从广义话题结构考察汉语篇章话题认知复杂度*

卢达威[1]　宋柔[1]　尚英[2]

1.北京语言大学　语言信息处理研究所，北京，100083

2.北京语言大学　预科教育学院，北京，100083

摘要：语言理解问题从认知的角度已有大量的研究，但针对汉语的研究却很少。由于认知实验操作复杂，不容易大规模复制，因此难以量化其结论的普遍性以及对语言事实的覆盖度。本文尝试模拟人补足汉语篇章片段中话题-说明信息的过程，建立广义话题结构认知机模型，并通过认知机对大规模汉语语料进行定量分析，考察汉语标点句的话题认知所需的记忆资源及认知局限性。用作统计特征量的广义话题结构特征有标点句的深度、话题结构内折返度、话题栈深度、话题栈折返度、搁置区使用量。统计数据可从认知行为的视角得到合理解释。本文一方面揭示了说汉语者的话题认知能力的表现和局限性，另一方面又说明了广义话题结构认知机是话题认知的合理模型。

关键词：广义话题结构；认知机；认知复杂度；标点句；话题自足句；汉语篇章；

The Cognitive Complexity of Topic in Chinese Text Based on Generalized Topic Structure Theory

LU Dawei[1], SONG Rou[1], SHANG Ying[2]

1. Center of Language Information Processing, Beijing Language and Culture University, Beijing, 100083, China

2. College of Preparatory Education, Beijing Language and Culture University, Beijing, 100083, China

Abstract: There have been a lot of researches on language understanding from cognitive perspective, but few of them are about Chinese language. Since cognitive experiments are too complicated to replicate on a large scale, it is difficult to quantify their generalizability and the degree of coverage in all language facts. We constructed a *Generalized Topic Structure Cognitron (GTSC)* by simulating human's cognitive process on complementing the topic-comment information of Chinese *Punctuation Clauses (P-clause)*. With quantitative analysis of large-scale Chinese texts by the way of GTSC, we studied the required human memory resources and the cognitive limitation in P-clause understanding. The features adopted in *Generalized Topic Structure* analysis are depth of P-clause, returning degree within topic structure, depth of topic stack, returning degree of topic stack and the number of lay-down area. The statistical results of *Generalized Topic Structure* produced by GTSC can be explained reasonably from cognitive perspective. On one aspect, this paper reveals the cognitive ability and limitation of Chinese people in topic-comment information processing. On the other aspect, it proves that the GTSC is a reasonable model for cognitive processing of topics in Chinese language.

Keywords: Generalized Topic Structure; cognitive machine; cognitive complexity; Punctuation Clause; Topic Sufficient Sentence; Chinese text

0 引言

从上世纪中叶认知革命的兴起开始，语言科学领域已经累积了大量从认知的角度来探究语言理解问题的研究。有的从记忆的机制和过程来研究语言理解的认知机制，如 Baddeley[1]、MacDonald[2]、Just & Carpenter[3]、Traxler[4] 等研究了工作记忆对语句理解的影响；Kintsch[5]、McKoon[6]、Bransford[7] 等则从长时记忆的角度研究篇章的语言理解机制。有的把语言理解过程看作是表层结构到深层结构的信息加工过程，分为语音、词汇、句法、语义等加工阶段，并考察这些不同层次的信息何时及如何被加工，以及这些信息之间如何交互影响语言理解的过程（如 Cairns & Cairns[8]，Forster[9]，Lindsay & Norman[10]，Marslen -Wilson & Tyler[11] 等）。

* 本文得到国家自然科学基金（61171129）的资助。

有的从认知复杂性的角度，提出了可计算的概念学习布尔复杂度（如 Feldman[12]）。这些研究通常基于一定的假设，通过认知实验的手段获得支持或否定这一假设的证据从而得出结论。随着科学技术的发展，实验的手段和技术已越来越丰富和先进，如近年来流行的神经电生理学技术（如 Event-related potential，ERP，事件相关电位）和脑成像技术（如 functional Magnetic Resonance Imaging，fMRI，功能核磁共振）等，以及与之相适的实验范式和数据分析方法的发展，都为研究得出科学可信的结论提供了保障。

然而，这些研究得出的有关语言理解的结论还难以检验其普遍性。一者，实验的被试数量有限，难以涵盖所有人的特征。二者，实验的语料更为有限，通常只是精选几句到几十句的人造语言材料作为实验材料，不可能覆盖所有的语言现象。三者，某一假设的适用性难以量化。由于缺少对研究对象全体的把握（包括人和语料），也缺少简单、可操作的形式化表达，难以证明一个假设是否普遍适用，或者明确区分适用和不适用的情况并计算出其适用度。此外，现有研究大多是针对英语的研究，专门针对汉语的研究非常缺乏。

本文尝试通过与一般认知实验不同的方法对汉语篇章的认知复杂度进行研究。首先，基于认知的客观事实以及对认知过程的模拟，以广义话题结构理论为基础，构造认知模型——广义话题结构认知机（以下简称“认知机”）。第二，通过统计和分析认知机在处理大规模语料过程时的资源消耗，归纳出汉语使用者对于汉语篇章话题结构的认知规律。

汉语篇章理解是最终的目标，而本文所提出的认知机的任务仅是补足标点句的话题-说明信息，篇章理解还需在此基础上，完成指代消解、逻辑结构分析、宏观话题分析等工作。本文所研究的认知复杂度，仅指补足标点句的话题-说明信息所动用的计算资源。

