中央民族大学“985工程”中国少数民族语言文化教育与边疆史地研究创新基地文库
国家语言资源监测与研究中心少数民族语言分中心计算语言学系列丛书

◎ 主编　戴庆厦　赵小兵

现代汉语基本词汇自动识别方法研究

◎ 赵小兵　张志平　田寄远／著

中央民族大学出版社
China Minzu University Press

中央民族大学“985工程 地研究创新基地文库
国家语言资源监测与研究中心少数民族语言分中心计算语言学系列丛书

◎ 主编 戴庆厦 赵小兵

现代汉语基本词汇自动识别方法研究

◎ 赵小兵 张志平 田寄远／著

中央民族大学出版社
China Minzu University Press

图书在版编目（CIP）数据

现代汉语基本词汇自动识别方法研究/赵小兵等著. —北京：中央民族大学出版社，2012.8

ISBN 978-7-5660-0252-5

Ⅰ. ①现… Ⅱ. ①赵… ②张… ③田… Ⅲ. ①现代汉语—词汇—自动识别—研究 Ⅳ. ①TP391.43

中国版本图书馆 CIP 数据核字（2012）第 193208 号

现代汉语基本词汇自动识别方法研究

著　　者　赵小兵　张志平　田寄远
责任编辑　满福玺
封面设计　布拉格
出 版 者　中央民族大学出版社
　　北京市海淀区中关村南大街 27 号　邮编：100081
　　电话：68472815（发行部）　传真：68932751（发行部）
　　68932218（总编室）　68932447（办公室）
发 行 者　全国各地新华书店
印 刷 厂　北京宏伟双华印刷有限公司
开　　本　787×1092（毫米）　1/16　印张：15
字　　数　305 千字
版　　次　2012 年 10 月第 1 版　2012 年 10 月第 1 次印刷
书　　号　ISBN 978-7-5660-0252-5
定　　价　45.00 元

前　言

基本词汇是一个民族日常使用的、不容易变化和比较稳固的词，它们一般都具有较强的构词能力，是语言中派生新词的基础。基本词汇是语言词汇系统的核心，基本词汇的研究对语言教学、词典编纂以及语言信息处理等领域应用都具有重要的意义。然而由于基本词汇所具有的“全民常用性、历史稳固性及构词能力强”三大特性，而且概念宽泛、评判标准模糊，其量化标准受限于计算语言学的发展程度，因而以往对基本词汇的研究及认定大都限制在语言学家例证性的范畴内，极少进行定量分析与研究。

本书以2002—2006年大陆地区发行的六种主流报纸（《人民日报》、《北京青年报》、《北京晚报》、《法制日报》、《环球时报》、《羊城晚报》）的动态流通语料库作为考察对象，动态跟踪和考察词汇在大众媒体中的使用情况，提出了词语通用度的计算公式，进而考察语言学家例证所获得的基本词汇先验集所具有的统计特征类型，确立基本词汇的特征描述向量，采用遗传算法构造基于动态流通语料库的“语言工程现代汉语基本词汇”CBVE[①]自动识别及提取模型，实现CBVE的自动提取，为现代汉语基本词汇研究提供了一种量化考察途径。本书的主要贡献体现在：第一，在大规模的动态流通语料库中，考察了大众媒体报纸的词汇真实使用情况。处理考察的语料规模庞大，覆盖文本数632255个，词次总数247257749，不同词种数8750105。第二，首次提出了在动态流通语料库中定量分析和考察CBVE基本词汇特征的一种方法，为今后基本词汇从定性研究过渡到定量研究提供了一种途径。第三，提出了一种词汇通用程度的计算方法，为词汇统计特征考察提供了一种新的计量指标。第四，借鉴了模式识别领域的研究方法，依据遗传算法搜寻特征向量空间范围广、收敛速度快、鲁棒性强等特点，将其应用到对CBVE自动识别及提取模型的参数训练上，获得了令人满意的结果。

