第十章

语言和计算机

计算机是根据一组存储在内存里的指令来处理数据的电子设备，它可以在很短的时间内处理复杂的任务。随着微机的普及，计算机为个体用户提供了越来越多的方便，我们今天已生活在一个以计算机为主导的社会，并使用“计算机操作能力”（computer literacy）这个词组，特指那些能使用计算机并具有大量计算机软件知识和技巧的人。在本章里，我们涉及的范围仅限于语言和计算机之间的关系，这种关系构成一门新的学科，即计算语言学。

计算语言学可以看作是应用语言学的分支，即通过计算机处理人类语言（Johnson & Johnson，1998/1999）。计算语言学包括：对语言数据进行分析，建立一个序列，通过它，语言学习者可以获得各种语法规律或某一特定词项的出现频率（程序教学）；人工言语的电子生成（言语合成）和人类语言的自动识别；它也包括不同自然语言之间的自动翻译和语篇处理；人与计算机的交流。本书还不能够覆盖这个新学科的每个层面，以下题目是我们目前关注的重点。

10.1 计算机辅助语言学习（CALL）

10.1.1 CAI/CAL 和 CALL

当讨论计算机辅助语言学习（CALL，computer-assisted language learning）的时候，我们首先需要区分 CAI（computer-assisted instruction，计算机辅助教学）和 CAL（computer-assisted learning，计算机辅助学习）。

计算机辅助教学，就是在教学过程中使用计算机（Richards，et al.，1998）。这包括：

a. 通过计算机按顺序显示学习项目的教学程序。学生在计算机上作出回答，计算机显示该回答正确与否。

b. 通过计算机监测学生的学习进程，指导学生选择合适的课程、材料等。这又称为计算机管理教学。

和 CAI 相对应，CAL 是计算机辅助学习。CAI 注重教师方面的问题，CAL

强调在教和学两方面使用计算机，通过学生自己的推理和实践，帮助学习者达到教学目标，这反映了最近提出的自主性学习的理念。第一种开发出的 CAL 程序体现了和程序化教学相类似的原则。计算机指导学生一步一步地完成学习任务，并用提问的方式来检查学生的理解。根据学生的正确反应，计算机向学生给出下一步练习，提供新的学习材料。在最近的 CAL 课件中，学生能够与计算机互相交流，探讨一个主题或问题时也能够执行更高级别的任务。

如果说 CAI 或 CAL 是处理一般的教与学，那么 CALL（computer-assisted language learning）则是用来处理语言教学的。在 20 世纪 80 年代，有些坚持老传统的教授和讲师都鄙视 CALL，认为“我已经学习和教授英语二三十年了，这证明没有计算机，我一样可以是一位合格的学习者，现在我同样是一位合格的教师。我们为什么要在语言学习中为计算机操心呢？”今天有不少人已改变了看法。

CALL，计算机辅助语言学习，特别指将计算机运用到第二语言或外语的教学中去。根据 Richards 等人（1998），它可以采取以下形式：

a. 通过其他媒介进行与学习并行的活动，但使用计算机设备（例如，使用计算机来显示阅读语篇）；

b. 以印刷或教室为基础的课堂活动的延伸或改变（例如，教学生写作技巧的计算机程序，它帮助学生逐步展开一个题目或主题，并从词汇、语法和主题展开等方面来检查学生的作文）；

c. 对 CALL 而言的独特活动。

10.1.2 CALL 的发展阶段

CALL 的发展过程中主要有四个阶段。

阶段一 在这个阶段，计算机限制在研究机构中运用，其主机非常大。它们是为大规模的教学方案设计的，例如伊利诺伊大学的 PLATO（自动化教学操作的程序逻辑）。无论是语法解释，还是听说教学，教学方法通常都是传统的。在进行语法解释时，学习者看到显示器上出现了一个要点的解释，随后还有实际材料。在进行听说语言教学时，以书面形式出现的语言点被一遍又一遍地操练。存在的问题是人们只能在某些大学网址的终端上接触这些程序。

阶段二 小型计算机出现了，比以前的要便宜。这使得新一代程序的产生变为可能。这些程序能被储存在磁带或软盘上，并且整个系统都是便携的。这些程序有的是多用途的、实用的以及以学生为定位的，而不是设想大块的语

言模式或教学理论。

阶段三 学习主要不是通过语篇自身的语言，而是通过认知的处理问题的技术和小组中学生之间的相互交流。因此计算机被用作学生之间交流的激发点。结果，对每个学生而言，计算机不再是个人资源，它逐渐被视为小组学习的焦点，这就使得许多被称为“交流性”的活动变得可行，例如交互的多角色游戏。

阶段四 文字处理已不再是为语言教学编写具体的程序，而是适应语言教学，使学生用一种非永久性形式写作并能修改他们的作品。从更为学术性的方面考虑，专家们已经开始采用多媒体技术使各种各样的信息类型在计算机上可以同步利用，以至于在屏幕上不再只是显示书面语句，也能同时产生口语语言和移动的图像。由于新技术已经进入普通家庭，使用这种方式进行语言教学无疑将变得更广泛。这使 CALL 与更多的关于学习的传统观点结合成 ICALL（智能 CALL）的工作已经开始。

10.1.3 技术载体

多年来，基础的操练软件程序占领了 CALL 市场。这些程序着重于词汇或离散的语法点。大量的操练—实践程序仍然被利用。然而，创新和交互性程序正以不断增长的数量在开发。下面是 Higgins（1993）总结的一些程序。

1. 定制、模块和编程

使用 CALL 进行教学的最大弹性在于编程的领域。教师们能够就他们自己的材料，运用这些程序来创造简单的或精心制作的软件程序。通过这种方式，教师们能够设计适合他们自己课程计划的程序（Garrett，1991）。编程的范围是从简单模块程序到十分复杂的编程语言。这使得多媒体有了发展的潜能，也使不太复杂的编程成为可能。