认知机对人的认知模拟的有效性建立在以下 2 个假设上：

1）若某一语言特征在语料中出现频率低，则人对该特征认知复杂度高。

2）若处理某一语言特征认知机调用资源多，则机器对该特征处理复杂度高。

假设 2）的合理性是显而易见的。假设 1）也是有道理的。熟能生巧是人所共知的学习规律，重复对于学习的重要性是认知心理学已经认定的。在话语认知过程中，某种语言现象出现频率高，意味着认知者会多次重复对这种现象的认知过程，其结果是降低了再次认知该现象所需的代价。反之，低频现象的认知未经过多次重复，其每次认知的代价就会高。基于以上假设，若统计数据表明假设 1）和 2）的前提高度正相关，则可推得认知机处理复杂度和人的认知复杂度具有相似分布，认知机就能够对人的认知行为有效模拟。

1. 广义话题结构基本概念

广义话题结构 [13] 是认知机的理论基础。广义话题结构揭示了汉语篇章微观话题层面的组织形式，是汉语篇章的结构单位。其理论的高覆盖性和可操作性在大量的语料标注中得到了证实 [14]。

标点句是逗号、分号、句号、叹号、问号、直接引语的引号以及这种引号前的冒号所分隔出的词语串，是广义话题结构处理的基本单位，也是本文所研究的认知机处理的基本单位。

例 1

突然， 他听到洗手间有流水声， 警官与特警踢开门， 将洗手间内的人猛地摔倒在地并铐住， 经辨认， 正是叶成坚。

例 1 是新闻语料中的一段话，共 6 个标点句。就每个标点句看，均代表了一定的意义，但除了第 2 和第 3 句，其他都不是完整的句子。下文中为了俭省，有时也把标点句称为句子。

例 2

广义话题结构流水模型（堆栈模型）	话题自足句
高松年发奋办公， 　　夙夜匪懈， 　　精明得真是睡觉还睁着眼睛， 　　　　戴着眼镜， 　　　　做梦都不含糊的。 　　摇篮也挑选得很好， 　　　　在平成县乡下一个本地财主家的花园里， 　　　　面溪背山。	高松年发奋办公， 高松年夙夜匪懈， 高松年精明得真是睡觉还睁着眼睛， 高松年精明得真是睡觉还戴着眼镜， 高松年精明得真是做梦都不含糊的。 高松年摇篮也挑选得很好， 高松年摇篮在平成县乡下一个本地财主家的花园里， 高松年摇篮面溪背山。

经过大量语料标注发现，对汉语而言，大多数的成分缺失位于标点句首，通过补全句首成分可以使标点句成句。从语义上看，这些句首缺失成分大体上属于传统语言学中所说的话题，如事物话题和环境话题（时间、处所），不属于传统话题的部分可归纳为状性话题、谓性话题和推理前提，故统称为**广义话题**。

广义话题结构**流水模型**（以下简称“流水模型”）是广义话题结构在汉语文本上的形式化体现。标点句通过换行缩进形式排列后，用线条相连，形似流水，互不交叉穿越，故称为流水模型。本文第 3 部分将结合认知机的设计对流水模型的各子模型逐一介绍。

标点句补充缺失的广义话题和说明后，通常能够形成语法通顺、语义清楚的句子，称为**话题自足句**。例 2 是《雪山飞狐》的一段话，左边原文按标点句换行并按话题缩进排列，右边是每个标点句补充广义话题后对应生成的话题自足句。

2. 认知机的设计原则

广义话题结构认知机是从机器角度模拟人理解汉语篇章话题-说明结构的计算模型。设计认知机时，我们尽量模拟大脑对汉语篇章的认知过程，考虑了以下设计思想和原则：

- **输入**。说-听的篇章给予人的输入是具有一定时长的停顿所隔开的片段。这样的停顿，从认知的角度看，既给予听者认知加工的时间，也给予说话人组织表达形式的时间，反映在汉语文本上，就是标点符号，其隔开的片段就是标点句。因此，认知机使用标点句作为系统输入的基本单位。

- **输出**。人对篇章的彻底理解涉及话题说明关系、逻辑关系、指代关系、宏观主题等。理解后大脑中的组织形式和内容目前不得而知，但若篇章片段话题-说明信息不完整，分析指代关系、组成逻辑关系树或提取宏观主题等都将有困难，因此，我们猜测，补全话题-说明信息的篇章片段是篇章理解的基础。对于认知机来说，标点句补全话题-说明信息后的结果就是话题自足句，故认知机以话题自足句作为系统的输出。事实上，补充完整信息后的篇章片段在大脑中形式上不一定是整句，可能只是某些概念单位。但是从信息结构角度讲，这种概念单位和话题自足句之间应该存在紧密的对应关系。

对每一个标点句的处理过程，认知机在设计上遵循以下原则：

- **无长距离回溯**。人对于篇章的认知过程，如果排除逻辑关系等高层语义的理解，应该是无长距离回溯的，如听小说广播。即使在对话时，除非听不清发音或无法消解歧义，极少要求对方重复说话内容。因此，认知机必须做到对标点句输入只能读取一次，且仅能顺序读取。

