时该音可以不发。

3. 单词前面的数字，表示该词常用程度的级别。从0到9，共10级。词表Ⅰ，从0到4，共5级，约5000词；词表Ⅱ，从5到9，共5级，也约5000词。

4. 单词的词性，用斜体英语缩写表示：

vi. 不及物动词	*ad.* 副词
vt. 及物动词	*num.* 数词
v. 动词	*pron.* 代词
aux. v. 助动词	*art.* 冠词
n. 名词	*prep.* 介词
a. 形容词	*conj.* 连词

注 *v.* 时，表示该词既可作及物动词，也可作不及物动词。

5. 单词或词组后均有释义。有多个义项时，用①、②、③排列。

6. 名词中注有[pl.]或[常 pl.]时，表示仅用复数或常用复数。

7. 符号用法

圆括号（ ）：

（1）用以表示对词义的补充，如“(美国的)州”，“充分利用(机会、时间等)”；

（2）归并义项，以节省篇幅，如储藏、储藏量合并为储藏(量)；

（3）表示动词、名词、形容词等的后接关系，如 prevent

(from), suitable (for);

(4) 表示不同的拼写方法，如 prgram(me), favo(u)r。

方括号[ ]:

用以注释词语的用法，如[引导从句]、[表示时间]。

斜线/ :

(1) 表示国际音标;

(2) 在词组中表示斜线前后的词可以互相替代，如 come/go into operation 表示该词组可以是 come into operation 或 go into operation。

分号; :

用以分隔动词的变化形态，如 go (went; gone)。

目　录

英语常用词汇的统计分析

雷秀云

《英语常用词 5000＋5000》建立在对约 350 万词学术英语语料的统计基础上。选择统计结果中选词指数居前的 10000 词作为词汇内容。分为十级，每级 1000 词左右。这一万词是英语的常用词。该词汇可供大学生、科研人员、工程技术人员及其他感兴趣的人学习英语时使用，也为英语教师及词汇研究人员提供重要参考数据。

一、语料的组成

JDEST 语料库始建于 1983 年。当时的规模是 100 万词，共分 2000 个语料单元，每单元平均长度为 536 个词。语料是严格按照随机抽样原理，从十个理工科专业文献中选择的。每个专业约 200 个语料单元，10 万词。所有材料均为英语国家正式出版物。选材时尽量保持了语料的完整性。语料库建成后，先后在 1984 年，1986 年进行过两次词汇统计。统计结果对理工科大学英语教学大纲词汇表的制定，及科技英语特征的研究工作均起过积极作用。

最近，JDEST 语料库在华中理工大学、大连理工大学、哈尔滨工业大学、南通医学院、江汉石油学院和上海交通大学等校共同努力下，增补了文科如：语言学、历史学、社会学等，及医学：如基础医学、临床医学等内容。原有的理工专业数也有增加，如新增了地理学、生物学等。新增部分的语料来源、抽样方法及处理原则仍照原语料库的做法。现在的 JDEST 已有文、理、工、医等 4 大类，33 个专业，超过 350 万词。以后还拟进一步扩建。由于文、医科内容的加入，原来的“科技英语语料库”现已成为初具规模的“学术英语语料库”。表 1、表 2、表 3、表 4 分别从不同角度说明现在的 JDEST 语料库的构成情况：

表 1　JDEST 语料库的语料构成

专　业	语料单元数	总词数	词汇量	最大词长	平均词长
电子	213	101614	5845	27	4.86
机械	198	105336	6067	25	4.75
冶金	194	107258	6404	36	4.83
电工	188	111375	6529	25	4.85
土建	194	100960	5895	31	4.83
造船	204	115644	7077	27	4.72
航空	197	108694	7388	38	4.86
原子能	201	108691	6417	28	4.96
物理	196	113284	7187	32	4.71
化工	206	113853	7825	26	4.97
语言学	200	108443	7693	27	4.79
历史学	200	111049	11412	35	4.65
社会学	200	112670	8427	30	4.99
文学理论	198	107813	10784	40	4.73
教育学	200	110548	7460	34	4.93
经济学	200	104417	8132	26	4.87
地理学	200	120830	11496	26	4.81
心理学	200	107669	8305	29	4.92
生物学	200	105651	9351	35	4.97
天文学	200	132712	8756	26	4.75
航天	194	92730	6739	27	4.88
通讯	199	107156	6537	32	4.82
铸造	194	117282	8026	32	4.87
焊接	182	89278	5629	28	4.80
地质	200	138740	11152	34	5.04
自动化	200	132118	8878	33	4.92
石油	200	152558	7452	31	4.88
环境科学	177	99439	9120	41	5.07
管理科学	189	98866	6912	24	5.02
基础医学	200	118928	10039	30	5.28
预防医学	193	110885	10012	29	5.04
临床医学	199	114846	12032	39	5.06
合　计	6316	3581307	77802	41	4.89

