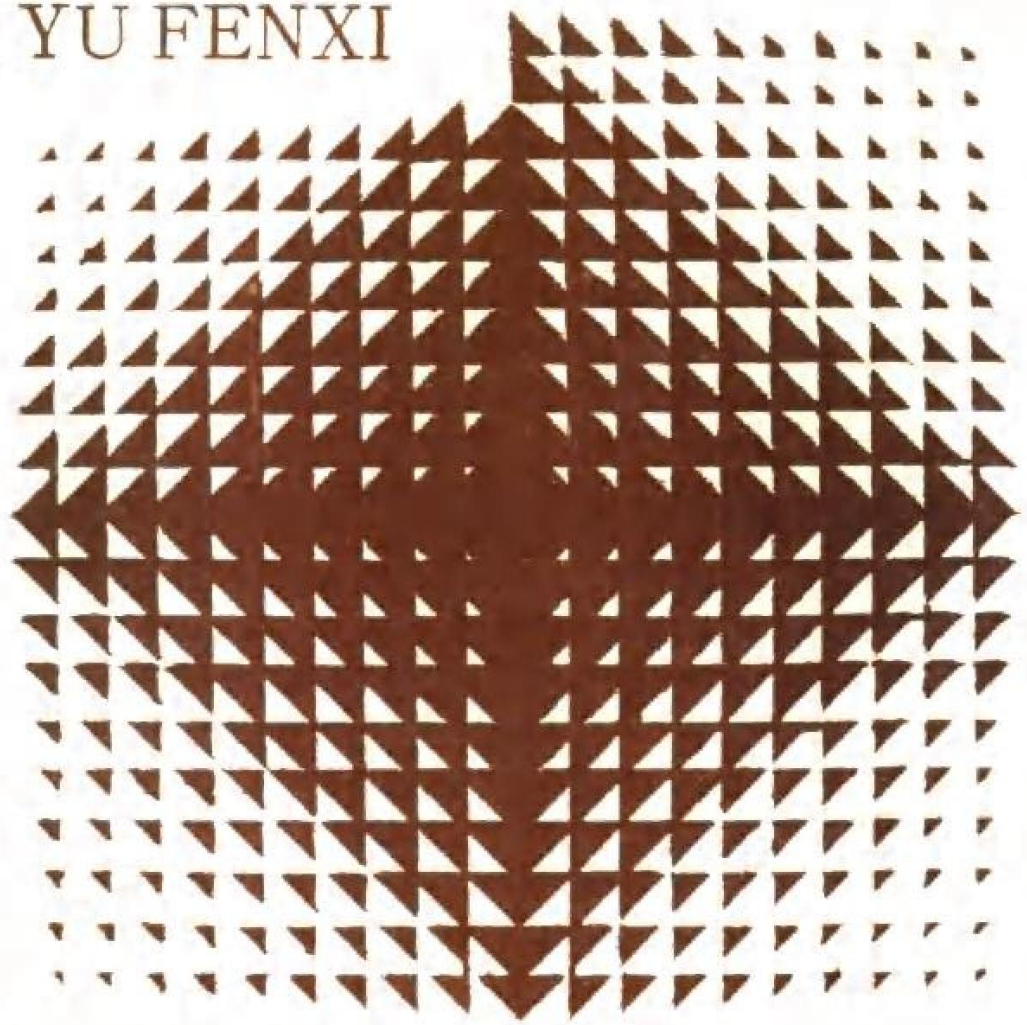
汉语词汇的统计与分析

HANYU CIHUI DE TONGJI YU FENXI

北京语言学院
语言教学研究所



外语教学与研究出版社

汉语词汇的统计与分析

北京语言学院语言教学研究所

语言教学与研究出版社

汉语词汇的统计与分析

北京语言学院语言教学研究所编

外语教学与研究出版社出版

（北京外国语学院23号信箱）

北京昊海印刷术咨询公司印刷

新华书店北京发行所发行

全国各地新华书店经售

开本 787×1092 1/16 68.25 印张 1,133 千字

1985年 4 月第 1 版 1985年 4 月北京第一次印刷

印数 1-7000 册

书号：7215·89 定价：12.50元

前　　言

在汉语教学中，特别是在汉语作为第二语言教学的基础阶段，究竟应该选择哪些词语、汉字最先教给学生？现代汉语中哪些词是最常用的？哪些是次常用的？我们的教师在设计教学大纲、编写教材时对这些问题往往多凭主观经验。因缺乏科学根据，所以众说纷纭、莫衷一是。汉语的最低限度词汇量有多大？一个外国留学生至少要掌握多少词汇，才可以同一个中国的高中毕业生大致相当，能够适应在大学听课、进行专业讨论和书面阅读的需要？这些更是长期没有解决的难题，也是我国乃至全世界汉语教学界共同面临的、亟待解决的问题。

为了在科学的基础上选择和确定现代汉语常用词语，以避免词汇教学的主观盲目性，提高教学效率，保证教材、辅助读物和工具书的质量；同时，也为了给从事汉语语音、文字、词汇等各方面研究的语文工作者以及从事中文信息传输、文献检索、机器翻译、人工智能等多门学科研究的专家和工作人员提供一批有价值的基础数据资料，北京语言学院语言教学研究所于一九七九年把“现代汉语词汇统计研究”列为重点科研项目，决定采用抽样统计的方法，对汉语词汇的使用情况作一次比较全面的调查。从一九八〇年开始，我们动员了相当数量有经验的教师，按照统一规定的原则和要求，对不同