- **同步处理**。人对篇章的简单层面的理解几乎伴随输入同步完成。即说话者逐句说，听话者逐句理解，当对方的话说完后听者基本能够理解完毕。因此，认知机对处理性能有较高的要求。若把同步看作一个时间概念，输入和输出直接的时间差应该很短，且新的输入不会干扰理解。

- **词序不变**。汉语形式标记不丰富，特别是口头表达时（包括对话、朗诵、广播等），除了停顿和语调之外基本没有别的形式标记。因此，汉语特别注重词序，词序通常反映的是

一种意义结构。有些计算语言学的模型对词序是不敏感的，但在认知机的处理过程中，不仅要保留了篇章片段序列之间的顺序，还要保证内部词序不变。

- **有限存储空间**。人的长时记忆的空间相当大，但短时记忆空间组块存在 7±2 的限制[15]，工作记忆空间非常有限。这决定了认知机所需空间必须为一个不随篇章长度变化的常数定值，且该定值必须严格限制大小。
- **分步处理**。我们设想，人补充篇章片段话题-说明信息的过程分解为两步。第一步是分析。当遇到停顿，即听完一个篇章片段时，首先判断：1）该片断是否缺话题；2）如果缺话题，所缺话题在前面还是在后面；如果在前面是哪一个成分；3）该片断是否缺说明。第二步是生成，补足标点句缺失的话题和说明。因此，认知机的设计分为两部分：第一部分是分析系统，第二部分是生成系统。分析系统负责对标点句序列进行流水模型标注，包括标点句缩进和添加流水模型的其他相关标记；生成系统负责生成对应的话题自足句序列。

分析系统较为复杂，涉及语法、语义、语用等语言学知识，甚至百科知识，目前还没达到自动化程度。但是可以说明，分析系统所使用的上下文知识，通常只限于两个话题自足句。也就是说，在已知前两句分析结果的条件下，通常可以直接分析出当前标点句应加的广义话题结构标记。关于这个问题，将有另文讨论。本文只讨论生成系统部分。本文以人工标注的广义话题结构流水模型汉语语料，代替分析系统对标点句的分析结果，作为认知机生成系统输入。通过认知机生成系统在大规模语料中表现出的特性来考察人理解标点句的认知复杂性。

根据以上原则，认知机采用递推的方式运行，每一步递推读入一个标点句，多数情况下每步输出一个话题自足句，但对于话题后置的标点句，需要待后置的话题出现后输出话题自足句；对于汇流语段内的标点句，需要待聚合部分出现后输出话题自足句。以下结合流水模型的各子模型，阐述在以上设计原则下，认知机生成系统对不同子模型的实现方法。

3. 广义话题结构流水模型及认知机生成系统实现

流水模型以堆栈模型为基础。为覆盖更广泛的语言现象，扩展出节栈模型、后置模型、汇流模型、封闭语段等子模型。

3.1. 堆栈模型

堆栈模型处理标点句仅使用换行缩进排列，无需使用其他辅助标记（如例 2），是流水模型的基础模型。生成话题自足句时，需要两个参数，一是本标点句，二是上一个标点句的话题自足句。把本标点句缩进的部分用上一句的话题自足句补充，就形成了新的话题自足句（如例 2-1，例 2-2）。

例 2-1：例 2 的第 4 句生成话题自足句图示：

Φ：高松年精明得真是睡觉还睁着眼睛，（第 3 句的话题自足句）
Ψ：　　戴着眼镜，（第 4 个标点句）

↓

Φ：高松年精明得真是睡觉还戴着眼镜，（第 4 个句的话题自足句）

例 2-2：例 2 的第 5 句生成话题自足句图示：

Φ：高松年精明得真是睡觉还戴着眼镜，（第 4 句的话题自足句）
Ψ：　　做梦都不含糊的。（第 5 个标点句）

↓

Φ：高松年精明得真是做梦都不含糊的。（第 5 句的话题自足句）

例 2-1 和例 2-2 分别展示了例 2 的第 4、第 5 个标点句生成话题自足句的过程。认知机用 Ψ 表示本标点句，用 Φ 表示上一句的话题自足句，生成结果放在 Φ，作为下一句生成话题自足句的参数。从话题自足句生成的角度看，Φ 就像一个堆栈，退掉原句右边的一部分，再接上后一个标点句，因此称为堆栈模型。广义话题结构理论在通过大规模语料的实践证明，原 Φ 中清除掉的内容不会再被后续标点句使用。将新的话题自足句放在 Φ 中，能够实现认

知机生成系统的递推处理。

堆栈模型仅使用 2 个话题自足句空间，其递推机制实现了无回溯原则；对标点句整句存储实现了词序不变。由于进栈出栈操作简单，故在认知机生成系统层面上输入输出是同步的。

3.2. 节栈模型

例 3

广义话题结构流水模型（节栈模型）	话题自足句
顾炎武在城中买了一份邸报， ‖ 上面详列明史一案中获罪诸人的姓名。 却见上谕中有一句说:	顾炎武在城中买了一份邸报， 一份邸报上面详列明史一案中获罪诸人的姓名。 顾炎武却见上谕中有一句说:

例 3 的第 2 句生成话题句时和一般堆栈模型不同，缺失的话题并不是缩进的全部，只是“一份邸报”，因此在前面加一道“节”，称为节栈模型。我们用“‖”表示节的位置，节左边的部分在生成话题自足句时并不输出，认知机生成系统用一个专门的缓存区——话题栈 Π 临时保存（例 3-1）。生成第 3 个话题自足句时需要从 Π 中取出暂存的话题（例 3-2）。堆栈模型成为了节栈模型中话题栈 Π 为空时的特例。

例 3-1：例 3 的第 2 句生成话题自足句图示：

Φ: 顾炎武在城中买了一份邸报，
Ψ: ‖ 上面详列明史一案中获罪诸人的姓名。
Π: ∅

↓

Φ: [顾炎武在城中买了]一份邸报上面详列明史一案中获罪诸人的姓名。
Π: [顾炎武在城中买了]

例 3-2：例 3 的第 3 句生成话题自足句图示：

Φ: [顾炎武在城中买了]一份邸报上面详列明史一案中获罪诸人的姓名。
Ψ: 却见上谕中有一句说:
Π: [顾炎武在城中买了]

↓

Φ: 顾炎武却见上谕中有一句说:
Π: ∅

例 3 的第 3 句使用了话题栈 Π 的内容，并清空了话题栈。

3.3. 封闭语段

例 4

广义话题结构流水模型（封闭语段）	话题自足句
魏队长又说: “天塌下来找魏天贵替你撑着， ‖ 顶大不当这个骨泉队长。 这条狗嘛， 你就宰了算了， ……”	魏队长又说: “天塌下来找魏天贵替你撑着， 魏天贵顶大不当这个骨泉队长。 “这条狗嘛， “这条狗你就宰了算了， ……

直接引语之内的标点句生成话题自足句时不需要共享直接引语外的成分，故称为封闭语段。这些封闭语段被直接引语的引号括了起来。有些标点句由“心想”、“认为”等引出，虽未使用引号括起来，但功能上相当于直接引语，也看作封闭语段的内容，语料中用“【…】”标注起始和结尾位置。认知机处理封闭语段内的标点句时，封闭语段外的成分暂保存于话题栈 Π 中，待封闭语段结束后有可能被当作话题恢复出来。例 4 中，从第 2 标点句开始进入

封闭语段，话题栈 Π 要进一层；第 3 标点句既处于封闭语段内，又另成节栈，话题栈 Π 又进一层。话题自足句生成过程如例 4-1 所示。

例 4-1：例 4 的第 3 句生成话题自足句图示：

Φ: [魏队长又说]“天塌下来找魏天贵替你撑着，
Ψ: ‖ 顶大不当这个骨泉队长。
Π: [魏队长又说]

↓

Φ: [魏队长又说][“天塌下来找]魏天贵顶大不当这个骨泉队长。
Π: [魏队长又说][“天塌下来找]

例 4-1 中的话题栈层次为 2。当处理第 4 句时，根据标点句的缩进所示，需要要退掉 Π 的最外层（如例 4-2）。

例 4-2：例 4 的第 4 句生成话题自足句图示：

Φ: [魏队长又说][“天塌下来找]魏天贵顶大不当这个骨泉队长。
Ψ: 这条狗嘛，
Π: [魏队长又说][“天塌下来找]

↓

Φ: [魏队长又说]“这条狗嘛，
Π: [魏队长又说]

3.4. 后置模型

例 5

广义话题结构流水模型（后置模型）	话题自足句
╠到了镇上， ╠投了村店， ╠开发了车夫， 四个人 脱下鞋子来，	四个人到了镇上， 四个人投了村店， 四个人开发了车夫， 四个人脱下鞋子来，

例 5 前 3 句缺失的话题并不在上一个话题自足句中，而在第 4 个标点句中。处理这种现象的模型称为逆向堆栈模型，也称为话题后置模型，简称后置模型。话题后置的标点句用“╠”标记。由于认知机生成系统生成话题自足句时，每次仅读入一个标点句，不能回读，且仅有上一句的话题自足句被保留在 Φ，故前 3 句需要暂时被搁置起来，等待生成系统读入第 4 句后，再把后置话题补充完整。因此，认知机生成系统需要有一个缓冲区搁置这些未完成的话题自足句，这个缓冲区是未完成话题自足句的队列，这里称为搁置区，记作 Σ。当生成系统读入第 4 句，发现前 3 句的话题后再到 Σ 中把话题补充完整，然后输出 Σ 中的所有完成了的话题自足句，并把它们从 Σ 中移除。例 5-1 展示了 Φ 中的第 4 句补充 Σ 中的后置话题的情况，其中 Out 代表话题自足句输出。

例 5-1：例 5 后置话题补充图示：

Φ: ╠到了镇上，
Ψ: ╠投了村店，
Σ: ∅

↓

Φ: ╠投了村店，
Ψ: ╠开发了车夫，
Σ: ╠到了镇上，

↓

Φ：　　╠开发了车夫，
Ψ：四个人脱下鞋子来，
Σ：　　╠到了镇上，
　　╠投了村店，

Φ：四个人脱下鞋子来，
Ψ：……
Σ：　　╠到了镇上，
　　╠投了村店，
　　╠开发了车夫，

Σ：四个人到了镇上，
　四个人投了村店，
　四个人开发了车夫，
　四个人脱下鞋子来，

Out：四个人到了镇上，
　　四个人投了村店，
　　四个人开发了车夫，
　　四个人脱下鞋子来，
Σ：∅

3.5. 汇流模型

例 6

广义话题结构流水模型（汇流模型）
我们深切怀念『为中国革命、建设、改革， 　　为中国共产党建立、巩固、发展』作出重大贡献的老一辈无产阶级革命家，
话题自足句
我们深切怀念为中国革命、建设、改革作出重大贡献的老一辈无产阶级革命家， 我们深切怀念为中国共产党建立、巩固、发展作出重大贡献的老一辈无产阶级革命家，

