第 1 章　概　　论

对语言进行统计研究，是语言学研究的一种重要的方法。传统的语言学研究，往往以个人内省为主要依据——语感，从而对语言的词汇、句子、语法等进行考察，进而归纳出纷繁复杂的言语现象背后的语言规律。然而，语言研究的科学化，一直是现代语言学家的主要努力方向和重要目标。对语言的科学化研究，即是要求通过一些实验方法和数学模型来对语言进行考察，并且得出的结论是可验证的，而且对于语言规律和发展具有较强预测能力(刘海涛 2012)。

统计——通过实验方法和数学模型对于研究对象进行考察从而得到可验证的结论，对于现代自然科学的大发展起到了极为重要的作用。统计已广泛应用于语言学、文学、历史研究、心理学、社会学、人类学、教育学、生态学、考古学、农业、动物学、审计学、牙医学、工程、流行病学、金融、遗传学、地理学及工业等各个领域中。

统计是一种极为有用的研究手段和方法，因此将统计方法引入语言学研究并发展，是实现语言研究科学化的重要途径和手段。

1.1　统计语言学

使用概率论、数理统计等统计学的方法来对语言进行研究，称为统计语言学(冯志伟 2012)。

统计学可以帮助语言学者对语言学研究中的数据进行分析和推断。通过统计数据来揭示反应语言内在的和固有的一些规律性。

自然语言是其统计和分析的对象，概率论和数理统计等统计知识是其统计的理论基础，计算机是其可以实现统计的工具。因此，对语言进行统计不仅要有语言学方面的知识，而且还要有数学和计算机科学方面的知识。

1.2　统计语言学与其他学科

1.2.1　计量语言学

计量语言学，是指使用计量的方法，对于真实语言交际活动中产生的各种语言现象、语言结构、结构属性以及其相互之间的关系进行研究，从而找出语言现象背后深层的数学

规律，描绘出语言现象的数学面貌，并对其进行原因分析和理论阐释的一门学科。计量的方法，通常包括，概率论、统计学、微积分、随机过程等数学定量方法。

计量语言学以真实语料为基础，用计量的方法研究语言的结构和发展规律，其目的在于探索语言的数学面貌并发现隐藏在语言现象中的内在的数学规律。

计量语言学的根本任务是从真实文本中抽象出来的数量规律来描述与理解语言的各组成成分的关系。其发现的语言规律对于精确地描写与解释相应的语言现象以及构建现代科学意义上的语言学理论都具有极为重要的作用和意义。

计量语言学研究的主要内容有：语言成分统计规律、字频、汉字熵、词频与词序的秩的关系、词语的频率分布、词的音节构成及分布规律、英语动词的义项分布规律、名词分布规律、类符形符比、词语的使用度和通用度、文本中词长数量的分布、描述语言结构在语言系统和语言使用中的分布定律、描述不同的语言结构（及其属性）间相互关系的函数定律、描述语言演化规律的演化定律、基于短语结构树库的计量研究、基于依存树库的计量研究、概率配价模型、基于词汇系统的协同语言学模型和基于句法系统的协同语言学模型等等（刘海涛 2012）。

计量语言学广泛应用于音位学、形态学、句法学、词汇学、语义及语用学、地理语言学及方言学、类型学与语言的历时研究等领域。

1.2.2 计算语言学

计算语言学，也称自然语言处理或自然语言理解，它是研究如何利用计算机来分析、处理和理解自然语言。

计算语言学主要研究语音的分析和生成、分析型语言的分词、印欧语系语言的形态分析、词性标注、句法分析、语义分析、篇章分析等。

计算语言学主要用于语音的自动识别和自动生成、自动文摘、自动校对、信息自动检索、人机对话、自动问答、自动分类、信息自动抽取、机器翻译等领域。

计算语言学是利用计算机来分析和处理自然语言，其目的在于建立自然语言的应用系统，其研究方法有统计和规则两种方法。

1.2.3 语料库语言学

语料库语言学是利用计算机强大的检索、统计和处理语料的能力，从大规模的语料库中检索符合研究问题的实例，对其进行统计。在大量实例和统计数据的基础上，对研究问题进行定性分析，从功能上对其进行语言学解释。利用计算机和语料库可以对语言各个层面的特征（单个特征或多个特征）进行分析和研究。

