

## 论语言学研究中的战略转移

教育部语言文字应用研究所/大连海事大学 冯志伟

**提要:** 本文作者根据对计算语言学中战略转移的分析后指出,传统语言学的方法正面临着经验主义的严峻挑战,一种新的“基于语料库”的研究方式或者“语料库驱动”的研究方式将逐渐地代替传统的依靠“内省”和“诱导”的研究方式。但是,这种基于语料库的研究方式或者语料库驱动的研究方式离不开语言学家对于语言现象的“洞察力”,我们决不能忽视理性思维的重要作用;把理性主义与经验主义结合起来,是当前语言学战略转移的正确方向。

**关键词:** 战略转移、计算语言学、语料库、理性主义、经验主义

[中图分类号] H030

[文献标识码] A

[文章编号] 1003-6105(2011)01-0001-10

### 1. 计算语言学中的战略转移

计算语言学兴起于二十世纪五十年代。二十世纪九十年代以前,从事计算语言学研究的绝大多数学者,都把自己的目标局限于某个十分狭窄的专业领域之中,他们采用的主流技术是基于规则的句法-语义分析。尽管这些应用系统在某些受限的“子语言”(sub-language)中也曾经获得一定程度的成功,但是,要想进一步扩大这些系统的覆盖面,用它们来处理大规模的真实文本,仍然有很大的困难(冯志伟 1983)。因为从自然语言处理系统所需要装备的语言知识来看,其数量之浩大、颗粒度之精细,都是以往的任何系统所远远不及的。而且,随着系统拥有的知识在数量上和程度上发生的巨大变化,系统在如何获取、表征和管理知识等基本问题上,不得不另辟蹊径。这样,就提出了大规模真实文本的自动处理问题。

1990年8月,在芬兰赫尔辛基举行的第13届国际计算语言学会议为会前讲座确定的主题是:“处理大规模真实文本的理论、方法和工具”。这说明,实现大规模真实文本的处理将是计算语言学在今后一个相当长的时期内的战略目标,计算语言学正面临“战略转移”(strategic transit)的关键时刻。为了实现这样的战略转移,需要在理论、方法和工具等方面实行重大的革新。1992年6月,在加拿大蒙特利尔举行的第四届机器翻译的理论与方法国际会议(TMI-92)将会议主题定为“机器翻译中的经验主义和理性主义的方法”。所谓“理性主义”(rationalism),就是指以生成语言学为基础的方法;所谓“经验主义”(empiricism),就是指以大规模语料库的分析为基础的方法。从中可以看出当前计算语言学关注的焦点(冯志伟 1996)。

当前语料库的建设和语料库语言学的崛起,正是计算语言学战略目标转移的一个重要标志。随着人们对大规模真实文本处理的日益关注,越来越多的学者认识到,基于语料库的分析方法(即经验主义的方法)至少是对基于规则的分析方法(即理性主义的方法)的一个重要补充。因为从“大规模”和“真实”这两个因素来考察,语料库才是最理想的语言知识资源。但是,要想使语料库名符其实地成为自然语言的知识库,就有必要首先对语料库中的语料进行自动标注,使之由“生语料”变成“熟语料”,以便于人们从中提取丰富的语言知识。可以看出,计算语言学当前面临的这一场战略转移的关键是知识的获取方式和方法:从依

靠“自省”方式转向依靠“语料”的方式，从“基于规则”（rule-based）的方法转向“基于语料库”（corpus-based）的方法，也就是“基于统计”（statistics-based）的方法（冯志伟 2010）。

随着战略转移的深入，统计方法已经逐渐成为计算语言学的主流方法。

2003年7月，在美国马里兰州巴尔的摩（Baltimore, Maryland）由美国商业部国家标准与技术研究所 NIST/TIDES（National Institute of Standards and Technology）主持的评比中，德国亚琛（Aachen）大学年轻的博士生奥赫（Franz Josef Och）获得了最好成绩。他使用统计方法，在很短的时间之内就构造了从阿拉伯语和汉语到英语的若干个机器翻译系统（冯志伟 2010）。

过去我们研制一个机器翻译系统往往需要几年的时间，而现在采用奥赫的方法构造一个机器翻译系统只要几个小时就可以了，研制机器翻译系统的速度已经大大地提高了。这是机器翻译技术的一个史无前例的重大进步。

早在1949年，美国科学家韦弗（W. Weaver）在他的以《翻译》为题的备忘录中，提出了使用解读密码的方法来进行机器翻译。他说，“我可以这样说，一本用中文写的书实际上是用英语写的，只不过它是用中文的代码符号编了码而已，这样的说法是很有诱惑力的。”（“It is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the Chinese code.”（Weaver 1949: 2）<sup>1</sup>。

在六十多年以前的这段话中，韦弗首先提出了用解读密码的方法进行机器翻译的想法，这种想法成为后来“噪声信道模型”（noisy channel model）的滥觞。

这种所谓“解读密码”的方法实质上就是一种统计的方法，韦弗是想用基于统计的方法来解决机器翻译的问题。但是，由于当时尚缺乏高性能的计算机和大规模的联机语料库，采用基于统计的机器翻译在技术上还不成熟。韦弗的这种方法难以付诸实现的。现在，这种局面已经大大改变了，计算机在速度和容量上都有了大幅度的提高，也有了大量的联机语料可供统计使用，我们有可能从大规模的真实语料库中获取机器翻译的知识。因此，在二十世纪九十年代，基于统计的机器翻译又兴盛起来。

