DOI:10.16499/j.cnki.1003-5397.1992.01.015

计算语言学对理论语言学的挑战

冯志伟

计算语言学是采用计算机技术来研究和处理自然语言的一门新兴学科。计算语言学对自然语言的研究和处理,一般应经过如下三个方面的过程:

第一,把需要研究的问题在语言学上加以形式化(linguistic formalism),使之能以一定的数学形式,严密而规整地表示出来;

第二,把这种严密而规整的数学形式表示为算法(algorithm),使之在计算上形式化(computational formalism);

第三,根据算法编写计算机程序,使之在计算机上加以实现(computer implementation)。

因此,为了研究计算语言学,我们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。这样,计算语言学就成为了一门介乎语言学、数学和计算机科学之间的边缘性的交叉学科,它同时涉及到文科、理科和工科三大领域。

计算语言学的研究是从机器翻译开始的。1946 年电子计算机刚一问世,人们在把计算机广泛地应用于数值运算的同时,也想到了利用计算机把一种或几种语言翻译成另外一种语言或几种语言。从 50 年代初期到 60 年代中期,机器翻译一直是计算语言学研究的中心课题,当时采用的主要是"词对词"翻译方式,这种不是建立在对自然语言理解的基础上的简单技术,没有得到预期的翻译效果。60 年代中期,人们开始转入对自然语言的语法、语义和语用等基本问题的研究,并尝试着让计算机来理解自然语言。许多学者认为,断定计算机否是理解了自然语言的最直观的方法,就是让人们同计算机对话,如果计算机对人用自然语言提出的问题能作出回答,就证明计算机已经理解了自然语言,这样,就出现了"人机对话"(或"自然语言理解")的研究。计算语言学的理论和方法也就在这些具体的研究中逐渐形成、成熟并完善起来。目前,除了机器翻译和自然语言理解之外,计算语言学的研究领域还扩展到了自然语言人机接口、语音自动识别与合成、自然语言情报检索、术语数据库、风格学研究等领域。计算语言学已经在世界范围内引起了广大学术界的瞩目,成为了一个独立的学科。它象一股强劲的东风吹进了传统的理论语言学的许多部门,使这些部门面目一新。

计算语言学对传统的形态学提出了新问题。在机器翻译和人机对话的研究中,都要进行形态分析,这就促进了形态学的研究。传统的形态学都要区分屈折(inflection)和派生(derivation)。如英语的 amend/amended 是屈折,amend/amendment 是派生,前者作为词形变化看待,后者作为构词法问题看待。然而对于计算机来说,并没有必要一定要作这样的区分。通

常是把 amended 和 amendment 都归入 amend 进行统一的处理。一个自动形态分析方案可包括一部词干词典和一套描述词形变化和构词的规则系统，其中既有派生，也有屈折。这样，在分析时，给出词干，计算机就可以自动地列出它的所有的变化形态，而给出一个变化形式，计算机就可以自动地把它切分为词干、词缀和词尾。计算机还要求区分各种同形现象，例如，英语 frighten 中的－en 要与 oven 中的－en 区别开来，reaped 中的－ed 要与 reed 中的－ed 区别开来。另外，还要考虑一些特殊的现象。如 perform、give、go 等动词的过去时形式分别为 performed、gave、went，city 的复数形式 cities 在去掉词缀之后，还要把词干的形式作些改变，编写形态分析程序时，应该设法使这些各不相同的情况条理化。在机器翻译欣欣向荣的 50 年代末和 70 年代初，学者们曾经对俄语、德语这样一些屈折变化丰富的语言进行过严格的形态分析，编制过相当精细的自动形态分析规则。目前，在机器翻译和人机对话中的自动形态分析技术已经十分成熟。

计算语言学对于传统的句法学冲击最大，各种立足于自然语言自动处理的句法分析理论和方法犹如雨后春笋应运而生，形成了百花齐放的局面。

在机器翻译研究的早期，苏联数学家库拉金娜(О. С. КУ ЛАГИНА)就用集合论方法建立了俄语句法的数学模型，精确地定义了一些语法概念，这一模型成为了苏联科学院数学研究所和语言研究所联合研制的法俄机器翻译系统的理论基础。著名数理逻辑学家巴希勒(Y. Bar－Hillel)提出了范畴语法(category grammar)，建立了一套形式化的句法和演算规则，通过有穷步骤，可以判断一个句子是否合乎语法。这些，都大大地推动了传统句法分析方法向精密化、算法化的方向发展。

乔姆斯基的形式语言理论是影响最大的早期计算语言学的句法理论。乔姆斯基定义了 0 型语法、上下文有关语法、上下文无关语法和正则语法 4 种类型的形式语法。其中的上下文无关语法又叫做短语结构语法(phrase structure grammar，简称 PSG)。这种短语结构语法广泛地应用于自然语言的自动分析和生成中。但是，人们不久就发现，短语结构语法的分析能力不高，难以区分大量的歧义句子，短语结构语法的生成能力过强，往往会生成大量的不合语法的句子。就是乔姆斯基本人，也认为短语结构语法不能充分地描述自然语言。于是他提出转换语法来克服短语结构语法的这些弱点，后来转换语法逐渐发展成为转换生成语法。不过，这种生成转换语法的分析效率也不高，并没有在实际的自然语言处理系统中受到欢迎。由于短语结构语法结构清晰，易于操作，计算语言学的学者们抛弃了转换生成语法，又转向短语结构语法，于是出现了各种增强的短语结构语法。例如，受限语言(restricted language)和扩充转移网络(augemented transition network，简称 ATN)。受限语言的表层结构分析和深层结构生成是分别进行的，而 ATN 的表层结构分析和深层结构生成是同时进行的。60 年代后期，查斯特里(Chastellier)把程序设计语言的 W－语法引进了自然语言处理中，他证实了英语和法语的转换语法都可以通过这样的 W－语法来重写。

美国语言学家布列斯南(J. Bresnan)主张建立面向词汇的非转换的语法，她和卡普兰一起，于 1983 年提出了词汇功能语法(lexical－functional grammar，简称 LFG)。马丁·凯依于 1983 年提出了“合一语法”(unification grammar，简称 UG)，于 1985 年提出了“功能合一语法”(functional unificational grammar，简称 FUG)。盖兹达(G. Gazdar)、克莱因(E. Klein)、沙格(I. Sag)和普鲁姆(G. Pullum)等人于 1985 年提出了“广义短语结构语法”(generalized phrase structure grammar，简称 GPSG)。珀拉德(C. Pollard)于 1984 年在他的博士论文中，提