---

① “语言工程现代汉语基本词汇”CBVE的概念界定，请参见1.2节内容。

目　录

第一章 绪 论

从一个民族的语言系统来说，词汇是承载语言信息的基本载体，它是语言系统中最活跃、最具生命力的元素。假如没有了词汇，语音发挥不了作用，语法也无法建构起来，三要素中词汇占有十分重要的地位。

然而，语言是随着人类经济及社会生活的发展进步而不断更新和发展变化的，是具有生命力的“活”的语言，是一个“语言生态系统”。新词新语不断产生、旧词旧语也在不断消亡。同时，各民族语言也由于交流、表达的需要，在不断地相互借鉴、融合和发展。如中国大兴安岭地区的鄂伦春语，主要词汇都是狩猎词汇[①]，当他们从山上走下来，进入到以农业为主体的社会中时，原有词汇的表达能力就显得十分贫乏，必须借鉴周边大量的汉语及蒙古语词汇，并发展成为其日常用语。然而同时，其本民族语言也面临着逐步消亡的命运。[②]

当今，世界经济一体化的趋势越来越明显，各国家民族不仅在经济上，而且各民族在日常生活、政治、教育等方方面面都相互影响、制约、渗透。所有这些都反映在了我们所使用的汉语言词汇中，如2002年新版《现代汉语词典》（简称《现汉》）补充收录的新词语“黑客”（意“1. 指精通电子计算机技术，善于从互联网中发现漏洞并提出改进措施的人；2. 指通过互联网非法侵入他人的电子计算机系统查看、更改、窃取保密数据或干扰计算机程序的人”，来源于英语词“hacker”，见《现代汉语词典》，2002）、当今很流行的网络词语“粉丝”（意“某个人物或事物的忠实拥护及支持者，即‘……迷’”，来源于英语单词“fans”）、大量的字母词如“CD盘”（即“光盘”）、“卡拉OK”等以及纯英文单词夹杂在汉语中被直接使用。所以，汉语言的词汇正在受到前所未有的冲击和挑战。那么，如何更好地规范汉语言的使用，如何能使汉语言与中国经济发展同步，使其对世界文化产生更深刻的影响，就成为摆在我们面前的一个重要课题。我们需要跟踪考察语言的发展，掌握语言的变化规律，在不断更新且广泛流通的动态流通语料库中观察词语的真实使用情

① 鄂伦春族以前的主要生活方式是狩猎。

② “现在，鄂伦春族中60岁以上人口中90%以上的人都能说能听懂鄂伦春语，40—60岁的人口中，只有60%的人能够做到这一点，20—40岁的人约一半能听懂鄂伦春语，但已不能说，而0—10岁的人100%根本不懂鄂伦春语。”内蒙古社会科学院人类学专家、鄂伦春族研究员白兰语（2006）。

况，考察和研究汉语词汇系统中动态及稳态两个部分的特征表现，如处于动态部分的更新发展相对迅速的新词、新语的特性以及处于稳态部分的构成新词新语基础的变化缓慢的“基本词汇”的特性，可以帮助我们更好地发现和寻找汉语词汇的演变规律。本研究的重点是构造一种定量分析和考察稳态部分的“基本词汇”的研究方法，它可以让我们更好地了解并掌握汉语言词汇的核心——“基本词汇”。

同时，推广和发展汉语言的一条重要途径就是语言教学，无论是以汉语为母语的基础语文教育，还是以汉语为第二语言的语言教学，汉语词汇学习都是最重要的一个环节。West 曾明确指出：“学习一种语言最重要的就是词汇的习得以及练习如何使用它”（陆俭明引用，1984）。然而汉语字、词以难学著称于世。本世纪初编纂的《中华大字典》就收录了 4.5 万个汉字，即使是近期编写的《现代汉语规范词典》（李行健主编，2004）上也收录单字 1.3 万个，词目 6.8 万余条，所以任何人都不可能完全掌握，而且也无此必要。“在以往的语言统计研究也证明了一个事实，即人们在日常的语言交流中，最经常使用的词语，除专业名词外，只有几千个”（常宝儒，1984）。所以我们有必要对汉语词汇的层级进行研究，考察汉语基本词汇与一般词汇、专业词汇与通用词汇的关系、现代汉语基本词汇、通用词汇的特征属性及提取方法，排除罕用词语，以便更好地服务于汉语教学、词典编纂及语言信息处理等领域。