在韦弗思想的基础上，IBM公司的布劳恩（Peter Brown）等人提出了基于统计的机器翻译的数学模型。基于统计的机器翻译把机器翻译问题看成是一个“噪声信道”问题，如下所示：

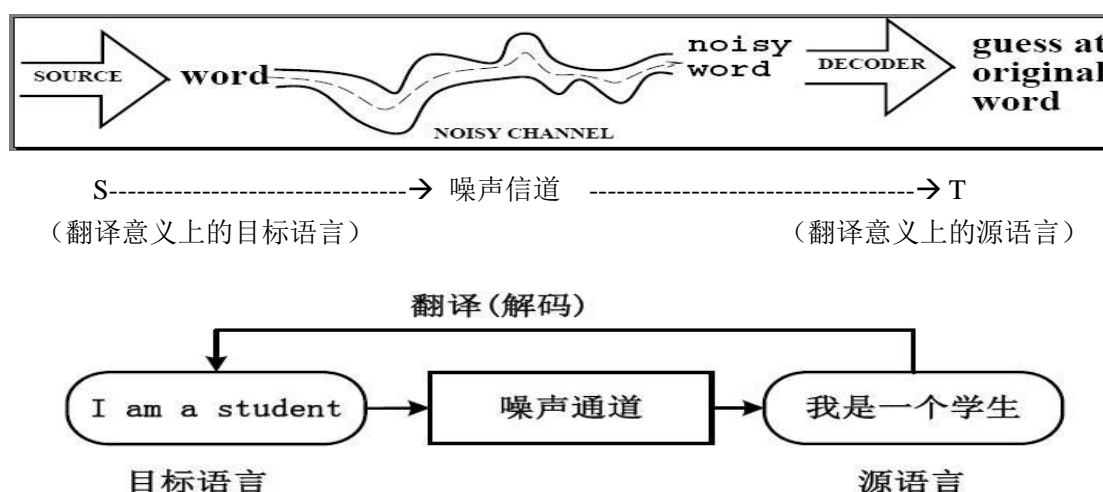


图1 噪声信道模型

<sup>1</sup> Weaver的这段名言在国内外的很多论著中都引证得不确切，一些学者把 Chinese 误引为 Russian，经核对，应为 Chinese，现把原文照抄出来，供读者参考。

可以这样看机器翻译：一种语言  $S$  由于经过了一个噪音信道而发生扭曲变形，在信道的另一端呈现为另一种语言  $T$ ，翻译问题实际上就是如何根据观察到的语言  $T$ ，恢复最为可能的语言  $S$ 。语言  $S$  是信道意义上的输入，在翻译意义上就是目标语言，语言  $T$  是信道意义上的输出，在翻译意义上就是源语言。从这种观点看，一种语言中的任何一个句子都有可能是另外一种语言中的某几个句子的译文，只是这些句子的可能性各不相同。机器翻译就是要找出其中可能性最大的句子，也就是对所有可能的目标语言  $S$  计算出概率最大的一个作为源语言  $T$  的译文。由于  $S$  的数量巨大，可以采用“栈式搜索”（stack search）的方法。栈式搜索的主要数据结构是表结构，表结构中存放着当前最有希望的对应于  $T$  的  $S$ ，算法不断循环，每次循环扩充一些最有希望的结果，直到表中包含一个得分明显高于其他结果的  $S$  时结束。但栈式搜索不能保证得到最优的结果，它会导致错误的翻译，因而只是一种次优化算法。

可见，统计机器翻译系统的任务就是在所有可能的目标语言（翻译意义上的目标语言，也就是噪声信道模型意义上的源语言）的句子中寻找概率最大的那个句子作为翻译结果。其概率值可以使用贝叶斯公式（Bayes formula）得到（注意：下面公式中  $T$  是在翻译意义上的目标语言， $S$  是在翻译意义上的源语言）：

$$P(T|S) = \frac{P(T)P(S|T)}{P(S)}$$

由于等式右边的分母  $P(S)$  与  $T$  无关，因此，求  $P(T|S)$  的最大值相当于寻找一个  $T$ ，使得等式右边分子的两项乘积  $P(T)P(S|T)$  为最大，也就是说：

$$T = \arg \max P(T)P(S|T)$$

其中， $P(T)$  是目标语言的语言模型， $P(S|T)$  是给定  $T$  的情况下  $S$  的翻译模型。根据语言模型和翻译模型，求解在给定源语言句子  $S$  的情况下最接近真实的目标语言句子  $T$  的过程就相当于噪音信道模型中解码的过程。

统计机器翻译系统要解决三个问题：（1）估计语言模型概率  $P(T)$  – 译文的流畅性；（2）估计翻译概率  $P(S|T)$  – 译文的忠实性；（3）设计有效快速的搜索算法来求解  $T$ ，使得  $P(T)P(T|S)$  最大 – 把  $P(T)$  和  $P(T|S)$  结合起来，综合考虑译文的流畅性和忠实性。

