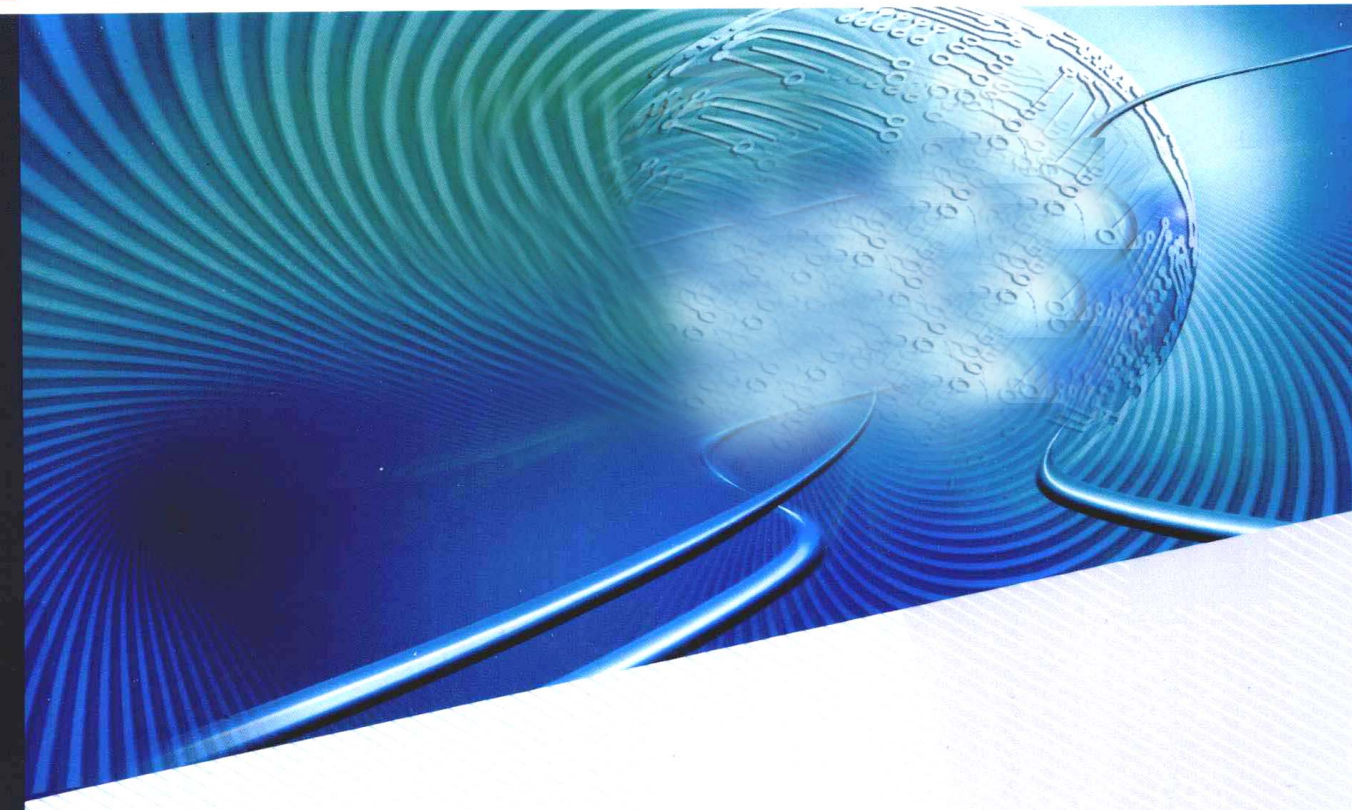
智能 科/学/技/术/著/作/丛/书

智能信息处理
——汉语语料库加工技术及应用

郑家恒　张　虎　谭红叶　钱揖丽　卢娇丽　著



科学出版社
www.sciencep.com

智能科学技术著作丛书

智能信息处理——汉语语料库加工技术及应用

郑家恒　张　虎　谭红叶
钱揖丽　卢娇丽　著

科学出版社
北京

内 容 简 介

本书以作者主持的国家项目、省部级项目及合作项目等为依托，以课题组近年来的研究成果为基础，重点介绍语料库深加工中的若干技术和方法，涉及分词、词性标注、句法分析、语义标注以及相关加工中的自动校对和一致性检验技术。同时，对语料库加工质量的评价技术和语料库的相关应用做了详细介绍。各章节的顺序展示了语料库加工中由浅入深的发展过程。