1.1 基本词汇的来源及争论

1.1.1 基本词汇的来源

基本词汇是指在一种语言体系中，该民族群众在日常生活的语言交流中需要经常使用的、不可缺少的那部分词汇。基本词汇是语言词汇系统中的核心部分。

1950 年，斯大林在《马克思主义和语言学问题》一文中，对基本词汇进行了论述：“语言的词汇中的主要东西就是基本词汇，其中就包括成为它的基础的全部根词。基本词汇比语言的词汇少得多，可是它的生命却长久得多，它在千百年的长时期中生存着，并且为构成新词提供基础。”

“基本词汇”这一术语并不是斯大林最早提出的[①]，但斯大林文章的发表却对基本词汇的研究产生了重大影响，在 20 世纪五六十年代掀起了一股研究基本词汇的热潮，对基本词汇概念的界定，也是众说纷纭。其中有代表性的包括：

① 最早使用“基本词汇”这一术语的论文是孙伏园在 1947 年发表的《基本词汇研究述要》一文。

“普通话词汇里，有些词是全民族使用得最多的，一般的生活当中最必需的，意义最明确，为一般人所共同理解，几乎用不着什么解释的。这样的词是词汇当中最主要的成分，叫做基本词。”（胡裕树，1962）

“基本词汇是词汇的核心部分，是构成全部词汇的基础。”（马国荣，1990）

“基本语汇是全民族使用最多、日常交际最必需、意义最明确的语汇成分的总和”，“基本语汇有三大特点：全民性、常用性和稳固性”；“基本语汇的核心是核心语汇，核心语汇多是一些单音语素，它们除具有基本语汇的一般特点外，还具有能产性，是构成新语汇的基本材料。”（邢福义，1991）

“基本词汇中的基本词使用频率高、生命力强、是一般生活交际中最必需、为一般人所共同理解的”，“基本词又分为根词和非根词两部分，根词是基本词汇中生命最长久、构词能力强的词，是基本词汇的核心。”（周建设，2001）

《现代汉语词典》对“基本词汇”是这样界定的：“词汇中最主要的一部分，生存最久、通行最广、构成新词和词组的能力最强，如‘人、手、上、下、来、去’等。”（《现代汉语词典》，2002）

尽管对“基本词汇”的界定，说法不一，但是对基本词汇所具有的“全民常用性、历史稳固性和构词能力强（或者说‘能产性’）”三大特点，语言学界已达成基本共识。在基本词汇的三大特性中，“全民常用性”是共时标准，“历史稳固性”是历时标准，而“构词能力强”是就它在词汇发展中的作用而言的，实质上也是历时标准。

1.1.2 有关基本词汇本质特点的争论

关于基本词汇，人们至今的观点和认识，基本上还是斯大林在《马克思主义和语言学问题》中所谈的几点，归纳为“全民常用性、稳固性和能产性”三大特点。但是关于基本词汇的最本质特点，至今仍有争论：如在《语言学引论》（1958）一书中对基本词汇的特性有这样的描述：“一般来说，全民性这个特征是一切基本词汇的词所共有的。”与之相对应，还有观点认为基本词汇的最本质特点是构词能力，符淮青先生的观点具有代表性，他认为在“普遍性、稳固性和构词基础”三大特征中，“构词能力是最重要的一个标准。有很强的构词能力，说明它是稳固的，因为它构成那么多词，要在一个长时期中才能陆续完成。构词能力强，有稳固性，往往又能显示它的普遍性”（符淮青，1985）；“这是基本词汇最主要的属性”，“词的有无构词能力是判断一个词是否基本的先决条件”（苏新春，1992）。然而，这样也使得一些构句的基本元素，在日常生活中十分常用，但缺乏构词能力的虚词被排除在了基本词汇之外，这从理论和实际应用角度都是无法被接受的。

