语言文学研究

句法结构认知难度的计算指标分析

方　昱　刘海涛

[摘　要]　句法结构认知难度是语言学和认知科学共同关注的话题。认知难度最直接的测量方式是心理实验。近年来,计算认知科学领域出现了一些计算指标,可以基于大规模真实语料,采用数理统计方法和自然语言处理技术测度认知难度。本文梳理了其中五个指标:基于工作记忆容量限制的存储成本、整合成本和依存距离,以及基于经验预测的惊异值和概率配价,探究这些指标应用于语言学研究的可能性。学科间的交叉和融合已成为当代科学研究的一个重要趋势,对语言学研究而言,学习和借鉴其他学科的研究成果,有助于更好地了解语言结构模式和演化规律,提升研究的科学化水平。

[关键词]　句法结构;认知难度;计算认知;语言学研究;跨学科

一、引言

句法复杂度是二语口笔语教学与研究中的一个重要构念,常用来衡量学习者写作和口语等语言水平的发展(李茜,2013;Lu,2011)。目前常用的句法复杂度指标包括平均子句长度、并列句比例和从属句比例等。这些指标涵盖了句子结构的众多方面,但大多没有涉及语言学意义上的句法。准确来说,这些指标关注的是句子结构的复杂度。

句法复杂度还可以从认知角度出发来测量。语言使用者在句法加工过程中会耗损一定的认知资源,消耗的资源越多,句法复杂度越高。这种句法复杂度又可称作认知难度。衡量认知难度最直接的方法是实验。借助自控速阅读(self-paced reading)、眼球追踪(eye tracking)和事件相关电位(ERP)等技术,获取反应时、注视时间和 N400 认知电位等数据,便可以直观描述认知难度。只要实验设计合理,结果就较为可信。但实验会耗费大量人力物力,研究者难以招募很多被试,所用实验材料的数量也较为有限,因而实验结果常常难以复制和推广。

自然语言处理技术的发展,尤其是句法分析技术的完善,使得大规模、高精度的自动句法标注成为可能。在这一背景下,带有句法标注的语料库(树库)越来越多,如宾州树库、布拉格依存树库和哈工大中文依存树库,这为基于真实语料衡量认知难度提供了新思路。研究者从这些树库中提取句法关系、词汇共现和共现频率等信息,借助数理统计、信息论和计算机建模技术,构建认知难度的计算指标,进而探究符合人类普遍认知规律的句法加工特点。如果计算指标的预测结果与实验结果一致,便可考虑用计

方昱,语言学博士,同济大学外国语学院助理教授(上海 200092);刘海涛,语言学博士,浙江大学外语学院教授、长江学者特聘教授,通讯作者(杭州 310058)。

算指标补充心理实验,而这也是计算认知科学(computational cognitive science)的初衷。

按研究者对句法加工内部机制的不同理解,现有计算认知指标可分为两类。第一类为基于工作记忆容量限制的指标,包括存储成本(storage cost,SC)、整合成本(integration cost,IC)、依存距离(dependency distance,DD)等。研究者认为句法加工过程需要工作记忆的参与,加工难度越大,工作记忆负荷越高。而人的工作记忆容量是有限的,因而高工作记忆负荷会造成较大的认知难度。另一类为基于经验预测的指标,包括惊异值(surprisal)、概率配价(probabilistic valency)、熵(entropy)等。这一派研究者认为在句法加工过程中,语言使用者会依据以往经验预测之后出现的内容,这些经验包括但不限于句法结构形式、句法结构使用频率和词的语法特征。使用者拥有的经验越丰富,预测成功的几率越大,认知难度便越小。相较于传统的句法复杂度指标,这些指标更加关注句法结构,衡量的是句法结构认知难度。

这两类指标自提出以来,已得到心理语言实验的验证,可以用来解释部分语言现象。这说明,虽然这些指标源于认知科学、信息论和自然语言处理领域,但同样有助于语言规律的探寻。在促进多学科交叉和深度融合的今天,语言学研究者需要更加积极地学习和借鉴其他学科的研究成果。鉴于此,本文将梳理上述两类计算指标,对比指标预测结果与实验结果,探究将这些指标应用于语言学研究的可行性,以期更好地了解语言结构的特点,揭示认知机制对语言结构的制约作用。

二、基于工作记忆容量限制的指标

第一类指标以工作记忆负荷为基础来衡量认知难度。研究者认为,句法加工是一个逐词递增的过程,人们会即时解析已出现的词与新出现的词之间的句法关系。如果句法关系出现在两个非相邻词 w_n 和 w_{n+i}之间,工作记忆负荷就会增加,而人们的工作记忆容量是有限的,认知难度由此产生。本文介绍三种基于工作记忆容量限制的指标:存储成本、整合成本和依存距离。

(一) 存储成本和整合成本

存储成本和整合成本是基于短语结构语法提出的,它们一起构成了依存局域理论(dependency locality theory,DLT)(Gibson,1998,2000)。这一理论认为,一个词出现后,语言使用者需要完成两种句法加工任务。其一,在工作记忆中保存当前尚不完整的句法关系,由此产生的认知难度用存储成本来衡量,以记忆单位(memory unit/MU)计;其二,从工作记忆中回溯与该词相关的句法信息,将其融入之前尚不完整的句法关系中,由此产生的认知难度用整合成本来度量,以能量单位(energy unit/EU)计。Gibson(2000,p. 102)认为整合成本更为重要,多数情况下可以只用整合成本描述认知难度。

整合成本又可分为话语处理成本(discourse processing cost,DPC)和结构整合成本(structural integration cost,SIC)。以图 1 为例,该图为句子 S 的短语结构句法分析结果,以 h_2 为中心词的投射 XP 与以 h_1 为中心词的投射 Y_1 之间存在句法关系。当 h_2 出现时,一方面需要为 h_2 的投射 XP 构建话语结构,当 h_2 为名词或动词时,便会出现话语处理成本;另一方面需要建立以 h_2 为中心词的投射 XP 与投射 Y_1 之间的联系,这一过程消耗的结构整合成本由 h_1 与 h_2 之间的名词和动词数量决定。