体裁和内容的200万汉字语料进行了切分和统计，并与中国社会科学院语言研究所合作，借助电子计算机完成运算和排序。全国中小学通用教材语文编写组1978—1980年编写的十年制《语文》课本（试用本），就是这200万字语料中的一部分。

作为语料的精心编选的全国中小学统一使用的语文教材，无论在文字上，还是在内容上都有一定代表性，同时也具有一定的完整性和系统性，因此我们有意识地对这部分语料做了专门的统计和分析。从统计结果中，不仅可以得到关于汉语单词的平均长度、语料的平均词长、基本词汇的大致数量及其复盖率、汉字的构词能力等基本数据，也可以了解到这套课本中出现的全部词语、它们各自的使用频率和在不同年级的分布情况等，由此可以推知我国受过初等或中等教育的毕业生在掌握汉语语文词汇方面的大致水平，为制定不同类型的（国内外）汉语教学大纲、编写不同程度的教材提供重要的选词依据；同时，又可以通过与200万字语料综合统计数据相对比，对这套教材选词的科学性做出客观评价，以进一步改进我国的中小学语文教学。我们相信，这批资料在科研和教学上的作用决不仅限于此。我们发表它的目的，正是要提供给广大语文工作者和有关的科研人员，根据各自的需要，进一步开发利用，使历时数年的劳动成果创造更多的价值。

语言单位的计量分析是语言研究的一项基本建设，在理论和实践上都具有很重要的意义，在国际上，近几十年来数理语言学及其所属的重要分支——词汇统计学有了巨大的发展。可惜的是，由

于种种原因，以往的大规模汉语统计工作还只停留在以字为单位的阶段，而词汇或者词语单位的计量研究却长期无人问津。为了尽快填补这项空白，我们不得不花费较长的时间边实践、边探索，力求在这个领域摸索出一条新路。但是由于可资借鉴的前人经验不多，加之我们的水平有限，有些问题至今不敢说解决得全然合理，如词语单位的确定和切分就是个很棘手的问题，为了统一起见，我们不得不按照语言教学的实际需要做了某些硬性规定。《语文》课本中的绝大多数课文应该说是比较规范的作品，但细加切分，仍然可以发现极少量误写或生造的词语或汉字。对这些，我们基本上采取客观描述的方法，作了如实反映。尽管我们对这批资料反复进行了认真细致的查对、校勘，恐怕错误仍然在所难免，诚恳希望国内外汉语教学界的专家、同行和其他人士热忱指正，以利改进和提高。