表 2　语料的文体

文　体	语料单元数	%	文　体	语料单元数	%
期刊	1323	20.95	文摘	129	2.04
教科书	964	15.26	手册	129	2.04
专著	1901	30.10	书评	72	1.14
论文	820	12.98	科技新闻	43	0.68
科普文章	328	5.19	其他	607	9.61

表 3　语料作者的国籍

国　籍	语料单元数	%	国　籍	语料单元数	%
美国	3403	53.88	新西兰	9	0.14
英国	1794	28.40	非英语国家	339	5.37
加拿大	93	1.47	国籍不明	644	10.20
澳大利亚	34	0.54			

表 4　语料的年代构成情况

出版年代	50 年代	60 年代	70 年代	80 年代	90 年代	年代不明
语料单元	39	222	2058	2279	1189	529
%	0.62	3.51	32.58	36.08	18.83	8.38

二、统计过程及统计项目

我们对这 350 多万词的语料进行了词汇统计，并制成各种统计词表：如按词频顺序、分布率顺序、选词指数顺序等编制的词表。还对各专业分别进行了统计，以便于对不同专业的英语词汇进行比较。本文将对这次统计结果做一个简单的介绍和分析。

统计时，首先把连续的文本切分成单个的词，然后把词的各种变化形态复原为词的原形(lemmatization)，再求出各种统计量并输出各种词表。统计量有频数、篇章分布数、专业分布数、大类(文、理、工、医)分布数及选词指数等项目。频数是指在语料中某单词出现的总次数，是累计的。篇章分布数是指出现某单词的篇章总数，与词在某篇文章中出现的次数多少无关。同理，专业分布数，大类分布数分别为出现某词的专业、大类的数目。选词指数是编写词表，尤其是常用词词表时选择合适的单词的选词标准。它比频数更能有效地综合反映某个词的常用

程度。当然频数也是选词的重要指标，可以根据词的频数的大小编制频数词表。但是频数相同的词，其常用程度并不一定一致。因为某些词可能因频繁出现于某一篇章或一个专业，造成较高的频数。而其专业分布数及篇章分布数如果不高，则其常用程度会因此而打折扣。如在对 JDEST 的第一次统计时就有 sometimes，policy 和 neutron 的频数同为 269，而 neutron 的分布数远少于另外两个词。显然 neutron 的常用程度也小于另外两词。可见常用程度还与分布数有关系。考虑到频数及分布数两方面的因素，通过试验我们设计了一个能综合反映频数及分布数情况的选词标准的经验公式：

$$I = [a \log F + b \log Dt + c(Ds - 1)^{0.5}] \times 1000‰;$$

其中：I：选词指数；Dt：篇章分布数；F：频数；Ds：专业分布数。

系数 a，b，c 由语料库的大小，篇章数和专业数决定。语料数量变化后需重新统计时，a，b，c 要做相应的调整。上式求得的选词指数在 0～1000 之间。下表列出的是利用该公式计算出的四个频数相同的词各自的选词指数。

表 5　频数相同的 4 个词的选词指数

单词	频数	篇章分布数	专业分布数	大类分布数	选词指数
besides	127	120	31	4	555.05631
wait	127	99	28	4	540.71358
carbonate	127	62	12	3	465.31671
annulus	127	28	5	2	391.34556

四个词虽然频率相同，但其他数据不同。选词指数反映出了每个词的常用程度上的差异。从 besides 到 annulus 选词指数依次降低，说明他们在学术英语中的常用程度也依次下降。这与我们的直觉也是相符的。

三、扩建前后语料库统计结果的比较

1. JDEST 总体及其中文、理、工、医各大类的词汇统计：

统计全部 6316 个语料单元，得出总词数为 3581307，词汇量为 77802。语料中最长词为 41 个字母，平均词长为 4.89。理、工、文、医四大类的统计结果见表 6。

词长与信息密度有关。较长的词与较短的词相比，一般有

表 6　总体统计结果

	专业数	总词数	词汇量	最大词长	平均词长
理	6	670145	27803	35	4.89
工	18	2015980	43807	41	4.87
医	3	344059	21393	39	5.13
文	5	550523	23608	40	4.82
合计	4(大类)	3581307	77802	41	4.89

