关于乔姆斯基和统计学习的两种文化

原文：http://norvig.com/chomsky.html

詹卫东 编译 （2013.09.25）

一 背景：

2011 年是麻省理工学院（MIT）建校 150 周年。2011.5.3-5 日，MIT 举办了“大脑、心智与机器（Brians，Minds and Machines）”专题研讨会（属校庆系列活动之一）。网址：http://mit150.mit.edu/symposia/brains-minds-machines

研讨会期间有一场主题讨论会（Keynote Panel），题为：黄金时代——人工智能、认知科学与神经科学的发端巡礼

讨论会全程视频：http://techtv.mit.edu/videos/13200-keynote-panel-the-golden-age-a-look-at-the-original-roots-of-artificial-intelligence-cognitive-science-and-neuroscience-

讨论会主持人是哈佛大学心理系教授 Steven Pinker。

参加讨论的人有：

Sydney Brenner, 索尔克生物研究所高级研究员（2002 年诺贝尔奖得主，在基因编码领域有突出贡献）
Marvin Minsky, 麻省理工学院媒体艺术与科学教授
Noam Chomsky, 麻省理工学院语言与哲学系教授
Emilio Bizzi, 麻省理工学院脑科学研究所教授
Barbara H. Partee 麻省大学语言与哲学系教授
Patrick H. Winston 麻省理工学院人工智能与计算机科学教授

在讨论会最后，Pinker 向 Chomsky 发问，如何看待概率模型近年来在认知科学领域到处开花的趋势。概率方法在人工智能、认知科学的黄金时代（上世界 70-80 年代）并不是主角。
http://languagelog.ldc.upenn.edu/myl/PinkerChomskyMIT.html

Chomsky 的回应：
（1）确实有许多研究工作在尝试用统计模型来解决各种各样的语言学问题。其中有一些取得了成功。但是大多数是失败的。

（2）那些取得成功的应用，是因为把统计方法跟语言的基本属性（比如普遍语法的属性）结合起来使然。比如在连续语篇中如何识别单词的边界。

（3）如果不考虑语言的实际结构就应用统计方法，那么所谓的成功不是正常意义

下的成功。就科学研究的历史经验来说，这种意义上的成功并非主流。这就好像研究蜜蜂行为的科学家只是对着蜜蜂录像，通过记录蜜蜂的历史行为，加以统计分析，来预测蜜蜂未来的行为。也可能统计方法可以预测得很好，但这算不上科学意义上的成功。研究蜜蜂的科学家并不关心这种预测。

二 Google 科学家 Peter Norvig 对 Chomsky 的上述看法发表评论

考虑如下 5 个问题：
1） Chomsky 的主要观点是什么？ 他是正确的吗？
2） 什么是统计模型？
3） 统计语言模型取得的成绩到底怎么样？
4） 在科学研究的历史中，有类似这样的成功吗？
5） Chomsky 不喜欢统计模型的到底是什么？

Norvig 逐一回答了这些问题。主要内容如下：

1

Chomsky 的主要观点：
A. Chomsky 认为统计语言模型取得过工程意义上的成功，但不关科学的事。
B. 为语言事实建模就像收集蝴蝶标本。科学(尤其是语言学)想要的是基本原则。
C. 统计模型无法理解，并不是关于研究对象的洞见。
D. 统计模型或许可以对一些现象做出精确的模拟，但这是迷途。人们并不根据前面出现的两个单词去预测后面一个单词。人们生成句子（词语序列）的方式是从内在的语义到树结构，再到表层的线性词语序列。
E. 统计模型已经被证实无法用于学习语言。因此语言必然是天生的。用语言模型去解释语言是浪费时间。