语料库语言学的主要研究内容是研究机器可读的自然语言文本在词汇、句法、语义、语篇各个层面的采集、存储、检索、统计和分析。

语料库语言学主要应用在词典编纂、语言（社会语言或方言等）调查、语域变异、历时语言研究、语言习得、作家作品风格分析和教育语言学等领域的研究中。

语料库语言学对语料库既有定量的统计，又有定性的功能解释，对语言的描写更加全面。

1.2.4　与三个学科的联系与区别

统计语言学、计量语言学、语料库语言学和计算语言学的研究都是涉及语言学、数学、统计学以及计算机科学等多个学科和领域，是典型的文理工交叉学科，具有鲜明的跨学科研究性质。这些学科的研究对象都是自然语言组成的大规模语料库，研究工具都是利用计算机的软硬件，研究的理论基础是数学的概率统计知识和语言学的语音、词汇、句法、语义、语篇和语用知识。几个学科有些研究内容可能存在交叉，不能完全割裂开来。

统计语言学、计量语言学、语料库语言学和计算语言学都可以对语言学的语音、词汇、句法和语义等层面进行统计和研究。但它们的研究方法和研究目标存在不同：

统计语言学和计量语言学都是利用统计方法来实现对语言成分的统计，计量语言学以发现语言成分或语言成分间的数学规律为目标。而统计语言学以所统计的语言特征在统计学上显著和不显著为目标。语料库语言学对大规模语料库进行词汇、句法和语义等统计，依据统计数据和实例上下文对所研究的对象进行语言学层面定性的分析，是定量分析和定性分析的结合，以研究语言的结构和运用为目标。计算语言学以语言结构的理解与生成为研究目标，以统计和规则为基本研究方法。1990 年以来，统计方法在计算语言学领域占据主流，但计算语言学的统计模型——隐马尔科夫模型、最大熵模型、条件随机场模型等和实现算法更复杂，利用这些模型进行语音识别与合成、汉语的切分、词性标注、句法消歧、语义消歧或机器翻译等时，利用的都是动态规划算法，来实现对语言的理解和自动处理(刘海涛 2012)。

本书所介绍的统计语言学包含了语言与统计的研究内容、计量语言学的一些研究内容以及计算语言学中的文本分类和文本聚类等。

1.3　使用统计方法研究的语言特征

随着语料库加工深度的增加和计算机软硬件的发展，目前，从语音、词汇、句子、短语、段落等层面都可以对自然语言进行统计。

1. 语音层面

可以统计的特征有：声母、韵母、句尾韵、声调以及上述几个特征的结合。

2. 字符特征

目前，可以用来统计的字符特征有：字母、汉字、大小写字母、数字、标点符号、词首字母、词尾字母、词首字、词尾字、空格等。

3. 词汇特征

可以用来统计的词汇特征有：特定词的分布、词频、虚词、高频词、一次词比例、句首词、句尾词、段首词、段尾词、词长、平均词长、词汇丰富度、词长离散度、词汇总量、词类比例、特定类词汇数量比例(比如方言词、新造词、语气词、专业术语、国际性词语或外来词语、古语词、成语、惯用语等)、关联词语的种类及其分布等。

4. 句子特征

可以用来统计的句子特征有：句长、平均句长、句长离散度、不同句型的比例、句式的

分类统计及其数量关系(比如主动句、被动句、把字句、倒装句等)、句子结构(比如主谓结构、方位结构、动宾结构等)、用句总量等。

5. 段落特征

统计段落长度,可相对划分为长段落与短段落,并求出段落的平均长度。

统计段落的起始句、开端词、终结句和末端词,以分析分段特点。

段落的离散度可看出不同段落与平均段落离散的程度。

6. 短语特征

统计的最简单短语特征就是 N 元语法(N 元字符串、N 元词串、N 元词类串等),还可以统计名词短语、动词短语和形容词短语等分布。

除此之外,还可以对语料库进行语用特征和篇章特征等的统计。

统计不同层面的语言特征,对语料库加工的深度要求不一样。对字符层面的统计、不需要对语料库进行加工和标注。对词类进行统计,需要对语料库进行词性标注。对汉语进行词频统计,需要对语料库进行汉语分词。统计名词短语的分布,就需要对语料库进行名词短语的标注。随着语料库加工深度的增加,对语言进行的统计特征会越来越多。