本书可作为计算机、语言学等专业高年级本科生、研究生教材，也可作为自然语言处理和计算语言学研究人员的参考书。

图书在版编目(CIP)数据

智能信息处理：汉语语料库加工技术及应用/郑家恒等著. —北京：科学出版社，2010

(智能科学技术著作丛书)

ISBN 978-7-03-029135-6

Ⅰ. 智… Ⅱ. 郑 Ⅲ. 人工智能-信息处理 Ⅳ. ①TP18

中国版本图书馆 CIP 数据核字(2010)第 191463 号

责任编辑：卜 新 王志欣 魏英杰 / 责任校对：李 影

责任印制：赵 博 / 封面设计：耕 者

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

新蕾印刷厂印刷

科学出版社发行 各地新华书店经销

*

2010 年 10 月第 一 版 开本：B5(720×1000)

2010 年 10 月第一次印刷 印张：20 3/4

印数：1—3 000 字数：401 000

定价：60.00 元

(如有印装质量问题，我社负责调换)

《智能科学技术著作丛书》编委会

《智能科学技术著作丛书》序

“智能”是“信息”的精彩结晶，“智能科学技术”是“信息科学技术”的辉煌篇章，“智能化”是“信息化”发展的新动向、新阶段。

“智能科学技术”(intelligence science&technology，IST)是关于“广义智能”的理论方法和应用技术的综合性科学技术领域，其研究对象如下：

• “自然智能”(natural intelligence，NI)，包括“人的智能”(human intelligence，HI)和其他“生物智能”(biological intelligence，BI)。

• “人工智能”(artificial intelligence，AI)，包括“机器智能”(machine intelligence，MI)与“智能机器”(intelligent machine，IM)。

• “集成智能”(integrated intelligence，II)，即“人的智能”与“机器智能”人机互补的集成智能。

• “协同智能”(cooperative intelligence，CI)，指“个体智能”相互协调共生的群体协同智能。

• “分布智能”(distributed intelligence，DI)，如广域信息网、分散大系统的分布式智能。

1956年，“人工智能”学科诞生，50年来，在起伏、曲折的科学征途上不断前进、发展，从狭义人工智能走向广义人工智能，从个体人工智能到群体人工智能，从集中式人工智能到分布式人工智能，在理论方法研究和应用技术开发方面都取得了重大进展。如果说，当年“人工智能”学科的诞生是生物科学技术与信息科学技术、系统科学技术的一次成功的结合，那么，可以认为，现在“智能科学技术”领域的兴起是在信息化、网络化时代又一次新的多学科交融。

1981年，“中国人工智能学会”(Chinese Association for Artificial Intelligence，CAAI)正式成立，25年来，从艰苦创业到成长壮大，从学习跟踪到自主研发，团结我国广大学者，在“人工智能”的研究开发及应用方面取得了显著的进展，促进了“智能科学技术”的发展。在华夏文化与东方哲学影响下，我国智能科学技术的研究、开发及应用，在学术思想与科学方法上，具有综合性、整体性、协调性的特色，在理论方法研究与应用技术开发方面，取得了具有创新性、开拓性的成果。“智能化”已成为当前新技术、新产品的发展方向和显著标志。

为了适时总结、交流、宣传我国学者在“智能科学技术”领域的研究开发及应用成果，中国人工智能学会与科学出版社合作编辑出版《智能科学技术著作丛书》。需要强调的是，这套丛书将优先出版那些有助于将科学技术转化为生产力以及对社会和国民经济建设有重大作用和应用前景的著作。

我们相信，有广大智能科学技术工作者的积极参与和大力支持，以及编委们的共同努力，《智能科学技术著作丛书》将为繁荣我国智能科学技术事业、增强自主创新能力、建设创新型国家做出应有的贡献。