随着统计机器翻译技术的发展，目前开发出来的基于语料库的自动分析软件大多数都是独立于具体语言的。只要有对齐的双语并行语料库（parallel corpus），即使研究者不懂有关的语言，仍然可以得出不错的分析结果。现在Google推出了超过40种语言两两之间的免费互译服务，这就意味着计算机可以在1,600个以上的翻译方向上随意地进行机器翻译。这是机器翻译研究历史上的一个重大进步，也是计算语言学战略转移取得的初步成果。多么令人鼓舞！

## 2. 传统语言学的方法面临严峻挑战

面对计算语言学的战略转移，我觉得语言学在获取知识的方式方法也应当进行一场战略转移。

与战略转移以前的计算语言学相似，传统语言学家获取语言知识的方法基本上也是通过“内省”进行。由于自然语言现象充满了例外，治学严谨的学者们提出了“例不十，不立法”（黎锦熙，1924：1）和“例外不十，法不破”（王力 1988：27）<sup>2</sup>的原则。这样的原则貌似严

<sup>2</sup> 王力（1988：27页）中指出，“所谓区别一般与特殊，那是辩证法的原理之一。在这里我们指的是黎锦熙先生所谓‘例不十，不立法’。我们还要补充一句，就是‘例外不十，法不破’。”

格，实际上却是片面的。在成千上万的语言数据中，只是靠十个例子或十个例外就来决定规则的取舍，难道真的能够保证万无一失吗？显然是不能保证的。因此，“例不十，不立法；例外不十，法不破”的原则只是一个貌似严格的原则，实际上是一个很不严格的原则。

在语料库出现之后，传统语言学的这个原则受到了严峻的挑战！

语料库是客观的、可靠的语言资源，语言学研究应当依靠这样的宝贵资源。语料库中包含着极为宝贵的语言知识，我们应当使用新的方法和工具来获取这些知识。当然，前辈语言学家数千年积累的语言知识（包括词典中的语言知识和语法书中的语言知识）也是宝贵的，但由于这些知识是通过这些语言学家们的“内省”（*introspection*）或者“洞察力”（*insight*）发现的，难免带有其主观性和片面性，需要我们使用语料库来一一地加以审查。辛克莱（*John Sinclair*）一针见血地指出：“生造的例子看上去不管是多么地可行，都不能作为使用语言的实例”（*Sinclair 1991: 4*）。他大声疾呼：“我们总不能靠造几朵人造花来研究植物学吧！”（*Sinclair 1991: 6*）。记得几年前在匈牙利的巴拉顿湖畔的美丽城市蒂哈尼（*Tihany*）开会的时候，在一次闲谈中，我对辛克莱说：“我们也同样不能根据赛璐珞造的玩具狗 *Rex* 来研究动物学”。当时他对于我的意见表示赞同。辛克莱甚至认为，对于语料库的标注也要采取非常谨慎的态度，因为标注在本质上带有标注者的主观色彩，他主张“要把标注与语言材料严格地区分开来”（*Sinclair, 2007: 433-434*）。

如果搞语言研究不使用语料库或概率，很可能就只能使用自己根据“内省”得到的数据，这是“第一人称数据”；或者使用根据“问卷调查”之类的“诱导”（*elicitation*）得到的数据，这是“第二人称数据”。在使用第一人称数据时，语言研究者既是语言数据的分析者，又是语言数据的提供者。在使用第二人称数据时，语言研究者不充当数据的提供者，数据需要通过“作为第二人称的旁人”的诱导才能得到。如果使用语料库的数据作为语言研究的数据来源，那么，语言研究者就不再充当数据的提供者或诱导者，而是充当数据的分析者了。这种“观察”得到的数据是“第三人称数据”。

这是威多逊（*Widdowson 2000: 1*）提出的看法，我觉得很有见地，值得我们思考。

当然，如果使用第三人称的观察数据，语言研究者同时也可以充当数据的“内省者”或“诱导者”，所以，第一人称和第二人称与第三人称是难以分开的。这也就是我不反对“拍脑袋”这种第一人称方法的原因。第三人称方法显然是比较科学的获取数据的手段。

乔姆斯基的生成语法采用的是第一人称方法，由于他具有非凡的智慧，也可以取得卓越的成就；心理语言学、实验语音学采用的是第二人称方法，也取得了不少的成果，而我们现在则提倡第三人称方法。当然，与此同时，我们仍然要充分尊重第一人称研究者和第二人称研究者的智慧和洞察力，我们并不反对第一人称的内省法和第二人称的诱导法。第一人称的“拍脑袋”方法固然会产生主观性，但是，脑袋拍得好也并不容易，聪明的脑袋可以拍出出类拔萃的优秀成果，前辈语言学家的智慧和洞察力仍然是值得称道的。

不过，我们认为，语言学的一切知识，不论是过去通过“内省”或“诱导”得到的知识，最终都有必要放到语料库中来“观察”（*observation*）和“检验”（*verification*），决定其是正确的，还是片面的或者错误的，甚至是荒谬的，从而决定其存在的必要性，决定其是继续存在，还是放弃其存在。

我们可以预见，语言研究中战略转移的时代必将到来！一种新的“基于语料库”的研究方式（*corpus-based approach*）或者“语料库驱动”的研究方式（*corpus-driven approach*）将逐渐地代替传统的依靠“内省”和“诱导”的研究方式（*Abeillé 2003; Hinrichs & Kibler 2005*）。“内省”和“诱导”的研究方式今后很可能只是基于语料库研究方式或语料库驱动研究方式的补充，而不能是语言学研究的主流。当然，这种基于语料库的研究方式或者语料库驱动的

研究方式仍然离不开语言学家对于语言现象的“洞察力”(insight)。我们决不能忽视理性思维的重要作用(刘海涛、冯志伟 2007)。

我们认为,传统语言学正在面临战略转移的重要时刻,我们应当从高度的历史责任感出发,敏锐地认识到这个战略转移的重要时刻或迟或早总会来临,为此而调整我们的研究方法和研究计划,从而为世界的语言学宝库做出我们中国学者应有的贡献。千万不可贻误这个良机啊!