已有研究者利用依存局域理论解释关系从句的加工难度。不同类型关系从句的研究中,以主语提取关系从句(subject-extracted relative clauses,SRC)和宾语提取关系从句(object-extracted relative clauses,ORC)最为常见(何文广、陈宝国,2011)。心理语言实验结果表明,英语 ORC 的加工难度大于 SRC(Grodner & Gibson,2005),与依存局域理论的预测一致。我们以 Grodner & Gibson(2005)使用的一组 SRC、ORC 为例,对比实验结果与依存局域理论预测结果,见表 1。例(2a)和(2b)分别包含 SRC

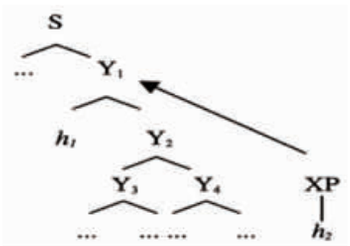

图 1　以 h_2 为中心词的投射 *XP* 到以 h_1 为中心词的投射 Y_1 的整合成本

和 ORC,用黑斜体标识,SC 表示存储成本,IC 表示整合成本。

表 1　依存局域理论对英语 SRC 和 ORC 认知难度的预测

(2a)	The	reporter	***who***	***sent***	***the***	***photographer***	***to***	***the***	***editor***	hoped	for	a	story.
SC	2	1	3	1	2	1	2	2	1	1	1	1	0
IC	0	1	0	1	0	1	0	0	1	1 + 3	0	0	1
派摄影师去见编辑的那个记者希望能有一篇新闻报道。													
(2b)	The	reporter	***who***	***the***	***photographer***	***sent***	***to***	***the***	***editor***	hoped	for	a	story.
SC	2	1	3	4	3	1	2	2	1	1	1	1	0
IC	0	1	0	0	1	1 + 2	0	0	1	1 + 3	0	0	1
被摄影师派去见编辑的那个记者希望能有一篇新闻报道。													

就(2a)而言,当第一个词 *the* 出现时,其后至少还需要一个名词和一个动词才能构成完整的句子,它的存储成本为 2 MUs;*the* 是冠词,不消耗话语处理成本,整合成本为 0 EU。第二个词 *reporter* 出现后,只需一个动词就能形成完整的句子,存储成本变为 1 MU;*reporter* 为名词,消耗一个话语整合成本,同时 *reporter* 与 *the* 相邻,构成名词短语,这一过程不会消耗结构整合成本,因而整合成本为 1 EU。类似可推知每个词出现后存储成本和整合成本的变化。句子的存储成本由句中最大的存储成本决定,整合成本由最大的整合成本决定。因而,(2a)的存储成本为 3 MUs,整合成本为 4 EUs;(2b)的存储成本为 4 MUs,整合成本为 4 EUs。就关系从句部分而言,SRC 的存储成本为 3 MUs,整合成本为 1 EU;ORC 的存储成本为 4 MUs,整合成本为 3 EUs。存储成本和整合成本均预测 ORC 的认知难度更大。

自控速阅读实验结果同样显示 ORC 的认知难度更大。表 2 给出了该实验的结果,表中数字表示每个词(词组)的阅读时间,单位为毫秒。为了方便对比实验结果与指标预测结果,我们将每个词(词组)的存储成本和整合成本以括号的形式加在阅读时间后面,括号中的第一个数字表示存储成本,第二个数字表示整合成本。由表 2 来看,ORC 中 *sent* 的阅读时间最长,是引起 ORC 认知难度的主要原因,这与整合成本的预测一致(*sent* 的整合成本最大)。但存储成本预测 ORC 中主语(即 *the photographer*)的认知难度更大,与实验结果存在差异。

表 2　英语 SRC 和 ORC 的阅读时间

	who	the photographer	sent	to the editor	总阅读时间
SRC	350(3,0)	360(2,1)	355(1,1)	356(2,1)	1421
ORC	343(3,0)	353(4,1)	422(1,3)	398(2,1)	1516

(二) 依存距离

基于工作记忆容量限制的另一指标是句法距离(syntactic distance)。在依存语法框架下,句法距离(依存距离)指句子中两个有依存句法关系的词之间的线性距离(刘海涛,2009)。在计算依存距离前,需要对句子进行依存句法分析。图 2 为例(2a)和(2b)的依存句法分析结果。

图 2 中带有箭头的弧线连接了两个有句法关系的词,箭头从支配词指向从属词,表示这两个词之间

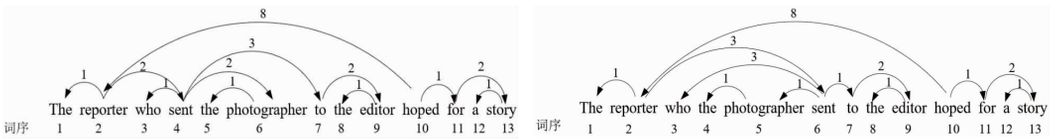

图 2　例(2a)(左)和(2b)(右)的依存句法结构

的二元非对称关系,弧线上方的数字表示依存距离。如 *the* 与 *reporter* 之间由一条弧线相连,*the* 为从属词,*reporter* 为支配词,它们的依存距离为支配词词序减去从属词词序(2 - 1 = 1)(Liu,2006,2008)。在这一句法关系中,支配词位于从属词后,依存距离为正值。但某些句法关系中,支配词位于从属词前,如 *hoped* 和 *for*,这时依存距离为负值(10 - 11 = - 1)。依存距离的正负体现了语言类型的差异(Liu,2010),与认知难度关系不大,因而这里只考虑依存距离的绝对值。多数情况下,依存距离越长,认知难度越大。基于依存树库的大数据分析结果显示,依存距离最小化是自然语言的一个普遍特征(Futrell,Mahowald & Gibson,2015;Liu,2008),这可能是人们为减轻交流过程中的认知负担而做出的选择。