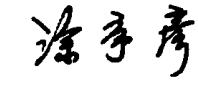
祝《智能科学技术著作丛书》出版，特赋贺诗一首：

智能科技领域广

人机集成智能强

群体智能协同好

智能创新更辉煌



中国人工智能学会荣誉理事长

2005 年 12 月 18 日

前　言

从20世纪90年代开始，国际自然语言处理领域发生了一些重大变化，重要特征之一就是转向对大规模真实文本的研究和处理。以大规模真实文本为基础的语料库研究和知识自动获取受到高度重视。显然，大规模真实文本的处理是计算语言学今后一个时期的战略目标，建设高质量的大规模语料库是中文信息处理领域的基础性工程。基于语料库的语言研究是计算语言学的一个重要领域，语料库的建立为语言学的研究提供了丰富的语言现象，为计算语言学学者从加工的语料库中获取语言知识、建立语言模型、研究语言信息处理技术提供了翔实的语言信息数据。作为研究资源的语料库的价值是通过对语料的加工来体现的，对语料库加工的层次越高，语料库的应用价值就越高。希望本书的出版能促进语料库加工方法和技术的发展，为基于语料库的相关研究和应用提供支撑。

作者及其课题组从事语言信息处理的教学与研究已有二十多年。近年来，作者有幸承担了若干国家863计划项目（中文文本自动切词和词性标注软件及其评测技术研究（863-306-03-09-4）、大规模中文文本语料库深加工质量检验技术研究（2001AA114031））、国家自然科学基金项目（大规模中文文本语料库分词与词性标注一致性检验技术研究（60473139）、基于中文文本的计算机中介通信中欺骗检测研究（60775041））、省部级项目及横向合作项目等。这些项目的研究成果为本书的编写提供了关键性支持。多年来，刘开瑛、黄昌宁等诸位学术前辈都为作者的相关研究思路和方法提供了许多指导。本书编写过程中，山西大学梁吉业、李德玉、李茹、王文剑、王素格等教授为作者提供了多方面的支持。魏善德、任玉、魏莉、魏丽霞、樊勇、王振宇、刘博、张剑锋、何苑、温艳霞、毋菲等同学也为本书的出版做了许多文字校对方面的工作，谨在此一并表示深深的感谢。

本书由郑家恒教授统稿，谭红叶编写第1、5章，钱揖丽编写第2、6章，张虎编写第3、7章，卢娇丽编写第4章。

由于作者水平有限，本书遗漏和不妥之处在所难免，恳请读者批评指正。

作　者

2010年9月

目　　录

第1章 绪 论

1.1 语料库的定义和作用

1.1.1 什么是语料库

关于语料库(corpus)的定义主要有以下几种：

(1) McEnery 和 Wilson[1]指出："总体来说，多篇文本的集合就是语料库，但在现代语言学中使用语料库这个术语时，更倾向于包含更多的内涵，主要有采样(sampling)收集、有代表性(representativeness)、规模有限(finite size)、机器可读(machine-readable)、标准参考数据(a standard reference)等内涵特征。"

(2) 语料库就是某种语言在实际运用中的大量实例集合，这些例子可以是书面文本，也可以是语音形式的文本[2]。

(3) 语料库是根据外部原则选择的电子形式的文本或文本片段的集合。该集合能够代表一种语言，或一种语言的分支，或一种语言的变体，并可作为语言学研究使用的数据源[3]。这里外部原则(external criteria)是指通过文本的交流功能来选择文本的原则。与外部原则相对的一个概念就是内部原则(internal criteria)，具体指按照文本反映的语言细节来选择文本。

在上述的几种定义中，定义(1)使用最多，认为语料库不是简单收集的文本集合，而是通过采样收集，具有代表性，规模大小可以确定，是机器可读的标准数据。但是 Kilgarriff 和 Grefenstette[4]提出了异议，认为 McEnery 和 Wilson 混淆了"什么是语料库"和"什么是好的、适合于某项语言研究的语料库" 这两个问题，他们认为语料库就是文本的集合。

然而在具体使用中，有些研究者认为有许多文本的集合并不一定是语料库。最具有争议的莫过于万维网(WWW)了。WWW 刚出现时，人们因为不了解搜索引擎，也不清楚对 WWW 如何采样，觉得 WWW 相当神秘。因此，文献[3]指出："WWW 不是语料库，因为其维度未知且不断变化，而且 WWW 最初也不是从语言学角度来设计的。"

尽管 WWW 缺少规律，其使用也不是计算语言学熟悉的范畴，但是基于 WWW 的工作却开始不断增多，主要原因为：①公司希望他们支持的研究可以直接与所要处理的 WWW 内容相关；②传统语料库有版权限制，而 WWW 可以满足研究需求；③人们希望探索使用更多不同类型的数据和文本。

