歧义结构理解中的依存距离最小化倾向

赵怿怡[1]，刘海涛[2]

（1.厦门大学 人文学院，福建 厦门 361005 2.浙江大学 外语学院，浙江 杭州 310058）

摘要： 本文用依存句法分析汉语歧义结构发现人脑在句法加工时倾向选择最小化依存距离的句法结构。该发现从依存理论角度解释了以往依照短语结构句法分析潜在歧义结构"VP+N1+的+N2"无法说明心理学实验结果的原因，找到了歧义结构实时阅读过程中倾向选择特定句法结构的语言学依据。最小化依存距离的认知机制是降低言语工作记忆成本的有效方法，是言语理解过程中的重要机制之一。

关键词：依存句法；依存距离；言语工作记忆；歧义结构；句法分析

中图分类号：TP391　　**文献标识码：**A

Minimizing Dependency Distance in Understanding of Ambiguous Structure

Abstract: Human beings tend to choose the structure with the minimum Dependency Distance during ambiguous structure understanding in order to reduce the burden on working memory. This paper reanalyzes the psychological experimental results within the framework of dependency grammar. The measurement of dependency distance provides the linguistic criteria for why is the potential ambiguity structure "VP + N1 + the + N2" considered as the modifier-noun construction (MNC) rather than narrative-object structure (NOS). Minimizing dependency distance is an important mechanism during natural language understanding and an effective way to reduce the memory cost.

Key words: dependency grammar； dependency distance; working memory; ambiguous structure; syntactic analysis

1 引言

言语工作记忆在句子理解中的机制和作用是认知心理学研究的热门课题。在众多的言语理解试验中，歧义结构是重要的实验材料，是言语工作记忆中的焦点问题。

从计算语言学角度来看，歧义作为任何语言中普遍存在的现象是自然语言处理中的难点。计算语言学发展的历史就是与歧义做斗争的历史（刘海涛，2009）。冯志伟（1996）提出的"潜在歧义理论"明示了汉语中存在潜在歧义格式，潜在的歧义格式可以产生两种以上的合理解释，消除歧义往往需要上下文来辅助理解。传统语言学和计算语言学试图通过句法规则和上下文约束来限制合理句法结构的生成，以实现计算机对自然语言的理解。

而心理学的研究关注人在言语理解过程中句法结构选择的过程和机制。张亚旭、张厚粲、舒华（2000）从心理学实验角度对潜在歧义格式进行研究，注意到均衡歧义结构的存在，这种结构被分析成歧义结构中的任何一种都是合理的。该文以歧义结构"VP+N1+的+N2"为例（如"关心学校的老师），发现在实时阅读过程中，人们往往按偏正（而非述宾）结构来分析均衡的偏正/述宾歧义短语，而以往的针对花园幽径句（garden-path）的解释原则（最小附加和迟关闭）并不能对这一现象进行很好的解释。该文猜测潜在歧义结构"VP+N1+的+N2"多被分析为偏正结构的分布"很可能是某种机制的结果，而这种机制也是被试实时阅读中按偏正结构分析均衡型歧义短语的原因。"

那么，这种言语理解过程中的机制是什么？它怎样运作？又是否存在合理的可计算的语言学依据呢？

本文从这些问题出发，尝试从语言学角度对已有的心理语言学实验成果进行深入挖掘，探索人在言语理解过程中的认知倾向。论文第二节，我们以依存句法为理论基础进行语言分析，以依存距离为衡量标准提出了“歧义结构理解中存在依存距离最小化倾向”的假设。第三节，我们利用心理语言学已有的实验材料与结果对假设进行验证与深入讨论，证明了在均衡歧义结构理解中人总是倾向选择依存距离较短的句法结构进行分析。结论部分，我们认为这种句法结构的选择是减小言语工作记忆负担的语言学表现，是经济（省力）原则的语言学体现，是言语理解的重要机制之一。

2 方法

依存句法是描述词间关系的句法。句法分析的三个要素是：从属词、支配词和词间关系（刘海涛，2009）。用依存句法分析潜在歧义结构的实例“关心学校的老师”，我们得到两个结构不同的依存图：图 1 的最终支配词是“老师”，表示出该短语被分析成名词为中心词的偏正短语，即潜在歧义结构实例“关心学校的老师”被实现为偏正结构的分析；图 2 的最终支配词是“关心”，表示该短语被分析成以动词为中心词的述宾短语，即潜在歧义结构实例“关心学校的老师”被实现为述宾结构的分析。