例 6 的第 1 句不是缺话题，而是说明部分不完整，不完整的部分用『』标记括起来，成为汇流语段，其中每一行的尾部都缺失说明。认知机处理时，将其搁置在未完成话题自足句队列 Σ 中，等待后续标点句把说明补充完整后输出，并从 Σ 中移除。过程如例 6-1 所示。

例 6-1：例 6 后置话题补充图示：

Φ：我们深切怀念『为中国革命、建设、改革，
Ψ：　　为中国共产党建立、巩固、发展』作出重大贡献的老一辈无产阶级革命家
Σ：∅

Φ：我们深切怀念『为中国共产党建立、巩固、发展』作出重大贡献的老一辈无产阶级革命家
Ψ：……
Σ：我们深切怀念『为中国革命、建设、改革，

Σ：我们深切怀念为中国革命、建设、改革，
　我们深切怀念为中国共产党建立、巩固、发展作出重大贡献的老一辈无产阶级革命家，

Σ：我们深切怀念为中国革命、建设、改革作出重大贡献的老一辈无产阶级革命家，
我们深切怀念为中国共产党建立、巩固、发展作出重大贡献的老一辈无产阶级革命家，

Out：我们深切怀念为中国革命、建设、改革作出重大贡献的老一辈无产阶级革命家，
我们深切怀念为中国共产党建立、巩固、发展作出重大贡献的老一辈无产阶级革命家，
Σ：∅

4. 特征统计分析

我们在实验中使用认知机生成系统处理了 30963 个标点句约 38 万字的带广义话题结构标记的语料。通过对机器处理实际语料过程的分析，推测人对于话题的认知规律。为了保证结论的一般性，语料包含三种不同语体类型：小说、百科释文（以下简称“百科”）和政府工作报告（以下简称“报告”）。其中，小说中包括普通当代小说、现代章回小说和古代白话小说；百科包括生物、地理、历史事件和人物 4 种题材。各项统计均以标点句为单位，统计每个标点句生成话题自足句时动用的存储资源，来模拟人对标点句话题信息的认知复杂性。

4.1. 单项特征的统计分析

4. 1. 1. 标点句深度和标点句字数深度

设有标点句序列{c_1，…，c_n}，c_m（1≤m≤n）的话题自足句是 s_m，s_m 中在 c_m 左边有 k 个话题串分别被 c_m 等 k 个标点句说明，则称 c_m 的深度为 k（如图 1）。封闭语段内，标点句深度从左括号算起深度为 0，节栈模型中最右节的话题所在的标点句深度为 0。

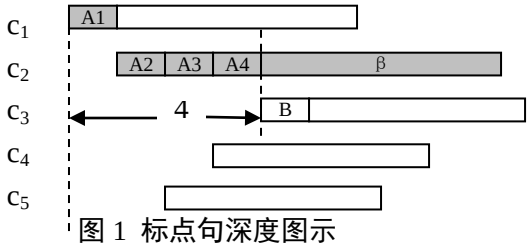

图 1 标点句深度图示

图 1 中，c_3 的话题自足句 s_3 为 $A_1A_2A_3A_4c_3$，c_3 最直接的话题串是 A_4，A_4 前的话题串 A_3 中的话题被 c_4 说明，A_2 中的话题被 c_5 说明，A_1 中的话题被 c_2 说明，所以 c_3 的深度为 4。c_1 的句首没有缺失话题，所以 c_1 深度为 0。

例 7 标点句深度

c_1 高松年发奋办公，（0）
c_2 夙夜匪懈，（1）
c_3 精明得真是睡觉还睁着眼睛，（1）
c_4 戴着眼镜，（3）
c_5 做梦都不含糊的。（2）
c_6 摇篮也挑选得很好，（1）
c_7 在平成县乡下一个本地财主家的花园里，（2）
c_8 面溪背山。（2）
c_9 这乡镇绝非战略上必争之地，（0）

例 7 中标点句后括号中的数字为该标点句的深度。语料库中不同深度的标点句分布如表 1 所示。

深度	0	1	2	3	4	5	总计
标点句数	12911	13457	4006	542	45	2	30963
占标点句总数的比例(%)	41.70	43.46	12.94	1.75	0.15	0.01	100
累计比(%)	41.70	85.16	98.10	99.85	99.99	100	

表 1 标点句深度分布

表 1 显示，标点句本身话题自足的（深度为 0）占所有标点句的 41.7%，即另外的 58.3%缺少话题（深度大于 0），可见话题缺省是汉语标点句的常态。而话题缺省的 18052 句中，13457 句深度为 1，占话题缺省的 74.5%，可见话题缺省中大部分仅围绕最外层话题展开。