Norvig 的主要回应：
A. 工程上的成功确实不是科学目标。不过科学和工程是比翼齐飞的。工程上的成功可以作为科学上成功模型的证据。
B. 科学是事实和理论的混合体。理论过分凌驾于事实之上并不可取。在科学史上，不断积累事实是科研正途，并非异类。关于语言的科学也不应例外。
C. 包含几十亿个参数的统计模型确实难以直观理解。个人确实无法核查每个个体参数的意义所在。但是，人们可以通过了解整个模型的特性而获得对于统计模型合理与否的认知：即一个统计模型是怎样有效的，或者为什么无效，它是如何从数据中学到模型函数的，等等。
D. 基于词概率的 Markov（马尔科夫模型）确实无法对所有的语言现象建模。这就像没有概率的简单树结构模型无法对所有的语言现象建模一样。我们需要的语言模型是可以覆盖词、树结构、语义、上下文、语篇等等不同层次语言现象的更复杂的概率模型。Chomsky 不能因为旧的统计模型的缺点就一概否定所有的统计语言模型。研究如何解释语言（比如语音识别）的人当中，绝大多数很容易发觉，解释是一个概率问题。当一个语音流到了我耳朵里，要把这串语音流恢复为说话者的意义，是一个概率问题。爱因斯坦说过，让事情变得简单，直到不能再简单为止。

许多科学现象都有随机性。最简单的模型就是概率模型。语言也是这样一种现象。因此概率模型是表达语言事实的最好工具。
E. 1967 年，Gold 定理指出，形式化的数学语言在逻辑推导上的理论限制。但是，这跟自然语言学习者面临的问题毫无关系。无论如何，在 1969 年，我们知道了，概率推理不受这一限制的约束（Horning 证明学习概率上下文无关文法 PCFG 是可能的）。我同意 Chomsky 所说的，人类具有学习语言的天赋。但是我们对如何获得概率化的语言表示，对统计学习，都还缺乏足够的知识。我认为很可能人类学习语言涉及到概率和统计推理，但是我们并不清楚细节。

2

统计模型是一种数学模型，通过给定的数据，训练得到。统计模型通常是概率模型，但并不一定如此。二者的区别很重要。

数学模型：一个数学模型是对变量关系的定义。可以用函数形式定义，即从输入到输出的函数。例如：y = mx + b。 也可以用关系的形式定义。例如：(x,y) 满足某种关系。

概率模型：描述随机变量的可能取值的概率分布。例如 P(x,y)。概率分布不再是严格的确定的函数关系。比如：y = f(x) 是确定性的函数关系。

训练模型：通过统计推断，在收集的数据基础上，选取最好的模型，通常也就是选取模型的参数。比如上面例子中 y = mx + b 这一函数中的参数 m 和 b。通过选取参数的方式来确定模型。

在 Chomsky 之前，Claude Shannon 提出了通信的概率模型，其基础正是单词的 Markov 链。如果你有一个 10 万词的词表，考虑一个二阶 Markov 模型（该模型刻画了一个单词出现的概率如何依赖其前面的两个单词），那么要确定这个模型的参数，你需要 10^15 这么多的数据（即 10 万*10 万*10 万的三维矩阵的数据量）。要学习获得这个模型，就必须收集数据，同时得想办法处理那些不存在数据的位置（即三维矩阵中值为 0 的那些位置）。大多数（但并非全部）概率模型都是通过参数训练获得的模型。许多训练模型（也并非全部）是概率性质的。

再看一个例子，牛顿的重力引力模型：两个物体之间的吸引力跟它们的质量和距离的关系为：
F = G m1 m2 / r^2

这里 G 是万有引力常量。这是一个训练模型的例子，因为 G 是由随机试验测定的结果决定的。同时，这又是一个非概率模型（确定性模型），因为它描述了一个明确的函数关系。Chomsky 大概不会反对这种意义上的“统计模型”。Chomsky 对统计模型的批评主要是针对 Shannon 那样的需要天文数字那么多的参数的统计模型，而不是只有一两个参数的模型。

万有引力模型还有一个区别意义。引力模型是连续和定量的。而语言学中的传统

模型往往是离散的、范畴化的、定性的。一个词要么是动词，要么不是，并没有关于它的“动词性”（verbiness）的量化程度描述。

还一个相关的概率统计模型是“理想气体定律”（ideal gas law）。这个定律描述了气压 P 跟气体分子数 N ，温度 T ，以及 Boltzmann（玻尔兹曼）常量 K 之间的函数关系：