出了“中心词语法”(head grammar),1985年又和他的同事们一起提出了“中心词驱动的短语结构语法”(head－driven phrase structure grammar,简称HPSG)。这些语法都采用了复杂特征结构来改进短语结构语法,采用合一运算来改进传统的集合运算,从而有效地克服了短语结构语法的缺点,保持了短语结构语法的优点。

理论语言学中的层次分析法实质上就是短语结构语法,因此,短语结构语法在计算机分析和生成自然语言时出现的各种问题,在层次分析法中也同样是存在的。上述的这些旨在改进短语结构语法的计算语言学理论,都带有很强的可操作性,具有强烈的方法论色彩,必定会有助于理论语言学中广泛使用的层次分析法的改进和完善。在这方面,我们应该提倡理论语言学家和计算语言学家进行经常的对话,互相学习对方的长处,共同来解决短语结构语法在应用中出现的各种问题。

计算语言学对句法学的如此巨大的影响,使我们想到了建立汉语产生式语法的问题。

不论那一种计算语言学的语法,其最根本、最关键的问题,是要指出各种语言形式出现和变换的条件,只有指出了条件,计算机才可能根据有关的条件,执行相应的动作,从而使整个系统成为一个可以动态地执行的过程。

不论那一种计算机,在执行有关程序时,总免不了给它指出条件,有了条件,并且让计算机知道究竟是什么样的条件,计算机才可能执行相应的动作。总而言之,计算机的任何操作,归根结底,可以归结为一个公式:

条件→动作

即在一定的条件下,执行一定的动作,在另一条件下,执行另一动作。这样的“条件—动作”偶对,是一切计算机工作的最基本的方式,因此,要使自然语言的语法规则成为可供计算机执行的形式,我们就必须指出各种语法现象出现的条件。

我国计算语言学的学者们多年来从事中文信息处理的研究工作,曾经提出了一些自然语言处理的算法,但是,在很长的时间内,由于我们对于自然语言形式化处理的关键问题不十分清楚,所以,这些算法,有的成功了,有的失败了,凡是成功了的算法,都是由于我们比较充分地研究了语言形式出现的条件,凡是失败了的算法,或者是由于我们根本没有提出语言出现的条件,或者是我们虽然提出了语言形式出现的条件,但是条件给得不具体、不精确,或者是条件给错了。积多年之经验,我们深知条件对于建立计算语言学语法的重要性,“条件—动作”偶对,确实是建立计算语言学语法的最基本、最关键的公式。由于汉语中单词或词组的种类与它们的句法功能之间没有明确的对应关系,语言成分的句法功能与它们的语义关系之间也没有明确的对应关系,所以,在汉语的计算语言学中,认真研究现代汉语的各种“条件—动作”偶对,就显得更加重要了。

我国汉语语法研究已取得很大的成绩,尽管过去的汉语语法研究没有专门考虑到计算语言学的需要,但是,汉语语法的许多研究成果都是自觉或不自觉地体现了“条件—动作”偶对这一公式的原则,因此,这些成果都程度不同地能够在汉语的计算语言学中得到运用。

例如,我们在进行汉语的自动生成时,起初以为“把字句”的作用是把及物动词的宾语提前,其实,这是一个极不严格的条件,我们把这样的条件写到程序中,凡是及物动词的宾语都用“把”字提前了,结果形成了通篇的把字句。实践使我们认识到,把字句的出现条件不只是及物动词的宾语提前,还有着更为严格的条件,进一步学习汉语语法研究的有关文献,我们加上了如下限制条件:①“把”字组成的连动结构,其中的动词不能是单纯的单音节或双音节动词,而

是一个比较复杂的动词组合；②“把”字的宾语在语义关系上是后边动词的受事，而不是一般的宾语；③“把”字的宾语在意念上是确定的、特指的。

根据这些规律对把字句的出现条件作了进一步的限制，结果计算机生成的把字句基本上正确了。

后来，我们根据汉语语法研究的有关结果，把上述条件进一步加以概括，得出这样更简练的规律：凡是受事主语句的主语之前，都可以加“把”字形成把字句。例如，“门开着”、“门关了”、“他免了职”等受事主语句，主语前加“把”字就可以形成“把门开着”、“把门关了”、“把他免了职”等把字句。找出了这样的概括性更高的条件，就能更好地通过简单的程序来有效地控制把字句的生成了。

其实，人学习语言的情况与计算机处理语言的情况有许多相似之处。一个学汉语的外族人，他必须知道汉语的各种语法现象的出现条件，才有可能去正确地使用它。现代英语语法对于动词的各种时态的出现条件作了比较确切的说明，因此，学习英语的人可以很快地掌握它，从而造出各种合乎规范的句子来。在学习英语时我们之所以觉得英语语法十分有用，非学不可，就是因为这样的语法是一种讲条件的语法。有的人之所以觉得汉语语法无用，是因为大多数汉语语法书只罗列现象，很少讲这些现象出现的条件。由此观之，不论从计算语言学还是从外族人的汉语教学来说，建立一套讲条件的汉语语法，就成为一件十分重要的事情了。

50年代末60年代初，美国描写语言学的方法介绍到中国来之后，我国现代汉语研究受到美国描写语言学的影响，比较注意语言现象本身的详尽描述，而不太注意对这些语言现象的解释。描述语言学现象是完全必要的，而且这是语言研究不可缺少的第一步。如果不详细地占有语言材料，不从各个方面、各个角度来描写语言现象，当然也就谈不上对语言现象的解释，语法研究就有如做无米之炊。但是，如果只停留在描述的水平，不进一步对这样的描述作出解释，那还不能算是探究了学问的根本。从应用的角度来看，如果不对语言现象的出现条件作出解释，这样的描写对人们学习语言的实践以及计算语言学的研究就很难发挥应有的作用。为了使汉语的研究更好地为我国的四个现代化事业服务，有必要把汉语语法研究的重点，逐渐地从描写的立场转移到解释的立场上来。语言研究者应该进一步钩深致远，尽其所能地把他们所描写的各种语言现象的出现条件说清楚，不但要说明语言现象是什么样，而且要说明条件，解释这些现象何以会这样，从而建立解释性的汉语语法体系。