更具体、更专业化的词义。较短的词使用频率高、义项多、意义不具体。对扩建后的 JDEST 的统计得出的平均词长为 4.89，比 1984 年，1986 年的统计值(分别为 5.113，5.154)短。其原因是统计的标准略有一些变化，前两次统计将字符串中非词符号，如代数符号 x、y、z，顺序符号 i、ii、iii 等全部删除不进入词汇统计。而这次统计没有删除这些符号。这些符号多数只有一两个字母而且在科技文章中出现频率相当高，因此对平均词长有一定影响。从表中可以看出，文科平均词长只有 4.82，是四大类学科中最短的。

2. 1986 年及此次统计出的频率最高的 10 个动词、名词及形容词的比较：

表 7　频率最高的 10 个动词、名词及形容词

动词	频率表次序		名词	频率表次序		形容词	频率表次序	
	1986	1998		1986	1998		1986	1998
use	1	1	system	1	1	new	1	5
make	2	2	time	2	2	large	2	2
show	3	3	result	3	4	high	3	1
form	4	7	process	4	3	small	4	6
work	5	6	problem	5	5	great	5	8
design	6	10	energy	6		different	6	9
give	7	8	material	7	6	present	7	7
increase	8	5	test	8		important	8	10
change	9	9	model	9		possible	9	
control	10	4	effect	10	8	general	10	
1998 年统计中出现的新词					case surface point			low same

从表 7 可以看出尽管语料从 100 万扩大到 350 多万，专业也从 10 个增加到 33 个。两次统计得出的频率最高的十个动词却完全相同，仅在前后顺序上有些变化，说明最常用动词的稳定性很好。名词中 case，surface，point 代替了 1986 年的 energy，test，model，但前 5 个词基本上是稳定的。形容词中 low，same 代替了原来的 possible，general，而且重复的词前后次序变动较大。

3. 1986 年及此次统计结果中选词指数最高的前 500 个词中，功能词、次技术词和技术术语所占比例的比较：

我们对两次统计结果中选词指数最高的前 500 个词中的功能词、次技术词和技术术语所占的比例做了比较，结果见表 8。这里功能词指介词、代词等主要表示语法意义的词，这是一个封闭的单词集(word set)。技术术语指仅在某一或某些技术领域内使用的专业词语，不包括那些随着科学技术进步已为大众熟知，并已进入词汇共核的词语。那些界于功能词与技术术语之间的词统称为次技术词。从表 8 中可以看出有功能词减少，而次技术词增加的趋势。考虑到语料增加了约两倍，这里的变化并不能说很大。说明在语料增加的情况下功能词的使用频率相当稳定，尤其是前 50 词中的功能词。技术术语仍旧没有能进入前 500 词。这可能是因为技术术语常常用于特定的领域，分布率较低，因而选词指数不会太高。

表 8　选词指数最高的 500 个词中功能词、次技术词和技术术语所占的比例

	1986		1998		1986		1998		1986		1998	
词数	功能词	%	功能词	%	次技术词	%	次技术词	%	技术词	%	技术词	%
前 20	20	100	20	100	0	0	0	0	0	0	0	0
前 50	49	98	46	92	1	2	4	8	0	0	0	0
前 100	77	77	67	67	23	23	33	23	0	0	0	0
前 200	102	51	89	44.5	98	49	111	55.5	0	0	0	0
前 500	130	26	117	23.4	370	74	383	76.6	0	0	0	0

4. 两次统计的前 5000 词的对比：

表 9 为两次统计选词指数居前的 5000 词中，相同词的数目。

表 9 1986 和 1998 年统计选词指数居前的 5000 词中相同词的数目比较

1998 年统计结果	与 1986 年 5000 词相同词数	累计词数
1—1000	991	991
1001—2000	984	1975
2001—3000	905	2880
3001—4000	802	3682
4001—5000	683	4365

从表中可以看出 1998 年统计的前 5000 词中，每千词与 1986 年统计前 5000 词相比，相同词数呈递减趋势，这表明越后面的词越不稳定。扩建前后两库相同的词有 4365 个，占 87.3%。说明语料在 100 万词级时，前 5000 词的统计结果已相当稳定，而这次统计的稳定和可靠程度将会更高。

5. 前一次统计时已发现，各专业频率最高的前几个实词很有专业代表性，从这次统计可以看出，其他专业同样有此规律。下表为文、理、医、工每大类各三个专业中频率最高的前五个实词一览表：