1.4 统计语言学基本研究方法

1. 统计描述法

研究如何取得反映客观语言现象的统计数据,通过图表的形式对所统计的数据进行加工、处理和显示,进一步通过分析和综合得出反映语言客观现象的规律性的数量特征。

实现描述的统计量主要有:频率、概率、表示数据集中的平均数、表示数据分散的标准差和方差、互信息、Z 评分、Dice 系数、Phi 平方系数、对数似然比、N 元语法、信息熵、极限熵、词语的实用度和通用度和 Yule 图等。互信息、Z 评分、Dice 系数、Phi 平方系数和对数似然比可用来描述两个字之间或两个词之间结合的紧密程度。熵是对语言符号不确定性的度量,表示该语言每一个字符所包含的平均信息量的大小。极限熵是将充分考虑上下文关系的情况下达到的最小条件信息量。Yule 图是用来考察文本中词汇丰富度的大小的一个度量。使用度是用于衡量词语常用性的重要指标。通用度是衡量词语在不同领域的使用情况的重要指标。

本书给出 R 语言中对数据直观展示和描述的几种图形方法——饼图、条形图、直方图、折线图、箱线图和散点图等。饼图用来描述同一类型具有比例关系的数值型数据。条形图适用于展示类别型变量的分布。直方图是将数据对象划分为一定数量的组。折线图是将若个数据散点连接起来形成反映数据变化趋势的图形。箱线图反映一组数据的最小值、两个四分位数、中位数以及最大值,描述连续型变量的分布。

2. 统计推断法

统计推断法包括根据统计数据得出的反应语言现象的数学规和根据样本数据对总体的均值、方差等进行的假设检验。

根据统计数据得出的单个语言特征或两个语言特征数量间满足的数学规律,主要介绍 Zipf 法则、Menzerath-Altmann 定律、Piotrowski-Altmann 定律和 Fuchs 公式。Zipf 法

则描述了词频和词的排序序号之间的反比关系。Menzerath-Altmann 定律描述了词所含音节数和音节的平均长度间的关系。Piotrowski-Altmann 定律描述了语言现象的演变规律。Fuchs 公式描述了不同语言中词的音节数目的分布规律。

假设检验是对未知的总体分布形式或总体的未知参数做出一定的假设，然后构造适合的统计量并根据样本信息进行计算，在设定的显著水平或置信度上判断假设是否成立。根据总体是否服从正态分布，可分为参数假设检验和非参数假设检验。参数假设检验主要介绍对总体均值 μ 进行检验的 U 检验、t 检验，以及对总体方差 σ^2 进行检验的 χ^2 检验和 F 检验。非参数假设检验主要介绍 χ^2 检验和秩和检验。

参数假设检验和非参数假设检验主要用于检验两个总体之间的均值或方差等是否存在显著差异，而方差分析主要用于检验更多总体均值之间差异是否显著。

3. 统计模型法

统计模型法是根据数学模型对语言中成分或文本之间关系进行推断的方法。主要介绍的数学模型是朴素贝叶斯模型、K-最近邻模型、支持向量机模型、层次聚类和划分聚类。层次聚类和划分聚类是无指导的机器学习方法，可以根据文本的一个或多个特征实现对文本的自动聚类。而朴素贝叶斯模型、K-最近邻模型、支持向量机模型是有指导的机器学习方法，根据训练语料库的文本分类来对未知的文本进行分类。

1.5　统计语言学研究的步骤

作为一门实证学科，统计语言学研究所遵循的思路和研究方法与其他实证学科基本相同 大致包括以下五个步骤(Reinhard Köhler 2005)。

如图 1-1 所示，统计语言学研究一般要经过五个步骤。

第一步，提出语言学假设。语言学假设是对自然语言现象的大胆推测。这个假设要满足一定的形式与内容，并且所提出的假设必须有相关的实证性和可验证性。对于统计语言学研究中使用很多的随机假设来说，一方面，要推翻该假设绝不能依靠个别的反例，而必须建立在足够大量的数据和进行充分的数理检验的基础上。另一方面，一个假设永远不能被认为是完全证实，即便已有的数据均支持该假设，但仍然有继续检验的必要。

