第一章　绪论：什么是计算语言学

先来看下面这一小段以汉语进行的对话。

甲：听说吴妈打赢了阿 Q。

乙：不错，阿 Q 确实被吴妈打败了。

甲：这个结果有些出人意料。

乙：阿 Q 是大意失荆州，怪不得别人。

这段对话也许从未在真实世界中发生过，但懂汉语的人很容易明白它的意思。如果这段对话是由两个说汉语的中国人讲出来的（比如两个相声演员），这一点也不稀奇，人们会觉得这“太过平常”了。但如果上面这段对话发生在一个人跟一台机器之间，甚至发生在一台机器跟另一台机器之间，可能人们就会“啧啧称奇”了。如果像这样的对话不仅仅只是说上四句就结束，而是滔滔不绝地说下去，那恐怕就要用“匪夷所思”来形容人们的感受了。

从某种意义上说，人类文明的发展历史或许可以表述为：将过去在人们看来“匪夷所思”的事情，变成现在人们容易理解，甚至“习以为常”的事情。那么，上面谈到的今日“匪夷所思”之事，是否能成为明日的现实？而要让这“匪夷所思”之事变成现实，人们又应该做怎样切实的努力呢？

上面提出的问题，实际上又可以转化成一系列相关的问题来加以考虑：

（1）人用来交际的“语言”具有什么样的性质？

（2）人用来交际的“语言”跟机器可以理解的语言有什么样的关系？

（3）人是如何运用“语言”进行交际的？

（4）人运用“语言”进行交际的“过程”是否可以描述为一个“机械的过程”？

（5）什么叫做“理解”一种语言？

（6）如何从“内在的交际意图”到“外显的语言表达”？

……

大致说来，正是对上述问题以及相关的延伸问题的探索，形成了计算语言学这一交叉边缘学科。如果要用相对严谨和概括的说法来表述的话，可以说：**计算语言学（Computational Linguistics）指的是这样一门学科，它通过建立形式化的数学模型，来分析、处理自然语言，并在计算机上用程序来实现分析和处理的过程，从而达到以机器来模拟人的全部或者部分语言能力的目的。**

上面对计算语言学的界说可以分解为以下几个方面来做进一步的阐述。

第一节 计算语言学的研究对象

从“计算语言学”这个名称上可以看出，这门学科的研究对象涉及“计算”与“语言”两个方面。计算语言学的研究工作一方面可以表述为是从“计算”的角度去看待“语言”的性质；另一方面也可以说是将“语言”作为某种特殊类型的“计算”对象，相应地来研究适用于这类计算的算法过程。这两个方面共同构成了计算语言学的核心研究内容。

一 从计算的角度来看语言的性质

所谓从计算的角度来看语言的性质，实际上就是要求将人们对语言的结构规律的认识以精确的、形式化的、可计算的方式呈现出来，而不是像传统的语言学研究那样，在表述语言的结构规律时一般采用非形式化的表达形式。比如：

在表述汉语中所谓“把”字句的结构规律时，传统的语法学可能会有这样的一些说法：

（1）汉语的“把”字句也叫处置式，表示处置的意义，通常是指主语所表示的人或事物将“把”后宾语所表示的人或事物置于某种状态。例如：张三把李四赶跑了。这个句子中，“张三”是主语，“李四”是“把”后的宾语，这句话表达的意思可以表述为：张三赶李四，李四跑了。

（2）汉语一般的主谓宾句式可以变换成“把”字句，通常也有对应的“被”字句。例如：“张三赶跑了李四”也可以说成“张三把李四赶跑了”或“李四被张三赶跑了”。

上面这样的描述当然揭示了有关汉语“把”字句的一些特点规律。但仅仅这样来描述，是不够精确的，同时也是非形式化的表述方式。就精确性要求来说，上述规则不能说明为什么“吴妈以前很喜欢阿 Q 的理论”这个主谓宾句式不能变换成相应的“把”字句和“被”字句（汉语中不说：“*吴妈把阿 Q 的理论以前很喜欢”，也不说“*阿 Q 的理论被吴妈以前很喜欢”）[1]。就形式化要求而言，上述对汉语“把”字句规律的说明是以自然语言（汉语）本身来描述的，没有采用符号化的规则形式来描述。