我们主张建立解释性的汉语语法体系，一点儿也不意味着要削弱汉语描写语法的研究，恰恰相反，我们还要进一步描写汉语的各种现象，揭示各种语言现象之间的细微差别。但是，我们作描写研究的目的，不是为描写而描写，而是要对这样的描写作出解释，说明其出现的条件，使这样的描写成为人们可以通过智能活动掌握的东西，成为计算机可以使之程序化的东西。

美国人工智能专家塞蒙（H. A. Simon）和列维尔（A. Newell）提出了“产生式系统”（production system），并论证了这种产生式系统与智能活动的关系。他们认为，智能活动可以分解为一系列最基本的单位，这些基本单位可以归结为两种：

第一，根据某种环境采取某种行动；

第二，根据某一前提作出某种结论。

所谓“人们有智能”就意味着，人们能够根据某种特定的环境产生某种行动，或者根据某一特定的前提产生某种结论。这种基本智能活动的单位就叫做产生式。

由这种产生式可以构成一系列比较复杂的认知过程。人类社会的许多科学和文学杰作，最

后都可以归结为这样的产生式。从人类进化的过程来看，是由一些简单的产生式系统发展成复杂的产生式系统；从一个人智能发展的过程来看，先由发展简单的产生式系统开始，逐渐发展到复杂的产生式系统，而这两条脉络的共同基础，就是上述的产生式。

把产生式系统的理论同前面提出的"条件—动作"偶对的那种观点相比较，可以发现它们之间是多么的相似！由此可以看出，既然人类的智能活动是建立在产生式的基础之上的，那么，要把语言这种人类的复杂的智能活动形式化，其最关键的问题，当然就是要为某种语言建立起一系列"条件—动作"偶对的产生式系统；要使语言便于人们学习或掌握，其最根本的问题，当然也就是要告诉人们如何根据特定的条件来运用语言中的各种规则。所以，解释性的汉语语法体系实际上就是汉语语法的产生式系统。建立汉语产生式语法，应该是汉语的计算语言学在句法研究方面的最重要的任务。

70年代以来，国外建立了一些立足于语义的自然语言理解系统，使长期不受重视的语义学得到了发展，计算语言学也影响到了语义学方面。

近数十年来，不少语言学家认为，语义学不应该是语言学的一个分支，他们只关心语言的形式研究，而把语义的研究推给哲学或其它学科去进行。但是，随着机器翻译和自然语言理解研究工作的进展，再加上语言学理论论战的需要，促使语言学家去研究语义学。通过研究的实践，学者们逐渐认识到，甚至句法的研究也是不可避免地与语义学纠缠在一起的，因此，他们又重新对语义学发生了兴趣，并且这种兴趣迅速地与日俱增。

哲学家们曾经提出过意义公设系统，它包括规则系统、蕴涵符号(—>)、逻辑连词(and、or、not)等，这样，便可以把词的意义分解为若干个基本意义组成的意义公设系统。在意义公设系统中，词的意义可以由一组语义公设来确定。哲学家们这些研究，为计算语言学中的语义研究打下了基础。在这种情况下，一些语言学家，如美国的弗托(J. A. Fodor)和玛考利(J. D. McCauley)等又把语言和逻辑相互关系这样的问题重新提了出来。乔姆斯基关于表层结构和深层结构的理论，把语义问题提到了相当的高度，卡茨(J. katz)和弗托等提出了解释语义学，采用成分分析法，利用语义成分、标记和关系来定义词符成分，并加上一些控制和选择限制来演绎地解释句子的语义。这样的研究对于计算语言学很有帮助。菲尔摩(C. J. Fillmore)提出了格语法(case grammar)，从句子的深层句法表示来推导句子的表层结构，较好地解决了句法与语义相结合的问题。格语法规则产生的结构，不仅与句法相关，而且与语义相关，给计算语言学的研究提供了方便，格语法在计算机上的分析效率也比较高，受到了计算语言学家的欢迎。玛考利等提出了生成语义学，他们一开始就用语义结构来描述句子，然后通过一系列的转换由这种语义结构产生出表层结构，而用不着对深层结构作任何说明。威尔克斯(Y. A. Willks)提出了优选语义学(preference semantics)，并把这种理论用于机器翻译系统的研究中。美国数理逻辑学者蒙德鸠(R. Montague)提出了蒙德鸠语法(Montague grammar)，美国计算机科学家杉克(R. C. Schank)提出了概念依存理论(conceptional dependency theory，简称CD理论)，美国人工智能学者西蒙(R. F. Simmons)提出了语义网络理论。这些理论都十分强调语义的作用，在计算语言学的应用中，有的理论(如CD理论)直接以语义模型制导，辅以句法检查，打破了以句法模型制导，辅以语义检查的传统格局，实现了自然语言处理的句法——语义一体化。

美国学者汉德雷斯(Handres)在描述一种语言的过程时，把大量的语义信息植入该语言的句法中，这样定义的句法系统叫做"语义语法"(semantic grammar)。语义语法提高了自然语言的处理速度，效率较高，后来被许多实时处理的自然语言系统所采用。

近年来，由于语义学与句法学的联系日趋密切，逻辑语法有了很大的发展。逻辑语法(logic grammar)是指用谓词逻辑来表达的语法，它是逻辑程序设计和计算语言学相结合的产物。在机器翻译和自然语言理解的研究领域里，经常使用谓词逻辑来描述知识和进行逻辑推理。70年代以来，逻辑以PROLOG语言作为形式被应用于程序设计，谓词逻辑就不再仅仅用于描述知识和逻辑推理的问题，还作为逻辑程序设计的工具来描述解决问题的过程。PROLOG语言使得逻辑和程序设计这两个相距甚远、完全不同的概念协调统一为一个单独的概念——逻辑程序设计。在用PROLOG语言来解决计算语言学的各种问题的研究过程中，逻辑语法日益成熟起来。目前主要有4种影响较大的逻辑语法：定子句语法(definite clause grammar，简称DCG)，外位语法(extraposition grammar，简称XG)，修饰成分结构语法(modifier structure grammar，简称MSG)，约束逻辑语法(restricting logic grammar，简称RLG)。这些语法巧妙地把逻辑和句法结合起来，使描述性的形式语法具备了推理的能力，这是计算语言学研究中应该注意的一个问题。