完成句法分析和依存距离计算后,便可统计每个词的认知难度。认知难度源于从工作记忆中提取已出现过的词,建立新出现的词与这个(些)词之间的句法关系。以(2a)中的 *sent* 为例,它与 *who*, *photographer* 和 *to* 之间存在句法关系,但只有 *who* 和 *photographer* 出现在 *sent* 之前,因此 *sent* 的认知难度由 *sent* 与这两个词的依存距离之和(1 + 2 = 3)决定。类似地,计算(2a)和(2b)中每个词的认知难度,结果如表 3 所示。

表 3　基于依存距离计算的(2a)和(2b)的认知难度

(2a)	The	reporter	***who***	***sent***	***the***	***photographer***	***to***	***the***	***editor***	hoped	for	a	story
	0	1	0	3	0	3	3	0	3	8	1	0	3
(2b)	The	reporter	***who***	***the***	***photographer***	***sent***	***to***	***the***	***editor***	hoped	for	a	story
	0	1	0	0	1	8	1	0	3	8	1	0	3

一个句子的认知难度由句中所有词的平均认知难度决定,即由句子的平均依存距离(mean dependency distance,MDD)决定(Liu,2008)。(2b)的平均依存距离为 26/12,大于(2a)的平均依存距离(25/12),(2b)的认知难度更大。如果只考虑关系从句,ORC 的平均依存距离为(13/7)也大于 SRC 的平均依存距离(12/7),与实验结果一致,说明依存距离能较好地预测句子的认知难度。具体到每个词,ORC 中 *sent* 的认知难度最大,同样与实验结果一致。

(三) 依存局域理论与依存距离对比

依存局域理论和依存距离对认知难度的动因有一致认识,即不断增加的工作记忆负荷与有限工作记忆容量的矛盾。存储成本测量的是理解过程中需要记住的不完整句法关系数量,需要记住的数量越多,认知难度越大。类似的指标还有瞬时信息块数(陆丙甫、于赛男,2018)、嵌入深度和 F + L - (van Schijndel & Schuler,2013)等。整合成本和依存距离则以词间距离衡量认知难度,距离越长,认知难度越大。由于整合成本是依据短语结构语法定义的,依存距离是依据依存语法定义的,它们对距离的测算存在差异。整合成本关注短语结构树中两个投射的中心词之间的距离,依存距离则关注具有依存句法关系的词与词之间的线性距离。

自然语言处理技术的快速发展使得指标的自动化计算成为可能。比如利用 Stanford Parser、MaltParser 和哈工大 LTP 等句法分析器标注依存句法关系,获取包括支配词和从属词词序在内的句法信息,便可计算依存距离。整合成本虽是依据短语结构语法定义的,但为了适应大规模语料的处理需求,在实际计算中,常常将短语结构树转换为线性结构,统计中心词之间的名词和动词数量,或直接

计算词与词之间的线性距离(Rajkumar,*et al.*,2016)。对于存储成本,虽然目前还没有直接可用的测量工具,但类似的指标,如嵌入深度和 F + L -,可以通过 ModelBlocks 计算得到。

总体上,三个指标都能较为准确地预测句子的认知难度,但它们却不一定能准确预测每个词的认知难度,如存储成本没有反映出(2b)中 *sent* 一词的认知难度。这可能是因为这些指标强调的是词与词之间的句法关系,没有考虑词本身的特点。不同词类的认知负担是有差异的,如人称代词比名词更容易理解,因而将 SRC 和 ORC 的主语换做代词时,SRC 和 ORC 的认知难度差异消失(Warren & Gibson,2002)。整合成本只统计名词和动词数量,反映了 Gibson 对这一问题的思考。但只做名词、动词与其他词类的划分显然是不够的。当英语 ORC 的主语分别为代词、人名、姓氏、带定冠词的名词、带不定冠词的名词时,认知难度逐步增加(Warren & Gibson,2002)。

这三个指标也不能准确预测所有句子的认知难度。比如,在德语句末动词前插入关系从句,动词的阅读时间不仅不会增加,反而会减少,与整合成本和依存距离的预测刚好相反。要解释这类现象,可能得借助基于经验预测的指标。

三、基于经验预测的指标

第二类指标以经验的丰富度为出发点来衡量认知难度。研究者认为,在句子加工过程中,语言使用者会根据经验预测接下来出现的内容(Levy,2008)。某个词或某种句法结构出现的次数越多,语言使用者所获得的经验越丰富,预测的准确性就会越高,这个词或这种句法结构的认知难度也就越小(Levy,2008)。这里主要介绍两种基于经验预测的指标:惊异值和概率配价。

(一) 惊异值

"惊异"源于信息论,用来描述某一观测事件的信息值。假设随机事件 X 出现的概率为 $p(x)$,其惊异值便为 $-\log_2 p(x)$。惊异值自 Hale(2001)引入心理语言学后,已被不少研究者用来评估句子的认知难度(Rajkumar,*et al.*,2016;Smith & Levy,2013)。如果将句中某个词的出现看作随机事件 X,这个词出现的概率越大,它的惊异值便越小,认知难度也就越小。假设一个句子的前 $n-1$ 个词为 $w_1 \cdots w_{n-1}$,第 n 个词 w_n 的出现受 $w_1 \cdots w_{n-1}$ 的制约,(条件)概率为 $p(w_n \mid w_1 \cdots w_{n-1})$。将这一概率代入惊异值的计算公式,便可得到 w_n 的惊异值 $surprisal(w_n) = -\log_2 p(w_n \mid w_1 \cdots w_{n-1})$。以例(2a)中的 *sent* 为例,这个词的条件概率可表示为 $p(sent \mid the\ reporter\ who)$,惊异值为 $-\log_2 p(sent \mid the\ reporter\ who)$。计算惊异值的关键是估算条件概率 $p(w_n \mid w_1 \cdots w_{n-1})$,可借助语言模型获取,如 N 元语言模型、神经网络语言模型、概率上下文无关语法模型和概率依存语法模型。前两种语言模型关注词的线性顺序,后两种语言模型除词的线性顺序外,还加入了句法关系。本文关注的是句法结构认知难度,接下来将重点介绍后两种语法模型。