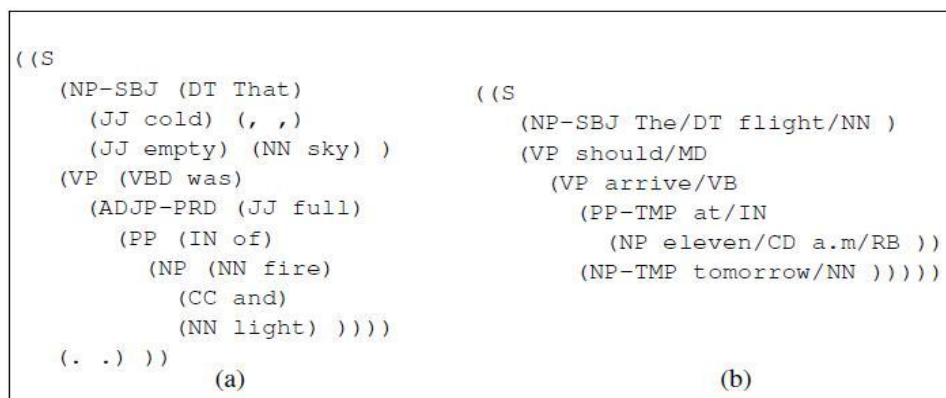
英国著名科学哲学家波普尔(Karl R. Popper)在《如何对待错误》(2008: 203)一文中说,“人们尽可以把科学的历史看作发现理论、摒弃错了的理论并以更好的理论取而代之的历史。……我不怀疑我们有许多科学理论是真实的;我所要说的是,我们无法肯定任何一个理论是不是真理,因而我们必须作好准备,有些最为我们偏爱的理论到头来却原来并不真实。既然我们需要真理,……我们除了对理论进行理性批判以外,别无其他选择。”<sup>3</sup>这段话非常具有启发性,最近被收入《宇宙简史——无限宇宙中的无穷智慧》一书中。正是本着这样一种对于传统的语言学研究结论进行理性批判的科学精神,我们需要大胆地对前辈语言学家的结论证实和证伪,在语料库中大量语言事实的基础上进行理性的审视。这样,我们就有可能提出不同的但更富于发展前景的学术意见。

在语言学研究中,我们尽最大的努力避免偏颇和错误。波普尔(2008: 204)在他的同一篇文章中还说:“科学是可以犯错误的,因为我们都是人,而人是会犯错误的。因而错误是可以得到原谅的。只是不去尽最大的努力避免错误,才是不可原谅的。但即使犯可以避免的错误,也是可以原谅的。”语料库给我们提供了极其丰富的客观语言事实,我们应当充分利用语料库给我们提供的客观语言事实,避免前辈语言学根据内省的研究方法做出的可能有片面性的结论或错误,从而推动语言学中的这个战略转移。

### 3. 经验主义方法的局限性

然而,在这个战略转移的重要时刻,我们还应当清醒地认识到,从语料库中挖掘和抽取各种各样的知识,并不是一件简单的事情,我们决不可掉以轻心。

我们来看一看从树库(Treebank)中挖掘语言学知识的经验。图2是宾州树库(Penn Treebank)中来自Brown语料库和ATIS语料库的句子的剖析树<sup>4</sup>。



注: (a) 中的句子 “That cold, empty sky was full of fire and light” 来自 Brown 语料库; (b) 中的句

<sup>3</sup> 哥白尼, 爱因斯坦, 霍金等. 《宇宙简史——无限宇宙中的无穷智慧》, 湖南科学技术出版社, 2008年, 第203-204页, 长沙。

<sup>4</sup> 宾州树库课题对于多种语言在不同的阶段发布各种树库; 例如, 发布的英语树库有树库 I (Marcus *et al.* 1993)、树库 II (Marcus *et al.* 1994) 和树库 III。我们的例子来自树库 III。

子 “The flight should arrive at eleven a.m. tomorrow” 来自 ATIS 语料库。

图 2 LDC 树库 III 版本中经过剖析的句子

在上面的 Brown 和 ATIS 两个语料库中，词类标记的格式有些差别，这样的一些小差别是常见的，有必要在树库加工的过程中进行处理。使用 LISP 风格的括号表示法来表示树的结构是很常见的。对于那些不熟悉括号表示法的读者，我们在图 3 中做出了一个用结点和线条来表示的标准的树形图，可作为对照。