这项工作得到了中国社会科学院语言研究所名誉所长吕叔湘先生、北京大学中文系朱德熙教授的支持和指教。中国社会科学院语言研究所副所长刘涌泉同志在语言统计学理论、汉字编码以及联系使用电子计算机和人力安排上给予了多方面的帮助。语言研究所林联合同志承担了机上运算的全部程序设计和实际操作工作。这项工作还得到了国家科委信息管理研究所的大力支援。信息所科研处和机房工作人员给予了我们无偿使用该所ND—500型计算机的优待，并且在机时安排上也给予了诸多照顾，至于参加统计工作的北京语言学院各部、系教师，更为之付出了大量辛勤的劳动。我们谨向关心、指导、支援和具体协助这项工作的单位和个人致以诚挚的谢意。

参加本项科研工作的人员有：王　还（以下按姓氏音序排列）

常宝儒　　李宜生　　林联合　　刘　杰　　孙印禄　　王正娃　　余云霞

编　　者

一九八四年九月

体例说明

一、语料及统计范围

本书采用中小学通用教材语文编写组编写，由人民教育出版社1978—1980年出版的全日制十年制《语文》课本（试用本）的课文作为现代汉语词语统计、分析的语料。这套课本共分二十册，其中：小学十册，初中六册，高中四册。

为保证统计资料的典型性，以下三种课文未予统计：

（1）诗歌韵文，包括课文中出现或引用的诗词、歌谣等。

（2）古代汉语，包括白话文文章中引用的成段文言文句子和段落。

（3）外国作品的翻译文章。

语料中出现的地名、人名等专名词、阿拉伯数码、外文字母、标点以及其它专用符号一律未作统计，秒语也未统计。

全部语料中的句子经过切分和逐一登记，共得出不同的词语单位18,177个，出现总词次为374,654，约等于52万个汉字。

二、词语单位的确定和切分

现代汉语词频统计工作最大的困难，莫过于词与词之间在书面上没有明显的界限；另一方面，现代汉语语素同词、词同词组之间界限的划分以及词的归类这些问题在理论上和实践上都尚未妥善解决，因此，要进行这项工作，首先必须制定出确定统计单位的原则和标准，把语料中的句子切分成一个一个单位。

出于教学和科研上的需要，我们在试点的基础上经过反复研究并征询校内外专家的意见后，拟出几项词语切分的原则规定，作为统计全部语料词频的统一标准，以便使较多的人参加工作时有共同遵守的准绳。我们的基本着眼点是，从实际出发，即针对外国人学汉语的特点进行切分，不在词的定义上纠缠，所切分的最小单位可能等于一个词，也可能大于一个词；注意结构成分在口语中能否单用，组合后是否等于成分意义的总合，意义有无显著改变，字序变动后是否能保留原义，也适当考虑长度因素。除结合紧密的成语或习用词组外，四音节以上单位尽量分成较小单位，如果切分影响原义，则不分。科学术语慎重对待，一般不硬性化小；某些常用的

前、后缀或类词缀也单独分出统计。

从上述原则出发，我们在确定词语统计单位的时候采取了意义同用法兼顾，侧重于用法，即侧重语境中的语法功能的作法，这种作法特别表现在同音同形词和多义词的处理上。

对于一般所说的多义词，即在普通汉语词典上列出不止一个义项的词目，我们主要依照它在具体语境中的用法，从词类归属上加以划分，分别立条目进行统计，一般不考虑同一类属或同一用法中不同意义的差别，如：极（名）、极（副），有些兼类的词在行文中难以确定或区分，为了统计的方便，把两类合并在一起，如介词“和”与连词“和”就归在一起统计。词汇意义单一、相同或相近，而语法功能或用法上难以区别的，一般确定为一个单位，不标注词性，如“发明机器”和“这是一项伟大的发明”，这两个“发明”就作为同一个词的不同用法归为同一个词汇单位，词汇意义明显不同，词类也明显不同的，分别立目，用标音、标注词类、简要注释或举例的方式加以区分，如：白（形）、白（副）；清明、清明〈节气〉；精神 jīngshén、精神 jīng·shen。