2. 计算机网络

除了单独的程序以外，计算机通过网络连在一起，扩大了我们外语教学的途径。LAN（Local Area Networks，局域网）就是指在教室、实验室或其他建筑里，计算机通过光缆连接在一起。它们为教师们提供了一种新颖的方式，来创造新的活动，对学生而言就是创造更多的活动，为目标语言提供更多的时间和经验。某些 LAN 设置允许学生和教师通过计算机互相通信，或指导学生用目标语言合作写作。学生们还能用目标语言进行互相合作的写作练习、会话，并且尝试着解决问题。教师能观察到学生的活动和进步，并从老师的网站对个别学生作出评价，这类似于在一个语言实验室里发生的情况。

远距离网络，或通过远距离连接在一起的计算机，发展了 LAN 的独特能力，促进了国内和国外的学生间的交流。通过调制解调器和电话线，计算机能运用电子交流软件在千里之外进行交流。国外和国内的用户能直接地交互地

进行言语交流。

3. 软盘技术

软盘技术在外语教育中用处很多，包括信息检索、交互音频和交互式多媒体程序。由软盘发展起来的光盘（CD-ROM）使得大量信息可以储存在一张磁盘上，以便快速获取信息。出版商已经把能够装满十来个软盘的百科全书的全部内容，放在一张光盘（CD）上。学生和教师能够快速有效地在课内或课后使用信息。近年来，许多外语计算机程序已经放在光盘上，淘汰了对过多软盘的需求。

4. 数字化语音

许多程序已增添了一个新的物理特性：数字化语音。利用数字化语音的激光盘，提供了快速自由的信息获取和优良的声音质量。例如，一个“Lingua ROM”软件有一个程序磁盘和各种各样的语言磁盘，它们能容纳数字化言语。有了这种程序，学生们能听到一个短语、单词甚至一个音节或声调的发音。然后模仿发音，录下他们自己的发音。他们可以再听原来的发音和自己的发音，并比较二者，还能再次录下自己的发音，比较二者，直到他们觉得自己的发音已经有了长进或已经正确。当然，数字化语音远比磁带录音优越得多，所以储存它的空间也相对大多了。然而，光盘技术的持续发展将缩小空间限制。

光盘技术的最新进展是CD-I（交互式激光盘）的开发。这种技术包括数字化语音、压缩视频、动画和可能的为交互式程序创造的多媒体平台的文本。

10.1.4 慕课

本世纪第一个十年前后，网络电子学习和远程教育发生了变化，这表现在网络课程、开放式学习和慕课发展的增加。根据维基百科（Wikipedia），慕课——大规模开放式网络课程（Massive Open Online Course，MOOC），首先由加拿大爱德华王子岛大学的Dave Comier在2008年就当时的一门课程“联通主义和联通知识”（*Connectivism and Connective Knowledge*）所命名。

作为网络课程的慕课，是让人们通过网络无限制地参与和开放式地学习。除传统的课程材料，如录像讲授、阅读和问题集外，许多慕课提供互动的使用者论坛，以支持学生、教授和助教之间的集体互动。

随着慕课的发展，人们又区别“联通慕课”和“延伸慕课”。联通慕课（cMOOC）英文名词中的“c”是“联通主义者”（connectivist）的缩写。它强调联通主义哲学、同行评审和小组协作。学习材料应当是聚合的（非预先选定的）、重新混合的、可转变用途的、前馈的（即展开的材料可在日后重新确定学习目标）。教学设计采用各种方法，使学习者彼此之间在回答问题时互相联通和/或对共同的项目互相协作。

对比之下，延伸慕课（xMOOC）的“x”是“延伸的”（extended）的缩写。延伸慕课有明确制订的教学大纲，事先录就讲课内容和供自我测定的问题。这类慕课是清楚明确的智能技术平台，对不同单位提供内容的分布。讲师是有关知识的专家，学生间的互动通常仅限于要求帮助，或就难点彼此商议。（Simons，2015）

至2014年5月，美国高等院校已开设900多门慕课。这些课程范围很广，涉及心理学和哲学、人文科学和社会科学。麻省理工学院和斯坦福大学在计算机科学和电子工程方面最早开设慕课。有的大学开设针对本科生和研究生的工科课程。

直至目前，人们对慕课的有效性尚有争议，如退学率高于传统的网络教育课程。其原因涉及缺乏对课程技术的掌握和必要的引导、数字文盲、技术笨拙、学习环境不佳、讨论区运用不当、未阅读价格昂贵的教科书、合格的讲师和助教不足等等。有些未完成学习者注册时目的就是“逛商场”，或者只是为了增长知识而不想取得证书。讨论或数据分析也不充分。

为了解决上述的某些问题，有人提出替代办法，如“分布式开放协同课程”（Distributed Open Collaborative Courses，DOCC），以应对有关讲师、层次、收费和规模等问题。DOCC确认不一定非得通过集中的单一的教学大纲也能更好地获取知识的观点，也确认专门知识分布于所有的参与者，而不是集中在一两个人。另一个选择是“自定学习进度网络课程”（Self-Paced Online Course，SPOC），其特点是提供高度的灵活性，学生可以决定自己的学习进度，从那一阶段开始学习。

10.2 机器翻译

机器翻译（Machine Translation，MT）指使用机器（通常为计算机）将语篇从一种自然语言翻译至另一种自然语言。机器翻译可以分为两类：不需帮助的和需要帮助的。不需帮助的机器翻译将语篇翻译后，不需人力参与，可直接译成另一语言。需要帮助的机器翻译在翻译后，有时在此之前，有待人力翻译者的加工，以获得较高翻译质量。通常，这个提高过程是通过使用词典和适用的句法来限定词汇的使用。

基于哲学、宗教、政治和经济学的原因，机器翻译总是受到计算语言学的关怀。从哲学和宗教来看，研究者喜欢验证圣经中所言“整个地球曾经只有一种语言，说一种话”；政治上，非英语国家不愿意看到自己的语言、身份和文化被另一种语言，如英语所替代；经济上，机器翻译可以减少雇佣一大批翻译者的费用。统计表明，一篇难度较大的语篇，翻译者每天只能翻译4－6页或2000词（Craciunescu et al.，2004）。

10.2.1 发展历史

机器翻译从20世纪50年代的最早阶段至今几乎没什么变化。当时使研究者意见产生分歧的那些问题至今仍然是争论的主要原因（Hutchins，1995，1999）。不过，我们仍然可以列出以下几个发展阶段：

