·学术讲座专辑·　　编号 0003

计算语言学研究现状的综述

（邀请报告）

美籍冀中田教授

1980.6.

计算语言学研究现状的综述*

（邀请报告）

美籍冀中田教授

提　　要

计算语言学和语言的机器翻译简史。适合计算机分析的自然语言的特点。自然语言与形式语言。文法的主要理论：直接成分分析，串分析，变换分析。计算机分析文法的困难：选择性、二义性、隐含性。子语言。关于计算语言学的成就和展望。

计算语言学是计算机科学和信息处理的一个领域，它试图分析和利用以自然语言形式所表达的数据。它不同于计算机科学的另一些领域，那些领域是利用高度结构化的，编码的和格式化的数据库中的数据。

由于使用算法的方法来分析自然语言材料的困难（下面将要论及），至今计算语言学的研究还不是很成功的。然而，在此领域中的最新的进展对于将来更大的成功呈现了乐观的前景。语言的机器翻译和情报检索将是从计算语言学的成就中得益最大的领域。

由于自然语言的高度复杂性，计算语言学的主要问题就是将自然语言的论述（书写的或口语的）简化为比较简单的规则形式（句法分析），由此就可能推导其含义或信息的内容（语义分析）。也就是说，句法分析处理的是确定句子的文法的或结构的成分，并且将句子转换成为在信息内容上等价的比较简单的句子，它们具有简单的标准格式。语义分析处理的是抽取和使用这些规则格式句子中的信息内容。

自然语言和形式语言

能够适合于计算机处理的自然语言的特点：

1936 年由 $A.M.$ 图林提出了关于描述计算和可计算性的理论方法。事实上，每一个实际的计算机都是“图林机”的实现。图林机包括来自字母表 A 的顺序排列的符号 a_i 的一条带。该机器可以处在一系列状态 s_j 中的任一状态。它具有读取带上一个符号的能力。并能以改变机器的状态和该符号作为响应。

状态 s_j 表示该机器的当前内部的图象和它的历史。图林机的三种可能动作可编码如下：

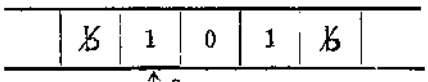
$a_i s_j a_i' s_j'$　　在 s_j 状态，如果读入符号是 a_i，则改变为 a_i'，并进入状态 s_j'。

* 美籍冀中田教授 1978 年 12 月在清华大学作此报告

s_i　　在 s_j 状态，如果读入符号是 a_i，则向右移动，并进入状态 s_j'。

$\overleftarrow{s_i}$　　在 s_j 状态，如果读入符号是 a_i，则向左移动，并进入状态 s_j'。

下述的图林机将对 0 和 1 构成的字符串求反码。



字母表　$A=\{0, 1, \not\Delta\}$

状　态　s_1：取反码

s_2：向右移动

s_3：停止

指　令　$0\ s_1\ 1\ s_2$

$1\ s_1\ 0\ s_2$

$\not\Delta s_1 \not\Delta s_3$

$a\ s_2 R s_1$（a 是 A 中字符）

$\not\Delta s_3$ 停止

因此，计算是由作用在离散对象上的一些相继的步骤所构成，而对象是以线性顺序，相邻接地排列着的。

自然语言具有计算所必须的三个基本性质，因此它是适合于计算的。

离散性　口语是连续的，但是通过倾听许多口语的讲述可以提取出许多音素，它是语言学上最小的区分单位，和词素，它是语义的最小单位。音素和词素是语言的离散的对象，无论是口语语言或是书写语言都是如此。

线性序列　口语语言是由词素（或音素）的线性序列所构成。这并不总是正确的，因为某些口语语言具有并列的音素，例如，英文句子“He went to the U.S to study”。按其语调，这可以是一个陈述句也可以是一个疑问句。这就是说，语法结构依赖于语调，而不依赖于词素的次序，这里对任何一种意思其词素的次序都是相同的。大多数的语言没有这个问题，同时在任何情况下，书写语言都绝不会出现这种问题，因为我们总是要插入“？”在词素的线性序列中以表示疑问句。然而，当输入是语言时，这个问题是自然语言处理极为困难的一个表征。这一方面的进展是缓慢的。现在，至多也就是大约几百个单词或简单的短语口语通过声音（声波纹或声音的图象）来识别，已经可以由计算机来完成。