这里特别值得一提的是定子句语法。这是瓦楞(D. Warren)和佩瑞拉(P. Pereira)于1980年提出的一种仅仅使用短语结构语法规则的逻辑语法。定子句语法的基本思想是：语法中所用的符号不仅仅是原子符号，还可以是广义的逻辑项。例如，短语结构语法的规则

sentence⇒noun－phrase，verb－phrase

表示一个句子由名词短语和动词短语两部分组成，在定子句语法中，同样的这个规则可以表示：如果存在一个名词短语和一个动词短语，那么，就存在一个句子的推理过程。短语结构语法的规则与定子句语法的规则在形式上虽然有许多相似之初，但是在本质上却有很大的区别，短语结构语法只是用于描述一种语言，而定子句语法则可以用来进行语言的推理。这样，定子句语法便实现了从描述性的形式语法到推理性的逻辑语法的转变，从而使短语结构语法产生了质的飞跃。

由于定子句语法的符号是逻辑项，这就使得定子句语法规则中的非终极符号可以携带有关上下文、转换、结构等多方面的信息，增强了短语结构语法描述自然语言复杂特征的能力。而且，定子句语法规则的右部不仅可以为终极符号和非终极符号，还可以携带测试条件的信息，便于描述自然语言的规律。这样，定子句语法虽然在形式上采用了短语结构语法，但它的描述能力已经相当于乔姆斯基提出定义的0型语法了。所以，定子句语法是对乔姆斯基短语结构语法的一个重大改进。这是计算语言学对理论语言学作出的又一贡献。

语言在实际使用时，总是以篇章或话语的形式出现的，省略和指代以及单词和句子的歧义问题一般要在上下文背景之下才能解决，而要在字里行间找出说话者的真正目的，则需要根据广泛的关于客观世界的知识和其它信息才有可能知其端倪。因此，计算语言学中还出现了一些关于篇章处理和话语分析的理论和方法，如脚本(script)、规划(plan)、故事语法(story grammar)、故事树(story tree)等。计算语言学对如何处理省略、指代、话题、照应关系以及篇章结构等问题，也进行了一些有益的探讨。这些都推动了语义学的发展，并且使语义学与语用学紧密地联系起来。1983年，美国斯坦福大学的巴威斯特(Barwist)和佩利(Perry)出版了《情景和态度》(Situation and Attitudes)一书，提出了情景语义学(situation semantics)的自然语言模型。所谓"情景"，就是个体、性质、关系和时空位置等构成现实世界(非语言环境和场面)的各种状况的集合，可以利用这样的情景来描述语言的语义。可见，情景语义学已经把一般的语义学和语用学紧密地结合起来，对自然语言的研究有重要作用。

情景语义学一提出就引起世人的瞩目。斯坦福大学为此成立了语言与信息研究中心(Center for the Study of Language and Information,简称 CSLI),专门在情景中来研究自然语言。CSLI 由 17 位来自斯坦福大学计算机科学系、语言学系和哲学系以及斯坦福国际研究所(SRI)的著名的老资格科学家组成,阵容十分强大。

CSLI 当前的任务是:

①把自然语言研究扩展到情景的领域。

②把计算机语言的研究也进一步扩展,使之能处理信息的内容和嵌套世界的情景。

③在整个科学哲学和数学基本原则的基础上,把传统的自然语言和计算机语言的理论融合为一个综合的整体,使自然语言和计算机语言的研究朝着统一的方向发展。

情景语义学在言谈分析和理解、上下文处理、照应关系、动词时态、话语焦点、篇章结构的研究方面都取得了可喜的成果。

计算语言学还促进了词汇学的发展。词典编纂历来是一件十分枯燥乏味而极为辛苦的工作。计算机使得这件工作变得简单易行、轻松愉快。计算机可以给词典提供足够的例句,免去了手工编纂时转抄大量卡片的麻烦;计算机可以通过单词频度和使用度的统计,确定常用词和通用词,编写出各种语言的基础词表和频率词表。近年来,还出现了各种形式的电子词典,这种词典中存贮着丰富的语言信息,为机器翻译和计算语言学其它部门的研究提供了基本的静态语言信息。日本成立了电子词典研究所,专门研究电子词典的理论和应用问题。现在,在许多国家,电子词典的编制已经成为了一种产业。

在计算语言学的推动下,文字学研究开始同图象识别的方法结合起来。因为文字也是一种图象,图象识别中采用的许多方法,如图象识别的句法分析方法,也可用到文字识别中去,这方面的工作,在美国和日本都取得了很大的成就,这也许会给古老的文字学研究开辟出一片新天地。

我国的汉字识别研究独具特色,采用选取汉字特征点和数学形态学的方法来提取汉字的结构特征,在印刷体汉字识别方面,已经研究出一批实用系统,部分系统已经商品化。这些系统一般都具有版面分析、文本识别、识别结果后处理、自动纠错、自动编辑、自动输出等功能。在联机手写体汉字识别方面,识别率正逐渐提高,已达到部分商品化的水平。

计算语言学还影响到了语言材料的搜集、整理和加工。理论语言学的研究必须以语言事实作为根据,必须详尽地、大量地占有材料,才有可能在理论上得出比较可靠的结论。传统的语言材料的搜集、整理和加工完全是靠手工进行的,这是一种枯燥无味、费力费时的工作。计算机出现后,人们可以把这些工作交给计算机去作,大大地减轻了人们的劳动。后来,在这种工作中逐渐创造了一整套完整的理论和方法,形成了语料库语言学(corpus linguistics),并成为了计算语言学的一个分支学科。语料库语言学主要研究机器可读自然语言文本的采集、存储、检索、统计、语法标注、句法语义分析,以及具有上述功能的语料库在语言定量分析、词典编纂、作品风格分析、自然语言理解和机器翻译等领域中的应用。现在,美国建立了布朗语料库,英国和挪威联合建立了 LOB 语料库。欧美各国学者利用这两个语料库开展了大规模的研究,其中最引人注目的是对语料库进行语法标注的研究。他们设计了自动标注系统 TAGGIT 来给布朗语料库的 100 万词的语料作自动标注,正确率为 70%。他们还设计了 CLAWS 系统来给 LOB 语料库作自动标注,根据统计信息来建立算法,自动标注正确率达 96%,比基于规则的 TAGGIT 系统提高了将近 20%。最近他们同时考察三个相邻标记的同现频率,使自动语法标注的正确率