1. MT研究者的独立工作

20世纪50年代初期，因为受到硬件局限性的限制，尤其是内存不足和存储速度慢，以及没有高级编程语言，研究的目标必须适中。除此以外，研究是在没有句法学家和语义学家必要的协助下独自进行的。结果，最早的机器翻译研究者只好求助于以未加提炼的词典为基础的方法，就是说，主要是逐词翻译，以及统计方法的运用。

在这样的背景下，早期研究者意识到，他们所开发的无论什么样的系统都只能产生低质量的输出。因此他们建议：输入文本的前期编辑和输出的后期编辑，主要采用人工翻译。他们还建议，先对有控制的语言进行翻译，并限定于某些特定领域。

2. 瞄准高质量的输出

1960年前后，由于受到早期演示系统的鼓舞，公众和机器翻译的潜在资助者开始乐观地看待事态。他们相信在几年之内就可以达到优质量的输出。这可能也是对大大改进的计算机硬件和最早的编程语言的回应。当然，编程语言首先是在句法分析上取得进步。由于从长远来看，哪种方法将取得最大的成功还不清楚，所以美国政府机构只好支持大批项目。同时，对机器翻译的热情传遍了全世界。因此，这个时期以如下假设为特征：机器翻译的目标必须是产生高质量翻译的全自动系统，而人工辅助的使用被视为仅是过渡的安排。系统越进步，后期编辑就应该越少。因此，研究的重点在于寻求“完美”翻译的理论和方法的研究。

1960年，Bar-Hillel对以理论为基础的项目强烈不满，特别是那些研究中间语（interlingua）的方法，他还证明了全自动高质量翻译（Fully Automatic High Quality Translation，FAHQT）在原则上的不可行性。作为替代，Bar-Hillel倡导在他称为“人机共生”的基础上特别设计的系统。

3. 翻译工具的发展

从20世纪70年代以来的持续发展有三条主线：供翻译者使用的基于计算机的工具，以各种各样的方式涉及人工辅助的操作性机器翻译系统，以及针对机器翻译方法改进的“纯粹”理论研究。

由于20世纪60年代以来已经可以利用实时交互的计算机环境，70年代又出现了文字处理，80年代生产了大量连接网络和大容量存储的微型计算机，翻译工具的发展成为可能，如词典和术语资料库、多语文字处理、词汇和术语

资源的管理、信息传递的输入和输出（如 OCR 扫描仪、电子传送、高级印刷）。最近又新添了“翻译记忆”设备，它能够存储和使用已经存在的译文，以便作为后来的（局部的）再利用或修订，或作为翻译范例的资源。

人们也已意识到，所有现行的商业的和可操作的系统所产生的译文都得编辑或修改。机器翻译在特殊领域或受控制的环境里运行较佳，这一点也被人们广为接受。在这方面，机器翻译开发者已经富有成效地采纳了 50 年代首先由那些先驱们提出的论题和建议。

10.2.2 研究方法

机器翻译研究方法可以从两个方面来叙述，一方面是语言学理论的应用，另一方面是 MT 研究者们实际所从事的内容。

1. 语言学的方法

机器翻译研究已经被看作一个试验新的语言理论或新的计算技术的领域。换言之，MT 已经被视为语言理论的实验基础，因为翻译和翻译的质量能由非专家来判断。

与之相关的理论有：20 世纪 50 年代和 60 年代的信息理论、范畴语法、转换生成语法、从属语法和层次语法；70 年代和 80 年代的人工智能，非语言知识基础，诸如词汇功能语法、广义短语结构语法、中心词驱动短语结构语法和蒙太古语法等形式主义理论；90 年代的神经系统网络、连接主义、平行处理和统计学方法以及其他理论。

人们还发现，那些以小样本为基础，在它们最初的试验中取得成功的新理论，最终被证实都存在一些问题。为了解决这些问题，必须检验一切有希望的方法并鼓励修正。

在 20 世纪 90 年代初期，由于以语料库为基础的方法的出现，机器翻译的研究有所增强，特别是引入了统计学方法和以实例为基础的翻译。统计学技术已经摆脱了以前专门以规则为基础（通常定位于句法）的方法的不足和越来越明显的限制。在以语料库为基础的技术的帮助下，歧义消除的问题、首语重复的解决和更多惯用语的生成都已经变得更容易驾驭。

2. 转移法

根据机器翻译转移理论的多数观点，在机器翻译系统中有某种转移成分。这种成分是特定的，从而一对语言可产生一个目标句子。转移成分有一个相应的词库，这是源语言的范式和短语映现于目标语言的包罗万象的清单（Napier，2000）。大量的工作取决于具体的对两种语言比较的信息。所以这种转移系统意味着，对每两种语言来说，都要翻译，因此这种方法主张翻译本质上就是一种比较语言学的练习。由此可见，必须建立一个对应的词库。

3. 语际法

语际法基于这样的理解：在一批语言的每两种语言之间，只要求将每一个成员语言翻译成中间语言，并从中间语言翻译成目标语言。如果有 n 种语言，就需要有 n 个成员被翻译成中间语言，然后再翻译成目标语言。Arnold 等（1995）对这种方法不以为然，认为使用中间语言会导致信息的丢失（Napier，2000）。

4. 基于知识的方法

近来的趋势是转向基于知识的机器翻译，这是由卡耐基梅隆大学与新墨西哥州立大学的语言研究中心倡导的。Arnold 等（1995）所持观点为，需要三种知识来改进机器翻译系统。

a. 不依赖语境的语言学知识（语义学）。研究者所要做的是将词语和语义特征联系起来，从而研究者能对其他出现的词语加以制约。

b. 依赖语境的语言学知识，有的成为语用知识（语用学）。有多种方法处理语用学问题，其中之一是学习句子焦点的概念。

c. 常识 / 真实世界的知识（非语言学知识）。

前两类是针对语言本身的，问题不是最大，但是一个含有双语词典和语法系统的知识不能保证优质的翻译。计算机缺乏真实世界的知识才使研究人员苦恼不已。计算机不懂得事物之间的关系或事物是如何放在一起的。例如，计算机不能发现一个可吃的苹果和个人计算机的“苹果”品牌之间的区别；计算机也不能说明汉语中“山顶”的“顶”，“顶风”的“顶”和“顶好”的“顶”的不同意义。多少年来，正是这样的问题不仅困扰机器翻译的理论家，也困扰神经网络和人工智能的研究者。