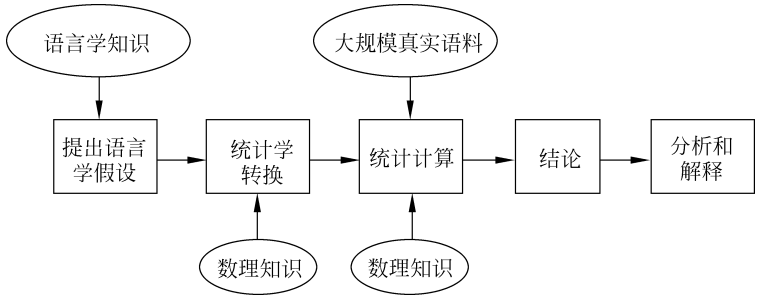

图 1-1　统计语言学研究步骤

第二步，将语言学假设转换为可以量化的特征。由于随机假设只能够通过数理方法的方法得到检验，因此，每一个假设，都必须要转换成为可以使用数理检验的形式。在转

换之前，必须要了解可能用到的数学模型以及这些模型使用的条件。

第三步，统计计算。首先对大规模语料库进行预处理，包括汉语分词、词性标注等。然后对经过预处理的语料库，利用计算机开源软件 AntConc 或编程进行语言学特征的统计。

AntConc 软件[①]是一款免费的语料库检索软件，由早稻田大学的 Laurence Anthony 教授开发（Anthony L，2004）。Antconc 提供了关键词索引、索引定位、文本查看、词丛、搭配、生成词表以及关键词列表等功能。

NLPIR[②] 汉语分词系统主要功能包括：分词（中英文混合分词）、词性标注、命名实体识别、新词识别、关键词提取、用户专业词典和微博分词等。

斯坦福句法处理器[③]可对汉语或英语等进行依存语法分析，斯坦福词性标注器[④]可对多种语言进行词性标注。

哈工大社会计算与信息检索研究中心研制了一整套高性能开放中文自然语言处理系统[⑤]，包括汉语分词、词性标注、依存句法分析、命名实体识别、语义角色标注在内的自然语言处理服务。

第四步，得出结论。对基本统计数据经过描述统计、统计推断或统计模型运算后，可以得出结论，即是否与假设相符。利用统计推断或统计模型，可直接给出结论。利用统计描述法，还需要对数据进行比较和分析再得出结论。

第五步，对于该结论进行分析和解释。根据最初研究的语言问题来解读统计检验的结果，即把数学形式的结果用自然语言的形式解读为对所研究的语言问题的回答。同时，针对统计结果，给出反应统计规律的语言学分析结果。

1.6 统计的语言学应用

统计语言学可用来发现语言学内在规律（词频与词排序之间的关系）、常用词和非常用词、词语搭配、短语获取、语言习得、语域变异以及作品风格分析等领域。

① http://www.antlab.sci.waseda.ac.jp/software.html.

② http://ictclas.nlpir.org/.

③ http://nlp.stanford.edu/software/lex-parser.shtml.

④ http://nlp.stanford.edu/software/tagger.shtml.

⑤ http://www.ltp-cloud.com/intro/#ltp.

第2章　语　料　库

2.1　语料库的定义

语料库(corpus)是语言材料的仓库,是计算机进行语言检索、比较、分析等处理的重要基础。(张普 1999)

语言学名词审定委员会 2011 年推出的《语言学名词》中,对语料库的定义、作用及应用领域的阐述为:

(语料库是)为语言研究和应用而收集的,在计算机中存储的语言材料,由自然出现的书面语或口语的样本汇集而成,用来代表特定的语言或语言变体。经过科学选材和标注,具有适当规模的语料库能够反映和记录语言的实际使用情况。通过语料库能够观察和把握语言事实,分析和研究语言系统的规律。语料库可以应用于语言学理论研究、语言应用和语言工程。

