汉语语料库的建设及应用

郭曙纶 著

上海外语教育出版社
外教社 SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS
www.sflep.com

汉语语料库的建设及应用

郭曙纶　著

上海外语教育出版社
外教社 SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS

图书在版编目(CIP)数据

汉语语料库的建设及应用 / 郭曙纶著. —上海：
上海外语教育出版社，2011
ISBN 978-7-5446-2379-7

Ⅰ. ①汉… Ⅱ. ①郭… Ⅲ. ①汉语—语料库—研究 Ⅳ. ①H1

中国版本图书馆 CIP 数据核字(2011)第 105885 号

出版发行： 上海外语教育出版社
（上海外国语大学内） 邮编：200083
电　　话： 021-65425300（总机）
电子邮箱： bookinfo@sflep.com.cn
网　　址： http://www.sflep.com.cn http://www.sflep.com
责任编辑： 周岐灵

印　　刷： 上海信老印刷厂
开　　本： 890×1240 1/32 印张 7.5 字数 228 千字
版　　次： 2011 年 10 月第 1 版 2011 年 10 月第 1 次印刷

书　　号： ISBN 978-7-5446-2379-7 / H · 1094
定　　价： 26.00 元

本书内容提要(前言)

本书的读者对象主要是对汉语语料库建设及应用感兴趣的大学生、研究生及相关领域的研究者。愿这本小书能为他们提供一些实例与启发。

本书探讨的内容大致分为两部分。第2章和第3章讨论汉语语料库的建设问题,包括汉语切词词典的加工、加工规范等,是对汉语语料库建设中碰到的一些有别于英语语料库建设的理论问题进行探讨,是面向中文信息处理的研究。第4章到第7章是本书的主体,讨论汉语语料库的应用问题,包括基于语料库的汉语字词统计与分析、语料库技术在对外汉语教材研究中的应用,是基于中文信息处理的研究。

第1章是语料库概述,介绍语料库的相关知识、国内汉语语料库的建设情况等。

第2章着重讨论汉语语料库加工中所用切词词典的加工问题,具体讨论了词短语和短语词及其应用的问题、结构化词表的加工问题等。

第3章讨论汉语语料库加工规范的问题,着重讨论词类标记和切词标注的规范,并具体讨论了汉语人名的标注问题。

第4章是基于语料库的汉语字词统计与分析,提供了HowNet的词语统计、结构化词表的统计、网络语料的用字统计、800万标注语料字词统计、500万标注语料校对记录统计等相关数据及分析。

第5章到第7章是讨论汉语语料库在对外汉语教材研究中的应用,其中第5章是基于语料库的对外汉语教材超纲词问题的专题研究,首次从多个角度比较全面细致地探讨了对外汉语教材中的超纲词问题,第6章和第7章分别讨论语料库技术在对外汉语学习词典编纂和在对外汉语教材编写中的应用问题,着重讨论如何利用语料库技术来为对外汉语教学中的教材编写与词典编纂服务。

目　录

第 1 章　语料库概述

1.1　语料库的定义

语料库(corpus 或 corpora, corpuses)研究的出现与语料库语言学(corpus linguistics)的诞生是语言学和计算语言学(computational linguistics)发展的结果,也是信息社会的需要。

计算语言学是随着计算机科学的诞生与发展而兴起的一门边缘科学。计算语言学有广义和狭义两种理解。广义的计算语言学几乎包括了与计算机(或计算机科学)和语言学相关的所有方面;狭义的计算语言学一般等同于自然语言理解,也就是通过建立形式化的计算模型来分析、理解和处理语言。不论是计算语言学还是自然语言理解都是边缘性学科。自然语言理解具体到汉语的研究中,也就是汉语的自然语言理解研究,人们通常又称之为中文信息处理或汉语信息处理(Chinese Information Processing)。

最初的自然语言处理系统,一般是基于规则的。以特定的例句或句型为基础,总结规律,逐步完善,以期实现自然语言理解。这也反映了传统语法——规范语法的语言观,即语言用法有正确和错误之分,一个句子要么是正确的,要么是错误的。