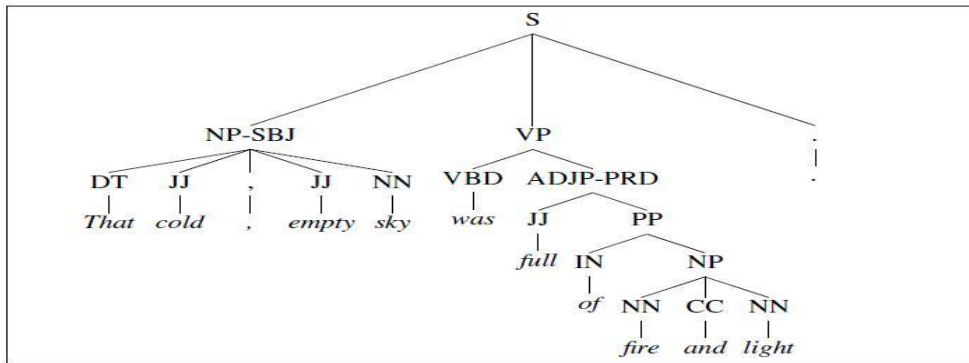


图 3 与上图中来自 Brown 语料库的句子 (a) 相对应的树形图

图 4 中的树形图来自 Wall Street Journal 语料库。这个树说明了宾州树库的另外一个特征：使用踪迹 (trace, 用结点-NONE-表示) 来标记长距离依存关系 (long-distance dependencies) 或句法移位 (syntactic movement)。例如，引语通常出现在诸如 say 这样的引语动词之后。但是在这个句子中，引语 “We would have to wait until we have collected on those assets” (我们必须一直等到我们收集了这些有用的东西为止) 却出现在 “he said” 等单词之前。只包含结点 -NONE- 的空 S 标示出在 said 之后引语句子出现的位置。在树库 II 和树库 III 中，这个空结点用索引号 2 来标示，表示引语 S 就在这个句子的开头。在某些剖析器中，这种相互索引的办法，可以使我们比较容易地重新找到那些已经提到句子前面或者已经主题化的引语原来就是动词 said 的补语这样的事实。类似的结点-NONE-标示的事实是：在动词 to wait 的直接前面没有句法主语，代替句法主语的是更前面的名词短语 (NP) We。这个结点用索引号 1 来标示，它与索引号为 2 的结点之间存在相互索引关系。

```
( (S ( ' ' ' ' )
  (S-TPC-2
    (NP-SBJ-1 (PRP We) )
    (VP (MD would)
      (VP (VB have)
        (S
          (NP-SBJ (-NONE- *1) )
          (VP (TO to)
            (VP (VB wait)
              (SBAR-TMP (IN until)
                (S
                  (NP-SBJ (PRP we) )
                  (VP (VBP have)
                    (VP (VBN collected)
                      (PP-CLR (IN on)
                        (NP (DT those) (NNS assets))))))))))
          ( , , ) ( ' ' ' ' )
          (NP-SBJ (PRP he) )
          (VP (VBD said)
            (S (-NONE- *T*-2) ))
          ( . . ) ) )
```

注意：空结点-NONE-的用法

图4 LDC 宾州树库中来自 Wall Street Journal 语料库部分的句子“*We would have to wait until we have collected on those assets, he said.*”

宾州树库 II 和宾州树库 III 还进一步增加了一些信息以便表示谓词和论元之间的关系。某些短语也可以用一些标记来标注，表示该短语的语法功能（例如，表层主语、逻辑主题、强调、非 VP 谓语），它们在特定文本范畴中的表现，或者它们的语义。例如，图 4 中的标记 -SBJ（表层主语）、-TMP（时间短语）。

树库中的句子隐含地构成了语言中的一部语法。例如，从图 2（包含 2 个句子）和图 4（包含 1 个句子）中的三个经过剖析的句子中，我们可以分别抽出上下文无关语法（context-free grammar, 简称 CFG）的规则。为了简单起见，我们可以不管规则中 NP, VP 等短语的后缀（如 NP-SBJ 中的 -SBJ 等）。最后我们可以得到图 5 中的语法（Grammar）和词表（Lexicon）。

Grammar	Lexicon
$S \rightarrow NP VP .$	$PRP \rightarrow we \mid he$
$S \rightarrow NP VP$	$DT \rightarrow the \mid that \mid those$
$S \rightarrow "S", NP VP .$	$JJ \rightarrow cold \mid empty \mid full$
$S \rightarrow -NONE-$	$NN \rightarrow sky \mid fire \mid light \mid flight \mid tomorrow$
$NP \rightarrow DT NN$	$NNS \rightarrow assets$
$NP \rightarrow DT NNS$	$CC \rightarrow and$
$NP \rightarrow NN CC NN$	$IN \rightarrow of \mid at \mid until \mid on$
$NP \rightarrow CD RB$	$CD \rightarrow eleven$
$NP \rightarrow DT JJ, JJ NN$	$RB \rightarrow a.m.$
$NP \rightarrow PRP$	$VB \rightarrow arrive \mid have \mid wait$
$NP \rightarrow -NONE-$	$VBD \rightarrow was \mid said$
$VP \rightarrow MD VP$	$VBP \rightarrow have$
$VP \rightarrow VBD ADJP$	$VBN \rightarrow collected$
$VP \rightarrow VBD S$	$MD \rightarrow should \mid would$
$VP \rightarrow VBN PP$	$TO \rightarrow to$
$VP \rightarrow VB S$	
$VP \rightarrow VB SBAR$	
$VP \rightarrow VBP VP$	
$VP \rightarrow VBN PP$	
$VP \rightarrow TO VP$	
$SBAR \rightarrow IN S$	
$ADJP \rightarrow JJ PP$	
$PP \rightarrow IN NP$	

