arXiv:2301.10595v1 [cs.CL] 25 Jan 2023

将文本提炼为电路

Vincent Wang-Ma´scianica[†‡]、Jonathon Liu[†] 和 Bob Coecke[†]

[†]Quantinuum，量子组合智能，牛津
[‡]牛津大学，计算机科学系

2023年1月26日

摘要

本文探讨自然语言中意义的结构。在 [13] 中，我们勾勒了一种名为 DisCoCirc 的自然语言框架，它具有以下特点：

(1) 它既是组合式的，也是分布式的（亦称向量式）；(2) 它适用于一般文本，而不仅仅是句子；(3) 它捕捉意义之间的语言学“联系”（参见语法）；(4) 随着文本推进，词义会得到更新；(5) 句子类型反映了其意义得到更新的词语；(6) 它能自然地容纳语言歧义。

本文中，我们针对英语的一个重要片段实现了 DisCoCirc。

此外，我们还表明，转向 DisCoCirc 文本电路时，可以消除一些“语法官僚主义”。也就是说，DisCoCirc 在以下方面表现出显著程度的：

(7) 语言间与语言内独立性。

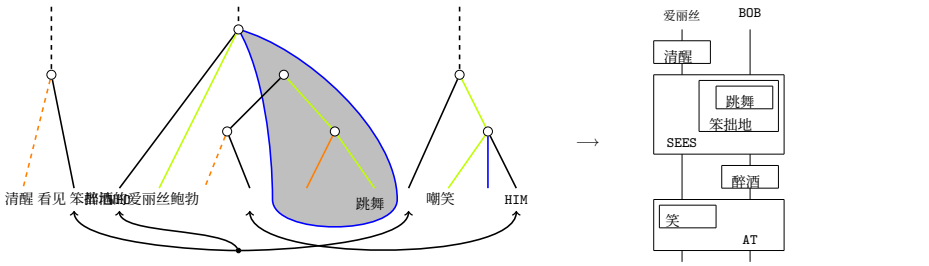
例如，语言间独立性指不受各语言之间不同的语序惯例影响；语言内独立性则指不受诸如使用许多短句而非少数长句之类的选择影响。这种语言间独立性意味着，我们的文本电路应能适用于其他语言，不同于范畴语法中特定语言的类型标注。因此，文本电路是一种精简的结构，用于呈现“文本的实际内容”，也就是文本中意义在多个表达层次（参见词语、句子、文本）上的内在运作，并可能捕捉到语法之下真正普遍的东西。

语法官僚主义的消除也解释了为什么 DisCoCirc：

(8) 适用于语言之外的领域，例如空间 [76]，视觉 [13, §4.7] 以及其他认知模式。

事实上，虽然人类无法通过文本电路进行口头交流，但机器可以；而抽象的文本电路可以被视为专为机器定制的语言结构。

具体而言，我们首先为英语的一个片段定义一种“混合语法”——这是一种专门构建的、最简的语法形式系统，是生成文本电路所必需的。然后，我们详细说明一个翻译过程，使得该语法生成的所有文本都能产生一个文本电路：



反过来，对于通过自由组合生成元而得到的任何文本电路，都存在一个（采用混合语法的）文本，可以生成该电路。因此：

(9) 文本电路能够生成文本。

目录

1 引言 **3**

2 文本的混合语法 **7**

2.1 预备知识.8 2.2 简单句结构.9 2.3 复合句结构 I：代词链接.11 2.4 复合句结构 II：短语作用域.15

3 文本图示 19

3.1 预备知识.19 3.2 作为文本图示的简单句.19 3.3 文本图示重写.21 3.4 文本图示中的代词连接.21 3.5 作为短语气泡的短语作用域.24

4 文本电路 **25**

4.1 定义.25

5 数学结果 **28**

5.1 主要文本电路定理.28 5.2 细化、扩展与惯例.29 5.3 文本图到文本电路.33 5.4 定理的证明.39 5.5 通过方程扩展语法.40

6 讨论 45

6.1 文本解析.45 6.2 任意文本.45 6.3 与其他句法和语义形式体系的关系.45 6.3.1 我们用于文本的混合语法.46 6.3.2 文本图示.46 6.3.3 文本电路.46

7 致谢 **48**

1 引言

本文探讨了一种组合框架，在该框架中，分布式（又称向量式）意义与语言结构能够进行有意义且透明的交互。此外，这一框架足够灵活，能够容纳超越弗雷格自下而上观念的组合性 [12]。例如，在自然语言中，词语的局部含义往往受到全局文本语境的影响。此外，现代机器学习方法也涉及这种自上而下的意义流动：全局训练会决定局部神经元的状态，尽管这些状态的含义通常不那么明确。

第一个将分布式含义与语法相结合的模型——在这个特定案例中，采用的是范畴语法 [2, 42, 72, 43] ——是 DisCoCat，该模型于 2008 [20]提出。DisCoCat 获得了实证支持[4, 28, 38, 82, 53, 83], ，旨在将自然语言处理中的组合式方法与数据驱动方法结合起来，此前这两种方法已各自走上不同的发展道路 [27, 11, 58]。具体而言，对语法的考量明确了词语和短语之间许多“语言联系”，因此在使用机器学习方法时，无需再学习这些联系。[1]

最近，DisCoCirc [13] 被提出，作为对 DisCoCat 的改进，旨在解决其三大主要不足。首先，DisCoCat 能根据词义生成短语或句子的含义，却无法根据句子的含义生成文本的含义。其次，在 DisCoCat 中，所有含义都是静态的，而词语的含义可能会随着文本的推进而改变。事实上，大多数当前的大语言模型也存在这一不足。第三，DisCoCat 并未明确界定句子含义的空间，因此其定义并非规范。在 DisCoCirc 中，该空间由相关名词构成。

DisCoCirc 实现这些改进的方式，是让文本呈现为文本电路的形式；文本电路的二维性对于实现上述每一项改进都至关重要。我们将在下文单独讨论这一维度问题。

[13] 中 DisCoCirc 的“幼儿版”已应用于多个问题，包括语言意义的时空模型 [76], 、自然语言中的逻辑否定和会话否定 [48, 66, 70], 、求解逻辑谜题 [24, 25], ，尤其是量子自然语言处理（QNLP），其中，文本电路与量子硬件的机器级代码相匹配 [14]。[2] 此外，还有一些关于 DisCoCirc 的进一步理论阐述 [18, 29, 22, 21]。

然而，[13] 中提出的 DisCoCirc 仅限于非常基础的主谓宾句，可能带有关系代词。在这些情况下，根据句子的（前群）语法 [43, 44], ，可以生成相应的电路，然后将其进一步组合成更大的文本。但当面对一般文本时，并没有生成文本电路的通用方法，因此也没有对何为“有效”文本电路的刻画。本文将全面解决这些问题，并阐明文本与文本电路之间的确切关系。我们首先会遇到以下一些难题：

- 首先，对于语法理论究竟是什么，人们并无普遍共识；现有多种截然不同的理论便是明证。仅举几例：类型逻辑语法，如组合范畴语法 [72], 兰贝克语法[42] 和预群语法 [44]；标记图方法，如依存 [73] 语法和链接[71] 语法；盒嵌套方法，如话语表征理论 [35] 和 [7]；以及各种形式的生成式方法，如转换生成语法 [64], 树邻接语法 [33] 等。

- 其次，自然语言的特性使得构建具体而完整的数学描述和算法十分困难，因此实践中人们依赖统计句法分析器。例如，要具体规定一个用于依存语法的句法分析算法，就必须完整列出某种语言中所有可能的语法关系，这一要求难以实现；因此，次优选择是使用 SpaCy [30]这类统计句法分析器。

- 第三，为了生成文本电路，需要消解代词并明确辖域，因为这些对于识别基本意义之间的连贯性也至关重要。

[1]特别要注意的是，DisCoCat 和 DisCoCirc 能够自然地处理涉及意义、语法或两者的歧义，方法是使用密度矩阵及相应的超算子 [61, 36, 60, 53]。[2]用于实现量子自然语言处理 [51, 50] 的工具包 `lambeq` 最近已开放使用 [37]。不过，目前 DisCoCirc 尚未纳入其中，但这种情况很快就会改变。