那么，要能够以精确的、形式化的方式来表述有关自然语言的知识，应该如何去做呢？一方面，可以用一定的形式系统来“显性地”、“概括地”表述，另一方面也可以用带标记的语料库来“隐性地”、“具体地”以统计数据形式表述。本书第二章（形式语法理论基础）和第三章（语料库）将分别对这两种方式做概要地介绍。

二 将语言作为计算对象来研究相应的算法

所谓将语言作为计算对象来研究相应的算法，是研究如何以机械的、规定了严格操作步骤的程序来处理语言对象（主要是自然语言对象，当然也可以是形式语言对象），包括一个语言片断（比如词组、句子或篇章）中大小语言单位的识别，该语言片断的结构和意义的分析（自然语言理解），以及如何生成一个语言片断来表达确定的意思（自然语言生成），等等。

现代的“算法”（algorithm）一词据说来自一位名叫阿尔·花拉子模（al-Khowārizmi）的波斯数学家兼天文学家的名字。现在人们用这个词指具有以下特点的计算过程：

（1）通用性：算法是针对一类问题的，而不仅仅是用于解决某一个具体问题。

（2）机械性：算法的每一个步骤都是确定的。

（3）有限性：算法必须在有限步内结束。

（4）离散性：算法的输入数据及输出数据都是离散的符号。

在计算机背景下来看所谓算法，则可以将“算法”定义为用计算机语言编制的一套程序，这套程序向机器发出指令，使得计算机能够机械地在有限步骤内完成任务。下面是一个算法的实例：对任意两个自然数 A、B，求其最大公约数。

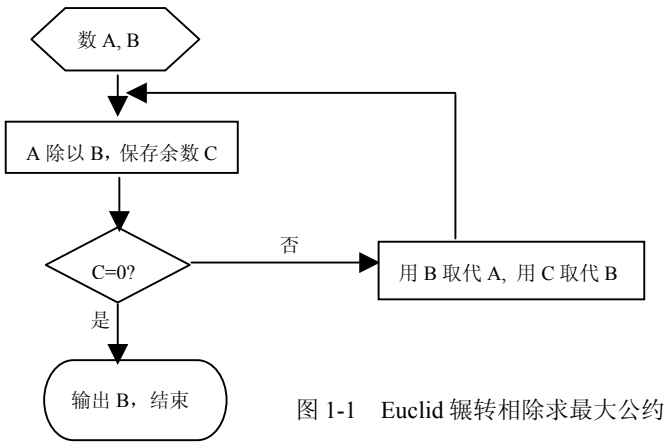

图 1-1　Euclid 辗转相除求最大公约数算法

[1] 本书通常用*置于句前表示该句不能说，或不被接受。

为了让计算机能够完成分析、处理自然语言的任务，人们已经从不同的角度、采用不同的处理策略，针对不同的任务提出了不同类型的算法，包括计算机科学中的编译理论所发展出来的用来分析程序设计语言的各种分析算法，针对形式语言的分析提出的各种分析算法，以及在处理大规模语料库过程中发展出来的各种基于统计模型的算法，等等。这些算法技术，为在不同水平上对自然语言进行不同程度的处理，提供了有效的手段和途径。本书第四章（词法分析技术）、第五章（句法分析技术）将对有关算法技术加以介绍。

三 对计算语言学研究内容多层次多角度的认识

上面讨论计算语言学的研究对象时是将语言笼统地作为一个整体对象来看待的。实际上，跟其他学科的情况一样，在计算语言学的研究中，通常会把“语言”这个大对象分解成一些相对独立的“部分”或者说是“层面”来“分而治之”。这不仅表现在计算语言学的发展历史过程中，不同时期，计算语言学特别关注的研究、处理对象有所不同，也表现在同一个历史时期，不同的学者对不同层面的语言对象的关注有所不同。这样，在计算语言学这顶大帽子下，按照所处理的语言对象的不同这个角度，又可以相对地区分出一些子研究领域，比如：在对自然语言的文本对象进行分析中，包括形态分析（着重在如何用计算机来自动分析语言中的基本组成成分——词）；句法分析（主要是以自然语言的句子为处理对象，分析句子的结构和语义）；自然语言篇章分析与文本生成（主要是以比句子更大的单位——篇章为分析处理对象开展的研究）。在对自然语言的语音对象进行处理中，又可分为语音识别与语音的自动合成技术等。一般来说，对大单位的语言成分进行处理，是建立在对小单位的语言成分进行处理的基础之上的。