随着语言学研究的发展,人们的语言观也发生了变化,由传统的规范语法转向结构主义的描写语法;到了五六十年代,以乔姆斯基的转换生成语法理论为代表,又由描写语法转向解释语法——从形式上来解释一个正确的句子是如何推导出来的。到了八十年代,认知语言学又试图从人的认知上来解释语言形式与语言意义的对应问题。从规范到描写再到解释,这说明了一种语言观的变化。正像有人说过的"存在即合理"一样,一种语言现象,未必是"是"和"非"的问题,有时候正误只是一个程度的问题。

随着计算机硬件和软件技术的发展与普及，近年来因特网发展迅速，大量的语言信息需要及时地处理。语言信息的大量增加，使得原有的基于规则的自然语言处理方法变得不能适用。因为数量的急剧增长，语言现象又是如此复杂多变，在有限的语言材料基础上归纳出来的有限规则根本不能涵盖所有的语言现象。

这样，如何收集、整理语言材料就成了一个相当重要的问题。

任何科学的研究都离不开对研究对象的收集与整理，语言学研究也不例外。汉语的语言学研究向来很重视语言事实的挖掘，有所谓“例不十，法不立”的说法。语言学家，尤其是词典编撰家们，曾通过制作大量的卡片来收集、整理语言材料。然而，这种靠手工收集、整理语言材料的办法，显然不能适应现代信息社会的需要，通过计算机来收集、整理语言材料就成为最好的选择。这样，语料库的诞生就成为必然的了。

语料库就是一个由大量在真实情况下使用的语言信息经过科学的收集和组织而集成的专供研究使用的资料库。传统上，语言学家用“语料库”这个术语表示作为语言学研究基础的、大量自然出现的语言数据。这些语料库可以由书面语和口语的样本组成，并通常用来表示一种特定的语言或语言变体。由于电脑语料库容量大，资料真实，信息提取准确，因此，语言学家借助语料库可以从多方面多层次描写语言并验证各种语言理论和假设。

语料库并非语篇的简单堆砌或集合，它应具有以下几个基本特征：

（1）样本代表性，是指某个语料库收集的样本应该能够代表该语料库所涵盖的特定语言或语言变体，即语料样本应该具有代表性。这就要求语料的收集需要根据语料库的类型按照事先设定的原则来进行。

（2）规模有限性，是指语料库的规模无论多么大，总是存在一定的限制，跟无限丰富的实际语言相比，它总是有限的。

（3）机读形式化，是指语料文本必须是以电脑可读的形式存在，现在大多数的语料文本是以文本文件的格式存储的，也有用 XML 文件格式存储的。这与过去用卡片记录存储的形式不同，主要是现在方便电脑处理，容易检索和统计。

语料库有不同的加工层次，加工的语料库一般指标有语言学标记的语料库。未加工的语料库称为“生语料库”，加工过的语料库称为

“熟语料库”。使用标注正确率高的熟语料库更有利于对自然语言的研究。

1.2　语料库语言学

语料库语言学是以语篇(text)语料为基础对语言进行研究的一门学科。过去,语料库中的材料由人工收集和整理;今天,由于使用了计算机的先进技术,语料库建设的效率和规模都大大提高,为语料库更为广泛的应用打下了坚实的基础。

关于语料库语言学,顾曰国在《语料库与语言研究——兼编者的话》中有过一段精彩论述:“语料库语言学”这个术语其实有两层含义。一是利用语料库对语言的某个方面进行研究,也就是说“语料库语言学”不是一个新学科的名称,而仅仅反映了一个新的研究手段。二是依据语料库所反映出来的语言事实对现行语言学理论进行批判,提出新的观点或理论。只有在这个意义上,“语料库语言学”才是一个新学科的名称。不过从现有的文献来看,属于后一类的研究还很少。本书讨论的语料库语言学也基本上是指前一类的研究。

语料库语言学的这一类研究包括两方面的内容:一是对自然语料进行加工、标注,二是对未经标注或已经标注的语料进行语言研究和应用开发。这两个方面的内容本书都会涉及到。本书的第 2 章和第 3 章讨论汉语语料库的建设,第 4 章至第 7 章讨论汉语语料库的应用。