- 最后，语法通常只处理句子，而不处理文本。

为了解决这些问题，我们在第2节概述一种专为文本设计的“混合语法”。我们的目标不是建立一套完整的新语法理论，而是构建足够的语法结构，从而能够在相当大范围的语言片段中，搭建起具备语法的文本与文本电路之间的桥梁。为了构建这一最简语法，我们采用短语结构语法，并加入若干额外要素：“代词链接”用于标识反复出现的指称对象，“短语区域”用于标识带句子补语的词语的作用域，而“转换语法规则”则让我们能够构造更复杂的句子并引入关系代词。

在第3节中，我们将混合语法文本转换为字符串图，从而得到文本图这一中间结构。文本图将混合语法中的语法现象——短语结构、代词链接、短语作用域和转换语法规则——整合进一个数学结构中。在文本图这一框架下，我们语法中为处理转换语法规则所需的伪影会消失。在第4节中，我们引入文本电路，它们能够生成文本，并展示如何从文本图得到文本电路。在第5节中，我们证明了主要结果：文本电路刻画了具备混合语法的文本的“本质意义连通性”；此外，我们还概述了如何在保持理论保守性的前提下扩展文本电路理论，以容纳更多语言特征。我们证明，每个符合混合语法的文本都会产生一个文本电路，也证明每个文本电路都源自某个文本。

有趣的是，从配有语法的文本到文本电路的映射是真满射，也就是说，不同的文本可能产生相同的电路。这反映了一个事实：DisCoCirc 消除了我们所说的“语法官僚主义”，抽象掉了承载相同含义的文本之间的差异。我们将在下文单独讨论这一特性。最后，在第6节中，我们将详细探讨一些实际考量，例如句法分析、扩展我们所建模的语言片段，以及文本电路与其他语法和语义形式体系之间的关系。由这一比较还可得出一条重要的结论：尽管本文选择从我们的混合语法中构造文本电路，但显然还有其他几种成熟的形式体系可以作为构造文本电路的替代起点。

总而言之，我们可以如下表示从文本中提炼出文本电路的过程：

$$\text{text} \xrightarrow{\text{引出}} \text{文本图示} \xrightarrow{\text{重写为}} \text{文本电路}$$

我们明确构造了这一过程，并证明了它的基本性质，例如（适当的）满射性。

然而，这一过程假设我们能够明确地为文本赋予语法结构。对于较短的句子，这通常是可行的，但较长的句子则不一定。我们还希望机器能够提炼出电路，因此一般来说，提炼过程更像是这样：

$$\text{text} \overset{\text{解析}}{\dashrightarrow} \text{文本+语法} \xleftrightarrow{\text{对应于}} \text{文本图} \xrightarrow{\text{重写为}} \text{文本电路}$$

其中，句法分析是自动为句子和文本赋予合乎语法的及其他结构的过程。本文不会明确展开句法分析这一步，因为每一种构成我们语法的要素都已有许多可用工具——见第6.1节。

我们的文本混合语法是生成式的：

$$* \xrightarrow{\text{generate}} \text{texts}$$

其中，$*$ 表示输入生成过程的数据。不过，根据我们的研究结果，现在我们也可以仅通过组合电路来生成文本，并使用将每个电路映射到生成它的文本的映射，这里将其记作文本化：

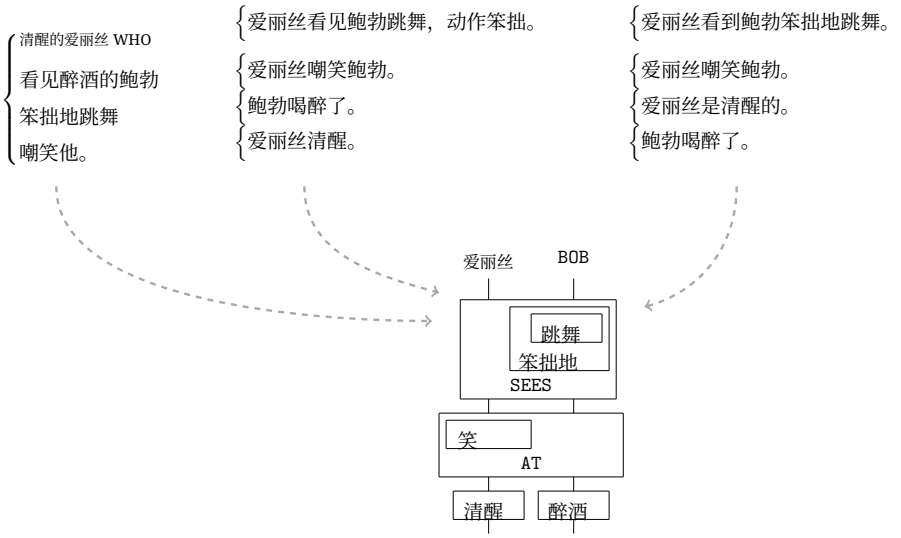
$$* \xrightarrow{\text{generate}} \overline{\text{text circuits}} \xrightarrow{\text{textualise}} \text{texts}$$

许多文本会生成相同的文本电路，这意味着生成的文本具有范式。像机器那样将文本作为电路处理时，我们自然可以直接生成这些电路，无须再做其他工作：

$$* \xrightarrow{\text{生成}} \text{文本电路}$$

我们还没有提到如何在电路中实际对意义进行建模。对于 DisCoCat，机器学习中的向量空间嵌入所对应的是向量空间范畴、线性映射和张量积， [19]对此有非常详尽的描述。此外，该框架的灵活性使我们能够转向密度矩阵和超算子，从而自然地处理词汇歧义 [61, 36, 60, 53]，和词汇蕴涵 [3, 47]。同一模型也可用于 DisCoCirc [13, 18, 22, § 6.1–6.3]，，不过根据本文的研究结果，还需加入高阶算子。我们将模型的详细讨论留待日后撰写。

语法官僚主义。本文所说的语言结构，并非完整的语法，而只是能够捕捉含义之间“关联性”的最简结构。例如，在含有及物动词、主语和宾语的句子中，主语和宾语都与动词相关联。在标准语法中，这种结构由它们的相对词序决定；然而，世界上的某种语言中都有这三种语法角色的各种可能排列方式 [74]。某种语言内部对词序所作的惯例，是我们所谓跨语言语法官僚机制的一部分。在特定语言内部，表达意义时还有其他选择，例如使用一个长句还是许多短句、相应地使用（关系）代词，以及文本的其他文体方面；因此，不同文本可能承载相同的含义。我们称之为语言内部语法官僚机制。不同文本可以共享相同意义这一事实，会在文本之间导出方程——我们在别处称之为“语法方程” [21]。转向语言电路时，这种语法官僚机制便被消除；也就是说，我们转而考虑等价类，只保留可称为文本“实际实质”的部分。以下是一些不同句子或文本片段产生相同语言电路的例子：



(1)

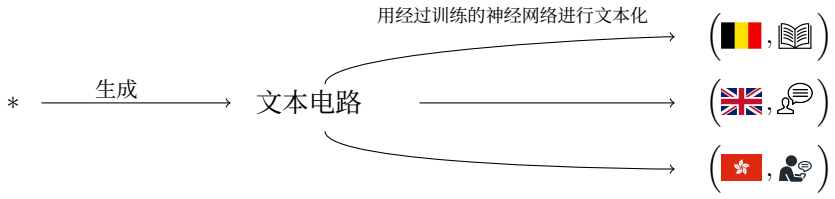
关于英语和乌尔都语这一情形下跨语言独立性的概述，可见 [78]。因此，我们的文本回路也许有助于回答以下问题：

语法之下真正普遍的是什么？

有人声称，语法是硬连线在我们人类大脑中的 [62]，；更强的说法则是，存在一种独特机制，能够解释我们的语言能力：一种“普遍语法” [10]。从帕尼尼到蒙塔古，历史上所有的语法演算在某种意义上都是试图揭示这种“普遍性”。然而，语法演算所刻画的语法官僚体系，与据称由“普遍语法”驱动、让我们毫不费力地使用语言的能力之间，存在着一种张力和距离。如果我们能够理解这种官僚体系——不是为了如实描摹一种语言，而是为了彻底消除这一官僚体系——那么我们或许就能更接近普遍性。