针对不同的处理对象，相应地，在计算机算法和适用的处理策略上也存在着差异。从算法和处理策略的角度，也可以相对地区分出一些子研究领域，比如用于词处理的自动分析算法和用于句处理的分析算法，以及用于篇章处理的分析算法等等。这些，在研究中都可以作为相对独立的研究内容开展研究工作。不过，研究人员仍然应该有整体的观念，独立只是相对的，任何一个相对独立的研究部分，都离不开也不应该离开将语言作为一个统一系统来考虑这一整体背景。

此外，对于计算语言学整个学科性质的认识，还需要注意下面三个方面的问题：

（一） 计算语言学的研究内容可以有广义和狭义理解之分。正如上文已经提及的，从计算的角度来看待语言的性质，以及以自然语言为对象来研究算法过程，都属于计算语言学的核心研究内容。狭义而言，计算语言学的研究就指这两方面的研究（两方面的研究互相影响）。但从广义来说，计算语言学的研究还可以包括以计算机作为工具手段对自然语言进行的所谓“计量研究”。从狭义的观点看，这部分研究属于**计量语言学**的范畴。研究内容大致包括语言符号的频度统计；基于统计特征的语言（方言）地理分布研究；基于统计特征的计量风格学研究；等等。计量语言学的研究目标跟狭义理解的计算语言学有所不同，前者主要是利用计算机这一现代计算工具来获得语言中隐含的数量规律；而后者是要用计算机来模拟人分析处理自然语言的行为能力。当然，对语言进行的计量研究结果，无疑可以为计算机模拟人的语言处理能力提供支持。也正因为如此，广义的计算语言学也把计量语言学包括在内。

（二）除了广义、狭义之分，不同研究者对计算语言学研究内容的认识也可能会有角度或者说是侧重的不同。以上述“语言”+“计算”的格局来看，如果明显地偏重“计算”或者说是专注于对自然语言进行各种类型的信息处理和加工技术这方面的研究，通常也可以用**“自然语言处理”**这个术语来称说；如果明显地偏重“语言”的形式化研究，专注于以数学方法来刻画语言的各种特点，从而形成表述严密的语言理论体系，通常也可以用**“数理语言学”**这个术语来称说。“自然语言处理”或者“数理语言学”这些术语，是从不同角度或侧

面来凸现计算语言学的研究内容。由于“计算语言学”除了注重语言的计算理论研究，同时也非常强调计算机实现，因此，在宽泛的意义上，有时候人们也径直以“自然语言处理”来指“计算语言学”。另外，从研究对象和目标的角度看，计算语言学跟宣称以人的思维活动为研究对象，并企图以计算机来模拟人的思维活动的“**人工智能**”显然有着密切的联系。人的“自然语言”能力可以说是思维活动的一部分或者说是外部表现形式，从这个意义上说，以“自然语言”为旨趣的计算语言学可以看做是“人工智能”的一个分支。