另外，深度越大，标点句数量越少，平均深度为 0.75，且最大深度不超过 5。从认知上看，深度越大，需要被记住的话题越多，越难被说出来或被理解。

4.1.2. 标点句话题结构内折返度

设有 3 个标点句 c_1、c_2 和 c_3 在篇章中前后排列，并且 c_2 和 c_3 紧邻。c_1 的句首没有成分缺失，深度为 0。如果 c_2 的深度为 d，c_3 的深度为 f，并且 $0<f<d$，则称 c_3 相对于 c_2 发生了话题结构内的折返，c_3 是折返句，其话题结构内折返度为 d-f（如图 2）。要求 $f>0$ 就是要求在 c_3 在 c_2 的话题结构内，而不是重新开始一个话题结构。不引起混淆的情况下，话题结构内折返度简称为折返度。

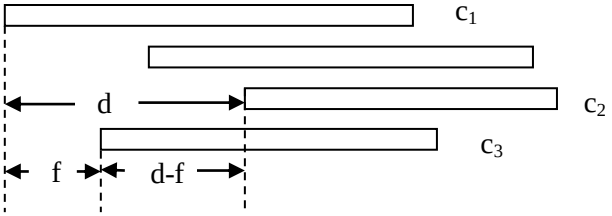

图 2 标点句话题结构内折返度图示

例 7’ 标点句折返度

广义话题结构换行缩进图式	深度	折返度
c_1 高松年发奋办公，	0	/
c_2 夙夜匪懈，	1	/
c_3 精明得真是睡觉还睁着眼睛，	1	/
c_4 戴着眼镜，	3	/
c_5 做梦都不含糊的。	2	1
c_6 摇篮也挑选得很好，	1	1
c_7 在平成县乡下一个本地财主家的花园里，	2	/
c_8 面溪背山。	2	/
c_9 这乡镇绝非战略上必争之地，	0	/

例 7’中，只有 c_5 和 c_6 存在折返。在语料统计中，折返度分布如表 2。

折返度	1	2	3	总计
折返句数	984	75	8	1067
占折返句总数的比例(%)	92.22	7.03	0.75	100
累计比(%)	92.22	99.25	100	

表 2 标点句话题结构内折返度分布

表 2 显示，所有发生话题结构内折返的标点句只有 1067 句，且最大折返度不超过 3。从表 1 的标点句深度分布得知，所有深度在 2 或以上的标点句，即可能发生折返的标点句共有 4595 句。就是说，实际发生折返的标点句 1067 句只占可折返标点句的 23.2%，占所有标点句的 3.4%，平均折返度为 1.09，可见标点句要发生折返还是有一定的困难。结合认知机，要发生话题结构内折返，相当于提取前一话题自足句 Φ 中靠前的词语串作为话题，从认知角度看，这对记忆时间有更高的要求，标点句折返有一定的认知难度。

4.1.3. 标点句话题栈深度

从认知机生成系统的话题栈 Π 的用法可知，标点句话题栈深度指的是标点句位于多少层嵌套的封闭语段或节栈模型的栈节内。

例 8

c_1 他费了许多唇舌，[0]
c_2 本想庄允城在一部明史之外，[0]
c_3 另有几百两银子相赠，[0]
c_4 ‖ 可是赠送的是他信口胡诌的“湖州三宝”，[1]
c_5 心下暗骂：[0]
c_6 “……，[1]
c_7 倘若我说湖州三宝乃是金子银子和明史，[1]
c_8 ‖ 岂不是大有所获？”[2]
c_9 气愤愤地回到客店，[0]

例 8 中每个标点句后方框内的数字表示话题栈深度。c_1~c_3 的话题栈深度都是 0，c_4 是节栈模型处理的标点句，话题栈深度是 1。c_5 退出节栈，话题栈深度为 0，并引出封闭语段（c_6~c_8），话题栈深度至少是 1。c_8 是封闭语段内的节栈模型处理的标点句，话题栈深度加 1，达到 2。c_9 分别退出前两层话题栈，话题栈深度为 0。在语料统计中，话题栈深度分布情况如表 3 所示。

话题栈深度	0	1	2	3	4	总计
标点句数	24151	6097	684	29	2	30963
占标点句总数的比例(%)	78.00	19.69	2.21	0.09	0.01	100
累计比(%)	78.00	97.69	99.90	99.99	100	

表 3 标点句话题栈深度分布

表 3 可见，标点句话题栈深度为 0 是占优势的，达 78%，若只使用 1 层话题栈，就能够覆盖 97.69%的语料。可见，在认知机生成系统中，话题栈 Π 并非任何时候都必不可少的部件，只在处理较为复杂的情况下需要调用。话题栈深度平均为 0.24，最大深度不超过 4，在认知上，话题栈深度过深，将难以理解。

4.1.4. 话题栈折返度

例 8 中 c_5 和 c_9 的话题栈深度都是 0，但他们的上句 c_4 和 c_8 话题栈深度均不为 0，且 c_5 和 c_9 本句的标点句深度也不为 0。我们称 c_5 和 c_9 发生了话题栈折返。其中，c_5 退出了 1 层话题栈（c_4 的话题栈深度是 1），话题栈折返度为 1；c_9 退出了 2 层话题栈（c_8 的话题栈深度是 2），话题栈折返度为 2。在语料统计中，话题栈折返度的分布情况如表 4 所示。

话题栈折返度	1	2	总计
话题栈折返句数	267	20	287
占所有话题栈折返句数的比例(%)	92.42	7.58	100.00
累计比(%)	92.42	100.00	