在美国有几个中心在进行语言的识别和综合的研究工作，在综合方面比起识别要有更多的成绩。贝尔电话实验室在这一方面的工作是很活跃的。

在处理书写语言方面已经有了更多的成果（虽然也是有限的）。所以本文的其余部分将集中讨论书写语言的处理问题。

邻接性　“依次邻接”这是两个字之间仅有的基本关系：对语言不存在长度的度量。

（如象在音乐中那样）。邻接性这一事实可以用来从音素中构成词素，也可用于从词素构造串。下面要详细讨论的串就是有意义的最小邻接语法单位。

由于词素（或单词）的某些组合而形成串和句子，虽然有语言的正则顺序的观念，然而大多数都不是由语法规则（即结构或文法）所决定。因此，文法可能用来分析句子并产生它的句法或结构的描述（即语法分析）

在这方面，自然语言和形式语言——逻辑的、数学的、或各种计算机的语言——很相似。然而这两种语言之间有着重要的差别，这就使得自然语言的研究十分不同于形式语言的理论和研究。这些差别是：

1. 规模： 自然语言的文法和词汇比形式语言要庞大得多。这就引起在存贮容量、文件结构的设计和处理速度等方面的困难。

2. 文法的不确定性和不完全性： 只有通过研究大量的自然语言的实例，才能看出自然语言的语法规则。自然语言是“活的语言”——它的规则在不断地变化，因此这一次所能接受的，另一次就可能不被接受。即便有一个庞大的文法系统，但往往总可以找到一些没有被概括进去的构造。1961 年，由 Mr. Jespersen 所描述的英语文法竟庞大到有七卷之多。

3. 不同的接受能力： 在数学语言和形式语言中，语句或是取真，或是为假，例如：

$$SIN(X+Y)=SINXCOSY+COSXSINY$$

它对所有的 X，Y 都是真的。

然而，自然语言文法规则就没有这种确切的意义。例如，形式

N　　TV　　N

〈名词〉〈时态动词〉〈名词〉

它可能是一个句子，也可能不是。这完全取决于具体的替换词。试比较：

“Men eat rice”（人吃饭）是一个句子，

“Men sleep rice”（人睡饭）就不是句子。

4. 二义性： 形式语言绝对不许可二义性，即一个语句不能有一个以上的结构分析。然而，自然语言却富于二义性，要消除它是非常困难的。例如，

“I saw the man in the park with a telescope”

这个句子有多种解释，本文后面将要详尽地加以分析。

文法的主要理论

直接成分分析（ICA—Immediate Constituent Analysis）是将句子的结构描述成为成分的序列，它们中的每一个，顺次地，又是低一级成分的序列，…如此等等，一直到最低一级（语言的词素是其成分）为止。ICA 分析法，最初是由 Leonard Bloomfield

在 1933 年研究提出的。每一成份是句子中的相邻的部分，在印欧语系中，初始成分是：$S = N_p + V_p$，即句子是由名词短语加上动词短语所构成。

串分析（String Analysis）是由 Zellizg Harris 在 1959 年专门为计算语言的应用而研究的，但是从这里已经开始了对于语言的理论研究。在串分析中，我们隔离出称为中心串的一个基本句子，再连接上其它的字串或序列就构成了句子。这些附加词可以放在中心串的右边或左边，或是放在中心串的某特定元素的右边或左边，或是放在另一个附加词或另一个附加词的特定的元素的右边或左边，如此等等，递归地定义。每一个词都可以属于一个或多个词类（名词、时态动词、…等等），而且如果用词的词类去替换句子中相应的词，这就得到了该句子的串公式。例如：

"The hungry cats always eat quickly"