P = N K T / V

这个公式是从统计力学的基本原理导出的。它是不确定的、不准确的模型。一个完全准确的模型应该是描述每一个个体的气体分子的运动。但这个模型忽视了单个气体分子的位置的不确定性。尽管它是一个统计概论模型，尽管它不能描述全然的真实情况，但是它对气体的整体状况提供了良好的预测 —— 这种关于气体的深刻洞察是无法通过了解单个气体分子的真实运动状况而获得的。

现在，让我们来考虑单词拼写的非统计模型。规则：I 在 E 之前，除非 I 在 C 之后（I before E except C，参见：
http://en.wikipedia.org/wiki/I_before_E_except_after_C）

描述这个现象的概率、训练的统计模型则是：

P(IE) = 0.0177 P(CIE) = 0.0014 P(*IE) = 0.163
P(EI) = 0.0046 P(CEI) = 0.0005 P(*EI) = 0.0041

这个模型是从英语万亿词级语料库（corpus of trillion words）中获取的统计数据（http://norvig.com/ngrams/）。
P(IE) 表示该语料库中的一个单词含有"IE"的概率。
P(CIE) 表示该语料库中一个单词含有"CIE"的概率。
P(*IE) 表示该语料库中一个单词含有"IE"但"IE"不在"C"之后的概率。
其余 P(EI), P(CEI), P(*EI)含义仿此类推。

统计数据表明："IE" 确实比 "EI" 常见（0.0177 : 0.0014），
"IE" 在 "C" 之后出现的情况确实相对少见，但是 P(CIE) > P(CEI)。
这是跟传统规则相反的。即便在"C"之后，"IE" 仍然比 "EI" 更常见 （0.0014 > 0.0005）。
包含"CIE" 的单词例子如：science，society， ancient，species 等等。

上述拼写规则的不足是它的精度（Accuracy）不够高。

Accuracy("I before E") = 0.0177 / (0.0177+0.0046) = 0.793
Accuracy("I before E except after C") = (0.0005+0.0163) / (0.0005+0.0163+0.0014+0.0041) = 0.753

更复杂的统计模型可以在拼写检查这样的应用中使精度达到现在的十倍。

（http://norvig.com/spell-correct.html）

再看最后一个例子，这个不是统计模型，但是是富于洞察力的一个模型。

高等法院法官握手理论（Theory of Supreme Court Justice Hand-Shaking）：

高院开庭时，所有法官都会和其他法官握手。法官参加人数为 n，取值范围 0-9。在给定 n 的情况下，总的握手次数 h 是多少？下面是三个可能的答案：

A. h = n * (n-1) / 2
B. h = Σi = 1 .. n (i - 1)
C. (n, h) 有如下对应表
（0,0) (1,0) (2,1) (3,3) (4,6) (5,10) (6,15) (7,21) (8,28) (9,36)

公式 A 背后的原理是：每个人跟其他人握手次数为 n*(n-1) ，但这样把“张三-李四”和“李四-张三”握手分别记了两次，所以总握手次数应除以 2

公式 B 背后的原理是：为避免重复记次，先对法官按年龄排序。只记岁数大的人跟岁数小的人的握手次数。

公式 C 背后的原理是，逐一遍历 n 从 0 到 9 的所有情况，把所有握手次数记录下来，构建 n 和 h 的对应表。

有的人可能喜欢 A 模型，有的人可能喜欢 B 模型，还有些不喜欢乘法和加法的人则可能钟意 C 模型。但其实这三个模型说的都是一回事 —— 其实是同一个理论 —— 都是从 n 到 h 的函数，可以覆盖 n 的所有可能取值。可能 A 跟 B 模型比 C 模型更有用。因为前两个模型更一般化，可以应付 n 值增大的情况。