同样，在词次的计算方法上也贯彻了以用法为主的原则。一般来说，某个词语单位在书面语料中出现几次，就等于使用了几次，但也有例外情况，如在实际语料中，会出现汉语中特有的“看看”、“想一想”、“高高兴兴”等动词、形容词的重叠形式。这一类重叠形式实际上是一种语法或修辞现象，因此不能因为“看”、“想”、“高兴”在字面上出现了两次就分别按各出现两词次来统计，而应当按这三个词实际只使用了一次来计算（第二例中的“一”要另外统计），这样才能保证某个单位的“词次”确实代表着它在这批语料中的实际使用频率。

如前所述，虽然本书是词汇统计资料，但所切分的统计单位不全是严格意义上的词，因此称为“词语单位”似乎更合适一些。在词表和说明材料中，为了行文方便一般简称为词，并由此产生其它术语“词条”、“高频词”、“双音节词”、“多音节词”等。这一点提请读者注意。

制定词语切分原则时，我们较多地参考了中国科学院语言研究所编的《现代汉语词典》确定条目的原则；在实际工作中吸收了吕叔湘先生在《汉语语法分析问题》中对划分语素、词、词组等问题的看法。

三、词表编排体例

本书的中心部分是以下两个频率词表：

表Ⅰ　按音序排列并注明年级分布的频率词表

表Ⅱ　按频率排列的词表（一）、（二）

现将两个词表的编排特点和各个栏目的内容介绍如下：

（1）表Ⅰ

表Ⅰ列入了语料中所出现的全部词和这些词使用情况的有关数据。词不论单音（单字）或多音（多字），一律作为条目用拼音和汉字两种形式书写，按汉语拼音顺序排列。轻声字，单独成词的按第五声处理；不在词首的，一般排在同形的非轻声字后面。读者可以借助表中标出的首字音节迅速找到所要查找的条目。

数据部分的栏目自左至右可以分为两组：

第一组包括——

词次：该词在语料中出现或使用的次数。

频率：该词的词次占全部词次总数的百分比。

年级分布：说明该词的词次分布在十个年级中的几个年级课本中。如果十个年级的课本都有这个词，那么它的年级分布即为10；若只出现在一个年级的课本里，分布即为1，其它依此类推。这个数字可以从一个侧面反映词语单位的常用程度。

第二组，反映出某词在小学和中学不同阶段各年级课本中分别出现或使用的次数、小学和中学两阶段的累积词次和频率。

利用表Ⅰ，读者可以根据教学和科研工作的需要了解、考察某个词的使用情况，例如可以对同义词、反义词、词缀等各种类型的词汇聚合现象进行专门的比较分析，或者考查属于不同词类的同形同音词各自的使用频率，也可以根据不同程度、不同年级的词语出现情况编写相应的教材和教学辅助材料。这个词表也可以为教师和研究者查找词语出处和例句提供方便。

（2）表Ⅱ

表Ⅱ是全部词按频率高低顺序排列的词表。词次相同的，按年级分布数的高低顺序排列，词次和年级分布都相同的同级词，按拼音字母顺序排列。本表分为（一）、（二）两部分。

第一部分包括词次最高的单位（“·de的”，20,184）直到出现4次的单位，共7,219条；第二部分收入的是其余出现3次以下的单位10,958条，为节省篇幅，按出现3、2、1次的顺序集中罗列，只注明是哪些词，不再分别标示每个词各自的数据。