随着语料库语言学的发展，最近研究者转向基于例句的机器翻译的研究。这种方法将正确的翻译作为信息资源，以建立新的翻译作品。

10.2.3 机器翻译和互联网

近年来，互联网的影响非常大。常常听到有人说，21 世纪是互联网的时代。自然而然地，我们已经看到互联网本身实时在线的翻译正快速成长。例如，近年来，已经出现许多特别设计的用于网页和电子邮件的翻译系统。很显然，人们对某种翻译系统有急切的需求。这种系统是特别为处理互联网上各种各样的口语化信息（常常是非正规的形式和拼写）而开发的。在这种情况下，一些语言学规则已显陈旧，我们不能只依靠基于规则的方法。虽然利用互联网上获得的大量数据并以语料库为基础来研究它们是合适的，但这样的系统研究还很少。

人们一致赞同：互联网的影响在未来将更加深远。这种影响自然也将改变机器翻译的前景。预言之一就是：带有独立文字处理系统、数据库、游戏等软

件的个人计算机，将在必要的时候被能从互联网上下载系统和程序的联网计算机所取代。在这种情况下，单独购买的机器翻译软件及字典等，将被远程存储的机器翻译程序、字典、语法、翻译成果、专门化的词汇等所代替。这种远程服务将根据用户所选择的内容来定价。

互联网的另一个深远影响与软件本身的特性有关。互联网的用户所寻找的是信息，无论它用哪种语言书写或存储，翻译只是其中一个手段。用户需要的是一个带有翻译功能的信息检索摘录和总结系统。因此，在未来几年里，“纯粹”的机器翻译系统将越来越少，基于计算机的工具和应用将越来越多，而自动翻译仅仅是其中一个部分。

10.2.4 口语翻译

新世纪人们最期待的发展，一定是口语翻译的发展。该研究项目在20世纪80年代末90年代初开始时，人人都知道实际应用不太可能。这些局限于小范围的系统对任何进步都是非常重要的。但更有可能的是将有为数众多的口语翻译系统被应用于小范围的自然语言交流，例如数据库的询问（尤其是财政和股票市场的数据）、商业谈判的交往、公司内部的交流等等。

10.2.5 机器翻译的质量

人们不得不承认：在所有现有的实际翻译中都会有错误。人们还能发现，有些错误是人工翻译中不会出现的，例如代词误用、介词误用、句法混乱、措辞不当、单数成了复数、时态错误等。Kay（1982）引用了一个著名的旧例，很好地说明了这个观点：

例 10－1

The police refused the students a permit because they feared violence.（警察拒绝给学生通行证，因为他们害怕暴力。）

假设此例被译成像法语那样的语言，其中 police 这个词为阴性。那么 they 这个代词也必将为阴性。现在我们用 advocated 来代替 feared，那么似乎句中的 they 是指 students 而不是 police 了。如果表示 students 的词是阳性的，则它又将有一个不同的译法。得出这些结论的常识没有哪一个是有关语言学的。这与 students，police，violence 这些日常现象以及我们所看到的它们之间的各种关系有关。

其次，翻译不是一项保留意义的工作。试考虑下列用法语陈述的问题：

例 10－2

Ou voulez vous que je me mette?

此句的字面意义是“Where do you want me to put myself?（你想让我把自己放在哪里？）”。但这是一个很自然的译法，因为 want 这种形式有一系列的英语提问形式“Where do you want me to sit/stand/sign my name/park/tie up my boot？”。在大多数情况下，英语“Where do you want me？”将被接受。但是，为了翻译得更流畅而增删信息，也是很自然和常规的事。有时这无法避免，因为有的语言像法语，代词表明数量和性别，日语中代词常常一起省略，俄语里没有冠词，汉语里的词不分单数和复数，动词也不分现在时和过去时，德语里词序的可适应性使得什么是主语什么是宾语变得不确定。

当然，如果翻译系统只作为研究对象，翻译质量低还没有什么问题，因为它毕竟没什么社会效应。但是，当商用系统出现时，整个机器翻译工业都将受到低质量翻译的打击。不幸的是，这种状况在近期内不会改变。没有任何迹象表明，基本通用的机器翻译机将有更大改进。

10.2.6 机器翻译和人工翻译

在新世纪之初，很明显，机器翻译和人工翻译能够且将会相对协调地同时存在。我们仍旧需要那些人工翻译者所作出的贡献。当翻译不得不讲究“可发行”质量时，人工翻译和机器翻译都具有各自的作用。对于大规模、快速地翻译令人厌烦的技术文件来说，机器翻译已被证明是合算的。（高度重复的）软件本地化手册的翻译和许多别的情况下，机器翻译加上必要的人工准备和校对的成本，或者（在没有计算机辅助时）使用计算机化的翻译工具的成本，明显比那些无计算机辅助的传统的人工翻译要低得多。通过对比可知，对非重复的语言复杂的语篇（例如文学和法律的语篇）以及个别高度专业化的语篇而言，人工翻译仍是也还将是无可替代的。

对语篇翻译来说，在对输出的质量要求不高的地方，机器翻译常常是一种理想的解决方法。例如，对科学和工业文件进行的“初略”翻译。如果人们通过这种翻译仅仅想找出重要的内容和信息，并不关心已翻译材料的可理解度，或是读者肯定不会因笨拙的语言或语法错误而有阅读障碍，这时机器翻译将逐渐成为唯一的选择。

对信息的一对一交换而言，人工翻译家很可能总是会有作用的，例如商业通信（尤其是如果内容是敏感的或有法律约束的）。但是对个人信件来说，机器翻译系统更可能逐渐被使用；并且，对电子邮件和网页的信息摘录以及以计算机为基础的信息服务而言，机器翻译是唯一可行的解决方法。

至于口语翻译，一定会是人工翻译者的市场。但是机器翻译系统正在开发一些新的领域，这些都是人工翻译从未涉及的。这些新领域包括：当作者用外语写草稿时，他需要源语篇的帮助；电视字幕的实时在线翻译；数据库里信