图 5 从图 2 和图 4 的三个树库的句子中抽出的 CFG 语法规则以及词表条目的一个样例

这样，我们就从语料库中挖掘到了语言学知识，包括语法规则的知识和词表的知识。这就是所谓的“文本数据挖掘”（text data mining）。

当然，我们还应当清楚地看到，直接从语料库或树库中挖掘知识并不是一件轻而易举的事情。例如，如果我们从美国的宾州树库中自动地提取语法规则，由于用来剖析宾州树库的语法相对地说来比较扁平，这就使得 CFG 语法的规则又多又长。例如，展开 VP 的规则就达 4,500 条之多，其中有很多带有任意长度的 PP 序列的规则，而动词论元排列的可能的规则也是各

色各样的:

VP → VBD PP

VP → VBD PP PP

VP → VBD PP PP PP

VP → VBD PP PP PP PP

VP → VB ADVP PP

VP → VB PP ADVP

VP → ADVP VB PP

还有一些很长的规则, 例如:

VP → VBP PP PP PP PP PP ADVP PP

在下面的例句中, 与这个 VP 规则中的各个 PP 相应的语言成分用斜体字母标示出来:

This mostly happens because we *go from football in the fall to lifting in the winter to football again in the spring.*

下面是宾州树库中数以千计的 NP 规则的一部分:

NP → DT JJ NN

NP → DT JJ NNS

NP → DT JJ NN NN

NP → DT JJ JJ NN

NP → DT JJ CD NNS

NP → RB DT JJ NN NN

NP → RB DT JJ JJ NNS

NP → DT JJ JJ NNP NNS

NP → DT NNP NNP NNP NNP JJ NN

NP → DT JJ NNP CC JJ JJ NN NNS

NP → RB DT JJS NN NN SBAR

NP → DT JJ NNS , NNS CC NN NNS NN

NP → DT JJ JJ VBG NN NNP NNP FW NNP

NP → NP JJ , JJ “ SBAR ” NNS

后面两个规则来自如下的两个 NP:

[<sub>DT</sub>The] [<sub>JJ</sub> state-owned] [<sub>JJ</sub> industrial] [<sub>VBG</sub> holding] [<sub>NN</sub> company] [<sub>NNP</sub> Instituto] [<sub>NNP</sub> Nacional] [<sub>FW</sub> de] [<sub>NNP</sub> Industria]

[<sub>NP</sub> Shearson's] [<sub>JJ</sub> easy-to-film], [<sub>JJ</sub> black-and-white] “[<sub>SBAR</sub> Where We Stand]” [<sub>NNS</sub> commercials]

如果我们把树库看成一个大规模的语法 (large grammar), 那么 Wall Street Journal 语料库的宾州树库 III 的语法就包括大约 100 万单词, 其中大约有 100 万条非词汇的形符 (token) 规则, 17,500 条不同的类符 (type) 规则。显而易见, 如果要从这样大规模的真实语料库或树库中获取有用的知识, 需要研究人员具有非凡的语言洞察力。否则, 必定会被淹没在数以百万计的数据的汪洋大海之中而迷失方向。在这个意义上说, 理性的思维 (所谓的“拍脑袋”) 在经验数据的分析和归纳中具有重大的作用。

乔姆斯基认为, 语言学研究应当遵循自然科学研究中的“伽利略-牛顿风格” (Galilean-Newtonian style)。“伽利略风格”的核心内容是: 人们正在构建的理论体系是确实的真理, 由于存在过多的因素和各种各样的事物, 现象序列往往是对于真理的某种歪曲。所

以，在科学研究中，最有意义的不是去考虑现象，而应当去寻求那些看起来确实能够给予人们深刻见解的原则。伽利略告诫人们，如果事实驳斥理论的话，那么，事实可能是错误的。伽利略忽视或无视那些有悖于理论的事实。“牛顿风格”的核心内容是：在目前的科学水平下，世界本身还是不可理解的，科学研究所要做的最好的事情就是努力构建可以被理解的理论，牛顿关注的是理论的可理解性，而不是世界本身的可理解性。科学理论不是为了满足常识理解而构建的，常识和直觉不足以理解科学的理论。牛顿摒弃那些无助于理论构建的常识和直觉。

因此，“伽利略-牛顿风格”的核心内容是：人们应当努力构建最好的理论，不要为干扰理论解释力的现象而分散精力，同时应当认识到，世界与常识直觉是不相一致的。

乔姆斯基的告诫是语重心长的，值得我们深思。在我们从语料库中挖掘知识的时候，应当特别小心，千万不要被现象序列的表面光彩所迷惑，从而导致对于真理的歪曲（冯志伟 2009）。