达到99.5%。这个指标已经超过了人工标注所能达到的最高正确率。

计算语言学不仅影响了传统理论语言学的上述部门，而且，还强烈地冲击着索绪尔(De Saussure)以来的普通语言学基本理论，以大量的新的事实和研究成果，严峻地考验着这些基本理论。

我们这里只是谈一谈关于语言符号的特性的问题。索绪尔在他的《普通语言学教程》一书中，曾提出语言符号具有如下两个重要的特性：

一、符号的任意性。语言符号的能指和所指联系是任意的，索绪尔认为，符号任意性的原则"支配着整个语言学，它的后果是不胜枚举的，人们经过许多周折才发现它们，同时也发现了这个原则是头等重要的"。〔1〕

二、能指的线条性。索绪尔指出，语言的能指属于听觉的性质，只在时间上展开，而且具有借自时间的特征：(1)它体现为一个长度，(2)这长度只能在一个向度上测定，它是一条直线。索绪尔认为"这是一个似乎为常人所忽视的基本原则，它的后果是数之不尽的，它的重要性与符号任意性的规律不相上下，语言的整个机构都取决于它"。〔1〕

索绪尔提出的语言符号的这两个特性，当然是十分重要的。然而，索绪尔以后现代语言学的发展，特别是电子计算机出现以后计算语言学的发展，严峻地考验着索绪尔的理论。

在我们看来，索绪尔提出的语言符号的任意性这一特征是无可非议的，但是，他提出的语言符号的第二个特征——能指的线条性就未必是正确的了。因为新的研究结果表明，语言的能指并不只是线条性的东西。英国著名语言学家弗斯(J. K. Firth)提出"跨音段论"(prosodic)，他认为，在一种语言里，区别性语音特征不能都归纳在一个音段位置上，例如，语调就不是处于一个音段位置上，而是处于前后相续的线条性的音段之外，笼罩着或管领着整个句子的东西。如果我们把语调这样的跨音段成分算进去，语言的能指就不宜看作是线条性的东西，而应该看作是立体性的东西了。

索绪尔是一个出色的天才的语言学家，他是名副其实的现代语言学的奠基人，他的语言学说，是语言学史上哥白尼式的革命，对于现代语言学的发展有着深远的影响。现代语言学的每一个领域，每一个流派，都直接或间接地受到了索绪尔语言学说的影响。他所说的语言符号的上述两个特性，是在当时的语言学和自然科学发展的水平下提出来的。在索绪尔的时代，还没有电子计算机，计算语言学这样的新兴学科还远远没有形成，语言学主要是与语言教学、文学、历史、考古学等学科有联系。在这种情况下，索绪尔当然不可能提出那些只有在电子计算机时代才能揭示出来的语言符号的新特点。

随着电子计算机的出现和发展，特别在计算语言学出现之后，普通语言学的理论也应该相应地发展。我们不能墨守成规，满足于旧有的结论，而应该站在前辈学者的双肩上，高瞻远瞩，吸取计算语言学的新成果，从新的角度，用新眼光，以新的方法来研究语言这一个极为复杂的符号系统。正是基于这样的认识，我们觉得，语言符号除了索绪尔所指出的那两个不尽完善的特点之外，还有着如下7个十分引人注目的特点。

第一，语言符号的层次性

弗斯的"跨音段论"已证明，语言符号并不是线条性的东西，而是立体性的东西。所谓立体性，就是说，语言符号具有分层结构，即层次性。

语言符号的层次性，在句子结构方面表现得特别明显。

美国描写语言学派的语言学家早就指出，英语的"The old men and women stayed at

home"(年老的男人和女人留在家里)这句话是有歧义的。如果我们把这一句话说给一些人听,很可能有的听话人会认为它的意思是"年老的男人和所有的女人(不论年龄大小)留在家里",另一些听话人会认为它的意思是"所有年老的男人和所有年老的女人留在家里",还有的听话人干脆不能作出决定,处于模棱两可的状态。

事实上,"old men and women"这个名词短语根据意义的不同有两种不同的层次结构。如果注意到层次的不同,那么,这种意义上两可的情况就可以得到解释。

一种层次结构是

old men and women

这时,这个名词短语的意义是:"年老的男人和所有的女人"。

另一种层次结构是

old men and women

这时,这个名词短语的意义是:"所有年老的男人和所有年老的女人"。

在计算语言学中,常采用树形图来表示语言符号的层次关系。计算语言学认为,任何一个句子的线性序列的表层之下,都隐藏着一个层次分明的树形图。当一个句子的线性序列之下隐藏着一个或两个以上的树形图时,这个句子就会产生歧义,就会得到不同的解释。

树形图由结点和连接结点的枝组成。树形图的各个结点之间,有两种关系值得注意:一种是支配关系,它反映了上下结点之间的先辈和后裔的关系,一种是前于关系,它反映了左右结点之间前位和后位的关系。语言符号的线条性只反映了前于关系,而没有反映支配关系,当然就有很大的局限。

树形图与计算语言学中广为应用的短语结构语法有着明显的对应关系。乔姆斯基的短语结构语法,既能描述自然语言,也能描述程序设计语言。短语结构语法可定义为一个四元组 G=(VN,VT,S,P),其中,VN 是范畴符号的集合,VT 是单词符号的集合,S 是初始符号,P 是重写规则。P 的规则形式为 A→ω,A 是 VN 中的单个符号,ω 是非空的符号串。如果有某个树形图满足下列条件,它就是短语结构语法 G 的推导树:

①每一个结点有一个标记,这个标记是 VN∪NT 中的符号

②根的标记是 S;

③如果结点 n 至少有一个异于其本身的后裔,并有标记 A,那么,A 必定是 VN 中的符号;

④如果结点 n1,n2,…,nk 是结点 n 的直接后裔,从左到右排列,其标记分别为 A1,A2,…,Ak,那么,A→A1A2…Ak 必定是 P 中的重写规则。

计算语言学建立的短语结构语法与树形图之间的这种联系,正是基于对语言符号层次性的认识的基础之上的。短语结构语法和树形图被广泛地使用于计算语言学中,几乎每一个计算语言学研究者天天都要与短语结构语法和树形图打交道,天天都要研究语言符号的层次关系。计算语言学的发展,进一步加深了我们对于语言符号的层次性的认识,语言符号的层次性、确实是一个比索绪尔提出的语言符号的线条性更为深刻的特性。