由此可见,语料库并不是语言材料的简单堆砌或随意集合。而是有着严格要求的有序的语料集合。其必须具有以下几个基本特征:(杨惠中 2002)

① 具有明确而具体的研究目标,并在系统语言学理论原则指导下进行设计和建设。

② 语料是按照明确的语言学原则选取,并采取随机抽样方法收集得来。所收集的语料必须是真实语言环境中出现过的自然语料。

③ 作为自然语言运用的样本,必须具有代表性。

④ 语料经过适当的加工或处理,如经过词语切分或者词性标注;并且以电子文本形式存储,即必须是机器可读的。

⑤ 语料文本是连续的文本或话语片段,不能是孤立的句子或词汇。

语料库广泛应用于语言研究、词典编纂、语言教学,以及自然语言处理等各个领域。对语言学、语料库语言学和计算语言学的研究和发展具有重要的意义。

2.2　语料库的类型

语料库根据不同的划分标准,可以分为不同的类型。常用的划分类型及主要的语料库类型有以下几种。

2.2.1 口语语料库与书面语语料库

根据语料的表达形式，分为口语语料库与书面语语料库。

1. 口语语料库

口语语料库(spoken corpus)，以口语语料为主要构成的语料库，是研究口语特征的重要工具。其常常包括由口语转写而来的文本，有时也包括语音文件。为方便对口语的研究，往往对口语语料库中的语音、语调、停顿、重复、修正等口语特征进行标注。(梁茂成等2010)如伦敦-隆德语料库(LLC 语料库)，美国英语口语语料库，北京语言大学北京口语语料查询系统。

2. 书面语语料库

书面语语料库(written corpus)，以书面语语料为主要内容的语料库。常常取材于报刊、书籍、书信、学术论文等各种形式的文本。目前语料库多为书面语语料库，并且容量一般比口语语料库要大。(梁茂成等 2010)如 Brown 语料库，兰开斯特－奥斯陆/卑尔根语料库(LOB 语料库)等。

2.2.2 单语语料库、双语语料库与多语语料库

根据语料所包含语种的数量，可分为单语语料库、双语语料库与多语语料库。

1. 单语语料库

单语语料库(monolingual corpus)即语料库仅收录了一种语言的语言文本。如英国国家语料库(BNC)、国家语委现代汉语平衡语料库等。

2. 双语语料库

双语语料库(bilingual corpus)，语料库收录了两种语言的文本。其下，又可细分为平行语料库(parallel corpus)与比较语料库(comparable corpus)。

1) 平行语料库

平行语料库中，两种语言的文本互为翻译，如北京外国语大学的英汉双语平行语料库；根据对齐的程度又可以分为词语对齐语料库、短语对齐语料库、句子对齐语料库、篇章对齐语料库。此类语料库多用于机器翻译、双语词典编撰等领域。如中国科学院软件研究所的英汉双语语料库，具有 15 万对英汉双语对齐句子。

2) 比较语料库

比较语料库是把表述相同内容的不同语言的文本收集在一起，文本间不存在翻译关系(语言学名词审定委 2011)。此类语料库多用于语言对比研究。

3. 多语语料库

多语语料库(multilingual corpus)，即语料库中收录的语言材料有三种或三种以上的语言。如欧洲平行语料库收录了欧洲议会的多种语言文集，将 11 种语言进行对齐处理。(梁茂成等 2010)

2.2.3 通用语料库与专用语料库

根据不同的应用层面，可分为通用语料库与专用语料库。

1. 通用语料库

通用语料库(general corpus),也称一般语料库。其要求收集的语料具有广泛的代表性,因此一般采用系统的方法进行采集,用于预先未指定的语言学研究。同时,要求其语料具有平衡性,因此,一般需要注意不同文类、语域、语式、体裁、主题、时间语料的平衡,口语与书面语均要具有。其能够充分反映和记录语言的实际使用情况。通用语料库也被称为系统语料库或平衡语料库,有时也称为"核心语料库"。(郭曙纶 2013)