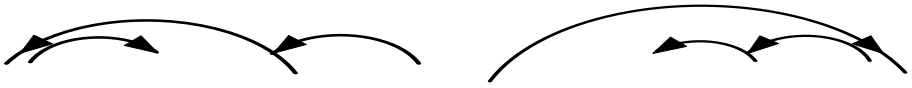


图 1 左：名词为中心词的偏正短语 右：动词为中心词的述宾短语

Lin（1996）用依存句法分析了英语的中心嵌套结构（Center embedding）和外置结构（Extraposition），试图用依存连接的总长度衡量句子结构复杂程度，解释句法变换的目的是降低句子的复杂程度。

Gibson（2000）从人脑计算资源的角度提出依存局部性理论（Dependency Locality Theory，DLT），他认为人类分析句子的过程包含两个资源的利用：结构的整合和结构储存。结构整合是把听到的词整合到已有的句法结构中；结构的储存是把接受的词储存在短期记忆中，以便整合时使用。这个过程也是计算机分析句子的过程。在这个过程中，句子处理的复杂程度和句法依存的长度相关：依存成分距离越长句子越难处理。与基于短语结构的句法理论相比，依存句法更为直接的描述了人脑接受单词并将其整合到已有的句子片段中的过程。用依存的方法分析语言结构，我们可以清楚的表示 DLT 所描述的句子理解的两个过程：短期记忆储存输入词并把输入词整合到已有的句法结构中，实现句子理解。

我们认为经过大量语言现象验证的 DLT 理论对解释人脑或计算机句子理解过程有着普遍性的贡献。那么，影响人类和计算机对潜在歧义结构理解的“某种机制”是否可以从依存成分的距离角度来解释呢？

Temperley（2007）基于DLT理论提出句子处理的复杂程度和句法依存（syntactic dependencies）的长度相关，句法依存越长句子越难理解。他针对宾州树库（Penn Treebank）中的部分语料充分分析了英语书面语中多种类型的语言结构，来验证其“英语书面语依存长度最小化”的观点。

刘海涛（2008）提出“依存距离[1]（Dependency Distance）”可以作为衡量语言理解难度的标准之一。他考察了20种语言，认为人类理解句子中存在最小化平均依存距离的倾向。为了考察在歧义结构理解过程中短期记忆的储存情况，我们使用这一指标来衡量两个歧义结构的的区别。

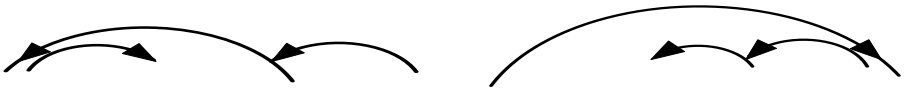
Liu, Hudson and Feng (2009) 提出了依存距离的计算方法。这种方法可计算依存距离的对象小到结构、句子大到样本依存树库。依存距离作为一种线性距离，首先定义词按线性顺序编号“W1...Wi...Wn ”，支配词Wa 和其从属词Wb的依存距离为*a-b*；相邻词对间具有依存

[1] 依存距离指支配词和从属词间的线性距离。

关系，依存距离为1。若*a*>*b*，依存距离大于0，表明支配词的线性顺序在从属词之后；若*a*<*b*，依存距离小于0，表明支配词的线性顺序在从属词之前。在依存距离的相关实验中，研究者往往考察依存距离的绝对值。整个句子（或短语）的平均依存距离计算公式为：

$$MDD(the\ sentence) = \frac{1}{n-1}\sum_{i=1}^{n} | DD_i |$$

公式中，*n*是句子中词的数量；DDi 是第i个依存对间的依存距离。在依存句法分析的句子中，原则上只有一个根节点没有支配词，它的依存距离被定义为0。这个公式可以被用来计算更大的句子集合（例如，树库）的平均依存距离。按照上述方法，我们实例“VP+N1+的+N2”为“关心学校的教师”，计算该短语按不同结构分析时，短语内部的平均依存距离。当短语被分析成偏正结构时，平均依存距离为1；而当短语被分析成述宾结构时，平均依存距离为1.25。



关心 学校 的 老师　　关心 学校 的 老师

2 1 1 0　　0 1 1 3

MDD=4/4　　MDD=5/4

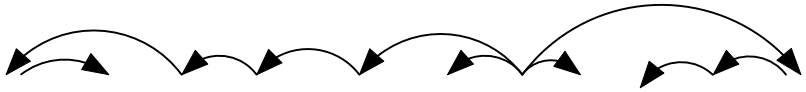
图 2 潜在歧义结构“VP+N1+的+N2”的平均依存距离 左：偏正结构 右：述宾结构