（三）跟计算语言学的发展过程相伴，人们对计算机以及人类语言能力的认识也在不断调整和发展之中，因而先后产生了不少跟计算语言学有密切关系的术语，包括“**机器翻译**”（Machine Translation），“**自然语言理解**”（Natural Language Understanding），“**自然语言处理**”（Natural Language Processing），“**人类语言技术**”（Human Language Technology），等等。这些不同的称说，多少也暗含了人们对计算语言学的认识和追求的旨趣变化。最早把计算机和自然语言联系在一起的是机器翻译（参见下文第三节），之后由于困难重重，人们意识到翻译的过程涉及到对语言含义的真正理解，而不是一个简单的机械过程，于是开始把注意力聚焦在自然语言的理解问题上。随着人们“理解”的程度日益加深，人们一方面对自身的局限性有了更多的了解，一方面也在这个过程中积累了许多副产品，拓宽了原先对计算机与自然语言结合的认识，从“理解”变为“处理”。而后更是为了名副其实，用“人类语言技术”这个更准确的术语，来称说人们围绕人类语言（而非动物语言或形式语言）所开展的处理技术的研究。一方面突出了这个领域的技术色彩（跟传统的有关人类语言的语文学研究相区别），另一方面也把以往形成的机器翻译也好，自然语言理解也好，自然语言处理也好，只要是跟人类语言有关的信息处理问题都纳入进来了。

以上简要说明了跟计算语言学有着这样那样联系的不同术语的所指为何。不同名称的存在，自有其研究内容本身或不同时代人们认识上差异的原因，而通过本书各章对计算语言学研究内容的详细介绍，可以帮助读者对这类“名实之辨”的问题进行更准确的把握。

第二节 计算语言学的研究方法

概括而言，计算语言学的研究方法可以区分为规则方法和统计方法两大类。在计算语言学的发展历史中出现过的各种具体的自然语言处理方法基本上都可以归入这两类中的一类或是这两类方法的融合。同时，在对各个不同层面的语言对象的处理中发展出来的不同方法，也同样可用上述这两大类方法去加以区分。

那么，什么是规则方法，什么又是统计方法呢？

回答这两个问题，最好是从了解计算语言学的一般研究模式（或研究过程）开始。

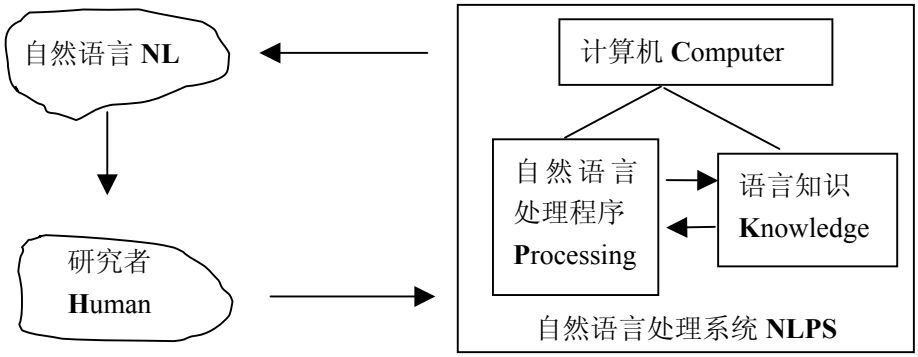

图 1-2　计算语言学的一般研究模式图解

图 1-2 至少传达了两层含义：

（1） 计算语言学研究中涉及到五大主要因素——这是静态地看——包括：

I. 自然语言 NL（包括语音、词汇、句法、语义、语用、篇章等等）；

II. 研究者 H（计算语言学家、程序员、数学家、语言学家、逻辑学家等等）；

III. 关于自然语言的知识 K（包括有关自然语言的各个方面的知识）；

IV. 计算机 C（包括硬件、操作系统以及其他相关的应用程序平台）；

V. 自然语言处理程序 P（包括算法/以程序设计语言实现的程序）；

（2） 计算语言学研究的一般过程——这是动态地看——可以描述为四个主要阶段：

S1：研究者以特定的方式对自然语言（NL）的规律进行抽象，以计算机能够处理的形式来表述关于自然语言的规律——得到所谓的语言知识 K；

S2：针对特定的语言知识表示形式，研制适合的分析和处理算法；

S3：根据算法编制计算机可执行的自然语言处理程序（P）。这样的程序加上语言知识，加上计算机硬件系统，共同构成一个自然语言处理系统（NLPS）；

S4：用这样一个自然语言处理系统对自然语言 NL_0 进行分析处理，根据反馈的结果调整原来的设计，改进 NLPS。

基于上述认识框架，所谓规则方法和统计方法的主要区别可以描述为：

（1）规则方法和统计方法在在如何认识 K，以及如何表示 K 上存在不同：规则方法主张以建立形式化知识系统的方式来表述 K；统计方法则主张搜集实际的语言材料形成语料库，将语料库本身视作 K。这样的 K 是统计意义上的知识。