其中，

"cats eat" 是中心串

"hungry cats" 是对 cats 加上左附加词 "hungry" 所得(即由 cats 扩展而来)。

"The hungry cats" 同样是对 "hungry cats" 用附加词扩展而得。

"eat quickly" 是对 eat 加上右附加词 'quickly' 而得。

"always eat" 是以 always 为左附加词，从 eat 扩展而成。

应用直接成分分析，可以得到：

名词短语 NP：The hungry cats

动词短语 VP：always eat quickly

以此作为最高一级的成分。

而其中，NP 的成分是 "The"，"hungry"，"cats"，VP 的成分是 "always"，"eat"，"quickly"。

由此可见，在一个句子的串分析中的基本串的后继词，就是该句子的直接成分分析中主成分的关键词。

变换分析（Transformational Analysis）是由 Zellig Harris 于 1959 年提出的，他发现除去某些不变的结构词或词缀外，文法形式不同的句子会具有同样的单词和相同的内容。这就彼此互称为变换。例如，被动语态的变换，将

'Men eat rice.'

变换为

'Rice is eaten by men'

这两个句子的信息内容是相同的，虽然着重点和风格不同。

变换分析把一个句子分解成许多基本句子，它是由特定的变换进行运[illegible]的。例如：

"The hungry cats eat quickly"。它的基本句子是：

"Cats eat"

"Cats are hungry"

"Eating is quick"

因此，变换分析消除了文法的意译，并且产生了对于句中信息的一致性的表示，即，如果两个句子包含有同样的信息，而形式不同，那么，在变换分析之下，他们的基本句子将是相同的。对于计算语言学，这是很有价值的。因为在计算语言学中，为了推断自然语言的信息内容，我们要试图将自然语言简化为简单的规则形式。

事实上，用一组简单的对象作出复杂对象的标准表示，这是所有数学系统的基本特点。例如，每个数都有分解成质数的唯一的因式分解。

ICA，串分析和变换分析都是文法的分析系统，它们通过把句子分解成为其组成分量的办法来研究句子。

变换生成文法 (Transformational Generatine Grammar) 是由 Noam Chomsky 在 1957 年首先提出的，是一个由组成分量来生成句子的系统。它使用“短语结构规则”（类似于生成句子的 ICA 规则）称之为句子的“深度结构” (deep structure)。接着，它使用变换分析对“深度结构”进行加工来产生“表面结构”。

计算语言学和机器翻译简史

这方面最初的工作开始于 50 年代初期，首先是在机器翻译方面 - 从俄语翻译为英语。早期的努力曾简单地假定翻译是一个字典查找问题，即句法问题被忽略了。这就导致错误的翻译，最熟知的例子是：

“The spirit is willing but the flesh is weak.” （力不从心）

当将该句子机械地翻译成俄语，然后再译回英语时，它却成为：

“The wine is good but the meat isn't well cooked”. （酒是好的但肉未煮好）

第一个英语语法分析程序是 1958 年～1959 年期间在宾夕法尼亚大学首先研制出来的。在该系统中，把 ICA 和串分析结合起来了。该系统取得成功的原因是不企图包罗整个英语语言，它限定所包括的有关内容，因此，所必须处理的仅仅是英语的子语言，即指处理棒球游戏的语言。这个方法是今天所有成功的系统中最典型的例子。也就是说，我们离开能够分析来自任何领域的文本的语言分析系统仍然是很远的。

60 年代早期，IBM 公司研制了另一个设计方案，它是以 Chomsky 变换生成文法为基础的。

在这个时期以后，计算机语言学名声扫地，并且为研究提供资金也变得非常困难。

在 60 年代后期和 70 年代，语言分析设计不那样野心勃勃了，因而有了一些较大的成就。许多设计都是有关很窄范围内的问题回答系统。在此情况下，分析问题的范围，在语法上仅仅限于问题的形式，而在语义上仅限于所给定的数据库的确切词类 --它不是以自然语言形式提供的，而是结构化的，格式化的数据库，即仅对问题必须应用自然语言分析，以便能从不是以自然语言给出的数据库中检索所要求的信息。