语料库语言学的出现,有助于解决目前存在的一些问题。(1) 解决以偏概全的问题。因为语言学家所关注的往往是一些并不常见的特例(在许多时候,特例往往受到更多的关注,因为人们总是对习以为常的现象视而不见)。由此总结出来的规则总是有限的,不能刻画所有的语言现象,甚至不能刻画一些常用的、基本的语言现象,因而很难用来处理真实文本。(2) 解决规则相互矛盾的问题。自然语言复杂多变,有限的规则难以处理大规模的真实文本。规则的增多又很难保证相互之间的相容性,可能导致规则之间发生矛盾和冲突。

语料库语言学通过对大规模真实语料的统计、分析来发现、归纳自然语言的规律,提取语言知识(可称之为语言目标知识,或简称为目标

知识)。要充分发挥语料库的作用,除了要保证语料的真实可信以外,还必须对生语料进行深加工。而要加工语料又必须先赋予计算机一定的语言知识(可称之为语言源知识,或简称为源知识)才能提高语料库加工的正确率与效率,更好地提取更多的语言知识。一般说,最初的源知识是靠语言学提供的,目标知识则是计算语言学提取的。提取后的目标知识又成为下一次运用的源知识。如此循环往复,呈螺旋式上升。

早期的自然语言处理系统,通常是基于规则的系统,其本质上都是解决"是"和"非"的问题,难以解释复杂多变的自然语言。即使增加规则数量,也无济于事。随着规则数量的增加,规则之间常常发生矛盾和冲突,不能保证语言学规则之间相容,更何况要获取语言学知识还有与语言学相关的世界知识还是一件非常困难的事情。正是因为这些原因,语料库语言学随之兴起,人们试图从大规模语料库中获取颗粒度较小的语言知识或者说更细致的规则来支持大规模真实文本的自然语言处理系统。近年来,语料库语言学有了很大的发展,其原因主要有两条:

(1) 计算机技术的迅速普及和应用为语料库语言学的发展提供了物质基础。现在获取电子文本不像当初那样必须通过人工录入或者扫描并且 OCR 识别,因特网上有许多现成的电子文本可供选择。有相当一部分研究者专门研究因特网上的语料。

(2) 对原有研究方法的深刻反思。转换生成学派等对语料库语言学的批评和否定,经过 20 年的实践验证,有的是错误的,如指责计算机分析语料是伪技术;有的是片面的,如对语料数据价值的否定;有的则是正确的,如乔姆斯基关于自然语言无限性的观点。对于乔姆斯基倡导的唯理方法,人们经过跟从、应用和反思之后,也逐步发现其不足,如其不可验证性等。1994 年 IBM 的 Adam L. Berger 等人发表了题为"The Candidate System of Machine Translation"的文章,初步的研究工作使译准率超过美国著名的 SYSTRAN 机器翻译系统,使国际计算语言学界为之震动。

1.3 语料库的类型

根据其选择的语料内容、选择的方式以及建设目的的不同,语料库

的类型可以有不同的划分方法，比如通用语料库与专用语料库、异质语料库与同质语料库、动态语料库和静态语料库、第一代语料库与第二代语料库、书面语料库与口语语料库，等等。下面对这些常见的语料库类型做简要介绍。

通用语料库（general corpus）：又称一般语料库，是文本的集合，为了保证收集的语料具有广泛的代表性，对语料采用系统的办法进行采集，用于事先未指定的语言学研究。如 Brown 语料库、LOB 语料库。前者是当代美国英语语料库，后者是当代英国英语语料库。SEU Corpus 也是一个通用语料库，它已被用于语法研究。通用语料库应有“平衡性”（balanced），即语料库要收集不同类型、不同领域的包括口头的和书面的文本。通用语料库还可称为系统语料库或平衡语料库，有时还被称为核心语料库。

专用语料库（specialized corpus）：又称专门用途语料库，指用于某种特殊研究的语料库。如 Helsinki Corpus of Historical English，用于研究古英语；JD 学术英语语料库，用于研究学术英语。它又可分为方言语料库、区域性语料库、非标准语料库和学习者语料库等。如由广东外语外贸大学桂诗春教授和上海交通大学杨惠中教授牵头开发的中国学习者英语语料库 CLEC（Chinese Learner English Corpus），就是一个学习者语料库。它还可分为书面语料库和口语语料库。如 the London-Lund Corpus、the Corpus of Spoken American English 就是口语语料库。口语语料库是研究口语特征的重要工具，如语音语调的规律，其研究成果在语音合成中有重要应用。口语语料库的建设涉及口语真实语料的采集及语音转录，工作量极大。