3

统计语言模型取得的成功到底怎么样？

成功在这里定义为：对世界做出准确的预测。

- 搜索引擎：100%的训练和概率模型
- 语音识别：100%的训练和概率模型
- 机器翻译：NIST 中排名靠前的系统 100%的使用统计方法。一些商用系统使用统计和规则混合的方法。在机器翻译系统可以处理的 4000 种语言对中，统计方法的系统都表现更出色，除了日语-英语之间的翻译。对于日-英机译系统，最好的统计机器翻译系统跟混合系统的表现相当。
- 问答系统：研究尚不成熟。多数统计和概率方法的系统使用搜索引擎来实现问答。IBM 的 Watson 系统完全是概率和训练模型。Boris Katz 的 START 系统是混合系统。所有的系统都至少使用了一些统计技术。

再看一些计算语言学家感兴趣，但不是用在终端用户的技术：

• 词义消歧： SemEval-2 比赛中排名靠前的系统 100%使用统计技术。多数是概率模型，一些使用概率模型加知识库（例如 Wordnet）规则的混合模型。
• 指代消解： 主要的系统都是统计方法。 Haghighi and Klein 的系统是混合系统，其中规则方法比训练模型更重要。该系统的性能跟统计方法的系统相当。
• 词性标注： 主要的系统格是统计方法。Brill 标准器是混合系统。它从统计数据中学习确定性的规则。
• 句法分析： 大多数成功的句法分析系统是统计方法的，主要是概率模型。

显然，说统计模型在语言处理方面所取得的成功有限，是不准确的。事实是，统计模型在语言处理的各项任务中，已经取得了压倒性的优势。

另一个视角是看研究人员的态度。在计算语言学家中，统计方法已经成为被接受的主流方法。（Norvig 自己在经历了 14 年的规则方法后转向概率方法）

上述理由可能会被视为是“工程视角”，那么，接下来，就来看“科学视角”吧。

4

在科学史上，统计模型有过成功吗？

Chomsky 认为，科学史上罕有统计模型的成功例子。
Chomsky 的意思是,“精确的模型化这个世界”是罕见的,科学史上的成功标准是，提供对世界的解释 —— 事物为何是它现在这个样子，而不是描述它怎样成为这个样子。也就是说，科学关心的是 why 的问题，而不是 how 的问题。

科学的词典释义是“通过观察和试验,对物理和自然界的结构和行为的系统研究”。就这个定义而言,科学对 why 和 how 的问题是并重的。看一看《科学》(Science)，亦可以达到以管窥豹的效果。随机地从《科学》中选取一篇文章的标题：

Chlorinated Indium Tin Oxide Electrodes with High Work Function for Organic Device Compatibility （具有高功函数的氯化铟锡氧化电极的有机元件兼容性）

（科学杂志文章地址：
http://www.sciencemag.org/content/332/6032/944.abstract）

这篇文章关注“精确地模型化世界”胜过“提供对事物的解释”。

Norvig 翻检了一期《科学》杂志的全部标题和摘要，另外也看了一期《细胞》杂志。还有 2010 年诺贝尔物理学奖、化学奖、生理和医学奖的工作。

结论是，这些研究工作 100%的重视“精确地模型化这个世界”胜过“提供对这个

世界的解释”。Norvig 同时也承认，分辨这二者并不容易，这是一个没有清晰定义的问题。
Norvig 甚至还考虑把这个问题抛给土耳其机器人（Mechanical Turk）来回答。不过有朋友告诉他这实在是太难为机器人了。

5

Chomsky 不喜欢什么样的统计模型？

统计模型跟概率模型常常难分彼此。Chomsky 反对的是概率模型。Chomsky（1969）写道：必须认识到,“一个句子的概率”是完全没有用的概念，不管在什么意义上，这都是一个没用的概念。”

Chomsky（1957）写道：
我认为……概率模型没有对句法结构的基本问题给出有意义的解释。

Chomsky 的依据可以用下面的例子来说明：

1） I never, ever, ever, ever, ... fiddle around in any way with electrical equipment.
2） She never, ever, ever, ever, ... fiddles around in any way with electrical equipment.
3） * I never, ever, ever, ever, ... fiddles around in any way with electrical equipment.
4） * She never, ever, ever, ever, ... fiddle around in any way with electrical equipment.