两个特别有成效的系统是 Bott，Beranck 和 Newman 的系统，用于月球岩石的数据检索和 IBM 的商业统计检索请求系统，(IBM'S Request System for business statistic retrieval) 。

纽约时报的数据检索系统是一个问题询问系统，它允许检索由当前报纸和杂志所包括的任何方面的信息。但在此系统中，也仅仅是提问用自然语言形式，而其数据库是人工编码的，它每天都要耗费大量的人工劳动。

在纽约大学，由 Naomi Sager 领导进行的语言学的串设计，一直从事于串分析文法的研究，这个串分析文法相当庞大，只要对所给定的课题附加上专门的规则，它就可用来分析英语中任何简单的领域。该语言串分析程序大约包含 1200 条文法规则和限制，并且有 100 个以上的不同词类（名词，时态动词，等等）。对此，我们还要增加另外的成份，它构成了专门的子语言的特殊的句法和词类，例如，药物学和医院病历记录等。

计算机语言学文法的困难

应用于语言学的计算方法的基本问题，在于自然语言是一个意义保持系统，而数学语言和形式语言却是真值保持系统，在数学中是无法表达意义和因果关系的。

正如前面所讨论的，大多数成功的计算机文法都是用于英语的子语言 — 是特定范围内的语言。比如，研究洋地黄的作用（一种草药）扩展这些子语言文法以包括整个语言，即或是一大部分，实际上都是不可能的。因为一个课题领域的处理会影响到其他领域如何处理，而且各种子语言是相互交叠的。因而我们必须从一个一般性的大的文法开始，而从这一设计的规模和范围来看又是很困难的。

正如前面已指出的，另一严重的困难是不同的可接受性。“Men sleep rice.”在语法上是不合适的，因为它不符合文法规则：“sleep 是不及物动词，它不能有宾语。在计算机的文法中，这个规则是很容易实现的　对及物和不及物动词加以区分。（当然，有些是可以重叠的。例如，“Men eat.”和“Men eat rice.”都是可接受的。）

然而，有着不同的可接受性的另外一些问题，这种不同的可接受性不是文法的而是语言的。字 W 的选择性是指在给定的语法关系中通常和 W 同时出现的那些字的集会。例如，如果 W = “red”，则它的选择性是描述那些可能具有颜色属性的所有名词。因此，

“red book”是可接受的。然而

“red idea”是模棱两可的，而

“red sleep 是不可接受的。

这就说明，对象的选择性不是由名词的全体所构成 -- 仅仅是它的一个子集。因此，虽然句子 ‘I had a red sleep’. 在语法上是合适的，但在语言上它是不可接受的。然而，“I had a good sleep.”是合适的。这一问题大大增加了计算文法的规模和复杂性。事实上，当两个字是确切的同义词时，它们才具有同样的选择性。因此，一个字的选择性与其含义密切相关　如果两个字的选择性有着很小的重合，那它们的含义是十分不同的。试比较：‘red’，‘blue’和‘necessary’。因此，包含整个语言的文法就必须具有区分全部词汇（除去确切的同义词）的正确的，足够的规则或限制。

对整个语言中的字而言，选择性是十分难以决定的。然而在特定的子语言中，特别是在科学领域内，它的使用是很有规则的。所以选择性的类别是较易确定的，而选择性的限制也是易于实现的。例如，考虑在计算机领域内：

"Computers have memories"

"Memories have computers"

这是关于子语言的计算机语言学为什么能比整个语言的计算机语言学有更多成就的重要原因。

二义性反映了计算机语言学的另一个严重问题。英语和大多数的其他语言都具有名副其实的语法二义性——这种二义性来自对句子有超过一种的正确的语法分析。

试看："The steel industry requires increases"。

存在着两种同样可接受的分析方法和含义：

1. 中心串是 industry requires increases，而 steel 是 industry 的左附加词，

主语是 steel industry

动词是 requires

动词的宾语是 increases

2. 中心串是 steel increases，industry requires 是 steel 的右附加词，

主语是 steel

动词是 increases，它是不及物动词。

这两种分析在语义上和语言学上都是可接受的。

另一类型的语法二义性可以通过例子："I saw the man in the park with a telescope" 来说明。对这一句子存在着三种正确的分析，主要取决于介词短语 with a telescope 是用来修饰：I，the man 或 the park.