另一方面，不难看出，要对这种代价高昂的语法官僚主义进行完整的数学刻画相当困难。标准的自然语言处理会以蛮力解决

语法官僚主义问题，方法是采用数据驱动的机器学习式训练来训练非常大型的模型；目前，这或许确实是处理这些语法官僚主义最实用的方式。因此，在我们看来，生成通用语言（包括语言的选择和风格）可以分为两个部分：构建文本电路——即不受官僚主义束缚的语言实质内容——然后重新引入官僚主义，将文本电路转化为实际文本：

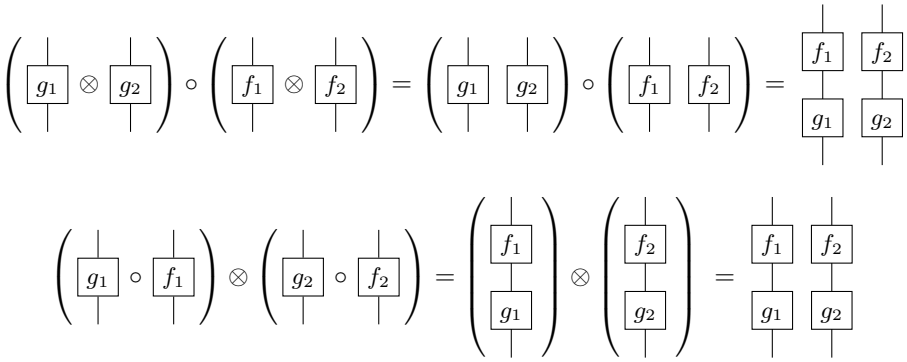


其中 fla gs 之类元素指向特定语言，其他图标则指向文体方面 .

二**维性**。我们认为，语法官僚主义在很大程度上源于这样一个事实：人类语言是一种承载高维内容的一维媒介；这种观点与乔姆斯基的最简方案相呼应。那么，这是什么意思呢？有时，事物被迫生活在并非其自然栖息地的世界里。数学中的一个例子是幺半范畴 [5]，也称为张量范畴。这种结构在量子理论 [16]、计算机科学 [1], 等诸多领域中发挥着核心作用，刻画了并行组合 $\otimes$（又称“同时”）和顺序复合 $\circ$（又称“之后”）这两种运算。通常，在教科书中，这种结构以代数符号的一维（1D）字符串表示，而要表达它们之间的相互作用，则需要借助方程，最著名的莫过于所谓的“双函子性”：

$$(g_1 \otimes g_2) \circ (f_1 \otimes f_2) = (g_1 \circ f_1) \otimes (g_2 \circ f_2)$$

另一方面 并用二维图表来表示它，按照从上到下的顺序阅读图示 o
方程是必需的 要的：

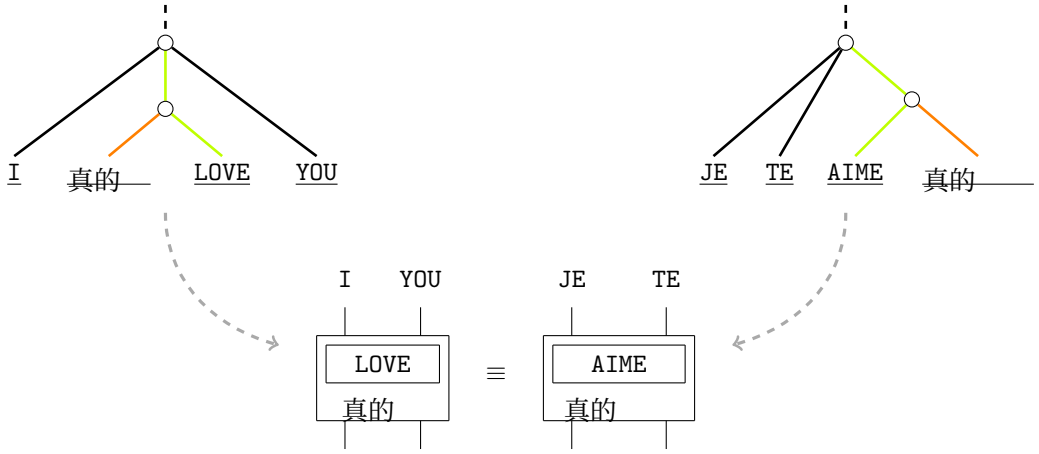


也就是说，借助图示的二维格式，定义性的符号方程被融入平面几何之中。符号表示之所以更加复杂，是因为要把二维内容强加到一维直线上，就需要引入人为的繁文缛节（例如括号标记），这也需要额外的规则。关于这一切的更详细讨论，请参见 [19, 16]。

语言也有类似的情况，不过方式更为微妙，如图（1）所示：将词语并排放置，会产生表达同一意思的不同方式，这也可能导致不同语言形成不同的惯例。本文所做的部分工作，就是将语言中的一维与二维之争形式化。

另一个较小的例子如下，由谷歌翻译生成，我们现在明确指出

两个句子彼此不同的树语法：



因此，不同的平面树仍会生成相同的电路。原因在于，虽然树看起来像一个二维数学实体，但词语仍排列在一条线上，而树实际上是对这串词语进行括号标记的结构。

组合意义。组合性一词源于语言学，通常与弗雷格的概念相关 [26], ，即整体的意义由以下因素决定：(i) 各部分的意义，以及 (ii) 这些部分如何组合在一起。具体而言：句子的意义源于：(i) 句中各词语的意义，以及 (ii) 句子的语法结构。正如 [55, 12] 中所详述的，这种组合性概念无法成立，而 [12]提出了一种更一般的组合性概念。我们的文本电路就是这种更一般的组合性概念的一个例子，在这种概念中，整体的意义也能影响各部分的意义。换言之，组合性构成了一种跨尺度的关系，其中较小和较大的部分地位平等。关于补充讨论，请参阅 [15] 。此外，语言转化为文本电路时所呈现的特定组合形式，也适用于语言之外的领域，例如空间 [76], 、视觉 [13, §47] ，其他认知模式也同样如此 [6]。从这个意义上说，以文本电路的形式呈现的语言，似乎是一种「世界的组合逻辑」。

2 文本的混合语法

现在，我们介绍一种混合语法。它是生成式的，旨在捕捉语言连接性。我们分三步来构建它。首先，从一种用于简单句的“上下文有关”[3] 语法入手，这些简单句只包含一个动词。其次，引入代词链接的概念，用以标识反复出现的名词，以及与其指称对象相对应的代词。此外，我们还将介绍一些规则，允许通过关系代词将含有重复名词的简单句融合起来。第三，引入接受句子补语的动词以及短语作用域边界的概念。这样，我们的片段便扩展到能够处理含有子短语的复合句，而这些子短语本身也是句子。

免责声明 2.1。如前所述，我们的目标是构建一种最简语法，使我们能够从文本生成文本电路，并得出明确的数学结果。为此，我们采用了融合不同形式体系基本思想的弗兰肯斯坦式混合体：我们使用乔姆斯基的转换短语结构语法，并以兰贝克预群的特征加以调整；代词链接的灵感来自话语表征理论，短语边界的灵感则来自依存语法。第6节将更详细地阐述这些联系。

[3]我们的语法可以表示为上下文无关语法，但为了更容易在数学上与文本图和文本电路建立联系，我们选择采用上下文有关的表示方式。我们的混合语法中只有两条规则是*potentially* 上下文有关的：一条用于引入及物动词的介词，另一条用于管理短语作用域。两者都可以改写为上下文无关形式，我们在脚注 4 和 6 中对此作了概述。有关上下文无关语法和上下文有关语法的定义，请参阅 [31] 。

免责声明 2.2。为求清晰起见，我们不处理某些语法现象（例如时态），省略某些语法模式（我们假设副词总是位于动词之前），并且只处理语言的一部分（这里不考虑限定词和量词，并假设双及物动词可以用带介词的及物动词作等效表达）。这种方法也存在形式语法通常具有的缺点——例如，没有考虑英语等语言中表示用途、来源等的形容词顺序——但这些问题可以用常见方法处理，即让语法与意义相结合。事实上，DisCoCat/DisCoCirc 的核心就在于结合语法与意义；我们（与许多人一样）认为，归根结底二者不应彼此独立，而应相互影响。实用自然语言处理已通过实证表明，这对于开发高效工具至关重要。