对于基本词汇的稳固性特征也有争论。有些词，如“天、地、日、月、山、水、年、上、下、人、口、手、牛、羊”等词，从上古时代已经存在，几千年来变

化不大；还有些词，如“皆、祭、祀、鼎”等，“是两汉时代的基本词，用在《尔雅》、《说文》中，作为训释词来解释前代的词”（张世禄，1984），它们也已经存在了上千年的历史，但现在已经很少被使用了；然而对另外一些词，如“飞机、电话、党”等词，虽然可能只有上百年的历史甚至更短，但“它们仍然是现代汉语的基本词汇”（《语言学引论》，1958）。

那么基本词汇是否应该具备所有上述三项特征，或者具备其中的一两项特征也可被称为基本词汇呢？语言学家在前期的基本词汇研究中，提出如下观点，如潘允中先生认为：“基本词汇……具有的特征是：历史稳固性、全民性、有构词能力。不过，并不是所有的基本词都同时具备这些特点，有的只具有其中的两个，例如亲属名词多半是这样。”（潘允中，1989）即亲属名词多半只具有历史稳固性和全民常用性特征，而它们的构词能力一般不强。赵振铎先生也提出：“我们决不能因为某些词只符合两个标准，而不符合另一个标准，不考虑它的原因就把它排斥到基本词汇之外去”，“能不能归入基本词汇，要看它是否具有基本词汇的条件，如果条件具备了而不把它列进去，那是不公正的”。因为“基本词汇的标准不应当理解成一个签条，而应当理解成表示它们历史发展特殊趋势的词的一定范畴的特性”（赵振铎，1959）。赵先生的这些观点对于我们判断基本词汇以及构造基本词汇集的自动提取模型具有重要的启发意义。

1.1.3 基本词汇研究存在的问题

1.1.3.1 确定基本词汇的三个标准比较宽泛，没有科学量化

大家都承认基本词汇的客观存在，也清楚这种存在对语言的重要性，遇到举例时，能很轻松地顺手拈来。可是，在某一时期，基本词汇不是一个动态变化的开放集，而是一个相对稳定的封闭集合。当然，这种封闭是相对而言的，是指在有限时段内的封闭，如果放到语言发展的历时空间里看，基本词汇是动态发展变化的，它和一般词汇也在进行着交流和互动，并不是静止不变的。然而，基本词汇不会像一般词汇那样总在不断地衍生、变化，更不会像流行语那样日新月异。相对而言，基本词汇的发展变化是相对比较缓慢的，这也正是其稳固性所在。遗憾的是，到目前为止，各位学者在介绍基本词汇时，也只是举例式的说明，从来没人指出过一个基本词汇的明确量化界定标准。

在20世纪90年代之前，人们对基本词汇的研究还停留在对基本词汇本质特点的争论上。刘叔新先生率先对基本词汇理论进行独到思考并给出衡量标准，他认为：“历史悠久”和“当代社会普遍使用”是区分基本词汇和一般词汇的标准，并给出解释：“只要存在大约七八十年，即民国初年‘五四’时期便存在而现今仍沿用的词语，便可以认为具备了历史悠久的特征。”“只要在当代大多数阶层和社会集团里通用，其余阶层、集团也能够了解，就算是用得普遍。”（刘叔新，1990）之后