异质语料库（heterogeneous corpus）：大量收集文字材料，尽可能广泛地接受各类材料而没有事先制定任何选材原则。收藏的文本在格式和内容上各异，而存储的格式和原来的出版物完全一样。例如牛津文本档案库 OTA。

同质语料库（homogeneous corpus）：一般用于专业语料库，例如美国政府的 TIPSTER 项目的语料库，专门用于收集军事文本；还有个别作家作者语料库也属此类。

动态语料库（dynamic corpora）：又称监控语料库，用于观察现代语言的变迁，如 COBUILD 语料库。与此相对的是静态语料库（static

corpora)，只收集某一固定时期的共时语言材料，语料库建成后，就不再扩充。

第一代语料库指的是20世纪60年代到80年代所建成的一批语料库，这个阶段是以电子语料库的兴起为主要特征。第一代语料库规模相对比较小，大多只在百万词级，如SEU Corpus(1959, pre-electronic corpus), Brown Corpus (1964, Brown University Standard Corpus of Present-Day American English), LOB Corpus(1970-1978, the Lancaster-Oslo/Bergen Corpus of British English), LLC 口语语料库(1975, London-LUND Corpus)。在这一阶段，语料库的发展以容量不断增加和种类的不断扩展为主要特征。

第二代语料库指的是从20世纪90年代中期开始建成的上亿词的大型语料库，如，COBUILD(1997, 3亿词), Longman Corpus Network(它包含三个主要的语料库：Longman/Lancaster Corpus〈LLELC〉, Longman Spoken Corpus〈LSC〉和Longman Corpus of Learner's English〈LCLE〉), British National Corpus(1995, 1亿词), International Corpus of English (1996, ICE)。

平行语料库(parallel corpus)：把两种语言中完全对应的文本(如法律文件)输入计算机，通过分析对比找出两者对应关系，可用于机器翻译研究。

1.4 语料库的规模

关于语料库规模的问题，有人认为语料库越大越好。其实要讨论语料库规模的大小问题，先要看语料库是给谁用的。

如果是给语言学家用，那么对语料库规模的基本要求是语料库覆盖绝大多数语言现象，并且每种语言现象出现一定次数以上。下面是一个粗略的估计：

对汉语来说，如果要研究一个10万词的词典(《现代汉语词典》有6万多词，但是有许多能见字知义的词没有列出，大多数中文信息处理用的词表，多在10万词左右)，每个词平均有2个义项，要求每个义项出现5次，则语料库中需要包含100000 * 2 * 5，即一百万个左右的句

子。若句子的平均长度是 30 个字,则语料库要有 3000 万字。

如果是给计算语言学家用,那么主要考虑词性(状态)转移概率和词性到词的转移概率。因为词性数远远小于汉语中词的数目,所以,词性(状态)转移概率不会受到数据稀疏问题的影响,用较小规模的语料库就可以获得较精确的结果。而词的转移概率会受到语料库数据稀疏的影响,因为语料库中有很多低频词。我们对 1995 年人民日报(2000 万字)进行切分,统计发现:我们切词词表的 10 万词中有 2 万词出现次数为 0,1 万词出现次数为 1,出现次数超过 5 次或 5 次以上的词只有 45000 词左右。

在语料库规模问题中有一个"水桶原理",即决定语料库规模的因素是使用频率低的词。理论上,语料库应具有使绝大多数低频词在语料库中出现次数超过某最低频次(如 30 次)的规模。从这个角度计算,语料库应具有 50 亿 ~200 亿的规模才能达到这一目标。

1.5　语料库的加工

语料库的加工可分为两个方面。一方面是语料库的标注,就是给语料库的某些单位(词、句、段落、篇章等)加上表示对这些单位的某种层次的"理解"的知识信息(标记符);另一方面是语料库的知识获取,指通过对语料库的处理,获得语料库所代表的普遍现象的语言知识。它独立于语料库中某特定单位,反映了语言中的某种普遍规律。