免责声明 2.3. 仅给出一个词语字符串而没有任何进一步的上下文时，无法根据文本唯一确定代词性连接和短语边界数据——不过，很多情况下，语法类型也同样无法唯一确定。之所以纳入这些内容，是因为它们对于获得可以进行数学推理的消歧后文本结构必不可少。

2.1 预备知识

我们通过提供一个由（一组通常有限的）产生式规则 [31]组成的字符串重写系统，来建模短语结构——即约束句子中词语顺序的结构：

$$\alpha \mapsto \beta$$

其中，α 和 β 是由符号组成的字符串。在我们的例子中，它们可以是短语成分（例如 NP），也可以是英语单词（例如 BOB）。我们对后者加下划线，以表示此类符号是终结符，即不能再通过产生式规则进一步重写。

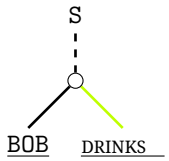
所有此类语法都按如下方式规定语言——即由符号构成的字符串集合。指定一个特殊的开始符号S，与该重写系统关联的语言就是所有可通过有限次应用可用产生式规则生成的终结符字符串的集合，例如，给定以下规则：

$$\mathtt{S} \mapsto \mathtt{NP} \cdot \mathtt{IVP} \tag{2}$$
$$\mathtt{NP} \mapsto \underline{\mathtt{BOB}} \tag{3}$$
$$\mathtt{IVP} \mapsto \underline{\mathtt{DRINKS}} \tag{4}$$

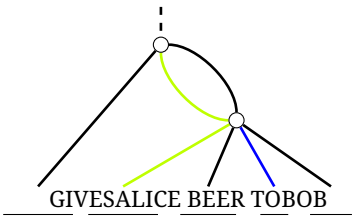
其中，$\cdot$ 表示字符串连接，我们可以构造如下简单句：

$$\begin{aligned} \mathtt{S} &\overset{(2)}{\mapsto} \mathtt{NP} \cdot \mathtt{IVP} \\ &\overset{(3)}{\mapsto} \underline{\mathtt{BOB}} \cdot \mathtt{IVP} \\ &\overset{(4)}{\mapsto} \underline{\mathtt{BOB}} \cdot \underline{\mathtt{DRINKS}} \end{aligned}$$

我们将用平面“树”来表示字符串的推导。图示从上到下阅读：



这些“树”中，父节点和子节点之间可以有多条边。

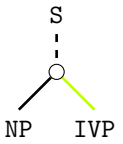


我们省略中间符号的符号标签，并用彩色边代替。例如，NP 用黑色边表示，IVP 用绿色边表示。在介绍规则时，我们也会遵循带类型的线的着色约定，这样之后即使从图示中省略 IVP 之类的类型标注，也不会造成混淆。

2.2 简单句结构

现在，我们通过给出各类合乎语法的树片段来介绍一种短语结构语法；首先讨论我们称为简单句子的句子，它们只有一个不带句子补语的动词。

不及物动词和及物动词。一个简单句可以包含一个不及物动词或及物动词。前一种情况下，句子由一个名词短语和一个不及物动词短语组成（例如，爱丽丝跑步。）。后一种情况下，句子由一个名词短语、一个及物动词短语和另一个名词短语组成（例如，爱丽丝喜欢鲍勃。）。它们对应的树片段如下：

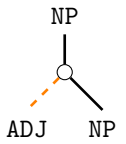
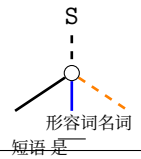
Rule	平面树片段
句子 $\mapsto$ 名词短语 · 不及物动词短语	
句子 $\mapsto$ 名词短语$_1$ ·及物动词短语 · 名词短语$_2$	 名词短语$_1$ 及物动词短语 名词短语$_2$

此外还有动词的终结规则，其中语法的终结符是相应类型的动词，例如不及物动词或及物动词：

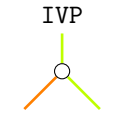
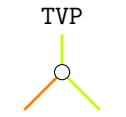
Rule	平面树片段
不及物动词短语 $\mapsto$不及物动词	
及物动词短语 $\mapsto$及物动词	

接下来，我们省略终结规则，改为给出完整推导的示例，从中可以推断出终结符。

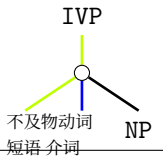
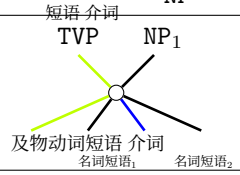
形容词。 形容词可以出现在名词短语之前（例如：醉酒 快乐 鲍勃）。或者，如果将系动词 IS 视为动词，单个形容词也可以出现在名词短语之后（例如：鲍勃 IS 喝醉了）。树片段如下：

Rule	平面树片段
名词短语 $\mapsto$ 形容词 · 名词短语	
句子 $\mapsto$ 名词短语 · 是 · 形容词	

副词。副词可以出现在动词之前（例如：爱丽丝 快速地 快乐地 RUNS .）：

Rule	平面树片段
不及物动词短语 $\mapsto$ 副词 · 不及物动词短语	
及物动词短语 $\mapsto$ 副词 · 及物动词短语	

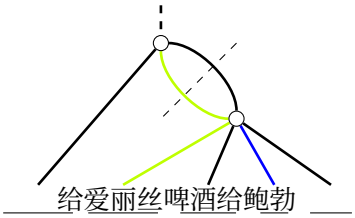
介词。介词可以出现在不及物动词短语的右侧，后接一个名词短语（例如，从 ALICE RUNS. 变为 ALICE RUNS TOWARDS BOB.）。如果及物动词短语右侧有一个名词短语，则还可以在其右侧添加一个介词，后接一个名词短语（例如，从 ALICE THROWS BEER. 变为 ALICE THROWS BEER TOWARDS BOB.）：[4]

Rule	平面树片段
不及物动词短语 $\mapsto$ 不及物动词短语 · 介词 · 名词短语	
及物动词短语 · 名词短语$_1$ $\mapsto$ 及物动词短语 · 名词短语$_1$ · 介词 · 名词短语$_2$	

例2.4。首先应用及物动词短语生成规则，得到 NP · TVP · NP，这对应于“爱丽丝给啤酒”这一合乎语法的结构。然后应用 TVP · NP 生成规则，将“给啤酒”（即一条绿色导线和一条黑色导线）转换为“把啤酒给鲍勃”。总而言之，我们得到

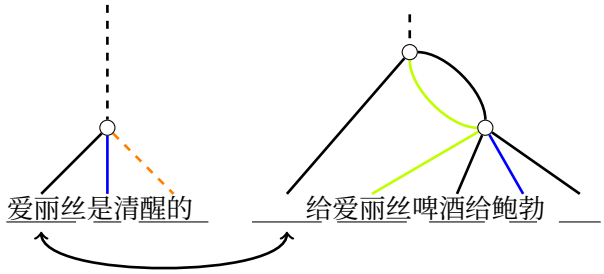
[4] 我们可以用一种等效的上下文无关方式来表示这条上下文有关规则。仔细查看混合语法的规则，可以发现 TVP 总是作为字符串 TVP · NP的一部分出现。假设我们用一个新的符号 TVP’ 替换这个字符串，并通过以下新规则 TVP’ $\mapsto$ TVP · NP将其“展开”。这样，我们就可以用上下文无关规则 TVP’ $\mapsto$ TVP’ · ADP · NP替换这条“上下文有关”规则。此外，我们还可以将涉及 TVP 的所有其他规则替换为涉及 TVP’的版本，从而得到一种等效的上下文无关表示。