概率上下文无关语法(probabilistic context free grammar,PCFG)属于短语结构语法的一种,是将概率引入短语结构语法形成的语法规则系统。概率依存语法(probabilistic dependency grammar)(Nivre,2006)是另一种将概率与语法规则相结合的语法,是依存语法的概率化扩展。PCFG 和概率依存语法的核心都是用数理统计的方法分析语言成分之间的关系,分析句法结构出现的概率。假设基于 PCFG 分析词串 $w_1 \cdots w_n$ 的句法结构,这 n 个词一起出现的概率 $P(w_1 \cdots w_n)$ 可以表示为 $\sum_T P(T, w_1 \cdots w_n)$,其中 T 代指短语结构树,$\sum_T P(T, w_1 \cdots w_n)$ 为由 $w_1 \cdots w_n$ 构成的所有短语结构树的概率之和。如果换成概率依存语法,$\sum_T P(T, w_1 \cdots w_n)$ 则表示由 $w_1 \cdots w_n$ 构成的所有依存树的概率之和。这

样，w_n 的惊异值可以按照下面的公式来计算：

$$surprisal(w_n) = -\log_2 p(w_n \mid w_1 \cdots w_{n-1}) = -\log_2 \frac{P(w_1 \cdots w_n)}{P(w_1 \cdots w_{n-1})} = -\log_2 \frac{\sum_T P(T, w_1 \cdots w_n)}{\sum_T P(T, w_1 \cdots w_{n-1})}$$

接下来应用 HumDep（Boston，*et al.*，2008）和 TdParse（Roark，*et al.*，2009）估算例（2a）和例（2b）的惊异值，以验证惊异值能否准确预测句子的认知难度。HumDep 是基于概率依存语法开发的，训练集只包含词性信息，只能输出非词汇化惊异值（unlexicalized suprisal）。TdParse 是基于 PCFG 开发的，训练语料包含词性和词信息，既可以输出非词汇化惊异值（在该软件中被称作句法惊异值，syntactic surpirsal），也可以输出词汇化惊异值（lexicalized surprisal）。例（2a）和例（2b）的分析结果如表 4 所示。

表 4　基于 TdParse 和 HumDep 计算的例（2a）和（2b）的惊异值

（2a）	The	reporter	***who***	***sent***	***the***	***photographer***	***to***	***the***	***editor***	hoped	for	a	story
SynS	0.88	0.73	6.27	1.36	1.55	0.08	2.38	3.04	0.41	4.45	0.56	1.02	0.12
LexS	0.96	9.58	0.06	6.25	1.19	4.53	0.00	0.44	8.79	8.00	2.40	1.60	4.46
PosS	0	0.03	0.53	1.48	2.20	0.79	1.46	2.20	0.92	1.11	0.75	0.99	0.51
（2b）	The	reporter	***who***	***the***	***photographer***	***sent***	***to***	***the***	***editor***	hoped	for	a	story
SynS	0.88	0.73	6.27	3.69	0.07	1.86	2.06	2.59	0.37	4.80	0.68	1.04	0.11
LexS	0.96	9.58	0.06	0.55	4.48	5.99	0.00	0.44	8.66	7.93	2.42	1.60	4.47
PosS	0	0.03	0.53	1.88	1.02	1.12	0.90	1.91	1.17	1.34	0.59	0.99	0.51

表 4 显示了（2a）和（2b）中每个词的惊异值，SynS 和 LexS 的结果来自 TdParse，分别表示句法惊异值和词汇化惊异值，PosS 的结果来自 HumDep，表示非词汇化惊异值。句中所有词的惊异值之和便是整个句子的惊异值（Fang & Liu，2021）。根据 TdParse 的估算结果，（2a）的句法惊异值为 22.85，（2b）的句法惊异值为 25.15；（2a）的词汇化惊异值为 48.26，（2b）的词汇化惊异值为 47.14。根据 HumDep 的估算结果，（2a）的非词汇化惊异值为 12.97，（2b）的非词汇化惊异值为 11.99。其中，SRC 的句法惊异值为 15.09，词汇化惊异值为 21.26，非词汇化惊异值为 9.58；SRC 的句法惊异值为16.91，词汇化惊异值为 20.18，非词汇化惊异值为 8.53。

由以上结果来看，只有句法惊异值的结果显示 ORC 的认知难度大于 SRC。此外，这三种惊异值的预测结果均未体现 *sent* 是造成 ORC 认知难度的主要原因。这一结果表明：一方面，惊异值能预测句子的认知难度，但预测的准确性受语法、词和词性的影响；另一方面，惊异值在预估具体词汇的认知难度时，可能作用有限。

（二）概率配价

另一个基于经验预测的认知难度指标是概率配价。配价是依存语法的核心概念，表示一个词（类）与其他词（类）的结合能力（刘海涛，2009）。当一个词（类）进入句子时，这种能力得以实现，多种可能的配价变为一种，此时词（类）与词（类）之间便形成了句法关系。因而，配价与句法关系是包含与被包含的关系，句法关系是实现了的配价。正如句法关系中存在支配词和从属词一样，词（类）的配价也分为支配和从属两类。支配表示它作为支配词（类）的能力，从属表示它受别的词（类）支配的能力。词（类）的配价模式可以借助图 3 表示。

其中，W 代表一个词（类），$G_1, G_2, \cdots G_{n-1}, G_n$ 为 n 种可以支配 W 的句法关系，$D_1, D_2 \cdots D_{m-1}, D_m$ 为 m 种可以受 W 支配的句法关系，带有箭头的线条表示支配方向。在语言使用中，不同句法关系出现的可能性不是均等的（Liu，2006）。假设 W 为动词，它可以支配名词形成主谓句法关系（D_3），也可以支配量词形成动补句法关系（D_2）。由语言使用经验可知，主谓句法关系比动补句法关系更常见。