表 10 各专业频率最高的五个实词

专 业	实 词
计算机	system, computer, program, datum, machine
机械	machine, tool, bearing, work, system
电工	system, current, motor, power, voltage
语言学	language, use, make, word, example
历史学	war, year, make, time, state
教育学	school, education, teacher, student, use
基础医学	patient, cell, study, use, disease
社会医学	study, use, case, age, disease
临床医学	cell, use, increase, result, study
天文学	star, time, energy, earth, field
生物学	cell, use, protein, DNA, plant
物理学	energy, wave, force, light, time

语料库应该多大才合适，一直没有定论。大家似乎已达成了多多益善的共识。但是语料是无限的，语料库再大，也不可能包含所有某一真实语言的总体，终归存在抽样问题。尽管电脑的存储及处理文本的能力越来越大，毕竟也有一个极限。而且对语言或其他任何无限总体采取抽样方法进行研究时，原则上都是希望样本容量尽量小而其代表性能尽量高，这样，得出的结论会有概括性、普遍性。从 JDEST 两次统计结果的简单比较可以发现，只要抽样方法正确，语料样本较大时，统计结果会有很多共同之处，当然同时也将有一些差异。样本容量增大，统计精确度显然会有一定程度的提高。不过两者不可能是正比例关系。

四、统计结果分析

我们知道在用词上学术英语与普通英语有些差异。但是究竟有什么不同，却不太容易说清楚。为此我们把 JDEST 的统计结果与普通英语语料库 LOB 的统计结果做了粗略的对比。

1. 通过对 JDEST 和 LOB 中频率最高的前 100 个词进行对比分析，我们发现：

(1) LOB 语料库前 50 个词中没有实义词，前 100 词中有 9 个，它们是 say，time，make，like，new，man，year，many，people；而 JDEST 前 100 词中有 39 个实义词，其中前面 10 个为：use，system，time，make，process，result，show，control，increase，work。从词义上看，JDEST 中的这些词都是学术英语中的次技术词。

(2) 人称和物主代词的数目，LOB 为 16 个。它们是：it，he，I，his，her，she，you，they，we，their，him，my，them，me，its，our；JDEST 为 8 个，分别为：it，they，their，its，we，he，his，I。

LOB 的 16 个词中有 her，she，you，him，my，them，me，our，而 JDEST 前 100 词中没出现这些词。学术英语一般强调客观性，较少个人色彩。因此这些代词在学术英语中频率较低是可以理解的。

(3) LOB 的前 100 词中有 13 个介词，JDEST 有 15 个，LOB 中除 over，JDEST 中除 from，between，through 外，其他 12 个是共同的。

表 11　16 个代词在 LOB，JDEST 中的前后排列位置

代词	LOB	JDEST	代词	LOB	JDEST
it	1	1	we	9	3
he	2	6	their	10	4
I	3	8	him	11	13
his	4	7	my	12	12
her	5	11	them	13	16
she	6	14	me	14	15
you	7	10	its	15	5
they	8	2	our	16	9

(4) 在前 100 个词中助动词和情态动词分别为 LOB 中 10 个；JDEST 中 7 个。LOB 的前 100 词中有 can，could，may，而 JDEST 中没有。这种表示“可以、会”的情态动词在学术英语中频率低一点也是与学术英语的特点吻合的。

(5) 在前 100 个词中关系代词 LOB 中有 that，which，as，when，who；JDEST 中有 that，as，which，when。LOB 中的 who 在 JDEST 前 100 词中没出现。

2. LOB 与 JDEST 中以字母 A 开头的部分词之间的比较：

我们还从两库中选取 abacus 及 adaptation 之间的以 A 开头的词进行了对比，发现有以下一些现象。比较时，排除了缩略词、人名、地名以及不可比因素。

(1) 两库各有一些特有的词。如 LOB 中有 abashed，abcess，aber，abhominable，abject，abondance，absentminded，absinthe，absinthium，absit 等 25 个词 JDEST 中没有。其中 9 个频数在 2 次以上，其余均为一次，即在普通英语中也只是偶然出现。JDEST 中有 abase，abaxial，abetalipoproteinemia，abietic，abiogenically，abiotropy，ablate，ablaze，abnegate，aboral 等 86 个词 LOB 中没有。他们几乎都是技术术语。其中 46 个频数在两次以上，其余均为一次。即大致一半词即使在学术英语中也只是偶然出现。两库共有的词前十个为：abacus，abandon，abate，abbey，abbot，abbreviate，abdicate，abdomen，abduct，aberrancy。这些词大部分在《新英汉词典》中都可以查得到，而两库独有的