3 结果与讨论

我们认为“潜在歧义结构“VP+N1+的+N2”多被分析为偏正结构的分布”可用上一节中提到的依存距离最小化来解释。潜在歧义结构“VP+N1+的+N2”按照偏正、述宾结构分析得到的结构内部的平均依存距离存在固定的差距。为了进一步证实在包含潜在歧义结构的句子理解过程中同样具备最小化依存距离的倾向，我们收集了 10 组经过心理学测试的句子[2]。这些句子中的歧义结构被心理学实验证实为均衡型歧义结构，即歧义短语两个可能的结构在语义或语用方面的比较是相当的；而这些歧义短语不同结构所对应的意义在日常生活中是典型合理的。相关心理学实验（张亚旭等，2000）已经证明：被试者在理解这些包含均衡歧义结构的句子时，歧义结构部分倾向按照偏正结构来解析；均衡型歧义短语按照述宾结构来解析容易出现加工困难。

我们对这些包含均衡型歧义结构的句子进行了依存句法分析，并在依存句法分析的基础上计算了句子的平均依存距离。以第一组句子为例，我们首先依照依存句法关系对 1a、1b 进行标注，并在计算依存距离时去掉了句末和句中标点，来减少句子非必要成分对依存距离的影响。在(1b)“保护小徐的战马不成，孙刚感到非常内疚”中，前后两分句各自表达完整的意思，在依存句法中两分句的支配词“不成”、“感到”应由承接关系连接，同样为了避免过长的依存距离我们把两分句视为句子单独处理。即“感到”句法上的支配词是上一分局的“不成”，依存距离为 2，剔除分句承接关系的影响后，“孙刚感到非常内疚”单独成句，“感到”成为第二分句的根支配词，依存距离为 0。

(1a)



保护 小徐 的 战马 头部 连续 中 了 三 颗 子弹　　MDD=1.36

2 1 1 1 2 1 0 1 1 1 4

[2] 句子来源：《汉语偏正/述宾歧义短语加工初探》（张亚旭等，2000）所采用的 10 组试验材料。

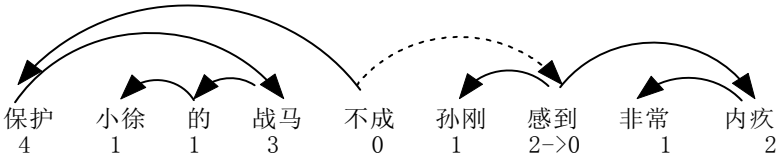


图 3 含均衡型歧义结构句的平均依存距离：(1a)偏正结构 (1b)述宾结构

经过依存句法标注和计算，我们得到了 10 组句子的平均依存距离（表 1）。在这 10 组句子中，除第 2 组外，所有均衡型歧义结构实例为偏正结构句的依存距离均小于实例为述宾结构句。而第 2 组的述宾结构句“抵制美国的政策不成，日本决定恢复谈判”包含 2 个分句，如果不考虑第二分句的依存距离，述宾结构第一分句的平均依存距离为 1.8，超出同组偏正结构句“抵制美国的政策实行半年以来收效显著”平均依存距离值 1.7。

表 1 10 组句子的平均依存距离

MDD	1	2	3	4	5	6	7	8	9	10
(a)	1.36	1.7	1.4	1.5	1.44	1	1.5	1.5	1.3	1.64
(b)	1.44	1.33	1.91	1.9	1.64	1.36	1.78	1.78	1.5	1.91

注：(a)：偏正结构句(b)：述宾结构句。

从语言学角度来看，我们发现在 10 组实验材料中，述宾结构句的表达形式多为两个分句。述宾结构位于第一分句中后接时间指示词“之前”、“之后”做事件型时间状语的比例较多，例“护理丽丽的养父之前”、“接触小陈的医生之后”。而这种语言现象在实际语料库中的数量有限，这说明在现实言语交际中此类语言现象的使用率并不高。这也从语料库的角度证明了：潜在歧义结构“VP+N1+的+N2”实例为述宾结构相对于实例为偏正结构，存在平均依存距离较大，导致句法复杂性增加，容易产生加工困难，不利于理解。