刘海涛和冯志伟（2007）将概率引入词（类）的配价模式，借助概率说明句法关系出现可能性的差

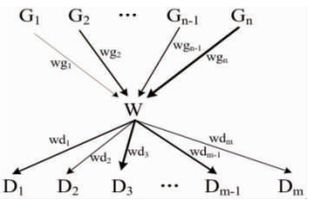

图 3　词(类)W 的配价模式

异,这便是概率配价。如图 3 所示,用 wg_1,$wg_2 \cdots wg_{n-1}$ 和 wg_n 分别表示句法关系 G_1,$G_2 \cdots G_{n-1}$,G_n 在 W 的整个被支配能力中所占的比例,$wg_1 + wg_2 + \cdots + wg_n = 1$;用 wd_1,$wd_2 \cdots wd_{m-1}$ 和 wd_m 表示句法关系 D_1,$D_2 \cdots D_{n-1}$,D_n 在 W 的总支配能力中所占的比例,$wd_1 + wd_2 + \cdots + wd_m = 1$。同时,用不同粗细的线条表示不同的比例,比例越大,线条越粗。具体到计算过程中,可以通过依存树库获取这些比例:首先从树库中提取所有 W 为从属词的句法关系,再分类统计每种句法关系的数量,便可得到每种句法关系所占的比例 $wg_1, \cdots, wg_n$。类似地,计算 $wd_1, \cdots, wd_m$ 的值。

概率配价表示两个词(类)形成句法关系的概率,应用到句法加工中,可以理解为当前出现的词(类)w_n 与已经出现的词(类)形成句法关系的概率。当 w_n 出现时,它可能与之前的某个词(类)形成 m 种句法关系,对应 m 个概率。人们一般倾向于按概率最大的句法关系理解。比如,高松(2013)以“小王研究鲁迅的文章发表了”为例,利用最大概率原则解释了花园幽径句理解困难的原因。当我们看到专有名词“鲁迅”时,会将其理解为动词“研究”的宾语,此时这种句法关系的概率最大。看到名词“文章”后,我们会修正之前的分析,将名词理解为动词“研究”的宾语,因为名词作宾语的概率大于专有名词作宾语的概率。动词“发表”出现后,我们又会推翻之前的分析,将前面出现的词理解为名词短语,作动词“发表”的主语。句法分析过程一波三折,正体现了概率对句法加工的影响。从计算角度来看,如果 w_n 与已经出现的多个词(类)都有句法关系,那就先按最大概率原则确定 w_n 与每个词(类)的句法关系,然后将这些句法关系对应的概率相乘,便是 w_n 在这个位置形成句法关系的概率。在计算句法关系的概率时,还需要考虑构成句法关系的词(类)在句中的相对位置,也就是要考虑两个词(类)之间的依存距离。

接下来以(2a)和(2b)为例,具体说明如何用概率配价理论分析句子的认知难度。配价既可以是词类的结合能力,也可以是词的结合能力,这里暂只考虑词类的配价。我们将从布拉格捷克英语依存树库(Prague Czech-English Dependency Treebank)中提取概率信息,因而依照该树库的标注准则对(2a)和(2b)进行词性标注和句法分析。按以下步骤从左到右依次分析(2a)中每个词类的认知难度,即分析每个词类与已出现词类形成句法关系的概率。

(1) 任意句子中,第一个词类的概率记作 0,因为它没有与任何词类形成句法关系。

(2) 确定当前词的词性,分析该词性与已出现词性的关系。如果这个词有多种词性,则需分别分析每种词性的情况。如 *sent* 既可能是动词的过去式,也可能是过去分词,它的前面有 *the*(冠词)、*reporter*(名词)、*who*(疑问代词),需要分别确定过去式(或过去分词)与冠词、名词、疑问代词的关系。

(3) 根据最大概率原则,确定每对词性组合的共现概率和句法关系。如前所述,计算词类的概率配价时,需分别考虑它作支配词和从属词两种情况。计算词性组合的共现概率时,同样需分开考虑。以过去式(*sent*)与名词(*reporter*)为例,第一种情况以过去式为支配词,名词为从属词,依存距离 2 为条件,检索到树库中共出现了 323 次,包含三种句法关系,其中主语句法关系(*Sb*)出现了 300 次,占比最大,为 0.93;第二种情况以过去式为从属词,名词为支配词,依存距离 -2[①] 为条件,得到 24 个

①布拉格捷克英语依存树库在计算依存距离时区分了正负,这里同样以支配词词序减去从属词词序作为它们这两个词的依存距离。

检索项，均为属性关系（*Atr*），概率为1。最终确定这对词性组合构成了属性关系，概率为1。

（4）得到当前词性与已出现词性的共现概率后，将这些概率相乘，作为当前词性形成句法关系的概率。如过去式（*sent*）形成句法关系的概率为1（过去式－名词）×1（过去式－疑问代词）=1（树库中没有检索到过去式与冠词的句法关系，说明冠词出现与否并不影响过去式的概率）。

（5）如果一个词有多种词性，它的词性由概率较大的那种词性决定。如 *sent* 作过去式时的概率为1，作过去分词时的概率为0.99，因而将其标记为过去式。

（6）依存树库中可能有一些标注误差，因而只有当检索项超过10个时，才认定两个词性的句法关系成立。此外，如果两种句法关系有冲突，选择概率较大的那个。比如步骤（3）中将过去式理解为名词的从属词，不仅因为该句法关系的概率为1，还因为如果用过去式支配名词，构成主谓句法关系，就同过去式与疑问代词构成的主谓句法关系产生了冲突，而后者的概率更大。