表Ⅱ除反映出表Ⅰ中每个词的词次、频率、年级分布等项基本数据外，还增加了一些栏目。

条目部分左起第一栏序号，表示词在这个词表中排列的顺序位置，便于读者考查高频词的累计词数和累计词次、频率之间的对应关系，同时也可以一目了然地看出某词在词表中排列在第几位。

数据部分新增加的项目有：

词次级别：本表将全部18,177个词语单位按不同词次分成388级。词次相等的算做同级，词次最高的（20,184次）为第1级，只有一个；最低的（1次）依次排为第388级，共6730个。词次级别反映出某词语在使用词次上所占的地位。

累计词次：表示前面所列词的词次累加数。

累计频率：累计词次占语料总词次的百分比。这个数字通常可以看作是前面所列的所有词在语料中的复盖率。如表中显示，频率最高的前10个词累计频率为17.6763%，说明在语料里，仅这10个词的复盖率就已经达到了近17.7%；同样，前180个词已经占全部语料篇幅的一半以上（50.8%）。（参见表Ⅶ）。

使用度：某词的使用度是由词次和年级分布两个相关因数决定的。这个数值大致体现出一个词的常用程度。即是说，只考虑词次高低，并不能完全反映出该词的常用性，还要看这些词次在10个年级的课本中的分布面广不广。假如两个词次同为100的同级词，一个分布在全部十个年级课本中，另一个只分布在一个年级里，那么使用度就极为悬殊。反之，在年级分布数值相等的情况下，词次越高，当然常用程度也就越高。

使用度的计算公式是：

某词使用度＝该词词次×年级分布／10

使用表Ⅱ，有助于从词次、年级分布和使用度的高低比较上来选择和评价常用词、次常用词和一般词等。

四、条目书写格式

拼音栏

（1）全部词语条目均采用汉语拼音字母注音，除轻声字和某些象声词外，一般只注原来的读音和声调，不注在具体语流中的变化。

（2）大写、轻声和儿化的拼音标写法参照《现代汉语词典》。

（3）双音词语拼音连写，但极常见的、结合紧密的“动宾”和“动补”性的组合，中间如能插入其它成分，音节之间加注〃号，如“看见”注作kàn〃·jiàn，“洗澡”注作xǐ〃zǎo；少数用在动词或形容词后的双音趋向补语一律不加〃，如“起来”（补）注qǐlái，“下去”（补）注xiàqù。

（4）由一个单音节词和一个双音节词组成的动宾结构，两部分拼音分写，如拜天地：bài tiāndì，打天下：dǎ tiānxià，除此以外，三音以上的多音词为了节约篇幅一律连写，如不相干：bùxiānggān，不是滋味儿：bùshìziwèir。

汉字栏

（1） 绝大多数词条都不标注词性、释义等。标明词性，只是为了区分同形同音的不同词。如bái白（形）／bái白（副）、bǐ笔（名）／bǐ笔（量）。同形的词，如在注音上能区别开的，一般也不标注词性，如dìdào 地道／dì·dao 地道；huì〃kuǎn 汇款／

huikuǎn汇款。在表Ⅱ中，这一点需要特别引起注意。

（2）（　）是标注词性的符号，少数包括两种词性的条目，词性同注在一个括号里，中间用顿号（、）隔开，如gēn跟（介、连）；bù部（名、尾）。

这里所谓词性，既包括名词、动词等词类，也包括补语、词尾等一些语法功能特征。简写形式如下：

（名）	名词	（动）	动词	（形）	形容词
（副）	副词	（介）	介词	（量）	量词
（助）	助词	（叹）	叹词	（数）	数词
（代）	代词	（象声）	象声词	（助动）	助动词
（补）	补语	（头）	词头	（尾）	词尾