一、加工的层次与原则

语料库有不同的加工层次。对语料库可以进行下列加工并形成不同加工层次的语料库,对语料库的加工还包括"预处理"。

被加工的语料库可以包含文本的全部(full-text corpora),也可以从文本中抽取一部分构成。

1. 索引(concordance)

逐词索引:提供在语料库中出现的每个词每次出现的相关信息。逐词索引记录了每个词在语料库中每次出现的相关位置,据此就可以提供每个词每次出现的上下文信息。

关键词索引:提供出现指定关键词的文本、段落等信息。

就汉语而言,可以是以字为单位的逐字索引和关键字索引。

2. 主题标引(subject indexing)

主题标引是指对文本内容进行主题分析、赋予主题词标识的过程。

3. 切词(segmentation)

切词就是从信息处理需要出发,按照特定的规范,对汉语按切词单位进行切分的过程。换句话说,就是将连续的字串按照一定的规范切分并重新组合成词串的过程。

4. 词性标注(POS tagging)

词性标注就是对已经切词的语料中的每一个词赋予一个词性标记。词性标注与切词经常是由同一个系统来处理。词性标注的主要问题是兼类词的处理,还有一个问题是未登录词的处理。

5. 句法成分标注(parsing)

句法成分标注就是平时常说的树库加工,对已经标注了词性的文本标注上句法成分的信息,也就是标注上主语、宾语、谓语、定语、状语、补语等是什么,一般同时标注上这些句法成分是由什么样类型的短语(如名词短语、动词短语、形容词短语、介词短语等)充当的。

6. 语义信息标注(semantic tagging)

语义信息标注可以有不同的理解。一种是词义标注,一般在标注词性之后进行,给每个词语标注上词义信息,往往是义项标注,也就是通常所做的词义消歧(WSD, Word Sense Disambiguation)。一种是语义角色标注,一般在句法成分标注之后进行,给每个句法成分标注上语义信息,如施事、受事等。

7. 语用信息标注(pragmatic tagging)

语用信息标注,就是对文本标注上相关的语用信息,如话题、述题、话轮、省略成分等,为语用分析服务。它可以在生语料的基础上进行,也可以在熟语料的基础上进行。

8. 特定语言模式的标注

特定语言模式的标注,就是根据研究需要,标注上研究者所需要的相关信息,如未登录词的标注、专有名词的标注、最大名词短语的标注等。

加工、标注语料库时应遵循一些基本的加工原则,对此,G. Leech

曾提出了有标记的语料库应满足的七条基本原则：

（1）所作标注可以删除，恢复到原始语料。这主要是为了保证语料的充分利用。语料库可用于不同的目的，可能需要采取不同的标注方法。

（2）所作标注可以单独抽出，另处存储。这一原则实际上与第一条原则基本一致，由此可知，语料库中语料的标注应该最大限度地增加语料使用的灵活性。

（3）语料库的最终使用者应该知道标注原则和标注符号的意义。因此，大多数语料库都配有详细介绍标注原则和标注符号意义的手册，供使用者参考。

（4）在语料的使用说明中，应该说明标注是何人用何种方法所作。如，是人工标注还是计算机标注，是一人标注还是多人标注。

（5）应向用户声明，语料标注并非绝对无误，它只是一种可能有用的工具。不论是人工标注，还是计算机自动标注，或者两者的结合，都有可能产生标注的分歧甚至错误，因为标注的过程实际上是对语料中语言单位的特征进行解释的过程，不同的人可能会有不同的解释结果。

（6）标注模式应不依赖于某一家之言，尽可能中立。在标注的过程中，为了方便语料库的使用，标注应该采用综合的使用范围广泛的语法理论，而不是按照使用范围狭窄的某一特定的语法理论。

（7）任何标注模式都不能作为第一标准。即使有，也只能通过实践在大量比较中得到。目前，世界上还没有一种被普遍接受的标注模式。

这七条原则，概括起来就是最大可能地方便加工者和使用者。语料的加工和使用始终是一对矛盾。正如丁善信所说："从用户的角度，语料标注得越详尽越好，而标注者则还需考虑标注的可行性。因此，任何标注模式都是二者之间求得的一种妥协的产物。"