按以上步骤，计算例（2a）和（2b）中每个词类出现后形成句法关系的概率，结果见表5。

表5　例（2a）和（2b）中每个词性形成句法关系的概率

（2a）	The	reporter	***who***	***sent***	***the***	***photographer***	***to***	***the***	***editor***	hoped	for	a	story
	0	0.89	0	1	1	0.88	1	0	0.71	0.77	0.71	0.41	0.52
（2b）	The	reporter	***who***	***the***	***photographer***	***sent***	***to***	***the***	***editor***	hoped	for	a	story
	0	0.89	0	0	0.89	0.89	0.99	0	0.71	0.77	0.71	0.41	0.52

一个词形成句法关系的概率越大，这个词（类）被理解的可能性越大，它的认知难度便越小。句子的认知难度由句中所有词的概率之和决定。这样，（2a）的概率为7.89，（2b）的概率为6.78，其中，SRC的概率为4.59，ORC的概率为3.48。ORC的概率越小，认知难度越大，与实验结果一致，说明概率配价能准确预测这两个句子的认知难度。

刘海涛和冯志伟（2007）认为，在配价模式中引入概率，有助于更好地解释语言的生成和理解过程，判定句法的合格性。目前已有研究者基于概率配价理论，阐释了花园幽径句的理解机理（高松，2013）。还有研究者基于概率配价理论考察了语言结构的隐现规律（徐春山，2015）。本文基于（2a）和（2b）的试验结果则说明概率配价也可以用来衡量句子的认知难度，但可能还需要通过更多语言、更多例句来验证这个指标的有效性。当然，对其他指标也需如此。

（三）惊异值与概率配价对比

研究者提出惊异值和概率配价等指标，是因为他们认为在言语交流过程中，我们会根据对方说过的话，预测他接下来会说什么。预测的准确性与使用频率密切相关。一个词如果经常与某些词（串）一起出现，当我们看到这些词（串），会自然预测到这个词，当它出现时，便不会觉得“惊异”。但是，如果一个词很少与这些词（串）一起出现，当我们看到它时，就会觉得“惊异”。也就是说，使用频率与认知难度呈负相关。惊异值和概率配价的出发点便是通过数学运算，建立使用频率与认知难度的相关关系。因而，这里的主要任务就是从真实语言数据中获取频率，为每种可能的预测标记一个概率。

惊异值解决这一问题的方法是建立语言模型获取词的条件概率。最早广泛使用的语言模型是 n 元模型，该模型的基本思想是，句中某个词 w_n 的出现只与它前面出现的 $n-1$ 个词有关。理论上，n 越大，条件概率越精确。但 n 越大，需要的训练文本也就越多。在实际操作过程中，不可能无限增加文本，只能将 n 限制在一个合理的取值范围内，二元和三元便是常见的两种取值。二元和三元模型简化了词的条件概率，会损失部分潜在有用的信息。PCFG模型和概率依存语法模型的出现解决这个问题。近年来，研究者又尝试将循环神经网络、长短期记忆神经网络等技术应用到语言建模中，利用神经网络模型估算惊异值。这些新方法进一步提高了惊异值预测的准确性（Frank & Bod 2011）。

概率配价将配价、句法关系和依存距离等概念融合在一起，借助依存树库提取句法关系的使用频率，从而确定新出现的词（类）与已出现的词（类）可能形成的句法关系。但同时，这一指标可能还存在一些问题。首先，我们认为句法关系的概率与认知难度呈负相关，但某些概率为 0 的词可能并不是很难理解，如（2a）和（2b）中的冠词 *the*。冠词属于虚词的范畴，通常不会造成太大的认知负担。其次，在句法分析过程中，我们会根据后来出现的词不断调整之前预测的句法关系（如花园幽径句的理解），但概率配价并未衡量这一修正过程对词类认知难度的影响。最后，鉴于目前还没有成熟的工具可以自动计算概率配价，很难在大规模文本中推广应用这一指标。

四、认知难度指标与语言研究

由以上结果可知：整合成本、依存距离、惊异值和概率配价等指标都可以较为准确地衡量句子的认知难度。借助数理统计方法和自然语言处理技术，就可以获取这些指标。由此打破了被试和材料对实验的限制，为依赖实验的认知研究提供了一种新范式，同时也为语言学研究提供了新方法和新路径。这一部分将探讨将这些认知难度指标应用于语言学研究的可行性。

首先，认知难度指标可以应用于二语习得研究，尤其是二语写作研究。我们在开篇已经提到，以往二语写作研究多关注句子结构的复杂度。现有的各种分析工具，如 Coh-Metrix、二语句法复杂度分析器（L2 Syntactic Complexity Analyzer，L2SCA），能够从大规模文本中自动提取平均句长、并列句比例等复杂度指标，保证了数据处理的规模和速度。但是，这种复杂度并不是语言学意义上的句法复杂度。从语言学的句法角度出发来衡量句子的复杂度，需要考虑词与词之间的句法关系。本文介绍的几种指标或是基于短语结构语法计算的，或是基于依存语法计算的，可视作句法结构复杂度指标。将这些指标引入二语写作的研究，或许有助于研究者从更多维度探讨二语写作的特点。

已有研究表明，依存距离可以用来衡量二语学习者语言水平的发展。Ouyang & Jiang（2018）对不同年级的中国英语学习者的作文进行了依存句法分析，探析依存距离的概率分布特点。他们发现各年级作文的依存距离均符合齐普夫－阿列克谢耶夫分布（Zipf-Alekseev distribution），但分布函数中的具体参数存在差异。随着学习者年级的增加（或者说随着学习者语言水平的提高），参数越来越趋近于本族语者作文的拟合结果。Li & Yan（2021）以日本英语学习者的作文为研究对象，同样发现依存距离的概率符合齐普夫－阿列克谢耶夫分布，分布函数中的参数也能区分日本学习者的语言水平。蒋景阳和姜茜茜（2021）则基于中国英语学习者的写作文本，考察了写作错误、依存距离与二语水平之间的关系。中低水平的学习者对长距离句法关系处理能力较弱，错误率较高。