（3）〔　〕用于表示同一词语单位的不同写法或异体字，如hán•hu含糊〔含乎〕、dàngzuò当做〔当作〕。

（4）〈　〉用于词义的补充解释。如黄元帅〈苹果〉、不想〈不料〉；个别条目亦用〈　〉表示学科、语体等类属，如喷丝头〈机〉。缩语形式参见《现代汉语词典》。

（5）脱离语境不易理解确切或用简要文字难以解释清楚的词，举简单的例子加以说明。例子中该单位用～号代替，如diǎnr点儿吃～饭；快～、fēnzhī 分之　三～二。

（6）除条目本身外，其它一切注释性文字均使用楷体印刷以示区别。

（7）书面语中必须儿化的，汉字一般加“儿”（拼音写作r）。可儿化可不儿化的，一般不加“儿”。但是为尽量忠实原文，少数在语料中只出现一种形式的词语在词表中仍然保持原状。

五、关于附录的说明

表Ⅲ—Ⅶ是从汉字构词、音节、年级分布、复盖率等不同角度对表Ⅰ和表Ⅱ所做的进一步统计和分析，兹分别说明如下：

1．表Ⅲ　前1000个高频词的汉字组词能力分析

表Ⅱ中的前1000个词共出现278,448次，占全部语料词次总数的71%强。了解这些词是由哪些汉字组成的，这些汉字在单字条目或在多字条目的词首、词间和词末的构词能力如何，不论对汉语字、词教学，还是对信息传输工程中编制机器词典等都有很重要意义。

表Ⅲ按频率高低顺序，列出了构成前1000个高频词（出现在 50 次以上）的共731个不同的汉字。所列汉字以字形为标准，如长cháng、zhǎng仅作为一个单位。表中还标明每个字在不同位置上组合成不同词的数目以及这些汉字出现的次数。

例如，前10个频率最高的汉字依序是：的、一、了、我、是、在、不、们、人、有。它们在前 1000 个词中，除以单纯词形式出现外，还组成60个复合词（包括它们之间相互组合以及同其它字的组合），这 10 个汉字共出现 77,782 次，占全部语料 520,934 字的 14.93125%。前100个汉字，在这 1000 个词中共出现 214,099 次，占全部语料总字数的 41.099 %。731 个字在前 1000 词中共出现 332,884次，占全部语料总字数的63.90176%。此外，它们还要在其它词中出现，所以，可以这样说，这731个汉字占全部汉字语料总数百分率远远超过63%。它们的常用程度和重要性是显而易见的。

在本表中有些术语需要说明：

总计组合数，指这个汉字按照我们的词条标准在前1000个高频词中构成的词数，包括单音节词和多音节词。比如“有”在前1000词中构成的词有：“有”、“没有”(副)、“所有”、“有些”(副)、“有些”(部分)、“有时”、“有的”、“只有”八个，“有”的总计组合数就是 8 。“占有”也是“有”的组合，但它不在1000高频词内，所以不算在内。

单音词下的组合数，指这个汉字以单音形式在1000词中构成的词的数目，如“把”构成的单音词有“把”（介）、“把”(量)两个，组合数就是2。

多音词下的组合数，指这个汉字在1000词中的词首、词间，词末的不同位置上和别的汉字组成的词的数目，如“地”组成的双音词有“地方”、“地球”、“地面”、“阵地”、“土地”、“草地”六个，所以它的多音节词的组合数是6，词首组合数是3，词末组合数是3。

次数，指每个汉字在组合词中出现的总频次。

另有一种格式，如“一…就…”、“连…都…”，因前1000高频词中只有四组这类条目，不另列专栏，也算在多音词中统计。如“一…就…”格式中的“一”归入词首，“就”算在词末统计。