因此，一些研究者认为 WWW 是语料库，他们认为 WWW 提供了大量的不同语种、不同类型的文本，而且已成为电子文档形式（网页），是语言数据的一个自然的来源，便于语料库研究。Kilgarriff 和 Grefenstette [4]指出："WWW 是语料库，尽管它包含错误或噪声，但其用处远远多于噪声。"

1999 年，WWW 进入 ACL 会议。Mihalcea 和 Moldovan[5]提出用搜索引擎查询的点击数对词义在文本中的出现频率排序，并将其作为词义排歧系统的输入。Resnik[6]证明并行语料库可以通过 Web 来建立，因为很多网页同时存在多种语言的版本。在 2000 年的 ACL 会议上，Jones 和 Ghani[7]尝试并表明如何用 WWW 建立特定语言的语料库。在自然语言处理的很多项目中，也开始逐步将 WWW 看做语料来源，如欧洲在词义消歧方面的 EU MEANING 项目将 WWW 作为一个数据源来进行词义消歧[8]。其主要工作前提为：词在一个特定的领域内通常只有一个意义，而这个领域可以通过 WWW 来确定。又如，Chklovski 和 Mihalcea[9]将 WWW 看做语料库，并在 Word Expert 网站上收集人工进行的词义标注。而且在 IR 领域的 TREC 评测任务中，也开始引入一项 WWW 任务，对应的语料库就是 WWW 上极大规模的文本[10]。WWW 还被英国的 Sheffield 大学和美国微软公司等用作问答系统（QA）应用中的答案来源，他们将搜索引擎和语言处理技术结合了起来[11,12]。此外，研究者开始利用 WWW 来解决利用语料库所遇到的数据稀疏问题和数据不平衡问题[13,14]。

因此，可以看出，随着 WWW 和搜索引擎的迅速发展，人们对 WWW 的使用和理解也进一步加深，WWW 成了语言工作者和自然语言处理研究者的一个值得关注的新资源。

除了 WWW 以外，文献[3]认为档案和引文集合也不是语料库。首先，收集档案这些属于个人信息的文本的目的不同，这一点导致了其属性与语料库迥然不同。其次，引文是用于支持、解释某个问题的，往往很短，缺少上下文连贯性，不符合语料库中的样例特征，而且引文的准确位置对于一个语料库研究者并不重要。因此，档案和引文不属于语料库的范畴。

随着自然语言处理范围的拓宽和计算机与网络技术的发展，语料库的范围也越来越宽泛，而且有很多研究者也开始建立适合自己研究的小型语料库。因此，上述具有争议的资源，可能在某个应用领域中恰恰是关键的语料数据资源。

综上所述，本书认为语料库就是文本的集合。

1.1.2 语料库的作用

1989 年在温哥华召开的 ACL 会议上，语料库开始进入计算语言学领域。但那时的语料库大而杂乱，明显缺少理论完整性，因此许多人对它们在学科中的作用持怀疑态度[4]。争论一时四起，人们还不清楚语料库工作是否能成为领域中可接

受的部分。直到1993年在计算语言学的一次使用大规模语料的专题讨论中，语料库才获得了高度成功，至此语料库与计算语言学的关系得到了极度融合。

(1) 对语言学研究的作用。有了语料库人们可以通过计算机来研究语言，利用语料库所反映出来的语言事实对语言的某个方面进行研究，并提出新的观点或理论。例如，语料库如果事先标注好了，则有助于许多的自动分析工作。又如，可以用标注了词性的语料建立词频表或带有语法分类的词频字典，在这个表中可以区分 leaves(动词)和 leaves(名词)这两种词形相同但词性不同的情况，并给出不同的频次。因此，语料库为语言的研究提供了新的研究手段和研究素材。

(2) 对自然语言处理的作用。事先标注好的语料对自然语言的后期处理非常重要。例如，要想进行自动句法分析，必须具备句子中词和词性的信息，因此，分词和词性的标注可以看做是句法分析的第一个阶段，在此基础上进行深入的句法分析。又如，在口语合成中，合成器需要事先具有词的发音信息，而这些发音信息往往与词和词性相关，如“和”字在“暖和/a”和“调和/v”这两个词中发音是不同的。随着语料库的发展，人们开始利用语料库对自然语言进行大规模的调查和统计，利用词的权重、词的组合或搭配等建立统计语言模型，研究和应用基于统计的(statistical-based)语言处理技术，推动了信息检索(information retrieval，IR)、文本分类(text categorization，TC)、文本过滤(text filtering，TF)、信息抽取(information extraction，IE)和机器翻译(machine translation，MT)等应用的进展。例如，在口语领域，语言模型被用来预测哪一种词组合可以用来解释声音流；在 IR 领域，语言模型被用来决定哪个词更能反映主题；在 MT 领域，语言模型用来识别好的候选译文。具体来说，语料库在自然语言处理中的作用有利于人对数据及需求进行分析；提供了一个标准答案来评价系统结果；为开创自然语言处理的机器学习方法奠定了基础。