苏培成先生在此观点基础上，又提出“基本词汇”时段性问题，“一个是先秦到现代，这是汉语的基本词汇；二是自六朝至本世纪初，这是近代汉语的基本词汇；三是自本世纪初至当前，这是现代的基本词汇”，同时他还提出了确定时段基本词汇的方法，“将时段划分为若干共时平面，确定共时平面的常用词汇，寻找其共同的部分”。（苏培成，1993）

“苏先生关于‘确定汉语基本词汇的操作方法’是迄今为止唯一的一个有关汉语基本词汇认定的可供操作的具体方案，虽然这个方案的科学性和严密性尚有待于进一步完善。”（曹炜，2004）

因此，基本词汇作为词汇中重要的部分，并非很容易就能够从整个词汇成员中划分出来。简单地给出几个难以精确衡量的标准对词汇成员做硬性划分既不科学，也没必要，同时还会遇到难以解决的问题。比如，一个词存在多长时间算有稳固性，它作为构词词素能够构成多少个词算有能产性，在使用上达到什么程度算有全民常用性，这都是宽泛的概念，同时具备这三个标准的词也为数甚少。因此，迄今为止，尚没有一个科学量化的汉语基本词汇衡量标准。

1.1.3.2 基本词汇成员混杂

顾名思义，基本词汇应该是基本词和语的集合。所谓“词”[①]，包括单音节词、双音节词及多音节词；所谓“语”，是指固定短语，包括成语和熟语。从许多学者的举例中，发现了一些问题，基本词汇到底应该包括什么？还有很多分歧。

首先是语素[②]的问题：词汇是语言系统中最活跃的因素，它几乎是在不停地变异、消长着。尤其是汉语词汇，随着“五四”时期白话文对文言文的沉重打击，词汇形式由单音节占优势向双音节占优势演变，许多历史悠久的文言单音词在现代汉语普通话中失去了词的资格，变成了构词成分。虽然不是词了，但它们在现代汉语中仍保持着旺盛的构词能力。

其次是词类的问题：对于基本词汇应该包含哪些词类，斯大林并没有明确指出，所以对这个问题仍存在争论。经过语言学家的分析发现，基本词汇的范围很大，各种词类都有，虚词也不例外（潘允中，1989）。赵振铎先生的《虚词不能归入基本词汇吗》一文中，以翔实的例证说明了虚词应被归入基本词汇的缘由（赵振铎，1959）。孙伏园先生也提出了类似看法，他对汉语基本词汇进行了词性归类，其中也包括介词、连词、助词等虚词（孙伏园，1947）。

1.1.3.3 对基本词汇的认识存在偏差

国外的基本词汇与汉语的基本词汇的概念还是有偏差的，尤其以英语基本词汇集为例，斯瓦迪士 Swadesh（1952）的 200 词表也有词根，数量很少，而且是一个

① “词”是语言中最小的、能独立运用的表意单位。

② 语言中最小的音义结合体是语素，词是由语素组成的，如“儿童”是由“儿”、“童”两个语素组成的（王宁、邹晓丽，1999）。

封闭的集合。汉语的词根和词根组合构成复合词。而复合词一旦具备基本词汇的特点，便进入基本词汇的集合。我们要研究的基本词汇集正是包括很多这类复合词的集合。从汉语的基本词汇概念来说，现在还没有人真正拿出一个基本词汇集，无论是国外还是国内。

国内外研究基本词汇的学者都遇到了同一个问题：到底基本词汇的范围在哪里？什么样的词才算是基本词汇？

由于基本词汇的三个界定标准存在概念模糊，不好实际操作的问题，迄今为止，还没有人提出一个可行的界定方法。

1.2 基于动态流通语料库的现代汉语基本词汇概念的界定

基本词汇的词是一个民族的人民日常都在使用的、不容易变化、比较稳固的词语，它们一般都具有较强的构词能力，是语言中派生新词的基础。

基本词汇中的词是语言词汇[①]的核心，它表达的是与人们世世代代的日常生活关系非常密切的事物，如自然现象，家畜、亲属名称，人的肢体、器官名称，表示方位、时令、数目、劳动工具的词及与日常言行有关的现象等词汇。