第二,语言符号的非单元性

基于对语言符号的层次性认识的基础之上的短语结构语法,在机器翻译和自然语言理解的研究中很快就暴露出了它的不少缺陷。这种语法分析能力不高,分析时难于处理歧义等自然

语言中普遍存在的问题，常常捉襟见肘，进退维谷；这种语法生成能力过强，往往会生成许多歧义的句子或不合语法的句子，使人误入迷津，扑朔迷离。后来，计算语言学研究者发现，引起这些缺陷的症结在于，短语结构语法是采用单标记来描述语言符号的，它把语言符号看成是不可分割的原子式的单元。如果把语言符号看成是可以分割的非单元性的东西，采用多标记函数或者复杂特征来描述，便可以从根本上克服短语结构语法的上述缺陷，大大地改善短语结构语法的功能，提高它过弱的分析能力，限制它过强的生成能力。这样，便提出了语言符号的非单元性问题。

其实，早在1936年，美国语言学家雅可布逊(R. Jakobson)在比利时根特城举行的第三届国际语音学会议上，就提出了能否以对分法为基础来分解元音、辅音等音位的问题。1951年，他在与范特(M. fant)、哈勒(M. Halle)等语音学家合写的论文《语音分析初探》中，提出了对分法理论以及区别特征学说。他们认为，一切的音(无论元音或是辅音)都是可分的，可以根据它们的生理的或声学的特性，用对分法分成一对一对的“最小对立体”(minimum pairs)。例如，元音的舌位有“高—低”的对立，辅音的发音方法有“清—浊”的对立。他们把这些最小对立体归结为十二对区别特征(distinctive features)，并且指出，世界上各种语言都可以用这十二对区别特征加以描述。这样，过去一直被认为是不可分的单元性的元音、辅音就变成由若干区别特征组合而成的、非单元性的结构体了。这种区别特征理论已成为现代语音学进行音位分析的基础。任何一个音位都可以用区别特征的集合来加以描述。如某一个音位具有二项对立中的前项特征，记以正号“+”，具有二项对立中的后项特征，记以负号“−”，就可以作成一个矩阵表，作为对每一个音位的区别特征集合的描述。这种音位理论，已经在语音自动识别和合成的研究中得到应用，证明是行之有效的。这是语言符号非单元性的有力证明。

计算语言学的理论和实践，加深了我们对于语言符号的非单元性的认识。为了改进乔姆斯基的短语结构语法，在计算语言学的许多理论中，都自觉地采用的“复杂特征”的概念，使用“特征/值”系统来描述句子的结构。

计算语言学还提出了非单元性的这种“复杂特征”进行运算的数学方法——“合一”运算，从而使我们对于语言符号非单元性的认识可以在计算机上进行实际的操作和演算。这种合一运算，并不完全服从于传统的集合论的运算。集合运算一般并不考虑运算对象的相容性，而合一运算则必须考虑运算对象的相容性。合一运算具有两种作用：

①合并原有的特征信息，构造新的特征结构，这与集合论中的“求并”运算类似。

②检查特征的相容性和规则执行的前提条件，如果参与合一的特征相冲突，就立即宣布合一失败。

可见，合一运算提供了一种在合并各方面来的特征信息的同时，检验限制条件的机制。这正是非单元性的语言符号在计算机运算时所需要的。所以，计算语言学不仅在理论上证明了语言确实具有非单元性，而且还在实践上使这种非单元性获得了计算机上进行运算的可能性。

第三，语言符号的离散性

我们平时说话时的语流似乎是连续不断的，但在实际上，这些连续不断的语流却是由许多离散的单元所组成的(当然，这些单元本身又是一个复杂项，可以由若干个复杂特征组成，具有非单元性，但就每个单元对其它单元的关系来说，它们又是彼此独立的，具有离散性)。在水平方向上，语流可以被分解为若干段落，一个段落又可以被分解为若干句子，一个句子又可以被分解为若干短语，一个短语又可被分解为若干单词，一个单词又可被分解为若干语素，一个语

素又可被分解为若干音节，一个音节又是由若干个元音和辅音音位组合而成的。在竖直方向上，语流中的各个成分又引起联想，引出与之属于同一类聚的若干个离散单元来。所以，在连续语流的水平方向和竖直方向上，实际上都是与若干个不同的离散单元联系着的。

语言符号的这种离散性，在语流的停延时表现得特别明显，人们往往可以利用语流停延的这种离散性质，来区别语流的不同含义。

汉语的书面语在词与词之间是连写的，不象印欧语的书面语那样留有空白，因此，在汉语书面语中，词与词之间的离散特点体现不出来。这种情况，给汉语的自动句法语义分析造成了极大的困难。在中文信息处理中，汉语自动句法语义分析的第一步便是自动切词，根据词与词之间的离散特征，把相互连在一起的词切开。可以说，语言符号的离散性，是汉语自动切词在语言学上的理论根据。

美国语言学家朱斯(M. Joos)早就指出了语言符号的这种离散性。他说："数学研究工具一般具有两种类型：连续分析(例如，无限小量的计算)或离散分析(例如，有限群理论)，而可以称为语言学的那个部门则属于后者，这时，它不容许与连续性有半点儿妥协，因此，凡是与连续性有关的一切，都得排除于语言学之外。"朱斯十分明确地把语言看成是"不可分解的语言学原子或范畴"离散地结合起来的，据此，他提出用离散数学来研究语言。他说："物理学家利用连续数学来解释言语，如傅利叶分解、自相关函数等，而语言学家则与此相反，他们利用离散数学来研究语言。"〔2〕

朱斯关于语言符号离散性的论述似乎有点儿矫枉过正。语言符号当然具有离散性的一面，但是，语言符号也有连续性的一面，特别是在语言的使用中。在语言的交际过程中，我们也可以利用一些连续数学的方法来研究它，而且实际上在这方面我们已经取得了不小的成绩。朱斯要把"凡是与连续性有关的一切"，"都得排除于语言学之外"，确实是太过分了。事实上，"离散性"和"连续性"都是语言符号本身所具有的性质，不过，在语言的使用的交际过程中，我们强调语言符号的连续性，用连续数学的方法来研究它，在语言结构的分析中，我们强调语言符号的离散性，用离散数学的方法来研究它，而语言本身则是离散性和连续性的统一体。