表Ⅲ有两个附表：(一)、(二)，其内容详见书后的附录。

2。表Ⅳ 词的年级分布和音节构成统计

它显示出分布在一到十个年级课文中的不同词的数目，占总数的百分比以及这些词中单音节、双音节、三音节、四音节、五个音节以上的词各有多少。可以看到一种有趣的现象：随着年级分布数的减少，不同词的数量大体呈递增的趋势，而且单音节词数和双音节词数的比例也在显示出有规律的变化，如分布在十个年级课本中的单音节词401个，双音节词167个，共568个，无疑它们应属于最常用词的，其中单音节词和双音节词的比例为2.4:1，这说明在最常用词中，单音节词还保持着优势。随着年级分布的减少，双音节词的数量迅速超过单音节词，而且差别越来越大。出现在四个年级中的1,145个词中，单音节词216个，双音节词882个，比例为1:4；而仅出现在一个年级中的8,616个单位中，单音节词634个，双音节词6,684个，比例为1:10.5，体现出双音节词占绝对优势的倾向。

3．表Ⅴ 各词次级别和同级单位的数量统计

词次级别的概念，请参看第三节表Ⅱ的有关说明。

表Ⅴ列出了全部388个词级中各级的词次和每级的同级词条数。它同表Ⅳ一样，反映出同级单位中不同音节的单位各有多少。这个表提供的数据除了象表Ⅳ那样，可用于常用词音节构成多方面的研究以外，对验证词汇统计学中的某些定理、公式（如著名的“齐夫定律”）也有重要的意义。

4．表Ⅵ 不同音节词的数量及其复盖率对比

本表列出了单音节、双音节、三音节、四音节、五音节以上词的数量、所占百分比和它们各自的词次累计、频率累计数，为解释词汇平均词长和语料平均词长的差别提供了有说服力的依据。

5．表Ⅶ 前若干词在语料中的复盖率

表Ⅶ以简略形式抽样展示了不同数量的高频词的累计词次在语料中的复盖率，对确定不同等级常用词的数量、范围有一定的参考价值。

六、若干基本数字

（1）**不同的词的总数** 18,177个

（2）**词次总量** 374,654， 其中：小学课本部分137,469；中学课本部分237,185

（3）**汉字总数** 520,934个，其中：小学课本部分194,209个；中学课本部分326,725个

（4）**平均词长** 词表条目平均词长1.98个音节或汉字；全部语料平均词长1.39个音节或汉字。其中：小学课本语料平均词长1.41个音节或汉字；中学课本语料平均词长1.38个音节或汉字

（5）**平均词次** 不同的词在全部语料中的平均词次为374,654÷18,177＝20.61143（词次）；高于平均词次的词共2,261个，占总数18,177的12.4387%

（6）前1000个高频词由731个汉字组成，它们在这1000个词中共出现332,884次，占全部语料总字数的63.901%

目　　录

词语单位		总计			小学阶段各年级词次分布							中学阶段各年级词次分布						
拼音	汉字	词次	频率	分布	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	累计	频率	Ⅵ	Ⅶ	Ⅷ	Ⅸ	Ⅹ	累计	频率
A																		
a																		
ā	阿（头）	94	.02509%	4				2		2	.00145%		30		12	50	92	.03879%
āgōng	阿公	1	.00027%	1						0	——	1					1	.00042%
āpó	阿婆	6	.00160%	2		5				5	.00364%	1					1	.00042%
āyí	阿姨	21	.00560%	5	4	6	1		9	20	.01455%				1		1	.00042%
á	啊〔啊〕	87	.02322%	9	2		18	16	20	56	.04073%	6	2	3	12	8	31	.01307%
aya	啊呀	2	.00053%	2					1	1	.00073%				1		1	.00042%
·a	啊	304	.08114%	10	28	29	31	20	50	158	.11493%	14	32	26	68	6	146	.06155%
ai																		
āi	哎	20	.00534%	7		1	1	4	3	9	.00655%		2	4	5		11	.00464%
āiyā	哎呀	20	.00534%	5	1	1		4	4	10	.00727%				10		10	.00422%
āiyō	哎哟	6	.00160%	4			1	1	2	4	.00291%	2					2	.00084%
āichóu	哀愁	1	.00027%	1						0	——					1	1	.00042%
āidào	哀悼	1	.00027%	1						0	——				1		1	.00042%
āiháo	哀号	1	.00027%	1			1			1	.00073%						0	——
āiqiú	哀求	4	.00107%	2						0	——		1	3			4	.00169%
āishēngtànqì	哀声叹气	1	.00027%	1				1		1	.00073%						0	——
āisī	哀思	1	.00027%	1					1	1	.00073%						0	——
āitòng	哀痛	3	.00080%	2						0	——			1	2		3	.00126%
āiyuè	哀乐	2	.00053%	2						0	——		1		1		2	.00084%
āi	挨	34	.00907%	9		1	2	9	5	17	.01237%	3	4	6	3	1	17	.00717%
āi'āijījī	挨挨挤挤	1	.00027%	1			1			1	.00073%						0	——
āigèr	挨个儿	4	.00107%	3				2		2	.00145%		1		1		2	.00084%
āijìn	挨近	1	.00027%	1						0	——	1					1	.00042%
āiyā	唉呀	2	.00053%	1						0	——				2		2	.00084%
ái'ái	皑皑	1	.00027%	1						0	——					1	1	.00042%