表 4 话题栈折返度分布

从表 4 可知，发生话题栈折返的标点句仅有 287 句。从表 3 得知，所有话题栈深度大于 0 的标点句，即可能发生话题栈折返的标点句，共有 1703 句，实际折返 287 句，只占 16.8%，小于话题结构内部折返发生概率 23.2%，占所有标点句的 0.9%，平均话题栈折返度为 1.07，可见，话题栈折返极为困难。

4.1.5. 搁置区 Σ 使用量

搁置区 Σ 使用量是一个动态的概念，指处理当前标点句时，Σ 中已搁置的未完成话题自足句数。后置模型、汇流模型使用 Σ。后置模型使用 Σ 搁置的是待补后置话题的标点句，汇流模型使用 Σ 搁置的是待补说明尾部的标点句。这些标点句对于认知机来讲是被搁置在 Σ 中，对于人来讲可看作保存在短时记忆中。因此，Σ 使用量的指标也反映人在认知复杂模型时记忆的能力。在语料统计中，标点句对于 Σ 使用量的分布如表 5 所示。

Σ使用量	0	1	2	3	4	5	6	7	8	总计
标点句数	27379	3022	419	95	31	12	3	1	1	30963
占标点句总数的比例(%)	88.42	9.76	1.35	0.31	0.10	0.04	0.01	0.00	0.00	100
累计比(%)	88.42	98.18	99.54	99.84	99.95	99.98	99.99	100	100	

表 5 搁置区 Σ 使用量分布

表 5 显示，88.42%的情况下都不需要使用 Σ，需要使用 Σ 的情况只占 11.58%。可见，在认知机生成系统中，Σ 不是任何时候都必要的部件，仅当汇流和后置模型中的需要搁置标点句等待后续补全信息的时候使用。如果使用 Σ 中 1 个未完成话题自足句空间，则能够覆盖 98.18%的语料，使用 2 个未完成话题自足句空间，则能够覆盖 99.54%的语料。Σ 平均使用量为 0.14，最大使用量不超过 8，Σ 使用量为 5 及以上的标点句不足 20 句，在 3 万多个标点句中，已显得极为偶然。这体现了 Σ 所表现的认知上的复杂性。

4.2. 特征交叉的统计分析

4.2.1. 话题栈深度和话题结构内标点句深度

话题栈深度和标点句深度存在层级关系，标点句深度是在同一话题栈深度下计算的，二者关系如表 6 所示。

话题栈深度	标点句深度	句数	汇总	话题栈深度	标点句深度	句数	汇总
0	0	9336	23968	2	0	276	691
	1	10831			1	338	
	2	3312			2	72	
	3	448			3	5	
	4	39		3	0	15	30
	5	2			1	11	
1	0	3284	6271		2	4	
	1	2274		4	1	3	3
	2	618		总计		30963	
	3	89					
	4	6					

表 6 话题栈深度和话题结构内标点句深度分析

从表 6 看出，话题栈深度为 0 时，最大标点句深度是 5；话题栈深度是 4 时，最大标点句深度是 1。话题栈深度和标点句深度之和均不超过 5。若把话题栈深度和标点句深度相加，称为标点句总深度，则其分布情况如表 7 所示。总深度为 1 的情况最多，99%以上的标点句总深度不超过 3。

标点句总深度	0	1	2	3	4	5	总计
标点句数	9336	14115	5862	1419	211	20	30963
占标点句总数的比例(%)	30.15	45.59	18.93	4.58	0.68	0.06	100
累计比(%)	30.15	75.74	94.67	99.25	99.94	100	

表 7 标点句总深度分布

4.2.2. 标点句深度和话题结构内标点句折返度

将标点句深度和折返度作为两个维度考察，标点句数目如表 8 所示。这里的深度和折返度不涉及跨话题栈的情况。

深度 \ 折返度	1	2	3	总计
2	900	/	/	900

深度 \ 折返度	1	2	3	总计
3	75	71	/	146
4	9	4	8	21
总计	984	75	8	1067

表 8 标点句深度和话题结构内折返度统计

由折返度定义可知，发生折返的标点句，其话题一定取自于上一标点句之前的标点句，折返度越大，则话题来自越早的标点句。从记忆遗忘的角度，较早标点句的话题遗忘率应该更高，即同一深度折返度大的句数应该比折返度小的少。但从表 8 中深度为 3 和 4 的两行看出，对于同一深度而言，不同折返度的分布相对均匀。这一数据似乎与认知的直觉违背，但是认知机模型可以给予解释。虽然折返所涉及的话题来自于不同的标点句，但都存储在Φ，即上一个话题自足句中。每一个标点句的处理，都相当于对话题自足句的话题复述了一次，因此来自不同标点句的话题记忆程度没有差别。这也反证了认知机模型的合理性。

4.2.3. 话题栈深度和话题栈折返度统计

表 9 将话题栈深度和话题栈折返度进行交叉分析。

话题栈深度 \ 话题栈折返度	1	2	总计
1	234	/	234
2	30	19	49
3	3	1	4
总计	267	20	287

表 9 话题栈深度和话题栈折返度统计

相比话题结构内部折返，话题栈的折返从认知上更为困难。表 9 显示，话题栈的折返主要集中在话题栈深度为 1 的情况。且话题栈深度同为 2 或 3 时，折返度 1 的标点句数量大于折返度为 2 的标点句数量。从认知机模型解释，由于话题栈的内容在生成话题自足句时被排除在外，相当于在每次生成话题自足句时不能得到复述，容易遗忘。话题栈折返度越大则表示话题来自于越早的标点句，对记忆时间保持要求高，故表现出话题栈难以折返。