另外，语料库中也会存在某些错误，特别是做过标注的语料库，难免带有人为主观色彩，需要使用者去辨别。在这种情况下，特别需要语言学家的知识和洞察力来帮助我们辨别语料库中的知识的真伪，完全依靠语料库往往会得出偏颇的、甚至是荒谬的结论。这时，理性主义的方法就显示出优越性，经验主义的方法就往往会出现问题（冯志伟 2010）。

#### 4. 交叉路口的决策

在上述文本数据挖掘中的理性主义和经验主义这两种方法在语言学研究的其它领域也是存在的。面对理性主义和经验主义者两种方法，我们应当何去何从？自然语言中既有深层次的现象，也有浅层次的现象；既有远距离的依存关系，也有近距离的依存关系；既有使用统计方法就能解决的问题，也有必须依靠语言学知识才能解决的问题。语言学研究既要使用演绎法，也要使用归纳法。因此，面对这样的交叉路口，我们主张把理性主义和经验主义结合起来，把基于规则的方法和基于统计的方法结合起来。我们认为，强调一种方法，反对另一种方法，都是片面的，都无助于语言学的发展。

在英语句子中，介词短语（preposition phrase, 简称 PP）有时附着在 NP 上，有时附着在 VP 上，出现句法歧义。这是英语句法剖析的一个难题，也是英语教学中的一个难题，叫做“PP 附着问题”（PP attachment）。

这个问题可以借助于概率上下文无关语法（probabilistic context-free grammar, 简称 PCFG）来解决。在 PCFG 中，介词短语附着关系的选择归结为在下面两个规则之间的选择：一个规则是  $NP \rightarrow NP PP$ （NP 附着），另一个规则是  $VP \rightarrow VP PP$ （VP 附着）。这两个规则的概率依赖于训练语料库。Hindle 和 Rooth（1999）在 1999 年报告，根据对 AP 新闻专线 1,300 万单词的语料库统计，NP 附着占 67%，VP 附着占 33%。同年，Collins（1999）也报告，在混合 Wall Street Journal 和 IBM 计算机手册的语料库中 NP 附着占了 52%。不论 NP 附着的优先性是 52% 还是 67%，当出现 PP 附着的歧义时，NP 附着总是优先于 VP 附着。这是经验主义的统计规律。

但是，在选择有歧义的介词短语附着的正确剖析时，语言学知识仍然起着重要的作用。例如，在“Moscow sent more than 100,000 soldiers into Afghanistan ...”这个句子中，介词短语 into Afghanistan 或者附着于 more than 100,000 soldiers 这个名词短语(NP)，或者附着于以 sent 为中心语的动词短语 (VP)，出现 PP 附着歧义。根据 PCFG 的统计结论，NP 附着总是优先于 VP 附着。因此，在这个句子中，介词短语 into Afghanistan 应当附着于 more than 100,000

soldiers 这个名词短语 (NP)。

然而，这显然是一个荒谬的分析结果，与我们的语感不符。根据我们的语感，在上面这个例子中，PP 的正确附着应当是在动词上，而不在名词上，是 VP 附着而不是 NP 附着。在这种场合，动词 send 的次范畴 (sub-categorization) 知识起着重要的作用。因为动词 send 的次范畴需要表示“方向”，而“方向”正好通过介词 into 来表示。这样，就把 PP 与 VP 紧密地联系在一起了。由此看出，PCFG 这样的统计方法应当与具体单词的语言学知识结合起来，才可能得出正确的剖析结果。

其实，关于动词 send 的次范畴的语言学知识并不是人的头脑里先天地存在的，而是人们对于 send 这个动词使用的实践在人的头脑中的反映，这样的理性主义的语言学知识仍然是与语言使用的统计事实有关系的。我们对宾州树库中 send 的使用情况进行过细致的统计后发现，在宾州树库 (Penn Treebank) 的 Brown 语料库和 Wall Street Journal 语料库关于动词的语料中，与主要动词 send 有关的介词 into 所有歧义的 PP 附着全都是附着于动词，这样的统计事实完全支持我们关于动词 send 的次范畴的语言学知识。可见，理性主义的语言学知识实际上与人们使用语言的经验主义的统计数据是密切相关的，这些知识实际上是客观的统计规律在人们头脑中的反映。基于这样的理由，在语言学研究中，把理性主义方法和经验主义方法相结合便是不可避免的了。这样的结合不仅是研究方法上的革新，也是语言客观事实的要求。

英国经验主义哲学家培根 (F. Bacon) 既反对理性主义，也反对狭隘的经验主义。他指出，由于经验能力和理性能力这两方面的“离异”和“不和”，给科学知识的发展造成了严重的障碍。为了克服这样的弊病，他提出了经验能力和理性能力联姻的重要原则。他说，“我以为我已经在经验能力和理性能力之间永远建立了一个真正合法的婚姻，二者的不和睦与不幸的离异，曾经使人类家庭的一切事务陷于混乱”<sup>5</sup>。他生动而深刻地说道：“历来处理科学的人，不是实验家，就是教条者。实验家像蚂蚁，只会采集和使用；推论家像蜘蛛，只凭自己的材料来织成丝网。而蜜蜂却是采取中道的，它在庭园里和田野里从花朵中采集材料，而用自己的能力加以变化和消化。哲学的真正任务就正是这样，它既非完全或主要依靠心的能力，也非只把从自然历史和机械实验收来的材料原封不动，囫囵吞枣地累置于记忆当中，而是把它们变化过和消化过放置在理解力之中。这样看来，要把这两种机能、即实验的和理性的这两种机能，更紧密地和更精纯地结合起来 (这是迄今还未收到的)，我们就可以有很多的希望”<sup>6</sup>。