无论句子中的 ever 重复多少次，都不影响 1、2 是合语法的，而 3、4 是不合语法的。因此，一个 n 元马尔科夫概率模型在碰到句子中的 ever 个数超出 n 的时候，就分不清 1 跟 3 或者 2 跟 4 的区别了。概率马尔科夫模型对英语的描写因而是有限的。

这个批评没错。但这只是对马尔科夫概率模型的批评，并不意味着所有的概率模型都因此而要遭到同样的批评。从 1957 年到现在，已经发展处许多概率模型，上面这 4 个例子，可以用有限状态模型来描述。此外 PCFG（概率上下文无关文法）可以有更强的能力。PCFG 比单纯基于范畴的上下文无关文法更容易学习得到。每一个概率模型实际上都是一个确定性模型的超集（superset）。后者只不过是将概率值严格地限定为 0,1 而已。对概率模型的合理的批评必然是因为它们表达能力过强，而不是因为它们的表达能力不够。

在《句法结构》一书中，Chomsky 提出了一个著名的例子，同时也是对有限状态概率模型的一个批评：

（a）colorless green ideas sleep furiously （无色的绿色思想狂怒地睡觉）

（b）furiously sleep ideas green colorless （狂怒地睡觉思想绿色无色的）

尽管 a、b 的任何部分都是未见于历史上的任何英文文献的，但 a 是合语法的，b 是不合语法的。

就整个句子而言，Chomsky 显然是正确的。但说到句子中的“部分”，则并不尽然。下面是一些部分出现的例子：

- ″It is neutral green, colorless green, like the glaucous water lying in a cellar.″ The Paris we remember, Elisabeth Finley Thomas (1942).
- ″To specify those green ideas is hardly necessary, but you may observe Mr. [D. H.] Lawrence in the role of the satiated aesthete.″ The New Republic: Volume 29 p. 184, William White (1922).
- ″Ideas sleep in books.″ Current Opinion: Volume 52, (1912).

撇开关于“部分”的争议不说，实际上，基于统计训练的有限状态模型可以区分上面 a、b 两例。Pereira（2001）就提出了一个这样的模型，在增加了词类信息后，对新闻语料进行期望最大化的参数训练，计算结果是例 a 的概率是 b 的概率的 20 万倍。为了说明这不是因为这两个句子在新闻语料训练得到模型中有如此区别，Norvig 用 Google 图书语料库（1800-1954）的训练模型重复做了计算，结果是例 a 的概率为例 b 的 10 万倍。如果可以在树结构的基础上计算，则对句子“合语法性程度”的估计效果会更好。而不是像 Chomsky 提出的基于范畴的语法那样，仅仅只是区分“合语法/不合语法”。

Chomsky 对统计模型的另一个异议是，儿童在只有 10^8 秒的时间里，如何学习 10^9 那么多的参数（实际上，现在的统计模型的参数已经远远多于 1960 年代的 10^9 这个数量级了）。

确实，没有人会提议，儿童学习这些参数是一个一个学的。正确的假设是，那些接近 0 的参数是批量学习的（就像割韭菜一样，一刀下去一茬尽在手中），而那些高概率值的参数则随着观察数据的不断增加而持续更新。没有人认为马尔科夫模型是对自然语言的一个严肃的模型。但是，概率化的训练模型可以比范畴化的无训练模型更好地表达自然语言。

一个自然语言的科学理论必须正视这样的事实，母语者对很多短语和句子的合语法性，也有拿不准的时候。因此，概率模型可以比范畴化的形式语法模型做的更好。

1） The earth quaked.
2) ? It quaked her bowels.

quake 这个动词一般在词典中都标记为不及物动词（intransitivie）。因此，根据基于范畴的形式语法模型，上面例 1 是合语法的，例 2 是不合语法的。

但是，例 2 这样的句子又确实有这样的用法。于是基于范畴的形式语法模型就陷入了两难的困境。接受例 2 和排斥例 2 都有问题。像这样的问题，在概率语法模型中，就不存在困难。只需要说 quake 的不及物用法是概率很高的用法，及物用法的概率很低就可以了。