（2）规则方法和统计方法在如何得到 K 上存在不同：前者通常采用所谓内省（也不排除使用小规模的语料）方式来检验、调整、改进 K，使得 K 在自然语言处理系统中能有更出色的表现；后者通常通过构造统计模型，由计算机对语料库中的语言现象进行统计，得到统计规律意义上的 K。

（3）规则方法和统计方法在如何使用 K 来构造 NLPS 上存在不同：规则方法自 20 世纪 70 年代以来发展出许多比较成熟的算法技术，包括在处理上下文无关语法中实际使用较广泛的 Tomita 算法、基于合一的线图分析方法等。统计方法自 20 世纪 80 年代中后期以来逐渐为人们所关注，发展出基于隐马尔可夫模型的自然语言处理算法，如 Viterbi 算法，基于转换的错误驱动的自然语言知识学习方法等。

以上是从 K 来看两种方法的区别，但同时也应该注意到，K 与 P 是紧密相连的。上述（2）跟（3）两方面的差异也就意味着两种方法下构造的自然语言处理程序 P 的不同。

此外，除了以上述两分的方式来描述计算语言学的研究方法，还可以从不同策略的角度来看待计算语言学研究。比如：由于 NL 是一个非常复杂的处理对象，有时需要对 NL 进行简化处理得到一个所谓受限的自然语言对象 NL'，来降低处理的难度。这也就是所谓受限自然语言（controled-language）的研究。

第三节 计算语言学的实际应用

随着计算机的日益普及，计算从桌面走向网络，从主要集中在办公环境走向人们日常生活的各个角落，伴随着计算机的应用领域的不断拓宽，计算语言学的应用也呈现日益广泛的趋势。而就计算语言学这一学科本身的发展历史而言，可以说，正是来自“应用”的契机才触发了这门学科的诞生，并且可以说，自计算语言学这门学科诞生之后，来自实际应用的需求也一直都是计算语言学前进发展的主要动力之一。

20 世纪 40 年代以来，几乎跟数字电子计算机的问世同步，人们就产生了利用计算机来代替人类完成一些非数值领域的工作的想法，这其中最引人注目的目标恐怕要算是用计算机

将一种语言自动翻译成另外一种语言了。1954 年，美国 Georgetown 大学与 IBM 公司合作，在 IBM 701 型计算机上进行了第一次机器翻译试验，将俄语翻译成英语。这大概是世界上首次将计算机应用到非数值计算的信息处理领域。此后的半个世纪，机器翻译作为自然语言处理的核心研究领域，潮起潮落，经历了并不平坦的发展之路。1966 年 11 月，美国科学院下属的语言自动处理咨询委员会（Automatic Language Processing Advisory Committee，简称 ALPAC，这个委员会是 1964 年成立的）向美国国家基金会提交了一份关于机器翻译的咨询报告。该报告对机器翻译下了一个否定性的结论，称机器翻译的目标是不现实的，在可预见的将来没有成功的希望。

此后一段时间内，机器翻译的研究跌到低谷。在这段时期，研究人员开始反思机器翻译失败的原因，由此也引发了对自然语言的性质本身更深刻的关注，70 年代先后提出了一些有关自然语言知识表示和处理的理论和方法。这些理论和方法除了在机器翻译研究中进行尝试，还将自然语言处理的研究扩展到更广阔的应用领域，如智能计算机人机接口、专家系统自然语言接口等。

由于计算机软硬件技术本身的发展，从 80 年代开始，个人计算机系统（PC）迅速普及，这也使得信息处理在人们的社会生活中的地位比以往任何时候都重要。与这一大背景相适应，机器翻译软件、自然语言人机接口软件等陆续从实验室走向市场。如果说从 50 到 70 年代，计算语言学的应用主要还是停留在实验室阶段，那么从 20 世纪 80 年代开始，计算语言学的应用就可以算是开始步入社会生活了。