如下的“树”，其中虚线表示这两棵“子树”：

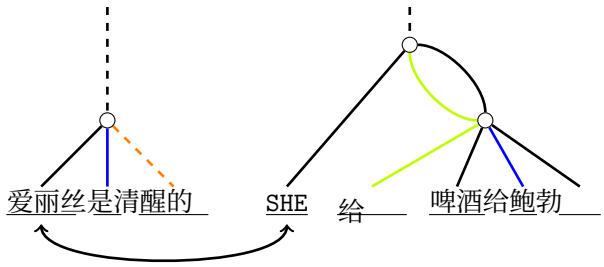


2.3 复合句结构I：代词链接

现在，我们将讨论范围扩展到单个句子之外。含有多个动词的句子称为复合句。我们构造这类句子的一种方式，是用关系代词将简单句融合在一起。我们用代词性连接来表示代词所承载的信息，它可以标识可能来自不同句子的名词。我们用树下方指向被标识名词的箭头来表示这些连接。例如，在文本“爱丽丝是清醒的。爱丽丝把啤酒给鲍勃。”中，我们可以像下面这样标识爱丽丝的两次出现：



存在代词性连接时，名词后续出现的位置可以用指代其先前出现位置的代词替换：



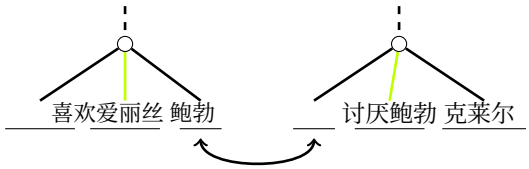
主语关系代词。 当代词链接以某些配置出现时，我们可以使用关系代词将两个或更多句法分析树中的句子融合在一起。首先，主语关系代词替代句法分析树 S_2，中的主语名词，并指向另一个先前句法分析树 S_1,中的一个名词，通常是宾语名词。图示如下：



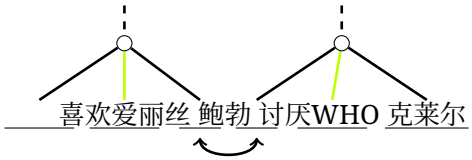
(5)

其中，右侧的大三角形是一种视觉惯例，用来表示这两个句子“融合”为一个句子。(5)这样的重写规则称为转换语法规则。变换会修改短语结构树。下面我们会看到更多关于关系代词的转换规则，并在第3节正式定义它们。

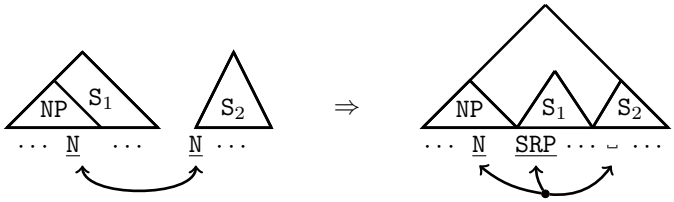
例2.5。 对于文本“爱丽丝喜欢鲍勃。鲍勃讨厌克莱尔。”，我们从两棵树开始。我们用代词性连接标出鲍勃的两处出现：



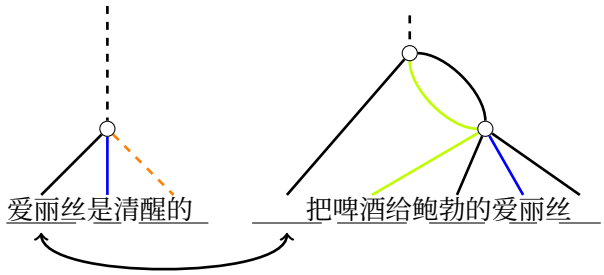
现在，我们可以将这些树融合：把第二个“鲍勃”替换为关系代词 WHO，得到：



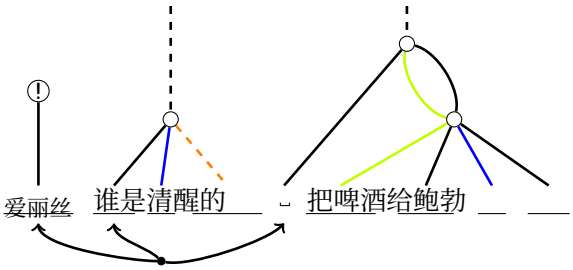
在英语中，主语关系代词还有一种特殊用法，可以让一个名词同时关联后续两个短语。给定依次对应句子 S_1 和 S_2 的句法分析树，当 S_2 的主语名词指向 S_1中的一个主语名词时，就会出现这种特殊情况。经过树变换后，结果依次为：S_1,的名词短语（以及任何形容词），接着是 S_1 （其中后出现的名词被关系代词替换），然后是 S_2 （其中指向的名词被删除）。我们用多重箭头指示两个以上由代词指代的名词，如下所示：



例 2.6。 我们回顾一个先前的例子：

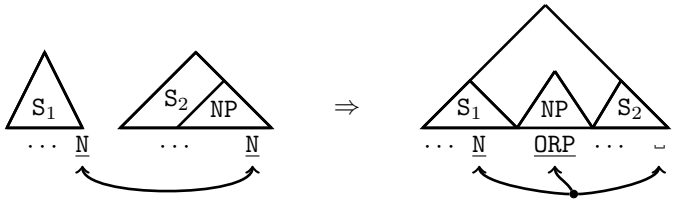


By 应用主语关系代词的特殊规则后，我们得到以下句子 e:

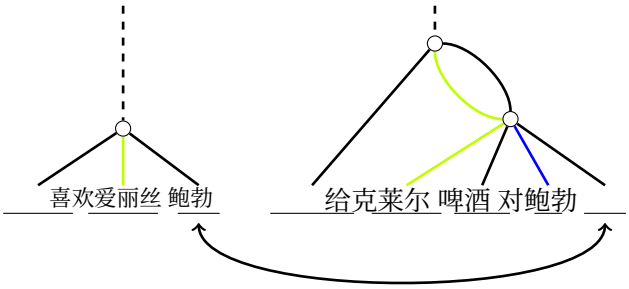


我们之前没有遇到过孤立名词——这里用 ! 表示——或空白标签；这些伪影在我们将在第3节介绍的文本图中会消失。

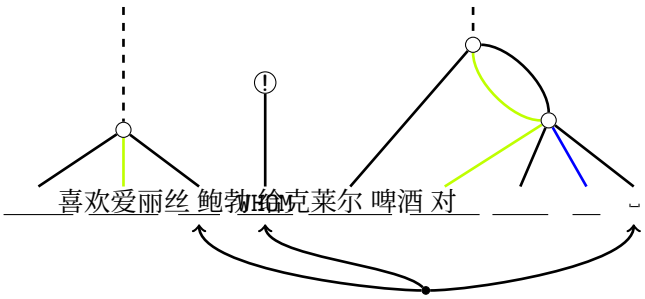
宾语关系代词。在连续的句子 S_1 和 S_2，中，如果 S_2 的宾语名词指向 S_1，的宾语名词，那么宾语关系代词出现在第一次出现之后，而名词的第二次出现则被替换为空白：



例2.7. 我们从文本“爱丽丝喜欢鲍勃”开始。“克莱尔给啤酒对鲍 B.

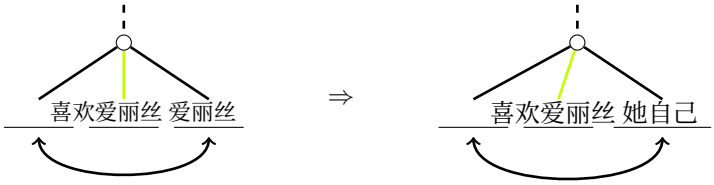


应用宾语关系代词规则后，我们得到：



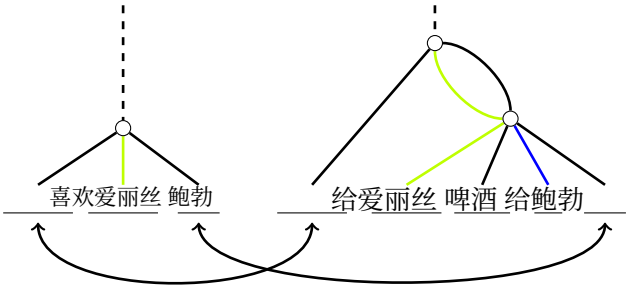
反身代　　名词 有时，同一个简单句中的两个名词在代词指代上相同　　e. 在这些情况下，我们可以引入一个反身代词。