二、加工所需的知识获取

1. 频率统计

这种统计是不基于语言模型的或是基于简单的语言模型的。统计的方法是简单的计数。这些统计包括：

基于 n 元模型、基于类的 n 元模型的参数统计；词频统计；双语对

译频率统计;共现统计;搭配对统计;基于 n 元模型的词义辨识等。

2. 基于概率语言模型的参数估计

这类知识获取方法的语言模型一般比较复杂。获取的概率知识不能通过简单的计数方法获得。参数估计一般是一个迭代的过程,具体可以包括:

马尔可夫模型的状态—状态转移概率,状态—观察转移概率的参数估计;上下文无关文法的产生式的概率的参数估计;基于有限状态自动机的机器学习;基于 IBM Brown 模型的统计机器翻译模型参数的估计;歧义消解知识的获取;词性标注中的消歧;句法分析中的消歧;词义消歧;词的聚类等。

3. 非概率型知识的获取

一般是通过"泛化",把对特殊情况适用的规律推广到一般情况,并在语料库中进行检验。如,语法知识的获取、词汇选择知识的获取、基于错误驱动的词性标注规则知识的获取。

三、加工的主要技术手段

在语料库加工的过程中运用到的主要技术手段包括:1. n-gram 模型;2. 马尔可夫模型;3. 概率上下文无关文法模型;4. 统计机器翻译模型;5. 互信息;6. 熵;7. 聚类;8. 共现统计;9. 分类;10. 平滑方法(解决数据稀疏);11. EM 参数估计方法;12. 韦特比(Viterbi)参数估计方法;13. 动态规划求解最优解的方法;14. 有限状态自动机理论和模型。

这些技术手段的具体操作都需要比较专门的知识与技术,在许多参考文献中已经有较多的论述,因不是本书的重点,这里就不展开了,感兴趣的读者可以阅读姚天顺等的《自然语言理解——一种让机器懂得人类语言的研究(第 2 版)》、王建新的《计算机语料库的建设与应用》、Christopher D. Manning 和 Hinrich Schütze 的《统计自然语言处理基础》、Daniel Jurafsky 和 James H. Martin 的《自然语言处理综论》、宗成庆的《统计自然语言处理》等,尤其是宗成庆的《统计自然语言处理》提供了许多技术细节的讨论。

语料库加工的主要困难有三个方面的问题:一个是数据稀疏问题,一个是歧义问题,还有一个是语言模型本身的精确度问题。围绕这些

问题的详细讨论，也请参考上面提到的专著。

1.6　语料库的应用

1. 词典编纂

语料库最直接的用途就是为编纂字典提供大量真实准确的例句。例如，根据 COBUILD 语料库编写的词典就有许多种。在语料库语言学这个学科问世以前，词典编纂一直依靠凭经验收集的语料来进行。现在，语言学家对语言的研究从传统的直觉经验方法转向基于统计的方法，运用语料库或其他机器可读的文本资料库，能在几秒钟之内从数以万计词次的文本语料中检索出有关一个词或词组的所有例句。

南京大学近年来开发了 NULEXID 语料库暨双语词典编纂系统，涉及英汉两种语言，在《新时代英汉大词典》的编纂过程中已经起到了重要作用。

2. 语义学研究

语义学是语言学的一个分支学科，它研究语词的意义和意义的变化。特别是着重从社会和历史的角度去探讨词语意义变化的原因和规律，例如语词意义的扩大、缩小、升格和降格，词义的转移，词与词之间的语义关系，语义与句法结构的关系等等。Dieter Mindt 证明了如何利用语料库为解释语言词语提供客观的标准。Mindt 指出，在语义学中，词语的意义常常是根据语言学家的直觉来描述的（理性主义的方法）。他认为，语义上的特征与篇章中可观察到的独特的语境相联系（如句法特征、词法、韵律特征等），通过利用语料库中的语料考虑语言实体的整个环境，就能对一个特定的语义上的特征作出客观的判断。

语料库在语义学研究中的另一个作用是更加牢固地确定了模糊范畴和渐变的概念。

3. 语言教学

语料库中的语料是人们实际运用的语言，所有的材料都取自真实的书面语和口语文本，提供的是语言实际使用的客观例证。对这种材料进行分析，有时可以发现现有的语言教学材料中存在的问题。Graeme Kennedy 调查了在 ESL（English as a Second Language）课本中