我们都承认，基本词汇在具有历史稳固性的同时，也是在缓慢变化的。根据西方年代语言学家的研究，“各种语言的基本词，每经过一千年，大致只能保持81%不变”（史存直，1958）。因此，我们认为，从上古时期一直演变到今天还一直存在，而且仍为汉民族语言所常用的基本词汇已较少见了（如“天、地、人”等）。而且，对斯大林所提出的基本词汇“历史稳固性”这一特征的考察也受到目前我们的研究手段、语料来源等条件的制约，尚不具备可行性。然而这并不妨碍我们在现代汉语的动态流通语料库基础上研究词汇的全民常用性、时间稳定性及构词能力等特征，考察在大众媒体中词汇的真实使用状况，为现代汉语基本词汇的自动提取提供依据。

为此，我们在参考《现代汉语词典》对基本词汇概念界定的基础上，对基于现代汉语动态流通语料库上进行研究的“基本词汇”（下称“语言工程现代汉语基本词”CBVE，Contemporary Chinese Basic Vocabulary of Language Engineering）给出了

① “词汇”是词语的总和。这里的“词语”指词和由词构成的、性质相当于词的固定短语。词和词汇不是同一概念。词是构成词汇的个体单位，而词汇是语言材料的总体和集体。一种语言中所有词和固定短语的总汇就是该语言的词汇（赵振铎，1959）。

如下界定："语言工程现代汉语基本词汇" CBVE 是现代汉语动态流通语料库[①]词汇系统的核心稳定部分，是当今汉民族人民在日常交流中普遍使用的、不容易变化，时间上使用稳定的词汇，它的词语具有较强的构词能力，是汉语言中派生新词的基础。

"语言工程现代汉语基本词汇" CBVE 具有全民常用性、使用的稳定性、构词能力强的特征。

对于"语言工程现代汉语基本词汇" CBVE 所具有的上述三大特性，我们将进行分级别考察，依据词语的不同表现特征，构造不同的 CBVE 自动识别及提取模型，这样的考察才更具有合理性和实用性，如"的"、"一"、"人"等词同时具备 CBVE 的三大特征，而"我"、"你"等代词以及"了"、"很"等虚词，尽管它们的构词能力较弱，但由于具有很好的全民常用性特征，因而也具有良好的词语表现能力，所以也将纳入到我们的 CBVE 范畴中来（详细内容请参见第五章、第六章内容）。

1.3 基本词汇的研究意义

1.3.1 基本词汇在词汇教学领域的研究意义

基本词汇是语言的意义基础，是不断变动的词汇系统中相对稳定的部分，是生成新词语的核心词汇。掌握了这把钥匙，即使去理解从未谋面的新词，也不会费很大气力。所以在中小学语文教学中，抓住了基本词汇，就等于抓住了词语教学的灵魂。同时，基本词汇集的研制，可以为教学用词汇等级大纲的制定，提供科学的参考依据。

基本词汇包含那些反映人类对世界最主要、最基本、最具体认识的词，而这些词反映的概念往往是在不同语言中共同存在的，所以基本词汇是第二语言教学的基础。对外汉语教学需要在短时间内教会学生应用汉语的能力，这些学生都是已具备母语认知能力的成年人。因而，具有构词能力高、常用性及概念共同性特征的基本词汇的熟练掌握，可以帮助学生更好地延伸到其他词汇，从而使学习达到事半功倍的效果，有助于他们快速提高运用汉语的能力。虽然目前我们已经有了面向对外汉语教学用的 HSK 词汇等级大纲，但随着时代的发展，人们日常交流的词汇发生了许多交替，特别是一些广泛使用的新词语并不包含在 HSK 汉语词汇大纲中。因此，不少人已经呼吁修改大纲。基本词汇集及通用词汇集的制定，可以为修订 HSK 词