例 2.8。 这里我们有爱丽丝喜欢爱丽丝：

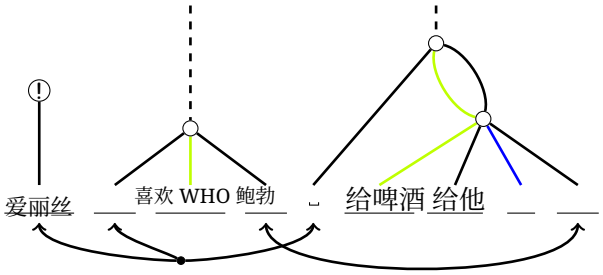


例 2.9. 我们注意到，在混合语法中处理关系代词和反身代词的代词链接时，会遇到各种限制，我们将在此展示这些限制。这些限制可以通过文本图的结构来解释。

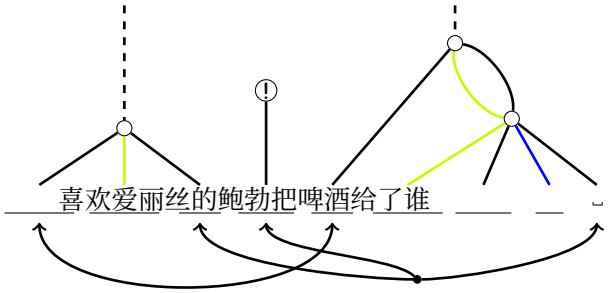
假设我们有以下带有代词链接的文本：



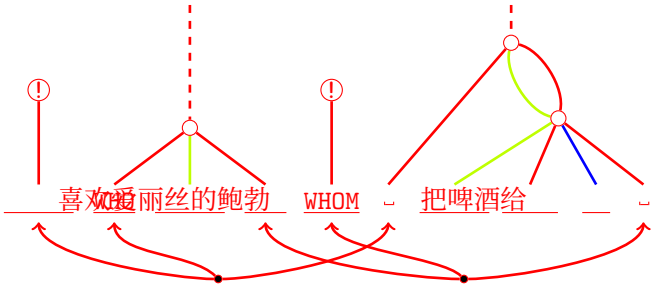
我们可以先根据主语关系代词规则合并 ALICE，然后将第二个 BOB 替换为普通代词 HIM[5]，从而得到合乎语法的句子：



或者，我们可以先应用宾语关系代词规则，将鲍勃合并起来，然后用常规代词她替换第二次出现的爱丽丝，从而得到合乎语法的句子：

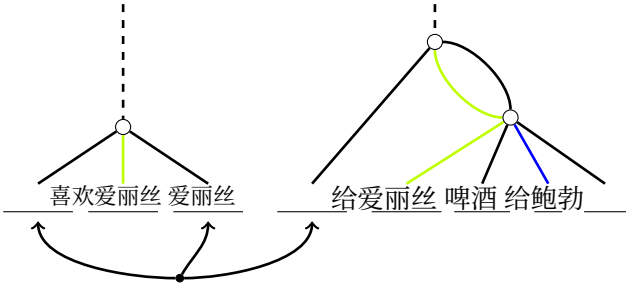


我们不能同时应用主语和宾语关系代词规则，除非这些句子最初是分开的，否则就会得出**不合语法的句子**：

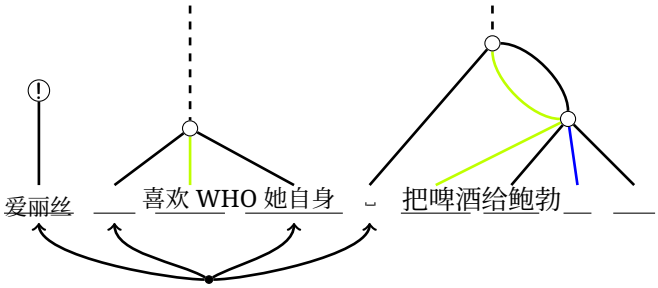


[5]为简便起见，我们假设代词采用任何“正确”的语法形式，并与性别、数等保持一致。我们主要关注指称结构。

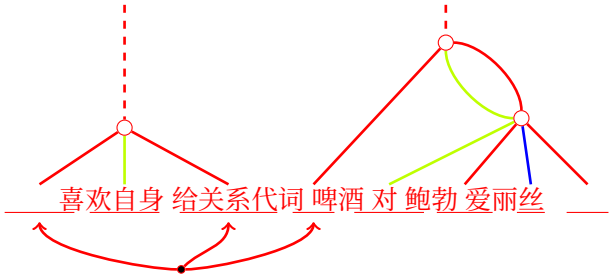
A si反身代词和关系代词也会出现类似现象。如果我们从这段文本开始 :



我们可以先引入一个反身代词，然后使用主语关系代词规则来协调爱丽丝的首次出现，从而得到一个合乎语法的句子：



除非初始句子是简单句，否则我们不能先应用主语关系代词规则，再应用反身代词规则；这样会得到**不合乎语法的句子**：



最后，我们指出，反转本小节中的例子后，包含关系代词的句子可以分解为简单句列表。

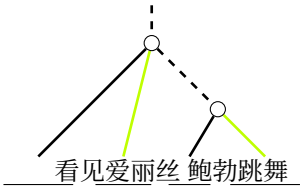
2.4 复合句结构 II：短语作用域

复合句也可以由包含子短语的句子构成，而这些子短语本身也是句子。我们将介绍两种这样的情况。我们先直观地介绍这些情况，然后引入对短语作用域的细化，以消除一种歧义来源。

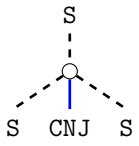
带句子补语的动词。 我们将带句子补语的动词（例如 SEE 或 THINK）视为一种独立的动词语法类别。一个句子可以由一个名词短语、一个带句子补语的动词，以及一个句子依次构成（例如，从 BOB DANCES 到 ALICE SEES BOB DANCE）。

Rule	平面树片段
句子 $\mapsto$ 名词短语 · SCV ·句子	

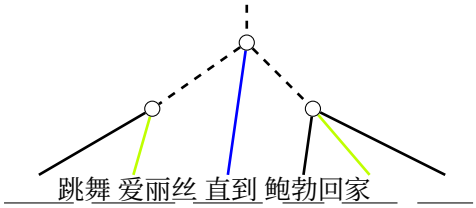
示例 2.10。 对于上面给出的例子，我们得到：



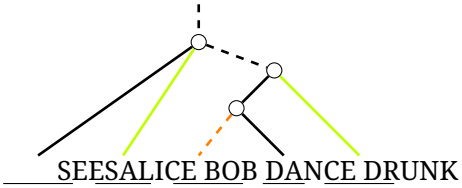
连词。我们对连词的处理方式与带有句子补语的动词类似，只不过连词两侧各有两个短语界定区域，而不是一个（例如，从 BOB DANCES. 和 ALICE LAUGHS. 到 BOB DANCES SO ALICE LAUGHS.）

Rule	平面树片段
句子 ↦ 句子 .连词 . 句子	

示例 2.11。如下：



上述表述中仍有一个歧义来源需要处理，这正是短语作用域发挥作用之处。对于句子 ALICE SEES DRUNK BOB DANCE.，只有一种短语结构：

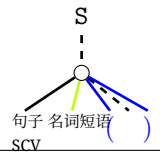
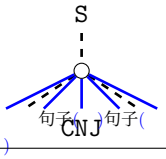


这使 DRUNK BOB DANCE(S) 全部处于爱丽丝所看见内容的辖域之内。不过，我们希望进一步区分以下几种情况：

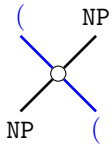
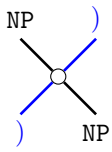
- 爱丽丝看见**既** 鲍勃喝醉了，也看见鲍勃跳舞。
- 爱丽丝**只** 看见鲍勃跳舞，却没有看见鲍勃喝醉。