根据语言符号的离散性，计算语言学中采用集合论的方法建立了自然语言的集合论模型，并把这样的模型应用于机器翻译中，获得了很好的效果。这意味着，语言符号的离散性这一特性，在自然语言计算机处理的实践中已经得到了证实。

第四，语言符号的递归性

语言的句子是无穷无尽的，而语法规则却是有限的，人们之所以能够借助于有限的语法规则，造出无穷无尽的句子来，其原因就在于语言符号具有递归性。

语言符号的这种递归性，在不同的语言里表现不尽相同。汉语的句法构造的递归性突出地表现为句法成分所特有的套叠现象。在汉语里，由实词和实词性词语组合而成的任何一种类型的句法结构，其组成成分本身，又可以由该类型的句法成分充任，而无须任何的形态标志。这种套叠现象在主谓结构、偏正结构、述宾结构、述补结构、联合结构、复谓结构中都是存在的。这是由语言符号的递归性导致的汉语语法的一个重要特点。

在计算语言学的研究中，语言符号的递归性起着很大的作用。机器翻译的实质，就是把源语言中无限数目的句子，通过有限的规则，自动地转换为目标语言中无限数目的句子。如果机器翻译的规则系统不充分利用语言符号的递归性，要实现这样的转换是非常困难的，甚至是不可能的。

乔姆斯基指出，早在19世纪初，德国杰出的语言学家和人文学者洪堡德(W. V. Humboldt)就观察到“语言是有限手段的无限运用”。但是，由于当时尚未找到能揭示这种理解所含的本质内容的技术工具和方法，洪堡德的论断还是不成熟的。那么，究竟应该如何来理解“语言是有限手段的无限运用”呢？乔姆斯基指出：“一个人的语言知识是以某种方式体现在人脑这个有限的机体之中的，因此语言知识就是一个由某种规则和原则构成的有限系统。但是一个会说话的人却能讲出理解他从来未听到过的句子及和我们所听到的不十分相似的句子。而且，这种能力是无限的。如果不受时间和注意力的限制，那么由一个人所获得的知识系统规定了特定形式、结构和意义的句子的数目也将会是无限的。不难看到这种能力在正常的人类生活中得到自由的运用。我们在日常生活中所使用和理解的句子范围是极大的，无论就其实际情况而言还是为了理论上描写的需要，我们有理由认为人们使用和理解的句子的范围都是无限的。”〔3〕

那么，怎样来刻画语言这个无限集的成分组成情况呢？

我们可以把语言中所有的元素列成一个表，进行简单枚举。例如，

$L=\{\Phi,a,b,aa,ab,\cdots\}$

这样的刻画办法，把后面一大部分东西省略掉了，后面未列出的部分，只好由我们根据给出的少量的元素去想象，这样的刻画办法显然是不好的。它不能体现“有限手段的无限运用”这一原则。

我们应该采用递归的方法来刻画语言，为此提出如下的公理系统的定义。

一个公理系统是一个有序三元组(A,S,P)，其中，A是符号的有限集，叫做字母表；S是A上的符号串的集合，叫做公理；P是在由A中的符号组成的符号串上的n位关系的集合，$n\geqslant 2$(即P中n元组至少必须是有序对)，P的元叫做生成式或推理规则。根据这样的公理系统，我们便可以从公理S出发，多次使用推理规则P，在符号集A上递归地生成各式各样的、无限的符号串，实现“有限手段的无限运用”。因而这个关于公理系统的定义是体现了递归的原则的。

如果我们把公理系统中的A想象成前面所述的短语结构语法中的非终极符号VN和终极符号VT的集合，把S想象成短语结构语法中的初始符号S，把P想象成短语结构语法中的重写规则P，那么，我们马上就可以发现，短语结语构法与公理系统是十分相似的。所以我们可以说，短语结构语法是采用体现了递归原理的公理化方法来描述自然语言的语法。

现在，计算语言学业已严格证明，乔姆斯基的形式语法实际上等价于数学上的一种公理系统——半图厄系统(semi－Thue system)，这种形式语法不过是数学中的公理系统理论在自然语言分析中的应用而已，语言的生成过程完全可以通过公理系统这一形式化的手段得到严格的描述。正因为如此，乔姆斯基的形式语言理论，才会既在自然语言的信息处理中，又在计算机程序语言的设计中，得到如此广泛的应用。

所以，我们认为，语言符号的递归性，是反映了语言符号本质的又一个特点的。计算语言学深化了我们对语言符号的递归性的认识，普通语言学的理论对此应该给以足够的重视。

第五，语言符号的随机性

索绪尔在《普通语言学教程》中，把语言现象分为言语活动(langage)、言语(parole)和语言(langue)三样东西，它们之间是彼此联系而又相互区别的。

他指出，“言语活动是多方面的、性质复杂的，同时跨着物理、生理和心理几个领域，它还属于个人的领域和社会的领域。我们没法把它归入任何一个人文事实的范畴，因为不知道怎样去

理出它的统一体。”因此,“言语活动的研究就包含两部分:一部分是主要的,它以实质上是社会的、不依赖于个人的语言为研究对象,这种研究纯粹是心理的;另一部分是次要的,它以言语活动的个人部分,即言语,其中包括发音,为研究对象,它是心理·物理的。”“把语言和言语分开,我们一下子就把(1)什么是社会的,什么是个人的;(2)什么是主要的,什么是从属的和多少是偶然的分开来了。”

在索绪尔关于语言和言语区分的理论的影响下,乔姆斯基提出,必须把说具体语言的人对这种语言的内在知识和他具体使用语言的行为区别开来,并把前者叫做语言能力(competence),后者叫做语言运用(performance)。我们认为,乔姆斯基的语言能力,大体上相当于索绪尔的语言,乔姆斯基的语言运用,大体上相当于索绪尔的言语。

在言语(或语言运用)中,当我们用语言来进行交际活动的时候,有的语言成分使用得多一些,有的语言成分使用得少一些,各个语言成分的使用并不是完全确定的,这种不确定性,就是语言符号的随机性。我们在学习语言时经常感到语言规则中总是有许多的例外,这些例外,就是由于语言符号的随机性造成的。所以,语言符号的随机性,也应该是语言的本质属性之一。