息的翻译。毫无疑问，随着全球交流网络的扩展，机器翻译的实际可用程度将被更为广泛的大众所熟悉，更多崭新的应用将在未来出现。

10.3 语料库语言学

在上一节，我们知道机器翻译的方式之一是建立以实例为基础的系统，也就是说，一个“能参考现有翻译存储”的系统。这就预先假设存在一个关于现有翻译的非常大的语料库。虽然美国结构主义语言学家都用过以语料库为基础的研究方法（Kennedy，1998），但还是由于计算机技术的发展，语料库和语言学之间的联姻才成为可能。在本节中，我们将主要讨论计算机语料库。

10.3.1 定义

关于“语料库”和“语料库语言学”有各种各样的定义。以下是两个同年出现的有代表性的定义：

> 语料库（corpus，复数形式 corpora）：一个语言数据的集合，可以由书面文本构成，也可以由录音言语的转写本构成。语料库的主要目的是鉴定关于语言的假说——例如，确定一个特定的语音、单词或句法结构的使用是如何变化的。
>
> 语料库语言学：论述语言研究中使用语料的原理和实践。一个计算机语料库是机器可读文本的重要躯干。
>
> （参见 Crystal，1992/1997：85）

> 语料库（CORPUS，13 世纪，来自拉丁语的 corpus 一词；意思是“body”[躯干；身体]。复数形式通常是 corpora）：（1）一个语篇的集合，尤其指完整的和自身需求的语篇集合；Anglo-Saxon 诗句的语料库。（2）复数形式也可写成 corpuses。在语言学和词典编纂学上，指语篇、语句或其他样本的集合，通常作为一个电子数据库储存。一般说来，计算机语料库可以储存上百万的流行词汇，其特征能通过标记的方式（为词和其他构型作标记，并加以确认和分类）和使用共现关系程序来分析。
>
> 语料库语言学：研究任何语料库中的数据。
>
> （参见 McArthur，1992：265－266）

10.3.2 语料库语言学受到的批判及其复兴

尽管语料库语言学对语言学上的美国结构主义的发展作出了贡献，但乔姆斯基（Chomsky）在相当短暂的时间内改变了语言学的方向，使其远离了实

证主义，而朝向理性主义。为此，他显然使语料库作为一种在语言调查中的证据资源而失效。乔姆斯基提出，语料库对语言学家而言不再是有用的工具，因为语言学家必须寻求概括语言能力而不是语言的使用。另外，叙述一种语言的语法的唯一途径是描述它的规则——而不是列举它的句子。乔姆斯基认为一种语言的句法规则是有限的。第三，即使语言是一个有限的构造，语料库方法论不一定是研究语言的最好方式，例如，在例 10－3 里，如何将不合语法的语句从那些仅仅是没有出现过的语句中区分出来呢？如果我们有限的语料库不包括句子 a，那么我们如何得出结论，认为它是不合语法的呢？而句子 b，c，d 又为什么是合语法的呢？

例 10－3

a. *He shines Tony books.

b. He gives Tony books.

c. He lends Tony books.

d. He owes Tony books.

除乔姆斯基所言，语料库语言学的实用性也有许多问题。例如，我们如何能想象，不用任何工具而光靠眼睛去搜索一个具有 11，000，000 个词汇的语料库？（Abercrombie，1963）

虽然乔姆斯基的批判使语料库语言学失去了权威性，但并没能阻止所有基于语料库的工作。例如，夸克（Quirk）在 1961 年计划并实现了他雄心勃勃的建设，即《英语惯用法调查》（*Survey of English Usage*，SEU）。1975 年，Jan Svartvik 在 SEU 和 Brown 语料库的基础上，开始建构 London-Lund 语料库。在计算机逐渐开始成为语料库语言学的支柱之后，Svartvik 就用计算机操作 SEU，结果产生了一些包括利奇（Leech，1991）仍然坚信的“（语料库）直到今天对研究口头英语而言仍是一种前所未有的资源”的成果。

10.3.3 共现索引

宣告语料库语言学的复兴可说是计算机创造的一个奇迹。计算机有能力搜索一个特定的词、词串，甚至一个语篇里的某一言语部分。计算机也能检索一个词所有的实例，以及带有该词的上下文，这对语言学家是有力的帮助。它还能计算一个词出现的次数，从而收集到有关这个词的频率信息。然后，我们可以以某种方式对数据进行分类——例如，按紧接被检索词词语的字母顺序进行分类。这通常被称为**共现索引**（Concordance）。我们可以先列出该词的所有实例，并从这样的序列中摘录出另一个序列，比如说可以是在该词的所有实例中跟被检索词最近的另一个词（比如：代词或所跟的标点符号）的全部实例。上面所描述的过程通常被包含在一个共现索引的程序里面。例 10－4 阐

明了 deal 这个词的共现索引（Biber et al.，1998）。

例 10－4

and secret plans preposed to	*deal*	with the mass sit-down	1
of companies and put one property	*deal*	through each. Mr.	2
In particular, a good	*deal*	of concern has been	3
hangs a tale and a great	*deal*	of money. Neville	4
where his new measures to	*deal*	with Britain's	5
just a matter of working a good	*deal*	harder before we really	6
I'm mixed up in a	*deal*	involving millions	7

这是在语料库语言学里检查语料库所常用的工具。且不论最终我们能从语料库中看到何种哲学高度上的意义，起码计算机已能够使我们更准确、快速地大规模运用语料库资源。

共现索引是语言学经常使用的工具，可用来研究语篇，如比较某词的不同用法、分析词频、寻找并分析短语和成语、编制索引表和单词表（对出版工作也是有用的）。

10.3.4 语篇编码和注解

无注解的语料库指未经加工的原始文本；有注解的语料库则人为扩充了各种语言学信息。毫不奇怪，当语料库被注解以后，它的可用性提高了，语言学信息不再只能通过文本含蓄地表现出来，而可以被认为是语言学信息的贮藏室，隐含的信息通过具体注解的方式变得明显。例如，形式“gives”包含词类的隐含信息“第三人称单数现在时动词”，在正常阅读里，我们仅能通过求助于预先存在的英语语法知识来检索它。然而，在一个已经注解过的语料库里，形式“gives”可能以“gives－VVZ”的形式出现，代码“VVZ”表示它是一个词汇中动词（VV）的第三人称单数现在时（Z）形式。诸如这样的注解，使检索和分析包含在语料库里的语言信息变得更快、更容易。