培根的以上主张值得我们深思。在语言学的战略转移中，我们不能采取像蜘蛛那样的理性主义方法，单纯依靠规则，也不能采取像蚂蚁那样的经验主义方法，单纯依靠统计。我们应当像蜜蜂那样，把理性主义和经验主义两种机能“更紧密地”、“更精纯地”结合起来，推动语言学的发展。我们认为，这才是当前语言学战略转移的正确方向。

当然，我们对于当前语言学战略转移的上述见解，仅仅是我们对于整个语言学的发展战略这个大问题的思考中的一个小小的过渡环节，如果我们的这种见解在检验中被否证了，那么我们就应当毫不犹豫地放弃自己的见解。我们始终坚信，任何的科学见解不管曾经获得何等的成功，也不管曾经受过何等严格的检验，都是可以推翻的。

#### 参考文献

- Abeill   A. 2003. *Treebank: Building and Using Parsed Corpora* [M]. Dordrecht: Kluwer.  
Collins, M. 1999. *Head-Driven Statistical Models for Natural Language Parsing*[D]. PhD thesis, University of

<sup>5</sup> 北京大学哲学系外国哲学史教研室，《十六——十八世纪西欧各国哲学》，第 8 页，商务印书馆。

<sup>6</sup> 培根，《新工具》，第 75 页，商务印书馆。

- Pennsylvania, Philadelphia.
- Hindle, D. & M. Rooth. 1999. Structure ambiguity and lexical relations [A]. *Proceedings, DARPA, Speech and Natural Language Workshop* [C]; 229-236. Hidden Valley, PA.
- Hinrichs, E. & S. Kübler. 2005. Treebank profiling of spoken and written German [A]. *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories* [C]. Barcelona, Spain.
- Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A., 1993. Building a large annotated corpus of English: The Penn treebank [J], *Computational Linguistics*, 19(2): 313-330.
- Marcus, M. P., Kim, G., Marcinkiewicz, M. A., Marcintyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B., 1994. The Penn Treebank: Annotating predicate argument structure [A], In ARPA Human Language Technology Workshop [C]: 114-119. Plainsboro, NJ. Morgan Kaufmann.
- Sinclair, J. M. 1991. *Corpus Concordance Collocation* [M]. Oxford: Oxford University Press.
- Sinclair, J. M. 2007. Intuition and annotation: The discussion continues [M]. In W. Teubert & R. Krishnamurthy (eds.). *Corpus Linguistics: Critical Concepts in Linguistics*. London/New York: Routledge, 415-435.
- Weaver, W., 1949, Warren Weaver's memorandum in 1949: Translation, Milestones in Machine Translation [A], In Locke, W.N. and Booth, A.D. (eds.) *Machine translation of languages: fourteen essays* [C], Cambridge, Mass.: Technology Press of the Massachusetts Institute of Technology, 1955.
- Widdowson, H. 2000. The limitation of linguistics applied [J]. *Applied Linguistics* 21: 3-25.
- 波普尔, 2008, 如何对待错误[A], 载: 哥白尼, 爱因斯坦, 霍金等, 《宇宙简史——无限宇宙中的无穷智慧》[C]. 长沙: 湖南科学技术出版社, 2008: 203-204。
- 冯志伟, 1983, 汉语句子的多叉多标记树形图分析法[J]. 人工智能学报(2): 29-46。
- 冯志伟, 1996, 自然语言的计算机处理 [M], 上海: 上海外语教育出版社。
- 冯志伟, 2009, 乔姆斯基《最简方案》[A]. 载《现代语言学名著导读》(萧国政主编) [C]. 北京: 北京大学出版社, 86-128。
- 冯志伟, 2010, 自然语言处理的形式模型[M]. 合肥: 中国科学技术大学出版社。
- 黎锦熙, 1924, 新著国语文法[M], 商务印书馆。
- 刘海涛、冯志伟, 2007, 自然语言处理的概率配价模式理论[J]. 语言科学(5): 32-41。
- 王力, 1988, 王力文集·第九卷·汉语史稿[M]. 济南: 山东教育出版社。

收稿日期: 2010-10-10

作者修改稿, 2010-12-25

本刊修订, 2011-01-02

通讯地址: 100010 北京市东城区朝内南小街 51 号 教育部语言文字应用研究所  
<zwfengde2010@gmail.com>

### On Strategic Transit of Linguistic Study

Feng Zhiwei

Institute of Applied Linguistics, Ministry of Education

Dalian Maritime University

After having analyzed the strategic transit in computational linguistics, the author proposes that traditional linguistics is facing the rigorous challenge from empiricism, the corpus-based approach or the corpus-driven approach shall gradually replace the traditional introspection or elicitation approaches, but these corpus-based approach or corpus-driven approach can not leave from the insight to language phenomenon of linguistics, we can not neglect the importance of rational thinking, the combination of rationalism and empiricism is the correct orientation of strategic transit in current linguistics.

**Keywords:** strategic transit; computational linguistics; corpus; rationalism; empiricism