① 动态流通语料库（DCC，Dynamic Circulating Corpus）概念的界定见 2.2 节。

汇大纲提供参考依据。

1.3.2 基本词汇在词典编纂领域的研究意义

一部高质量的词典不仅要为语言应用提供查阅参考工具，而且它也是实现词语规范化的重要途径。

词典是既往语言事实的定格，随着改革开放和现代化建设的进程，我们的社会生活也在发生着丰富的变化，这种变化必然要反映到我们日常所使用的语言及词汇中来，新词新义层出不穷，任何词典都显得相对滞后。“语言变化的速度使得任何人工编纂的词典（包括术语，下同）和语法规则都难以及时得到修订。在中国中型词典的修订需要10年以上的时间，大型词典和专业词典的修订周期更长。至今没有任何一部词典能够每年修订一次，更不要说即期修订。”（张普，2001）

然而，语言应用需要有及时更新的词典及动态更新的语言知识。“我们需要‘活’的词典去处理‘活’的语言事实。但是，我们无法依靠人工从‘活’的语料中随时寻找新的语言变化，也无法依靠机器自动搜寻以自动生成新的语言词典，因为机器不具备人的语感能力，不能自行评价和判断那些语言中的变化，不能自行进行吸纳和扬弃”（张普，2001）。

“高质量的词典编纂需要建立在科学研究的基础上，编撰者必须从现代汉语的语言事实出发。掌握第一手的资料，并应用先进的语言学理论为指导。”（沈家煊：《现代汉语词典》序，2002）

目前世界上能够提供的更新语言知识的最好的办法是“机器自动回收——专家进行评价”，即有人工后处理的计算机辅助更新，或者叫“协作性知识管理”（D. Vervenne，1999）。而提供“协作性知识管理”正是基于动态流通语料库基础之上所做的课题研究的优势所在，随着其上研究课题的不断深入和发展，将会揭开词典编纂领域新的一页。

基本词汇提取方法的研究，是基于真实文本的大规模主流报纸媒体上，动态监测处于核心地位的相对稳定的基本词汇，我们可以用计算机从大规模语料中辅助寻找词语释义及例句，这样就可以降低词典编纂人员的人力投入。该方法也可以推广应用到处于快速变化中的动态部分词汇——如“新词、新语”的研究中。因为词汇的动态部分和静态部分是同一事物的两个方面，在动态流通语料库中，我们不仅可以观测到处于稳态不变的词汇部分，而且同时我们也可以观测到处于不断变化中的动态词语部分。

研究汉语的基本词汇，还能够推动汉语的本体研究。基本词汇作为汉语词语的

基础，是词典编纂的释义原语。[①] “在释义描述时，希望能以最基本的概念与浅显易懂的词条将概念明确的表达。也就是说用来描述的词条是所谓的基本词，且具典型义的以达到典型的写法”（黄居仁，2006）。台湾语言研究所的黄居仁先生认为：“在词汇语义学与词典学中，释义基本语言（或语义元素 Semantic Primitive）的研究，一直是非常重要而具挑战性的。”（黄居仁，2006）

以英语词典为例，目前比较流行的学习型词典有《朗文当代高级英语词典》、《柯林斯合作英语词典》、《麦克米伦高阶英汉双解词典》、《建宏英汉多功能词典》等。

《朗文当代英语词典》Longman Dictionary of Contemporary English，用 2000 个释义词解释 56000 个词项，是当代词典用定量词汇释义的杰作。自 1978 年问世以来受到各国读者的欢迎并普遍被认为对外国学生学习英语很有帮助。朗文词典对 3000 最常用词的口语及笔语词汇标注词频，配有大量的“用法说明”，例句相当丰富，贴近现实生活，多取自英国国家语料库（Bank of English）和朗文语料库 BBC、CNN 和英美权威杂志，其词语搭配较牛津词典多，堪称学习型词典的典范。