同时，语言信息处理技术的发展也为语料库的建设提供了支持。从字符编码、文本输入和整理、语料的自动分词和标注，到语料的统计和检索这些自然语言处理的技术都为语料的加工提供了关键技术支持。

1.2 语料库的建立

语料库分为标注语料库和未标注的语料库。通常我们将未经标注的语料库称为生语料库(raw corpus)，标注过的语料库称为熟语料库。同生语料库相比，标注语料库通过丰富生语料的信息，使语料库在研究中更加有用。例如，进行了词性标注的英文语料库——Brown 语料库、LOB 语料库和英国国家语料库(British national corpus，BNC)为世界各地的研究者利用词性标记信息和对应的原始语料提供了良好的素材。

本书将着重介绍利用语料库的标注技术来建立和加工语料库，即建立标注语料库。

1.2.1 什么是语料库标注

语料库标注(corpus annotation)就是为语料库增加一些语言学信息[3]，有时称为语料库的加工。最普通的一种标注就是在原始语料库中增加一些标记(tag/label)表明文本中每个词的词类，这就是所谓的词性标注(part-of-speech tagging，POS tagging)。

根据文本分析的不同层次，语料库有以下几种标注。

分词标注(word segmentation)，就是在词定义和词表的基础上，运用一定的手段和技术在语料库中加入词的切分信息，是中文、韩文、日文等东方语言分析中特有的一个阶段。

词性标注，就是在原始语料库中增加一些标记表明文本中每个词的词类。这是最普通的一种标注。因为句法分析中需要用到 POS 信息，所以词性标注有时被看做是句法分析的一部分。例如，Brown 语料库(第一个百万词汇的美式英语语料库)和 LOB 语料库就是运用一系列类似的标注手段和技术，由 Brown 大学发起进行，后在 Lancaster 进一步开发为词性标注语料库。又如，BNC 语料库也是著名的词性标注语料库，和其相关的两个标记集合 C5 和 C7 非常有名。

句法标注(syntactic annotation)，就是按照不同的句法分析理论或句法分析模型给文本中的每个句子增加标注句法信息。句法标注主要有完全句法标注、浅层句法标注和依存句法标注。例如，SUSANNE 语料库是最早的、采用良好开发策略建立的完全句法分析语料库[15]。宾州树库(Penn Treebank)是目前最有影响的完全句法分析语料库，与其相关的成分结构分析策略(constituent structure analysis scheme)也成为语法中最有影响的策略之一[16]。依存句法标注主要采用的是依存语法模型(dependency grammar model)，尤其是 Karlsson 的约束语法模型(constraint grammar model)[17]，而不是成分结构模型(constituent structure model)进行句法标注。浅层句法标注也叫做部分句法分析或语块分析，是指对句子中的结构相对简单的成分，如非递归的名词短语、动词短语等进行标注。

语义标注(semantic annotation)，就是为原始语料库增加一定的语义信息。按照语言分析和处理的单位，语义信息可以分为词义、句义和篇章义信息。这里主要指词和句子级的语义信息。例如，SemCor 语料库就是由普林斯顿大学的 Miller 于 1993 年负责完成的词义标注语料库。该语料库采用 WordNet 1.6 语义标注体系，在完成词性标注后的 Brown 语料库上进行标注而成[18]。

语用和篇章标注(pragmatic/discourse annotation)，就是为原始语料库增加语用和篇章关系信息。例如，对文本中起连贯和衔接作用的信息进行标注。语料库

标注中，很难清晰地区分语用和篇章分析。又如，宾州语篇树库(Penn Discourse Treebank，PDTB)就是一个大规模的篇章关系和论元标注树库，规模超过100万个词[19]。