Steve Abney 还指出过，概率模型用于对语言变化建模，也更有优势。

Norvig 指出，看起来，合语法性并不是范畴化的、确定性的判断，而是概率性的。花时间观察真实语料中的例句是值得的，与此同时，通过内省自己的语言直觉来研究合语法性，也无不妥。观察和直觉在科学研究的历史中并不相悖。只不过，从来都是观察，而不是直觉，在科学研究中占据主流地位。

Chomsky 对统计模型的异议主要来自“精确描写”和“科学解释”之间的对立。达尔文对生物学的研究以富于深刻的洞察而著名。但他强调“精确描写”的重要性。达尔文说“错误的事实对科学研究的进程是有伤害的，而且会有长期影响。但有少量证据支持的错误的观点对科学研究的伤害要小得多。”物理学家费曼也说“物理学可以不需要证明而进步，但没有事实则不可能进步。”

三 两种文化

2001 年统计学家 Leo Breiman 发表了一篇文章《统计建模：两种文化》（Statistical Modeling：The Two Cultures）。(Leo Breiman 是加州大学伯克利分校统计学教授 http://www.stat.berkeley.edu/~breiman/）

一种是数据建模文化（data modeling culture）。
一种是算法建模文化（algorithmic modeling culture）。

前一种文化的要点是：自然可以被看作是一个黑盒子，有相当简单的模型把输入数据跟输出数据对应起来（在这个过程中，可能有随机的噪音掺入）。统计学家的任务是选择一个基本的模型，可以反映自然界的这种真实的数据对应关系。

后一种文化的要点是：自然界的黑盒子不一定能靠简单的模型来描写。复杂的算法（比如支持向量机 SVM、决策树、深度信念网）可以估算出从输入数据到输出数据的函数，但是，我们不能期望这样的函数形式可以反映自然界的真实本质。

大约 98%的统计学家是前一种文化的拥趸，2%的统计学家和许多其他领域的研究者（特别是研究复杂现象的学者）支持后一种文化。

Chomsky 着力反对的是后一种统计文化。不是仅仅因为这样的模型是基于统计的（或基于概率的）。而是因为这样的模型宣称是对现实的精确刻画，但却不易被人解读。同时这样的模型没有对自然的生成过程做出解释。换言之，算法建模只是描述了发生了什么，但没有回答为什么会这样的问题。

Breiman 在文章中解释了他为何反对第一种文化（数据建模）。基本上，基于数据

建模得到的结论是关于数据的，而不是关于自然本身的（Norvig 在 2000 年听火星登陆计划负责人 James Martin 说过，他作为太空工程师的工作不是登陆到火星上，而登陆到由地质学家提供的“火星模型”上）。问题是，如果模型对自然的刻画不够好，那么这个结论可能是错的。比如，线性回归（linear regression）是统计学家的百宝囊中最强大的工具之一。因此，许多分析都从“假设数据是由线性模型产生得到的……”开始。如果数据实际上不是由这样的模型产生，那么对实际模型的样子就会缺乏足够的分析。此外，对于复杂的问题，往往有许多不同的好的模型可供选择，它们对数据的适应性相差无几。统计学家如何做出选择呢？Breiman 想说服我们放弃这样的信念：我们可以得到关于自然的模型的唯一形式。如果我们能得到一个模型，该模型可以对观测数据做出很好的解释，并且能对未出现的数据做出不错的预测，我们就应该感到很满意了。Chomsky 则走上相反的一条路：他更喜欢简单的优美的模型，因此而放弃能很好地刻画数据的模型（这样的模型在数学上可能是很复杂的）。Chomsky 认为数据（他称之为语言表现 language performance）对语言学来说是有限，语言学真正面对的对象是语言能力（language competence）。