为了区分这两种情况，我们引入两种“形式”类型 (和)，分别表示短语作用域的左边界和右边界。我们重新表述带句子补语的动词规则

和连词规则，以便同时表示短语边界：

Rule	平面树片段
句子 ↦ 名词短语 ·SCV· (· 句子 ·)	
句子 (· 句子 ·) · 连词 · (· 句子 ·) ↦	

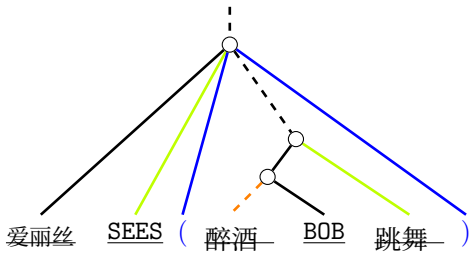
为了模拟名词可以部分置于短语辖域之外的许可条件，如上例所示，我们加入以下规则，允许一个名词短语“跨越”短语边界：[6]

Rule	平面树片段
(· 名词短语 ↦ 名词短语 · (	
名词短语 ·) ↦) · 名词短语	

例2.12。 为了消除句子中的歧义

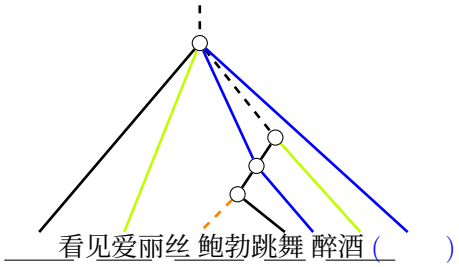
ALICE SEES DRUNK BOB DANCE.

爱丽丝看见 **两者都**鲍勃喝醉了和鲍勃跳舞这两个句子，我们有如下短语结构 ：



[6] 我们也可以用上下文无关的方式重新表述这些规则。我们可以不采用上下文敏感的辖域变换规则，而是使用标记变量 $\mathtt{X}^n$ 来表示它们处于辖域内，并用上标自然数表示辖域深度。举例来说，我们可以将句子补语动词规则 $\mathtt{S} \mapsto \mathtt{NP} \cdot \mathtt{SCV} \cdot (\cdot \mathtt{S} \cdot)$ 替换为 $\mathtt{S}^0 \mapsto \mathtt{NP}^0 \cdot \mathtt{SCV}^0 \cdot \mathtt{S}^1$。更一般地说 $\forall k_{\in \mathbb{N}}$： $\mathtt{S}^k \mapsto \mathtt{NP}^k \cdot \mathtt{SCV}^k \cdot \mathtt{S}^{(k+1)}$。通过将索引限制在一个较大的有限数值范围内，我们就能严格保持上下文无关。因此，我们可以采用一系列上下文无关规则，这些规则与之前相同，只是要求辖域索引匹配；再加上规则所生成图的平面性，这就确保了所有规则都在同一辖域内应用。例如，引入不及物动词的规则可以写作 $\forall k_{\in \mathbb{N}}$： $\mathtt{S}^k \mapsto \mathtt{NP}^k \cdot \mathtt{IVP}^k$。在这种设定下，我们可以将短语辖域规则——它们仅表示名词可以离开一个辖域——表述为 $\forall k_{>0}$： $\mathtt{NP}^k \mapsto \mathtt{NP}^{(k-1)}$。

To 消除歧义：爱丽丝**只**看见鲍勃跳舞，而没有看见他喝醉了，我们有 ：



例 2.13. 在图1中，我们展示了如何使用我们的混合语法生成文本。

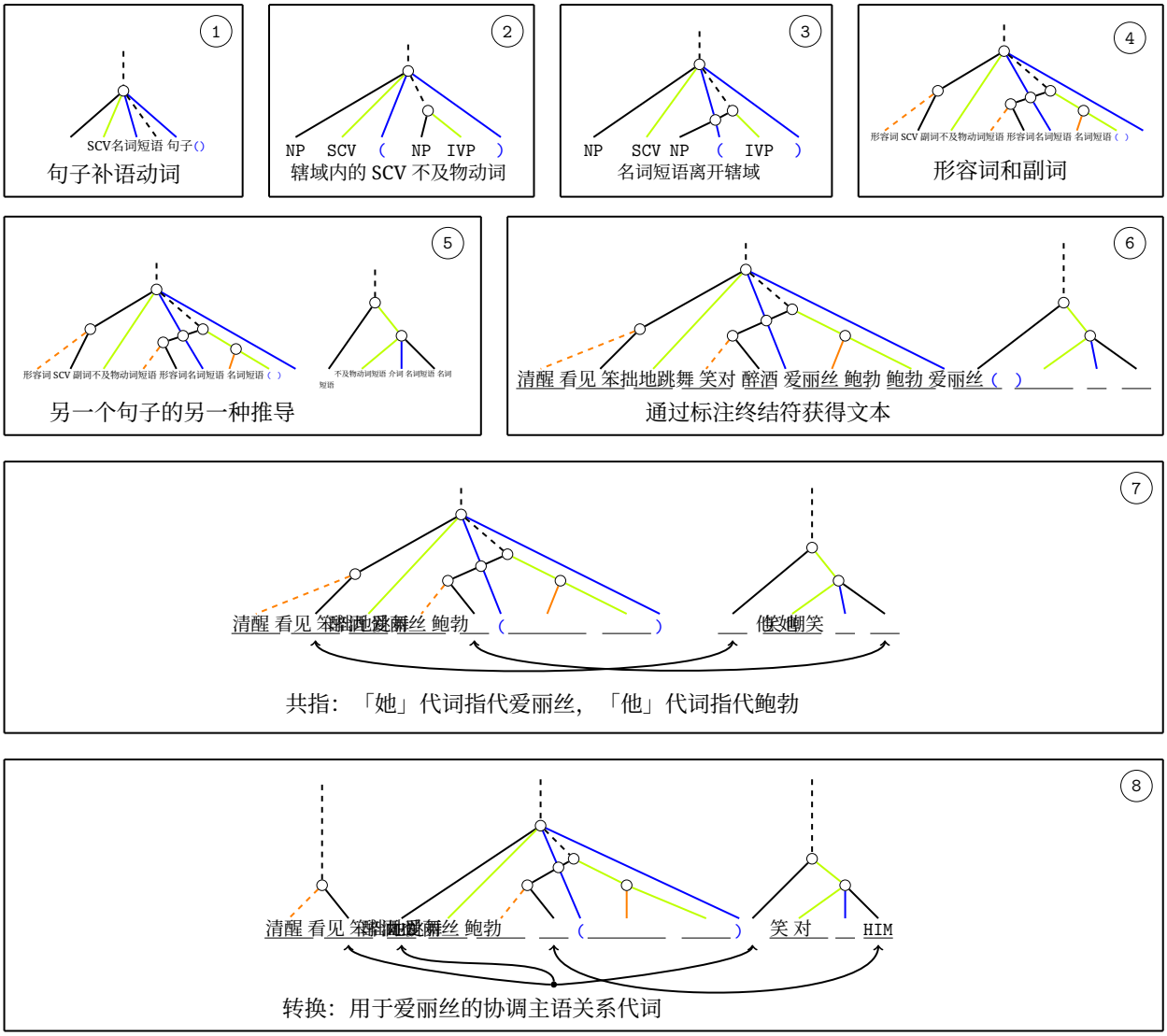


图1：混合语法文本的生成示例

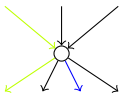
今后，在不会引起混淆时，我们将本节介绍的混合语法称为“混合语法”或简称“语法”。

3 文本图

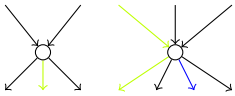
现在我们引入文本图，它将混合语法文本中的语法现象（参见代词链接和短语边界）整合到一个数学结构中。我们将通过从混合语法文本到文本图的系统性转换来实现这一点。在文本图的框架下，用于处理关系代词在混合语法中行为的伪影将不复存在。

3.1 预备知识

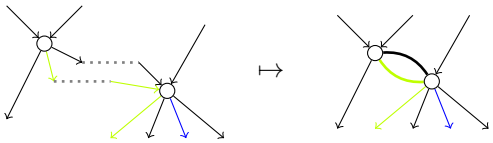
字符串图 [57, 34, 69, 16] 是一种用于复合输入输出框的图形化数学框架。对我们而言，这些“方框”将具有以下形状：