《柯林斯合作英语词典》Collins Cobuild English Language Dictionary，共收词 75000 条，实例达 10000 条，适合中高级读者使用。它以 2000 个常用词汇对所有词条进行释义，简明易懂；例句也来自于柯林斯出版公司和伯明翰大学联合开发的语料库。这个语料库里收集了各种类型的书面语口语资料，经由软件编制人员设计各种程序，使编纂人员可以方便地查找。这样，编纂的效率明显提高，词典的出版周期明显缩短，而且更加准确地反映了当代英语的特点，速度快、语料新、定义准、信息全，更具适用性，更适合读者的需要。这也是这些词典能够成为在世界范围内英语学习者中广泛流行的重要原因之一。

1.3.3 基本词汇在语言信息处理领域的研究意义

在信息化社会，人机结合实现自然语言理解与处理是我们新世纪信息化发展的主要目标。在机器翻译、语音识别与合成、文本分类、信息检索、信息过滤等自然语言处理领域以及语言教学的现代教育技术领域，都需要强有力的技术支持和扎实雄厚的语言学基础研究。

例如，基本词汇所具有的全民常用性特征，使得它们具有不同专业领域所共用的词汇特性，所以在进行文本分类时，是应该首先排除的不具备领域特征的停用词表。因此，基本词汇作为语言基础—词汇的核心，对其范畴、特性、发展变化规律的研究，必然能促进上述自然语言处理技术的进一步发展。

---

① 释义原语，即释义的基本语言，是指一种语言系统中，解释新词语含义的释义句所采用的基本词语（黄居仁，2006）。

随着网络时代的到来，现代化教学技术已经成为基础教学和第二语言教学的一个重要辅助手段，包括：CAI（计算机辅助教学）、语言的多媒体电化及远程网络教育等。但是因为我国网络教学发展相对较晚，汉语电化教育在教学内容和模式等均处在探索发展阶段，因而也缺乏与网络教学模式相适应的基础教学理论和大纲。基本词汇集是语言词汇系统中的核心，可以成为网络教学内容安排和软件设计的引导和标准。

总之，汉语的基本词汇研究，具有重要意义，它可以“帮助说明汉语的特点和它发展的规律性”，“有助于进一步学习和掌握汉语词汇”（赵振铎，1959）。

1.4 基本词汇研究及相关文献综述

1.4.1 基本词汇的语言学研究综述

1.4.1.1 国外研究现状综述

国外在英语词汇研究上很重视基本词汇的研究。美国英语作家肯尼迪（Kennedy）从外国学生学英语出发，经过多年调查研究，并根据大量文献记载，总结出100个常用英语单词，我国语言学家汪榕培也比较了英国英语和美国英语中各自使用频率最高的100个词（英式、美式和汉式英语100词中名词、动词、代词、介词等所占比例不同，反映了其中的差异）。

语言学家们发现了一个有趣的事实：100个常用英语单词覆盖了50%的书面语言。学会英语最常用的1000个词，就能理解任何一篇规范文字80.15%的内容；学会常用的2000个词，就能理解约89%的内容；学会常用的5000个词，就能理解约97%的内容。受此启发，1930年英语语言学家C. K. Ogden的850个《基础英语》分类词表问世。

目前通行的基本词汇表是美国学者斯瓦迪士（M. Swadesh）在1952年提出的，原为推算语言的发展年代而设计。一共筛选出200个词汇，以代表“任何语言中由根词、基本的日用概念组成的那部分词”。后来斯氏又对这个词表作了进一步的精选以适用于各种语言，成为“修正表（100词表）”（M. Swadesh，1952）。斯氏的词表实际上是根据语义原则建立的，它以概念上的必然性作为定义标准，是能够表达这些基本概念的最小词汇库。因此，该基本词汇库可以视为所有语言的核心词汇库①的交集。

① 核心词汇库（Core lexicon）：在不同语境中能被预测出现的最大词汇库（M. Swadesh，1952）。