正因为语言符号具有随机性,所以我们很难用确定性的规则来描述它。语言使用中大量的例外现象使研究语法的语法学家们伤透脑筋,有的语法学家甚至因此而误入迷津,以偏概全,得出了错误的结论。为了避免以偏概全的错误,我国前辈语言学者曾提出“例不过十不立,反例不过十不破”的原则来制定语法规则,这个原则常常作为判断语言学家治学态度是否严谨的准绳。其实,对于言语活动这样的随机现象来说,找出十个例子来立某条语法规则并不难,而找出十个反例来破某条语法规则也很容易,以十个例子或十个反例来作为某条语法规则破或立的标准,看来未必恰当。最好的办法还是采用统计数学的方法来对交际活动中所出现的各种语言现象进行描述。如果我们从语言学理论的高度,把随机性看成是语言符号本身的一种自然特性,并采用恰当的数学工具来描述这种随机性,使用计算机来进行一般手工操作所难于胜任的大量的统计计算和分析,那么,我们对于语法规则中的各种各样的例外情况,也就不会再感到迷惑和束手无策了,因为这些例外情况正是由于语言符号本身的随机性这一特点而形成的。

从计算语言学的角度看来,在语言成分的出现这一个随机事件中,随机事件 A 与条件组 S 之间虽然没有完全确定的联系,但是,它们之间却有着统计上的联系。尽管当条件组 S 实现一次时,事件 A 可能发生,也可能不发生。但是,如果条件组 S 实现多次,事件 A 的发生就有着某种规律性,这种规律性就是统计规律性。计算语言学认为,那些无一例外的必然的规律性,只不过是这种统计规律性的补充和表现形式罢了。

近年来,不少的语言学家开始认识到语言符号的这种随机性,自觉地使用统计方法来描述自然语言现象,这是令人可喜的。在计算语言学中,根据语言符号的随机性,已经在计算机上作了很多统计工作,成果累累。我国学者进行的汉字字频统计、汉字部件统计、汉字笔画统计、书面语词频统计、汉字熵值计算、汉字冗余度计算、汉语语音统计、汉语方言亲疏关系的分析和统计,为汉语的计算语言学研究提供了可靠的统计结果,推进了我国计算语言学的发展。这些事实说明,一旦我们在理论上自觉地认识到语言符号的随机性,就会产生出巨大的物质力量。语言学的理论对于语言研究的实践确实有着重要的指导意义。语料库语言学的研究,可以帮助我们从大量的经过标注的语言素材中,发现语言的统计规律,并把这些规律提炼为计算语言学的规则。这种研究生动地体现了索绪尔所指出语言和言语的相互关系。大量的语言素材相当于索绪尔定义的言语,语言学规则相当索绪尔定义的语言,通过对言语的统计研究,就可以发现

语言的规律。这是语言符号随机性的又一佐证。

第六，语言符号的冗余性

语言成分在交际活动中的出现是一个随机事件，语言成分之间彼此有着相互的影响和制约，也就是说，前后的语言符号具有相关性，我们根据前面出现的符号，常常可以预测后面的符号出现的可能性。当说话不清楚或文字有错落时，我们往往可以根据前后文来理解话语或文章的含义。就是当某个汉字或拉丁字母不清楚时，我们根据它们的残存部分常常就可以推断文字的全形。在有噪声或干扰时，我们仍然有能力根据已经听清楚的部分来识别那些不清晰的语音。

这些事实说明，并不是语言中的一切成分对于传达语言符号整体所包含的信息都是绝对不可缺少的，就是缺少了某些部分，语言本身有能力把这些缺少的部分补充和恢复出来。这意味着，语言符号具有冗余性。这种冗余性是必要的和有益的，它保证了不理想的环境下（如书面文章中有遗漏，谈话时有嘈杂声，书写的字母不清楚，发音不清晰），仍能发挥其交际功能。因此，我们不能认为冗余度就真的是语言中"冗余"的或不必要的东西，恰恰相反，这种冗余度是语言传递信息时必不可少的。没有冗余度的语言在实际上是无法理解的，因为日常语言总有很大的灵活性，要想理解句子的意思就必须考虑到字母在单词中的位置和单词在句子中的上下文关系。我国著名语言学家李荣教授建议把"冗余度"改为"羡余度"，这是很有道理的。事实上，只要语言有结构性就会有冗余度，语言符号的冗余度就是语言的结构性在语言使用过程中的体现。这样看来，语言符号的冗余性也应该是语言符号的一个重要特性，它与语言符号的随机性一样，无时无刻不在语言的使用中表现出来。

第七，语言符号的模糊性

索绪尔完全没有认识到语言符号具有模糊性。索绪尔认为，正是由于语言的作用，才使模糊的思想和声音的各个单位之间清晰起来。在索绪尔看来，语言本身是谈不上模糊性的。

关于语言的模糊性问题，在计算语言学出现之前，就有不少学者进行过探索和研究。著名哲学家罗素(B. Russell)于 1923 年写过一篇《论模糊性》的论文。1933 年，美国语言学家布龙菲尔德(Bloomfield)在《语言论》一书中，也指出了自然语言中存在着模糊现象。

由此可见，层次性、非单元性、离散性、递归性、随机性、冗余性、模糊性等 7 个特性也是语言符号十分重要的特性。索绪尔提出的语言符号的线条性可以用更为深刻的层次性代替，而他提出的语言符号的任意性确实是"头等重要的"、"支配着整个语言学"的原则。因此，我们认为，语言符号的特性除了上述的 7 个特性之外，还应该加上任意性，这样，语言符号就具有任意性、层次性、非单元性、离散性、递归性、随机性、冗余性、模糊性等共 8 个特性。计算语言学的发展，使我们对于语言符号的这些特性的认识和理解更为丰富、更为深刻了。在这种情况下，我们不得不修正索绪尔的旧理论，而代之以反映当前人类对自然语言符号认识水平的新理论。这是计算语言学在普通语言学的基本理论方面对理论语言学提出的挑战。

参考文献

〔1〕索绪尔：《普通语言学教程》，中译本，商务印书馆。

〔2〕F. Harady, H. Paper, Towards a general calculus of phonemic distribution, <Language>, Vol. 33, No. 2.

〔3〕乔姆斯基：《乔姆斯基理论介绍》一书序言，中文本，黑龙江大学出版社，1982 年。