利奇（Leech，1993）描写了适用于语篇语料库注解的七条准则。

1. 为了恢复到自然的语料，应当能从有注解的语料库里删去注解。

2. 应当能从语篇里单独摘录注解。

3. 注解方案应该以终端用户可利用的指导方针为基础。

4. 注解是如何并且由谁来完成这一点应清楚说明。

5. 终端用户应该知道语料库注解不是没有错误的，而只是一种潜在有用的工具。

6. 注解方案应尽可能地立足于普遍接受的和理论上中性的原则。

7. 任何注解方案都无优先权被视为是标准的注解。

必须指出：利奇的某些准则并不容易遵守。例如准则 1，有时它可以成为一个简单的过程 —— 例如在横线后面删除每个字符，如“Claire-NP1 collects-VVZ shoes-NP2”将变成 Claire collects shoes。然而，London-Lund 语料库的节律分析是在词的内部点标点 —— 例如“g/oing”表示在“going”这个词的第一个音节上有一个升调，这意味着原始词汇并不能如此简单地重建。

10.3.5 语料库数据的作用

语料库对语言学习的重要性与经验证据对语言学习的重要性一样。实证数据使语言学家作出客观的描述，而不是主观的描述，也不是基于个人固有的对语言的感知。从这点出发，语料库对语言研究的不同领域起着重要的作用，例如言语研究、词汇研究、语法、语义学、语用学、语篇分析、社会语言学、文体学、历史语言学、方言学、变异研究、心理语言学、社会心理学和文化研究等。考虑到语料库语言学研究的广泛性，我们这里仅选择几个例子。

1. 言语研究

口语语料库提供言语的大量样本，包括说话者的种种个体差异，如说话者的性别、年龄、阶层等。当语料库中的语料足够广泛且具有代表性时，我们就可以对口头语言进行概括。同时也可以对一种已知口头语言的某一变体进行研究。它也提供了自然言语的样本而不是从人为语境里抽取的样本。因为（标注的）语料库常常通过标注韵律和其他注解得到增强，所以比未经加注的数据更容易进行大规模的定量分析。在使用了不止一种类型的注解的地方，研究注解之间的关系，如语音注解和句法结构的相互关系，便成为可能。

2. 词汇研究

使用语料库，语言学家能在几秒钟内从一个有几百万词汇的语篇里抽取一个词或短语的所有实例。字典的编撰和校订将比以前快得多，语言的最新信息也可以得到及时体现。由于检验了大量自然实例，定义也将更加完整和精确。而且，从语料库中提取的是词汇组合而不是孤立的词，同时因为现在已有判断共现词关系的共用信息系统，因此我们比以前更有可能系统性地处理短语和搭配问题。一个熟语单位可以组成一条专门术语或一个成语，而词的搭配则是研究具体词汇意义的重要线索。

3. 语义学

语料库语言学确立了客观的研究方法，从而为语义学作出了贡献，因为语篇中的语义区别与特定的可观察到的上下文相联系 —— 或是在句法上，或是在词汇上，或是在韵律上。通过考虑语言实体的环境，我们就可得到某一特定语义区别的客观实证性标示。

语料库在语义学里的另一个贡献在于建立了更稳固的关于模糊范畴和渐变性的概念。在理论语言学里，范畴常常被认为是“死板的”——一个项目要么属于一个范畴，要么不属于。然而，对范畴化的心理研究表明，认知范畴并不常常是“铁板一块”，而是有模糊的边界。因此问题是为什么偏偏是属于这个范畴而不属于与之相对的另一个范畴。实证地观察语料库中的自然语言时，这种“模糊”的模式能很好地描述数据，这一点是很清楚的。完全清晰的边界是不存在的，反之，范畴的成员资格都有渐变性。这种渐变性是与其内涵出现的频率相联系的。

4. 社会语言学

虽然社会语言学是一个以实证为根据的研究领域，但是它并没有经常进行严格的抽样调查。有时，它的数据并不是从自然语料中抽取的。一个语料库能提供这种抽样数据所不能提供的东西——一个有代表性的自然数据的样本，这种样本能被量化。为了检验美式英语和英式英语的男性偏见，Kjellmer（1986）观察了阳性和阴性代词以及词语 man/men 和 woman/women 的出现情况。值得注意的是：在两个语料库中，阴性词语的出现频率都比阳性词语低得多；其次，在美式英语中，阴性词语的出现比英式英语普遍。

当考虑到方法时，Holmes（1994）指出：给词汇出现情况分类和计算时应该考虑到词的内容。例如，policeman/policemen 有一个无性别标记的替换词，即 police officer，而在 Duchess of York 的作品里，对 -ess 这个形式而言就没有可替代词了。因此，在检查书面语中的性别偏见时，后一个形式将被排除在性别后缀的考虑之外。Holmes 还指出了为一个正在经历语义变化的形式进行分类的困难。她认为：man 这个词可用于指单个男性（如 A 35 year old man was killed 里的 man），或指有类别意义的人类（如 Man has engaged in warfare for centuries）；而在 We need a right man for the job 里，很难断定 man 是指性别特性还是能够用 person 来代替。

5. 心理语言学

首先，在心理语言学领域，抽样语料库能为心理学家提供很多具体、可靠的有关频率信息，包括歧义词意义及其词性的频率。其次，语料数据能用于检验在自然会话中言语错误的出现情况。语料的第三个作用在于对语言病理学的分析方面。在此，为了假设和测试人类语言处理系统在什么方面会出现问题，必须先对异常数据作准确的描述。