2011 年 1 月份，电视名嘴 Bill O'Reilly 因“潮起潮落，你如何解释”而掀起争议。他是信上帝的。反对者嘲笑他不知道潮汐现象可以用太阳、地球、月亮的引力作用来解释。这个解释最早是 1776 年由 Laplace 提出的。当拿破仑问 Laplace 为何在他的解释中造物主没有一席之地，Laplace 说，“我不需要这个假设。”O'Reilly 似乎也不知道 Deimos 和 Phobos 等等其他关于太阳系的天文知识。不过，O'Reilly 知道，批评者认为他在天文学方面的无知并没有什么了不起，因为他的支持者们认为他直接触及到了更本质的问题：为什么（Why）？他不关心潮汐怎样（How）工作。他要问的是，他们为什么工作。为何月亮在恰当的距离来制造美妙的潮汐。为什么引力这样工作？等等等等。O'Reilly 是对的。这些问题只能靠编造故事、宗教或哲学来回答，科学干不了这个。

Chomsky 的哲学理念是：我们应该关注深层的“为什么（why）”，只是解释表层的现实是不够的。在这个意义上，Chomsky 其实跟 O'Reilly 是同路人。Chomsky 相信语言理论应该简单且可理解，就像线性回归模型那么简单，我们需要做的，只是估计斜率和截距。

举个例子，考虑著名的 pro-drop（代词脱落）语言（这个概念来自 Chomsky 1981）. 英语中，可以说 “I'm hungry” 但是在西班牙语中，同样的意思要说“Tengo hambre”（字面上相当于：have hunger），代词（主语）Yo 脱落了。Chomsky 的理论是，语言系统有一个 pro-drop 参数。该参数的取值，在西班牙语是“ture”（真），在英语是“false”（假）。如果我们可以找到描述所有语言的为数不多的参数，并且确定每个参数的具体取值，我们就真的理解了语言。

问题是，语言的现实比这个理论要杂乱得多。下面是英语中 pro-drop 的例子：

- ″Not gonna do it. Wouldn't be prudent.″ (Dana Carvey, impersonating George H. W. Bush)
- ″Thinks he can outsmart us, does he?″ (Evelyn Waugh, The Loved One)

- ″Likes to fight, does he?″ (S.M. Stirling, The Sunrise Lands)
- ″Thinks he's all that.″ (Kate Brian, Lucky T)
- ″Go for a walk?″ (countless dog owners)
- ″Gotcha!″ ″Found it!″ ″Looks good to me!″ (common expressions)

语言学家可以为如何解释上面这些现象争个没完没了。但语言的多样性似乎远比用布尔值（true or false）来描述 pro-drop 参数值要复杂。一个理论框架不应该把简单性置于反映现实的准确性之上。

从一开始，Chomsky 就把注意力放在了语言的生成性上。从这个方面来说，非概率性的理论是合理的。如果 Chomsky 把注意力放在语言的另一面“理解（解释）”上，如同 Claude Shannon 所关注的那样，Chomsky 或许会改变他的说法。在“理解”这一面，听话人需要对收到的信号进行消歧，决定哪种可能的解释概率最高。这很自然地会被看作是一个概率问题。语音识别的研究者如此看待对语音的解释。其他领域的研究解释的科学家也是如此。天文学家 Laplace 在 1819 年说“概率理论只不过是让人们的常识能够计算。”

Chomsky 不喜欢统计模型，还有一个原因。因为统计模型会让语言学成为一门经验学科，而不是数学。而 Chomsky 更喜欢把语言学看作是数学。Chomsky（1965）：“语言学理论是心理的，关心的是比实际行为更基础的心理现实。观察语言的实际应用或许可以提供一些证据，但是并不能构成语言学的主题。”

无法想象 Laplace 会说，观察行星的运动不能构成轨道力学的主题。

物理学家会研究理想的、从实际世界中抽象出来的力学（比如忽略摩擦力），但是这并不意味着摩擦力不能成为物理学的研究主题。

语言是复杂的、随机的、不确定的生理过程，受到进化和文化变迁的影响。构成语言的不是一个外在的理想实体（由少量的参数设定），而是复杂处理过程的不确定的结果。因其不确定性，用概率模型来分析语言就是必然选择。