4.2.4. 标点句总深度和搁置区 Σ 使用量

标点句总深度包括话题深度和话题结构内的标点句深度。标点句总深度与搁置区 Σ 使用量都是代表了一定的认知复杂度，表 10 对二者交叉对比，考察其复杂度叠加的情况。

标点句总深度 \ Σ 使用量	0	1	2	3	4	5	6	7	8	总计
0	9148	144	29	6	5	2	1	1		9336
1	11929	2061	91	27	4	2			1	14115
2	4770	789	244	34	19	5	1			5862
3	1172	135	70	31	4	5	2			1419
4	172	20	14	4	1					211
5	16	2	1	1						20
总计	27207	3151	449	103	33	14	4	1	1	30963

表 10 标点句总深度和 Σ 使用量的交叉分布

表 10 每列表示 Σ 使用量，最大为 8，每行表示标点句总深度，最大为 5。可以看出，标点句总深度和 Σ 使用量大致成反比，深度太深则难以搁置。相比之下，标点句深度增加比较容易，深度为 3 的标点句还有 1419 句，将其搁置在 Σ 中比较困难，有 1172 句不搁置，

搁置 2 句的情况只有 70 句。表 11 列出了表 10 中具有相关特征的标点句数超过标点句总数 1%的情况（表 10 的灰色部分）。

排序	标点句总深度	Σ 使用量	标点句数	占标点句总数的百分比(%)	累计比(%)
1	1	0	11929	38.53	38.53
2	0	0	9148	29.54	68.07
3	2	0	4770	15.41	83.48
4	1	1	2061	6.66	90.13
5	3	0	1172	3.79	93.92
6	0	1	789	2.55	96.47

表 11 标点句总深度和 Σ 使用量分布比重大于 1%的情况

表 11 显示，标点句总深度不超过 3，Σ 使用量不超过 1，二者之和不超过 3 的情况已经覆盖 96%以上的标点句，体现了说汉语时的认知局限性。

5. 结语

语言理解问题从认知的角度已有大量的研究，但针对汉语的研究却很少。由于认知实验操作复杂，不容易大规模复制，因此难以量化其结论的普遍性以及对语言事实的覆盖度。本文尝试模拟人补足汉语篇章片段中话题-说明信息的过程，建立广义话题结构认知机模型，并通过认知机对大规模汉语语料定量分析，考察汉语标点句认知所需的记忆资源及认知局限性。用作统计特征量的广义话题结构特征有标点句的深度、话题结构内折返度、话题栈深度、话题栈折返度、搁置区使用量。统计结果显示，特征统计频率低和认知机调用资源多呈高度正相关。同时，统计数据可从认知行为的视角得到合理解释。本文一方面揭示了说汉语者的话题认知能力的表现和局限性，另一方面又说明了广义话题结构认知机是话题认知的合理模型。

参考文献

[1] Baddeley AD. The episodic buffer: A new component of working memory?[J] Trends Cogn Sci, 2000, 4(11): 417~423

[2] MacDonald MC, Just MA, Carpenter PA. Working memory constraints on the processing of syntactic ambiguity [J]. Cogn Psychol, 1992, 23(1): 56~98

[3] Just MA, Carpenter PA. A capacity theory of comprehension: Individual differences in working memory capacity [J]. Psychol Rev, 1992, 99(`): 122~149

[4] Traxler MJ, Williams RS, Blozis SA, Morris RK. Working memory, animacy, and verb class in the processing of relative clauses [J]. J Mem Lang, 2005, 53(2): 204~224

[5] Kintsch W. The role of knowledge discourse comprehension: A construction-Integration mode [J]. Psychol Rev, 1998, 95(2):163~182

[6] Bransford JD, Marclay J, Frank J. Sentence memory: A constructive versus interpretive approach [J]. Cogn Psychol, 1972, 3(2): 193~209

[7] McKoon, Ratcliff R. Inference during reading [J]. Psychol Rev, 1992, 99(3): 440~466

[8] Cairns H S, Cairns C E. Psycholinguistics: A cognitive view of language [M]. New York: Holt, Rinehart and Winston, 1976.

[9] Forster, K. Level of processing and the structure of language processor [J]. In W. E. Cooper and E. C. T. Walker (Eds), Sentence Processing. Hillsdale, NJ: Erlbaum.1979: 27~85

[10] Lindsay PH & Norman DA. Human Information Processing: An Introduction to Psychology [M]. New York: Academic Press

[11] Marslen-Wilson, W. Tyler, L. The temporal structure of spoken language comprehension [J]. Cognition,1980, 8:1~72

[12] Feldman, J. Minimization of Boolean complexity in human concept learning [J]. Nature, 407(October), 630~633

[13] 宋柔. 汉语篇章广义话题结构的流水模型[J]. 中国语文，2013(6):483-494

[14] 尚英. 汉语篇章广义话题结构理论的实证性研究[D]. 北京语言大学博士论文，2014

[15] Miller, G.A. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information [J]. Psychological Review, 1956， 63(2): 81~97