10.4 计算机介入的信息交流

计算机介入的信息交流（computer-mediated communication，CMC）的

特点是突出语言与计算机网络环境中的语言使用的关系，并通过语篇分析的方法来谈论这个焦点（Herring，2001）。以语篇为基础的 CMC 有多种形式，如电子邮件、讨论组、实时聊天、虚拟现实的角色扮演游戏等。这些形式的语言学特征的变化，取决于所使用的信息系统的种类和包含某特定使用实例的社会文化语境。人与人之间通过计算机网络或互动网络的信息交流是晚近的现象（Herring，2001）。在这里，我们主要介绍以下几个题目，即邮件和新闻，PowerPoint，博客，聊天室、脸书和微信，表情符号和笑眯眯。

10.4.1 邮件和新闻

人们进入互联网络（Netscape 或 Internet Explorer），主要有两种浏览方式可选，一是搜索或网上旅行，一是信件邮箱。前者是信息检索，后者是邮件/新闻的获取和发送。过去如果你给朋友或亲戚写信，信件需要 2 至 3 天到达目的地，甚至要一两周时间才能到达另外一个国家，更不用说昂贵的邮资。长途电话可能会节省时间，但它是按照分秒收费的，如果同一个消息，你需要告诉不同的人，你就需要打多个电话。通过电子邮件，人们可以将同一个邮件同时发给许多相关的人或者通过附件发送文件和图片。

除此之外，信箱通过“电子邮件转发系统”（listserv 或 majodomo），可以帮助用户参与学术活动。用户所需要做的就是向一个电子论坛、学会或杂志订购。当订购得到确认后，用户会收到有关会议的信息、新书或杂志，甚至就业机会；它可以查询求助或参加学术事务的讨论，当讨论结束时，常可读到有关讨论的回顾和总结（胡壮麟，1997）。例如下面是笔者 2001 年 4 月 19 日在信箱中收到的 *The Linguist* 所提供信息的索引：

表 10-1 The Linguist 提供的信息索引

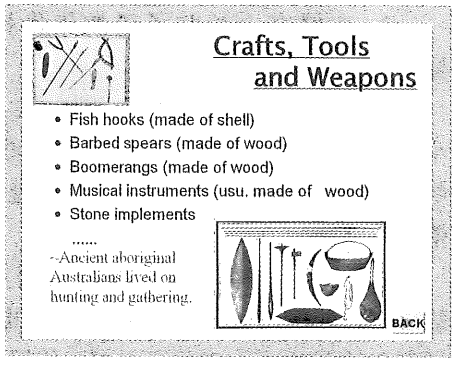
系列号	主题	收到日期
12.1085	FYI: Summer School	01－4－19 4:41
12.1084	Review: Corrections	01－4－19 3:54
12.1083	Review: Verbal Complexes	01－4－18 23:51
12.1082	Confs: Modality in...	01－4－18 3:30
12.1081	Qs: English speakers...	01－4－18 23:24
12.1080	Qs: DESSou un DEA	01－4－18 22:16
12.1079	Books: Syntax/Semantics	01－4－18 21:43

这里，“FYI”代表为你提供的信息（for your information），第 12.1085 是关于语言学的夏季学校的一个广告。第 12.1084 条和第 12.1083 条是对两本学术著作的书评，其中一篇是关于如何处理 WWW 的修正，另一篇则是有关

动词复合结构的讨论。第 12.1082 条是通知人们有一个关于情态的国际会议。往下，参加服务器的成员在第 12.1081 条和第 12.1080 条中提问，以便从其他成员那里寻求答案或帮助。最后一条是关于语法学和语义学的新书发布消息。

10.4.2 PowerPoint

PowerPoint（电子幻灯片）是在电子投影仪上演示的幻灯片。用户编制的幻灯片可以是书面语篇、图像、音响、动画和录像。当用户编制完成后，可以自己翻页，或让幻灯片自动翻页。



可见，PowerPoint 是一种可自行制作的演示软件，以文字和图像为主，音响可有可无。微软公司的 PowerPoint 软件是为 Windows 和 Mac OS 计算机操作系统开发的。现已为政府官员、商务人员、教育和培训工作者广泛使用，是最流行的演示工具，据微软公司推销人员的报道，全世界每天制作的 PowerPoint 演示文稿达 3 千万之多。

在谈论 PowerPoint 时，我们应分清作为工具的 PowerPoint 制作软件、作为语篇的 PowerPoint 演示文稿和作为语篇类型的 PowerPoint 演示方式。

PowerPoint 工具指用来编写幻灯片上的要点和创建相配视听材料的软件。这是微软公司科学家们设计的与 Office 一起打包的产品。它可以帮助用户传递信息，但本身不表示意义。

作为语篇的 PowerPoint 是指被广泛地用来在幻灯片上围绕一定主题制作各种具有内容的可视听的文字、图像、音像材料。这就是说，每一个 PowerPoint 演示，不论是报告、演讲还是课件，都是语篇。它可以是一张幻灯片，也可以是一套幻灯片，但它们都要有内容有主题。

作为语篇类型的 PowerPoint，指一种重复的活动或表示意义的形式，就像我们对书信、通知、故事、小说、诗歌、演讲、剧本等的认识。不论有多少种 PowerPoint 语篇，它的基本表现形式和功能是公认的。

10.4.3 博客

网络博客（Webblog）最初出现在 20 世纪 90 年代中期，是一种简便自由的出版形式，在上世纪末已广为应用。

Dan Gilmore 对网络博客（简称博客）的定义为：具有各种链接点和帖子

的网络杂志，按逆年代顺序编排，最新的帖子出现在网页的上端（BlogCanada，2005）。Matisse Enzer 的 *Glossary of Internet Terms* 是这么说的：一个博客基本上是一个在网络上可以找到的杂志。使博客不断更新的活动叫作 blogging（做博客），参与的人叫作 blogger（博客者）。博客的主要特点是逐日更新，其软件便于对技术知识知道不多的人掌握。

BlogCanada（2005）和 Walker（2005）对网络博客的特征有更具体的论述：

1. 网络博客是“以帖子为中心”的，不是“以网页为中心”的，因此帖子是基本单位。

2. 在博客上放帖子虽按年代排列，但最新增加的帖子具有最醒目的标识。

3. 网络博客是系列的和累增的，读者倾向于每次看少量的帖子，过了几个小时、几天或几周回过来再来看那些新增的帖子。这种系列或事件结构与那些书信体小说或日记相仿，不同之处在于网络博客是开放性的，直到写作者厌倦了才停止。