除二语习得研究外，认知难度指标也可以用来描述母语者的语言产出特征。基于多语种依存树库的研究表明，自然语言有依存距离最小化的倾向（Futrell，Mahowald & Gibson，2015；Liu，2008）。基于英语或汉语近义句式语料库的研究表明，当多种句式可以表达相近意思时，说话者倾向于选择依存距离小、惊异值小的那个句子（Fang & Liu，2021；Rajkumar，*et al.*，2016）。基于德语书面语依存树库的研究表明，德语句子的破框现象并非特例，破框句的依存距离缩小，降低了认知成本（李媛、黄含笑、刘海涛，2021）。

还有研究者利用依存距离分析翻译文本的语言特点。比如，以同声传译和交替传译译文文本为语料，研究者对比了这两种译本的依存距离，发现交替传译译本的依存距离更小（Liang，*et al.*，2017）。以英语翻译文本和英语母语文本为语料，研究者发现翻译文本与母语文本的依存距离存在显著差异，一定程度上证实了翻译语言为“第三语码”的观点（蒋跃、范璐、王余蓝，2021）。此外，通过一项英汉视译实验，研究者考察了依存距离的长短对口译流利度的影响，发现译者翻译依存距离长的句子时，流利度更差（蒋跃、蒋新蕾，2019）。

这些研究反映出语言学研究与认知科学相结合的趋势，说明将依存距离等计算认知指标应用于语言学研究是可行的。将认知科学领域的最新研究成果引入语言学研究，或可促进语言学研究的进一步发展，提高语言学研究的精确性和科学性。但与此同时，现有研究的不足也不可忽视。

第一，当前语言学研究多关注基于工作记忆容量限制的指标，较少应用基于经验预测的指标。这可能是因为后者需要借助语言模型计算，而这并不是语言学研究者擅长的领域。为解决这一困境，研究者可以尝试与计算语言学、自然语言处理等领域的学者交流合作，寻求技术上的支持与帮助。

第二，除以上提到的 Fang & Liu（2021）、Rajkumar 等（2016）的研究外，少有基于语料库的语言学研究综合考量这两种指标。值得一提的是，借助心理实验，研究者发现认知难度是这两种指标综合作用的结果（Husain，Vasishth & Srinivasan，2014）。未来的语言学研究可以更多关注两种指标的关系。

第三，这两类指标目前主要应用于二语写作、语言结构特征和翻译语言特征等研究，接下来或可探究这些指标在更多语言学研究中的适用性。比如，考察认知难度与文学作品质量（和读者接受度）的关系；讨论句子的认知难度是否与不同的话语策略和目的相关；从计算认知难度出发对比分析特殊人群使用的句子与正常人使用的句子。

五、总结与展望

本文梳理了计算认知科学中用来衡量句法结构认知难度的五个指标：存储成本、整合成本、依存距离、惊异值和概率配价。这些指标对认知难度的预测与心理语言实验结果基本吻合，说明除实验外，还可以尝试从计算角度出发探讨语言的认知机制。同时，我们也要认识到这些指标的局限。

第一，这两类指标或关注工作记忆容量限制对句法加工的影响，忽略了句法加工过程中可能出现的预测行为；或关注句法加工中的预测行为，忽略了工作记忆容量的限制。当前，已有研究者尝试整合这两种指标，构建新指标来量化认知难度。这些新指标或许有助于发现更多有趣的语言规律与特点。

第二，这两类指标的计算均依赖句法信息，而句法标注工作耗时耗力。尽管已有不少准确率较高的句法分析器，但这些分析器并不能保证百分之百的准确率，还需辅以人工校对。如果出于研究目的需要自建树库，便很难最大化语料规模，使得语料更全面地、更精确地反映语言现象。

不可否认的是，认知难度指标的出现和改进是自然语言处理技术应用于语言和认知研究的结果。与心理语言实验相比，指标的计算更加省时省力，结果的可重复率更高，可以帮助我们更好地基于大数据、基于真实语言材料发现语言与认知的规律。正如计算语言学学会（Association for Computational Linguistics）终生成就奖得主、词汇功能语法理论的创立者琼·布利斯南（Joan Bresnan）所说："我希望未来能加大对计算语言学理论、技术和资源的应用力度，以不断加深我们对人类语言和认知的理解（Bresnan，2016，p. 613）。"这是一种信息时代的语言观，是信息时代对语言研究提出的新要求，也是信息时代为语言研究提供的机遇。在大力推动不同学科融合发展的今天，语言学研究者更应该积极学习借鉴计算语言学的相关技术和资源，推进语言研究的科学化进程。

参考文献：

何文广、陈宝国，2011：《认知神经科学及多学科视域中的宾主关系从句》，《南京师大学报》（社会科学版）第3期。

高松，2013：《基于概率配价模式理论的花园幽径句研究》，《语言文字应用》第3期。

蒋景阳、姜茜茜，2021：《中国英语学习者写作中的错误、依存距离与二语水平的关系研究》，《语言文字应用》第1期。

蒋跃、范璐、王余蓝,2021:《基于依存树库的翻译语言句法特征研究》,《外语教学》第 3 期。

蒋跃、蒋新蕾,2019:《最大依存距离对口译中非流利度的影响》,《外语研究》第 1 期。

李茜,2013:《任务后语言形式聚焦对英语学习者口语产出的影响》,《外语教学与研究》(外国语文双月刊)第 2 期。

李媛、黄含笑、刘海涛,2021:《德语书面语破框现象是特例吗》,《现代外语》第 3 期。

刘海涛,2009:《依存语法的理论与实践》,北京:科学出版社。

刘海涛、冯志伟,2007:《自然语言处理的概率配价模式理论》,《语言科学》第 3 期。

陆丙甫、于赛男,2018:《消极修辞对象的一般化及效果的数量化:从"的"的选用谈起》,《当代修辞学》第 5 期。

徐春山,2015:《连词"而"的隐现规律研究》,《山西大学学报》(哲学社会科学版)第 2 期。

Boston, M. F., J. Hale, R. Kliegl, U. Patil & S. Vasishth, 2008, "Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus", *Journal of Eye Movement Research*, Vol. 2, No. 1, pp. 11 – 12.