多任务标注，就是对语料库进行两种以上的标注[20]。现代的NLP系统经常要使用不同任务下的标注结果。例如，机器翻译系统需要使用句法树标注和名实体标注的特征。对于这样的应用，我们显然需要同一语料库的不同标注。一个为不同任务标注的综合语料库是非常有用的。事实上，已经有人创建了不同标注任务的综合语料库，尽管每个标注任务的实现彼此相互独立。在宾州树库的《华尔街日报》部分，每个句子都标注了词性、句法树。后来，这部分语料又进行了语篇标注形成了宾州语篇树库。同时，语料中的谓语论元也被标注，该语料库被称为宾州命题库(Penn Propbank)。

1.2.2 语料库标注的原则

在语料库标注中，预先提出一套好的操作标准和原则非常重要。

语料库标注中常见的、有实际意义的原则主要有6项。

1. 标注应该是可分离的(separable)，而且不能引起原始语料的信息损失

由上可知，标注是研究者根据研究目标选择的附加在语料库中的信息，是可选择的附加物。因此，标注信息应该很容易从生语料分离出去，即标注之前的生语料应该能够被轻易抽取。另外，由于并不是所有用户都认为标注是有用的，所以标注不能引起原始语料数据的信息损失。

2. 提供详细清晰的文档

建立语料库时，应该提供其构成文档的相关说明。类似地，标注语料库应该提供清晰详细的标注文档。为了帮助用户准确地了解可获得的信息，标注文档应该包含以下信息。

(1) 标注的方式、地点、时间和标注者信息以及使用到的计算机工具，修改后的各种版本。

(2) 使用的标注大纲。标注大纲是根据语言学分析得到的对标注的解释，可以使用户知道各种语言现象是如何标注处理的，在语料库标注中非常重要。

(3) 使用的编码方式。这里的编码方式指标注使用的符号，编码应该与原始语料有差别。

(4) 标注结果的好坏。语料库的标注者总是声称他们的标注是好的，或具有某种优势或特点。因此，需要根据客观情况对语料做出客观全面的评价。准确率(accuracy)和一致性(consistency)是两个常用的衡量语料标注质量的标准。语料

库的标注者应当提供相关的标注质量信息。

3. 标注体系在理论上应尽可能保持中立

任何一种标注都预先假定了一个标注体系。但是，语言学可能对这种体系有异议。例如，在汉语中是否承认动词的兼类就存在明显的争议和分歧。在这种背景下，语料库标注不可能“绝对正确”。但是在很大范围内存在认识的一致性。例如，人们都同意，在一篇文本中一些词是名词，一些词是动词……只不过在少数情况下会产生分歧。

因此，要认真地对待语言学，尽量使标注大纲采用的标注体系基于语言学达成共识。这一点对语料库是否适用于更多用户，实现标注语料库的可重用性具有重要意义。如果标注者采用的标注体系面向某个特定理论，则标注语料库的共享性和重用性将会受到影响。

4. 标注应该尊重实际

这个原则可以看做是第 3 项原则的补充。前人在语料库标注中已经做了很多工作，一个新的项目应该利用这些工作。然而，一个新项目的目标或多或少都与以前的项目有些不同，如面向新的领域、采用了不同的数据、面向不同的语言等。因此，必须在以前工作的基础上，进行相应的调整才能建立新的语料库。

5. 实现标注的重用性(re-usability)

标注的语料库应该是一个可以被人类和科学研究者重用的、共享的资源。随着自然语言处理技术的发展，语料库的一些标注工作可以自动完成，但是不可能完全准确。诸如语义标注之类的标注工作还未达到完全自动化，因此自动标注的语料库需要人工后处理，有时甚至是大规模的。这种后处理往往需要花费相当多的人力和时间。因此，要保留语料库的各种标注，以便被更多的用户共享。

6. 实现语料的多功能性(multi-functionality)

进一步考虑重用性问题，会注意到标注可以有不同的目标或应用，即标注是多功能的。例如，经过词性标注的语料库，其中的词类信息可以用于词典编撰、语法分析、建立词频表、口语合成等应用中。语料库建立人员一般很熟悉这些功能，会预测语料库的用处。但是，语料库的实际用处往往会比语料库建立者想到的多。

1.2.3 建立语料库需要考虑的几个问题

语料库对研究者的用处取决于这个语料库在建立的过程中是否经过良好设计并严密实施。因此，对于语料库建立者，掌握和了解必要的操作原则至关重要。一