4. 许多网络博客条目是简短独立的叙述，有的则是清晰的或隐含的虚拟故事，虽然标准的博客期待的是非虚拟的内容。有的网络博客为个别帖子的微型叙述建立了大型页面分栏，使用一致的规则限定它们的结构或主题。

5. 鉴于任何具有网络联系条件者都可以发行自己的网络博客，网络博客的质量、内容和需求上呈现很大分歧。它可以是某个特定领域内当前事件的连续性评论，系列的个人沉思，纯粹是政治性的小册子，提供最新最伟大的袖珍机械的网址，由域名专家不时更新的“what's new”。

6. 许多网络博客一般都利用链接，使读者能按有关主题的条目之间的链接跟踪网络博客之间的会话。读者可以从网络博客之间的任何一个点开始，首先见到最新的条目，或者通过搜索引擎或另一个网络博客的链接接通一个老的网址。一旦进入一个网络博客，读者的阅读可以根据不同顺序，如年代、主题，或不同条目之间的链接，或搜索关键词进行。网络博客一般也包括网络登记表，这是一份能与其他网络博客链接的单子。许多网络博客容许读者对某些帖子提出他们的评论。

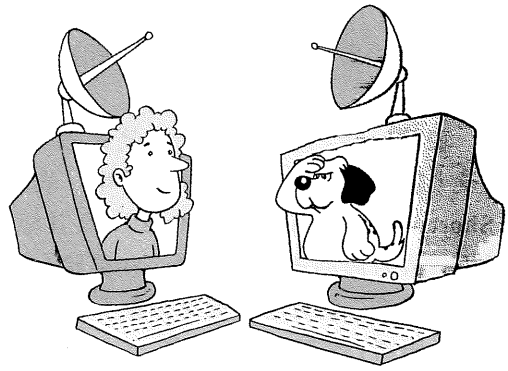
7. 网络博客是由个人发行的，因此他们的风格是私人的和非正式的。

8. 最好的个人博客共同具有的是声音——是人类以最大的激情清楚谱写的。

10.4.4 聊天室、脸书和微信

与博客相比，聊天室（chatroom 或 chat room）的概念要简单得多。用通俗的话说，一个聊天室意味着在互联网上一群人在讲话。因此，对聊天室最简单的定义是一个网址，用户们可以在这个网址里进行实时的信息传递（典型

的情况是围绕一个专题）。网络百科全书维基百科（Wikipedia）给我们提供了一个比较具体的描述："聊天室是一个网络论坛，人们可以在那里进行网络谈话（如与同一论坛的人们进行实时的谈话）。有时这些网址有人维护，比如，对讲话人加以限制（不常见），或由管理员在该网址巡游，提防破坏性的或不受欢迎的行为。"



在通常情况下，有意参加者需要注册，在聊天室留下名字，然后此人能够看到聊天室的其他内容。有的聊天室容许参与者只和一个人谈话，不让他人看到。虽然用户可以自由进入聊天室，自由谈论自己想谈的内容，但许多聊天室是有监控的，杜绝那些不受欢迎的、攻击性的、种族主义的、宣扬暴力的或色情的内容等。

根据维基百科，脸书（Facebook）网站始建于2004年2月4日，发起人为哈佛学院的Mark Suckerberg。自2006年起，年满13岁者均可注册参加。注册后，脸书用户可建立自己的用户概况，添上作为自己"朋友"的其他用户，从而与他们互通信息，刊登最新情况和照片，共享视频，使用种种应用程序（Apps），并接收到有关他人最新情况的通知。此外，用户可以参加由工作单位、学校或有关课题组所组织的兴趣相同的用户群，并可以将朋友分成若干小群，如"工作有关人士"或"密友"。在这些小群中，编者可将有些宣告置于页面顶部。用户也可对某些有负面看法的人士表述自己的不满，或切断与他们的联系。脸书对保护隐私有严格控制。至2015年1月21日，脸书的计算程序能清除虚假的或导向错误的内容。

由我国腾讯公司发展的微信（WeChat）公众平台提供即时信息服务，自2011年1月起运行。微信是当代最大的独立应用程序，为日夜需要信息的用户服务。至2016年5月，公众账号已有10亿个以上。微信平台主要提供文本信息、即时通话信息、广播信息、视频会议、视频游戏、照片和视频共享以及位置共享。照片也可能是经过装饰的乱码过滤器和字幕。该平台还提供机器翻译。

10.4.5 表情符号和笑眯眯

当人们通过电子邮件、博客或聊天室进行写作时，他们会注意到这种信息传递已经产生了巨大的反拨影响。如：用户倾向于使用很少量的标点符号；使用字母代替一些词语，例如"u"代表"you"，"4"代表"for"，"r"代表"are"，

“brb”代表“be right back”，等等。这样的变化也能在结构中找到，如短句和非正式表达形式增加了、直接称呼减少了，等等（张德禄，1998；董启明、刘玉梅，2001）。

表情符号（Emoticon）或笑眯眯（Smiley/Smilies）是人们使用在计算机键盘上能找到的字母组成的字符串。表情符号在电子邮件、聊天室、短信服务（Short message service，SMS），及其他使用计算机传递信息的途径中被广泛运用。最时尚的表情符号是一些笑脸（Smiley/Smilies），人们用来表示“别把我的话当真”的意思。如果你看不懂它表示一个笑脸，不妨仔细看看下表。冒号代表眼睛，连字符代表鼻子，右括弧代表嘴巴。最常见的和逗笑的组合如下：

表 10-2　常见的表情符号

: -)	标准笑眯眯（你在开玩笑；你满意了）
:)	给懒人设计的标准笑眯眯
, -)	眨眼的笑眯眯（即使你在开玩笑，你不是这个意思）
; -)	眨眼的笑眯眯，同上
: ->	后面是一个非常挖苦的评论
(- :	左撇子的笑眯眯
: -(	哀伤的笑眯眯（你不是开玩笑；你不满意）
:<	非常哀伤的笑眯眯
:C	非常哀伤的笑眯眯
: -*	亲吻的笑眯眯