要解决这种二义性是很困难的，通常仅能用将文法限于子语言和书写选择性限制以消除不需要的意义来实现。

计算机语言学的另一问题是虚拟性（或隐含性 Zeroing），即那些字在句中虽未出现，但却含蓄地反映出来，而且可以从上下文重新构造。在前面谈到的例句中，'which' 是以虚拟形式呈现的，如果将它插入到句中，则将消除其二义性：

The steel which industry requires increases.

虚拟性的另一个例子是：

Mao Tse Dung and Chou En Lai Fought in the chinese revolution

观察此句，你看不出什么是被隐含了。再看：

Mao Tse Dung And Chiang Kai Shek fought in the chinese revolution.

上述两个句子都是二义性的，因为在第一个句中隐含的成分是 together，而在第二个句中是 each other。

计算机语言学的成就和展望

…的经验，特别是纽约大学库伦特数学研究所（Courant Institute of …Science）由 Naomi Sager 所领导进行的语言学串的设计方面的经验，…个大规模的通用计算文法（在 CDC6600 上运行）能够分析串和决定句子…

…除语义上不正确的句子和未意识到的二义性，在语法规则上作许多限制是完…

…换分析应用于文法的分析输出，在将句子转换为标准形式方面已经取得成功。…得它们的信息内容或含义有一个一致的结构化的表示——当主要的问题是限于…域时，对该领域的信息设计一种格式是可能的。在此阶段，可通过对选择性的限…除二义性。

…因此，LSP 在二个领域内已经从自然语言的输入建立了结构化的数据库：医院病…录和洋地黄药物学。

现在，这些数据库能够用于回答问题的系统或统计分析。其主要问题是：建立该系…要花费大量人力和资金，特别是建立字典的工作（将单词按其语法用途和选择性分…）是手工完成的，而不是用计算机。为了完成一个单词的分类要花费一天以上的时…间，而当前在字典中我们要有 5000 个单词。而且，这些系统都是很狭窄的课题领域。

计算机语言学所存在的问题之一，常常是缺乏计算机可读形式的自然语言材料。要准备这样的材料成本很高。然而，今天在美国有一种趋势：这就是一开始就以计算机的可读形式来准备这些材料。例如，在一些现代化的医院中，对于病历的一切修改都由医生、护士或技术人员直接用终端打入到计算机中。这样做很大程度上方便了存贮，并易于输入和检索，并可以在任何设有终端的地方进行检索和修改病历，而无需将病历作任何实际的移动。因此，虽然并不是为了自然语言处理的目的，然而这些医院现在都有着为研究人员有效使用的大型自然语言数据库。

今天，在美国的许多机构中，政府、大学、图书馆、医院、商业部门、银行和工厂都设置了具有许多联机终端的大型计算机系统。他们热切地期待着自然语言处理系统的发展，这将使他们比现在更方便和更灵活地去访问和调整他们的数据库。我们希望在不久的将来，由于更快、更大和更便宜的计算设备的发展，将使其成为可能。然而，成功可能是逐渐来到而且是以小的步子前进。这是因为自然语言处理的复杂性和存在很多困难，问题已经了解到了，但不能期望有引人注意的突破。

杨德元　译

唐泽圣　校

A Summary of Computing Linguistics at Present Conditions

(Invitation Report)

By Professor Zi Zhong-tian

Abstract

Brief history of computing linguistics and machine translation of language. Characteristics of natural language which is appropriated for computer analysis. Natural language and formal language. Main theory of grammer: Direct component analysis, string analysis, transform analysis. Difficulties of computer analyzing grammer: selection, ambiguity, zeroing, sublanguage. Achievement and further prospect of computing linguistics.