到 20 世纪 90 年代以后，随着网络技术和 Internet 在全球范围内的飞速发展，可以说，人类社会开始以加速度的方式进入到信息时代的新纪元。在这样的潮流下，计算语言学的应用领域也随着全球网络的拓展而拓展，开始出现互联网上的在线机器翻译、跨语言的信息检索、多语通信系统等。此外，自然语言处理系统的开发者也呈现日渐重视用户使用的趋向。以用户易用性为主要考虑（定位）的自然语言处理软件，如机助人译系统、人助机译系统、计算机辅助写作系统等工具应运而生；随着无线通信网络技术的日趋重要，面向新型的小型移动计算设备（而非传统的 PC）的自然语言语音接口（包括语音输入、语音识别、口语翻译等）技术也日益受到研究人员和系统开发人员的注目。

不难看出，从盲目乐观到顾本务实，从实验玩具到上市产品，计算语言学的应用之路虽然崎岖不平，但终归是朝着让今日匪夷所思之事成为明日平常现实的方向在前进。

本书第六章和第七章将着重对计算语言学一些主要的应用领域作详细的介绍，包括：机器翻译、信息检索、信息提取、文本分类等。

第四节 小 结

以上只是勾勒了计算语言学这门学科的大致轮廓。对什么是计算语言学，只有深入到研究内容的细节，才能真正有所理解。为便于读者了解本书章节布局的用意，下面将本书总的结构安排归纳一下：

绪论 ：第一章（勾勒出一个有关计算语言学的粗线条的宏观图景）

第一部分 从计算的角度看语言（基础篇） ：第二、三章 ----- 理论

第二部分 以语言为对象的计算（算法篇） ：第四、五章 ----- 理论

第三部分 计算语言学的应用 （应用篇） ：第六、七章 ----- 实践

本书不敢说已窥得计算语言学堂庑之妙，但希望能充当一个好领路员的角色，为读者标引出一条登堂入室的路径。作为一门植根于多个学科土壤之上的交叉型学科，计算语言学的研究触角涉及到了包括计算机科学、语言学、数学、认知科学、逻辑学等多个学科在内的研

究范畴。交叉学科的特点决定了对一个问题的看法和研究可以从许多角度进行，同时也要求人们广泛地涉猎多学科的营养来丰富自身对“语言”—— 这一人类精神家园所栖身的土壤—— 的认识。

思考和练习

1. 什么是计算语言学？
2. 如何认识“机器翻译”、“自然语言理解”、“自然语言处理”、“人类语言技术”等概念之间的区别与联系？
3. 简述计算语言学的发展历史。
4. 计算语言学与理论语言学研究有什么联系和区别？

第一章参考文献

冯志伟 《自然语言机器翻译新论》，语文出版社，1995。

冯志伟 《自然语言的计算机处理》，上海外语教育出版社，1996。

刘开瑛、郭炳炎 《自然语言处理》，科学出版社，1991。

陆汝钤 《数学·计算·逻辑》，湖南教育出版社，1993。

陆汝钤 《人工智能》（上册），科学出版社，2000。

钱锋 《计算语言学引论》，学林出版社，1990。

姚天顺等 《自然语言理解 —— 一种让机器懂得人类语言的研究》，清华大学出版社、广西科学技术出版社，1995。

Dreyfus, Hubert L., *What Computer can't do*, Harper & Row, Publishers, 1979.

Grishman, Ralph, *Computational Linguistics: An Introduction*, Cambridge University Press. 1986.

Searle, John R., Minds, Brains, and programs, In *The Behavioral and Brain Sciences* Vol.3, 1980.

Turing, A.M., Computing Machinery and Intelligence. In *Mind* Vol.59, 1950.

Hutchins, W. John, Machine Translation over Fifty Years, In *Histoire, Epistemologie, Langage*, Vol. 22, No. 1, 2001.

Hutchins, W. John, Machine Translation and Human Translation: in competition or in complementation? In *International Journal of Translation* Vol. 13, No. 1-2, 2001.