词语单位		总计			小学阶段各年级词次分布							中学阶段各年级词次分布						
拼音	汉字	词次	频率	分布	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	累计	频率	Ⅵ	Ⅶ	Ⅷ	Ⅸ	Ⅹ	累计	频率
ái	癌	4	.00107%	2						0	——		1		3		4	.00169%
áizhèng	癌症	1	.00027%	1						0	——				1		1	.00042%
ǎi	矮	17	.00454%	6		7		5	1	13	.00946%	1	2		1		4	.00169%
ǎidūndūn	矮墩墩	1	.00027%	1					1	1	.00073%						0	——
ǎixiǎo	矮小	2	.00053%	1				2		2	.00145%						0	——
ǎizhuàng	矮壮	1	.00027%	1						0	——				1		1	.00042%
ài	唉	1	.00027%	1				1		1	.00073%						0	——
ài	爱	145	.03870%	10	15	8	8	18	32	81	.05892%	8	8	13	24	11	64	.02698%
àibùshishǒu	爱不释手	1	.00027%	1					1	1	.00073%						0	——
àidài	爱戴	5	.00133%	3				1		1	.00073%		1	3			4	.00169%
àifǔ	爱抚	4	.00107%	3				1		1	.00073%			2		1	3	.00126%
àihào	爱好	6	.00160%	4		1		1	2	4	.00291%		2				2	.00084%
àihù	爱护	7	.00187%	6	2	1		1	1	5	.00364%	1		1			2	.00084%
àimù	爱慕	1	.00027%	1					1	1	.00073%						0	——
àiqíng	爱情	13	.00347%	3						0	——			6	6	1	13	.00548%
ài·ren	爱人	14	.00374%	4						0	——	7	1	2	4		14	.00590%
àixī	爱惜	3	.00080%	2						0	——			2		1	3	.00126%
àizēngfēnmíng	爱憎分明	1	.00027%	1						0	——		1				1	.00042%
ài	隘	1	.00027%	1						0	——				1		1	.00042%
àikǒu	隘口	4	.00107%	1					4	4	.00291%						0	——
ài // shi	碍事	2	.00053%	2		1			1	2	.00145%						0	——
ài // yǎn	碍眼	1	.00027%	1						0	——		1				1	.00042%
àimèi	暧昧	2	.00053%	1						0	——			2			2	.00084%
an																		
ān	安	23	.00614%	9	1	2		3	5	11	.00800%	3	1	1	2	5	12	000506%
āndìng	安定	6	.00160%	4				2	1	3	.00218%			2	1		3	.00126%
āndùn	安顿	2	.00053%	2					1	1	.00073%				1		1	.00042%