Bresnan, J., 2016, "Linguistics: The Garden and the Bush", *Computational Linguistics*, Vol. 42, No. 4, pp. 599 – 617.

Fang, Y. & H. Liu, 2021, "Predicting syntactic choice in Mandarin Chinese: A corpus-based analysis of *ba* sentences and SVO sentences", *Cognitive Linguistics*, Vol. 32, No. 2, pp. 219 – 250.

Futrell, R., K. Mahowald & E. Gibson, 2015, "Large-scale evidence of dependency length minimization in 37 languages", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 112, No. 33, pp. 10336 – 11341.

Gibson, E. 1998, "Linguistic complexity: Locality of syntactic dependencies", *Cognition*, Vol. 68, No. 1, pp. 1 – 76.

Gibson, E., 2000, "The dependency locality theory: A distance-based theory of linguistic complexity", in A. Marantz, Y. Miyashita & W. O'Neil (eds.), *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, Cambridge, MA: The MIT Press, pp. 94 – 126.

Grodner, D. & E. Gibson, 2005, "Consequences of the serial nature of linguistic input for sentential complexity", *Cognitive Science*, Vol. 29, No. 2, pp. 261 – 290.

Hale, J., 2001, "A probabilistic Earley parser as a psycholinguistic model", in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pp. 159 – 166.

Husain, S., S. Vasishth & N. Srinivasan, 2014, "Strong expectations cancel locality effects: Evidence from Hindi", *Plos One*, Vol. 9, No. 7, e100986.

Levy, R., 2008, "Expectation-based syntactic comprehension", *Cognition*, Vol. 106, No. 3, pp. 1126 – 1177.

Li, W. & J. Yan, 2021, "Probability distribution of dependency distance based on a treebank of Japanese EFL learners' interlanguage", *Journal of Quantitative Linguistics*, Vol. 28, No. 2, pp. 172 – 186.

Liang, J., Y. Fang, Q. Lv & H. Liu, 2017, "Dependency distance differences across interpreting types: Implications for cognitive demand", *Frontiers in Psychology*, Vol. 8, 2132.

Liu, H., 2006, "Syntactic parsing based on Dependency Relations", *Grkg/Humankybernetik*, Vol. 47, No. 3, pp. 124 – 135.

Liu, H., 2008, "Dependency Distance as a Metric of Language Comprehension Difficulty", *Journal of Cognitive Science*, Vol. 9, No. 2, pp. 159 – 191.

Liu, H., 2010, "Dependency direction as a means of word-order typology: A method based on dependency treebanks", *Lingua*, Vol. 120, No. 6, pp. 1567 – 1578.

Lu, X., 2011, "A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development", *TESOL Quarterly*, Vol. 45, No. 1, pp. 36 – 62.

Nivre, J., 2006, "Inductive dependency parsing", in N. Ide & J. Véronis (eds.), *Text, Speech and Language Technology* Vol. 34, Heidelberg: Springer, pp. 87 – 120.

Ouyang, J. & J. Jiang, 2018, "Can the Probability Distribution of Dependency Distance Measure Language Proficiency of Second Language Learners?", *Journal of Quantitative Linguistics*, Vol. 25, No. 4, pp. 295 – 313.

Rajkumar, R., M. V. Schijndel, M. White & W. Schuler, 2016, "Investigating locality effects and surprisal in written English syntactic choice phenomena", *Cognition*, Vol. 155, pp. 204 – 232.

Roark, B., A. Bachrach, C. Cardenas & C. Pallier, 2009, "Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing", in *Proceedings of the* 2009 *Conference on Empirical Methods in Natural Language Processing*, pp. 324 – 333.

Smith, N. J. & R. Levy, 2013, "The effect of word predictability on reading time is logarithmic", *Cognition*, Vol. 128, No. 3, pp. 302 – 319.

Stefan L. F. & R. Bod, 2011, "Insensitivity of the human sentence-processing system to hierarchical structure", *Psychological Science*, Vol. 22, No. 6, pp. 829 – 834.

van Schijndel, M., A. Exley & W. Schuler, 2013, "A model of language processing as hierarchic sequential prediction", *Topics in Cognitive Science*, Vol. 5, No. 3, pp. 522 – 540.

Warren, T. & E. Gibson, 2002, "The influence of referential processing on sentence complexity", *Cognition*, Vol. 85, No. 1, pp. 79 – 112.

（责任编辑：张　升）

An Analysis of Computational Indicators for Measuring Cognitive Complexity of Syntactic Structure

FANG Yu, LIU Haitao

Abstract: The cognitive complexity of syntactic structure is a research topic shared by linguistics and cognitive science. The most direct way to measure cognitive complexity is to conduct psychological experiments. Recently, in the field of computational cognitive science, indicators have been proposed to measure cognitive complexity quantitatively with the help of mathematical statistics and natural language processing techniques. The present study is intended to introduce three memory-based indicators (storage cost, integration cost and dependency distance) and two experienced-based indicators (surprisal and probabilistic valency). Furthermore, this study explores the possibility of applying these indicators to linguistic research. Nowadays, interdisciplinary interaction and integration have become an important trend of scientific research. Linguistic researchers also need to learn from other disciplines, which could advance understanding of language structures and the patterns of language evolution, and could make linguistic research more scientific.

Keywords: syntactic structure; cognitive complexity; computational cognition; linguistic research; interdisciplinary

About the authors: FANG Yu, PhD in Linguistics, is Assistant Professor at School of Foreign Languages, Tongji University (Shanghai 200092); LIU Haitao is Distinguished Professor at School of International Studies, Zhejiang University (Hangzhou 310058).