通用语料库是描述语言全貌、编制工具书、核查语言用法等最理想的语料。

2. 专用语料库

专用语料库(specialized corpus),又称为专用目的语料库(special purpose corpus)。指用于某种特殊研究目的的语料库。语料库一般由某一特定领域的语料构成。如方言语料库、区域性语料库、非标准语料库以及学习者语料库,等等。(郭曙纶 2013)其既可以用来与通用语料库作对比,分析特定领域内的语言特点;又可以用来编纂特定领域的词典,如方言词典。(梁茂成等 2010)典型的专用语料库如温州口语语言资料库、密歇根学术英语口语语料库。

2.2.4　共时语料库与历时语料库

根据语料状态,可分为共时语料库与历时语料库。

1. 共时语料库

共时语料库(synchronic corpus),指语料库的语料来源于同一时代,是相对于历时语料库而言的。基于不同时代的语言所建的多个共时语料库可以构成历时语料库。(梁茂成等 2010)共时语料库常用来观察和研究某一时代的语言使用状况。Brown 语料库和 LOB 语料库就属于此类语料库。

2. 历时语料库

历时语料库(diachronic corpus),指收集不同时代的语料样本所构成的语料库。其主要用于研究和考察语言的历史演变。如赫尔辛基英语文本语料库(Helsinki Corpus of English Texts)就是一个典型的英语历时语料库。(梁茂成等 2010)

2.2.5　动态语料库与静态语料库

根据语料库动态更新程度,可分为动态语料库与静态语料库。

1. 动态语料库

动态语料库(dynamic corpus),又称为监控语料库(monitor corpus)。此类语料库在量级上和时间跨度上都没有限制,处于不断发展之中。可以用其观测语言的发展变化。(王建华 2002)典型的如北京语言大学 HSK 动态作文语料库。

2. 静态语料库

静态语料库(static corpus)。与动态语料库相对应,只收集某一固定时期的共时语言材料,语料库建成之后,就不再扩充。(郭曙纶 2013)第一代百万词级的语料库就属于此类语料库。典型的如 Brown 语料库,LOB 语料库等。

2.2.6 同质语料库与异质语料库

根据收集的语料内容，可分为同质语料库与异质语料库。

1. 异质语料库

异质语料库(heterogeneous corpus)，这是一种最简单的语料收集方法，其尽可能广泛地接受各种材料，事先并未制定任何选材原则，语料在格式和内容上并不完全相同。典型的如 ACL/DCI 语料库，以及牛津文本档案库(OTA)。(黄昌宁、李涓子 2002)

2. 同质语料库

同质语料库(homogeneous corpus)，与异质语料库相对立。指仅收录同一类内容的语料。一般用于专业语料库。(黄昌宁、李涓子 2002)如北京大学计算语言研究所的《人民日报》切分和标注语料库。

2.2.7 生语料库与标注语料库

根据语料加工程度，可分为生语料库和标注语料库。

1. 生语料库

生语料库，指收集语料之后，未经过任何人工处理的语料库，此类语料库未经过任何的加工和标注。

2. 标注语料库

标注语料库，即是对语料进行过标注等处理的语料库。根据标注的复杂程度，又可以分为(杨惠中 2002)：

① 不加任何处理的纯文本语料库，即是生语料库；
② 经过格式属性标注的语料库，如对段落、字体、字号进行标注；
③ 对识别信息进行标注的语料库，如作者、语域、体裁以及词性标注；
④ 特殊标注，如错误赋码；
⑤ 短语结构语料库；
⑥ 语义标注语料库；
⑦ 语篇标注语料库。

2.3 国内外主要语料库

2.3.1 国外的语料库

从 1959 年英国伦敦大学教授 Randolph Quirk 宣布建立英语用法调查语料库，即 SEU 语料库(Survey of English Usage)开始，语料库的建设和发展已经走过了半个多世纪。学界一般将语料库的发展划分为四个阶段，第一代语料库(1960—1980 年代)、第二代语料库(1980—1990 年代)、第三代语料库(1990 年代)、第四代语料库(目前)(韩效伟 2011)。

1. 第一代语料库(1960—1980 年代)

这一时期的语料库，包括 Brown 语料库(1964)、LOB 语料库(1970—1978)、伦敦—隆