Kimball(1975)在短语结构句法基础上提出的表层句法处理 7 策略来解释复杂句子的生成（这 7 条原则因为翻译的问题常被误读）。其中第二个原则是是：终极符号与最底层的非终极符号结合，被引申为右结合原则；第五个原则是：句法结构尽早关闭，除非下一个结点是该短语的直接成分,被引申为早关闭原则。这两条原则可以很好的解释歧义结构“VP+N1+的+N2”倾向被理解为偏正结构的原因。名词 N1 根据右结合原则被连接到前一结点动词 V 上，根据早关闭原则形成了一个述宾结构的“的”字短语。这样的表层句法分析原则的实质就是尽量减小工作记忆的储存量。Frazier(1987)为解决花园幽径句的句法分析问题，在花园幽径模型（Garden Parth Model）中提出了两个更为著名的句法分析策略，即迟关闭原则和最小附加原则。它们针对性的解决了花园幽径句句法分析常常需要回溯的难题，目的是实现花园幽径句的高效分析。如果用该句法策略来解释歧义结构“VP+N1+的+N2”就不十分有效了。我们注意到，这所有的句法分析策略都是在短语结构语法基础上进行自动句法分析的原则性规定，针对特定问题提出，但观点很不统一。用这些基于短语结构语法的计算机处理特定语言问题的句法规则来解释心理学实验发现的歧义结构理解中的差异并不充分的，没问发现问题的本质。而依存句法体系中依存距离最小化倾向是通过大量实际语料的统计得出的规律（见 Liu 2008），这可能正是人类言语理解的重要机制之一,是均衡型歧义结构“VP+N1+的+N2”倾向于被理解为偏正结构的合理解释。

研究者普遍承认，句子的句法复杂性影响对工作记忆的要求。而关于句法复杂性的探讨往往局限于关系从句等特定的句式、句型，这些语言使用中形成的习惯和语法规则属于传统语言学的范畴，往往缺乏可计算的性质。这直接导致研究者难以判断“一种语言的语言学特性，如何制约言语工作记忆过程在句子理解中的作用机制和性质（张亚旭，蒋晓鸣，黄永静，2007）”。如果我们把依存距离视为语言学特征的可量性指标，运用简单的句法分析就不难发现人脑对于特定句式、句型，特殊句法现象，包括对歧义结构的句法理解、语义选择都朝

着最小化该指标的方向发展。

4 结论

本文从语言学角度对认知科学领域普遍关心的言语理解中的工作记忆进行了探讨，结合心理学的实验结果和材料，证明了人类的言语理解机制与语言学可量性特征——依存距离存在关联，认为人在句法加工时存在最小化依存距离的句子理解倾向。最小化依存距离和人脑的短期工作记忆容量密切相关。本研究为心理学实验检测到的被试者在均衡型歧义结构“VP+N1+的+N2”理解时偏好以偏正的结构结构进行句法处理提供了一种合理的解释，也印证了 Gibson(1998)大脑运算系统中句子剖析理论中“整合成本”的语言学含义。最小化依存距离就是降低整合成本的方式之一。

鉴于实验用语言材料数量和形式的局限，依存距离作为语言的可量特征参与到言语理解机制的运作中仍旧需要专门、专业的心理学实验设计来证明。但本文结合语言学理论和心理学成果的研究方法，对人类言语行为模型和理论的探讨是有益的。

参考文献

[1] 冯志伟. 论歧义结构的潜在性[J]. 中文信息学报,1995,9(4):14-24

[2] Frazier L. Sentence processing: A tutorial review [C]. In: M. Coltheart (Ed.), The psychology of reading. Hillsdale, NJ: Lawrence Erlbaum Associates Inc, 1987:559-586

[3] Lin D.K. On the Structural Complexity of Natural Language Sentences [C]. Proceedings of the 16th conference on Computational linguistics-Volume 2, 1996:729-733

[4] Gibson, E. The dependency locality theory: A distance-based theory of linguistic complexity [C]. In: Marantz A., Miyashita Y. & O’Neil W.(Eds.), Image, Language, Brain. Cambridge, MA: MIT Press, 2000:95－126

[5] Kimball J. Seven Principles of Surface Structure Parsing in Natural Language [J]. Cognition, 1973, 2(1):15-47

[6] Liu H.T. Dependency distance as a metric of language comprehension difficulty [J]. Journal of Cognitive Science, 2008, 9(2):159-191

[7] Liu H.T., Hudson R. and Feng Z.W. Using a Chinese Treebank to measure dependency distance [J]. Corpus Linguistics and Linguistic Theory, 2009, 5(2):161-174

[8] 刘海涛.. 依存语法的理论与实践. 北京：科学出版社, 2009.

[9] Temperley D. Minimization of dependency length in written English [J]. Cognition, 2007, 105:300－333

[10] 张亚旭, 张厚粲, 舒华. 汉语偏正/述宾歧义短语加工初探[J]. 心理学报, 2000, 32(1):13-19

[11] 张亚旭, 蒋晓鸣, 黄永静. 言语工作记忆、句子理解与句法依存关系加工[J]. 心理科学进展, 2007, 15(1):22-28