对于复合，通常需要区分并行组合：



或通过将匹配的输出和输入连接在一起来进行顺序复合。



尽管“字符串”一词让人联想到一维性，但字符串图通常也能容纳更高维的结构，例如平面和立体 [65]。

3.2 将简单句表示为文本图

上一节中，我们的短语结构规则的初始符号是 S。从混合语法过渡到文本图的过程中，我们系统地进行如下处理：

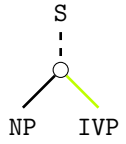
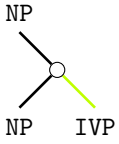
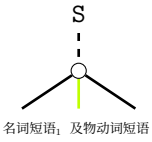
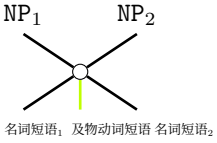
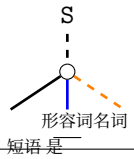
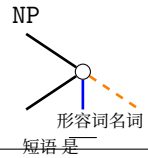
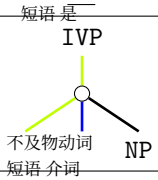
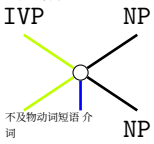
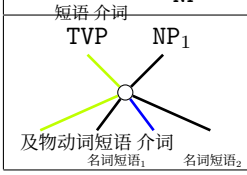
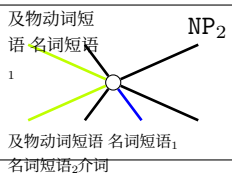
- 我们将用数量随句子而定的名词短语线替换句子类型 S。

这样一来，混合语法中的树形文本就变成了字符串图；文本中每个代词指代不同的名词短语标签，都对应一条输入名词短语线。这种视角上的几何变化使我们得以过渡到文本电路。为此，我们对第2节的规则作如下修改：

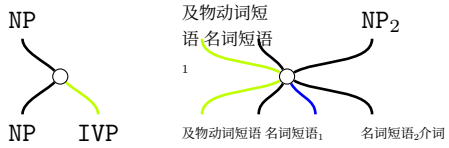
- 我们移除 S 类型。
- 每条规则都必须保持输入和输出名词线的数量不变，因此，引入名词短语类型的短语结构规则必须修改为同时接收数量相同的名词短语类型作为输入。

在下表中，我们仅列出以这种方式修改的规则。如表中所示，除了移除句子输入并添加名词输入、在有多个名词短语时为其添加下标外，我们还

将匹配的输入-输出对垂直对齐：

语法	Rule	图示
	不及物动词	
	及物动词	
	形容词	
	介词(不及物动词)	
	及物动词的介词	

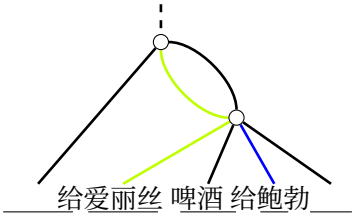
类似地，在绘制文本图时，作为一种视觉惯例，我们会始终将图示安排成匹配的输入和输出名词短语线彼此对齐，并将它们稍微弯曲，使其始终朝上或朝下，以更好地体现这样一种观念：在线图示的意义上，线应被视为输入和输出：



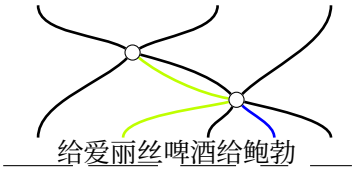
现在，我们也可以去掉其中一个重复出现的标签：



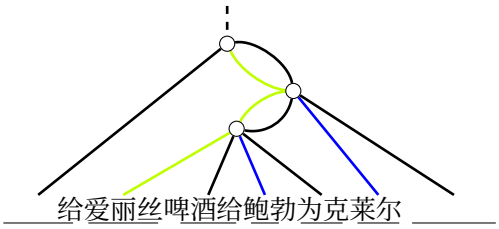
例3.1。 对于简单句“爱丽丝 把啤酒给鲍勃”，语法如下：



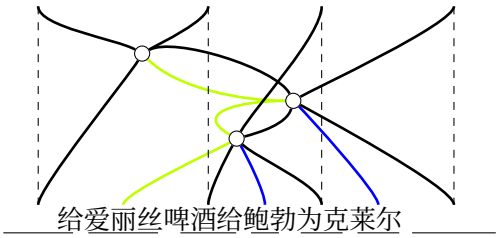
由此我们得到如下文本图：



例 3.2. 我们允许由语法到图示的转换得到的新输入名词短语线跨过其他导线，到达图示顶部。例如，在句子 ALICE GIVES BEER TO BOB FOR CLAIRE 中，语法如下：



fro由此我们得到以下文本图，其中名词线垂直对齐：

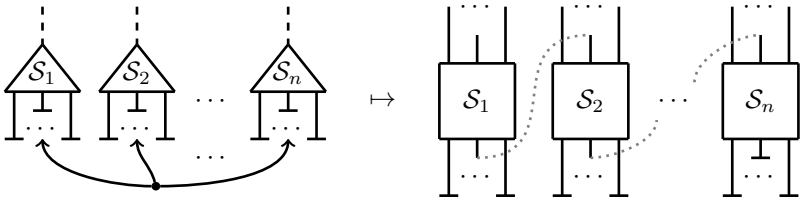


3.3 文本图示重写

处理关系代词时，我们见到了与乔姆斯基短语结构转换类似的东西：融合并更广泛地转换文本图的操作。我们将这些操作形式化为图示重写，每次重写都会用另一个图示替换原图示，且两者具有相同的输入线和输出线。接下来的章节将展示此类重写的示例。

3.4 文本图中的代词链接

在文本图中，代词链接的形式与语法中的不同。2.3 中的代词连接箭头变成了连接链中输出和输入名词线的导线[7]：



[7]代词链接也可以沿链从西北指向东南。我们选择这个方向，使得按从上到下、从左到右的顺序读取生成的文本电路中的门时，其顺序与文本中词语出现的顺序一致。

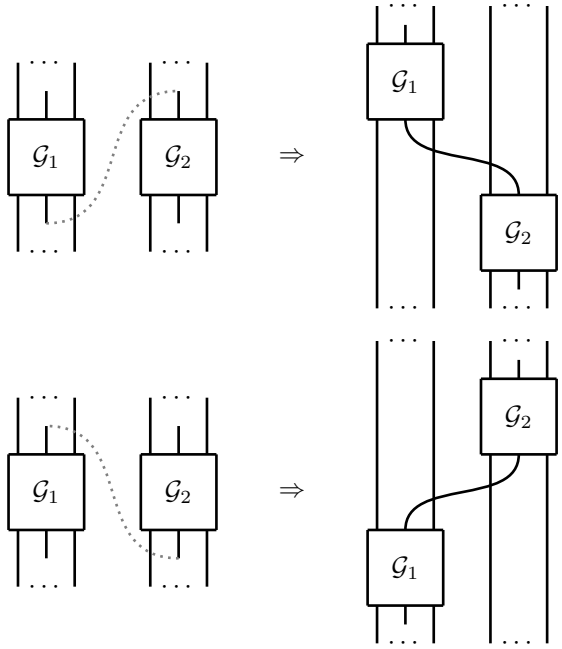
其中，以下部分展示了代词连接线与名词线之间的关系：



这些片段可以通过以下连接消去重写消失：



特别是，这意味着，当 $\mathcal{D}_1$ 和 $\mathcal{D}_2$ 是文本图时，我们有：



正如我们现在所见，线可以彼此交叉：这是文本图的一项重要特性。

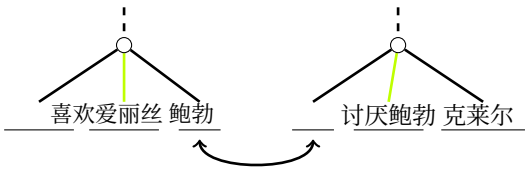
约定 3.3（在文本图中，只有连接关系才重要）。 混合语法图是平面的，而在文本图中，只要保持输入输出连通性，线和标签就可以自由交叉。

连接消去改写可能会使名词线发生扭结；这是一个更普遍原则的结果：只有输出到输入的连接关系才重要。因此，我们还有以下连接消去改写：

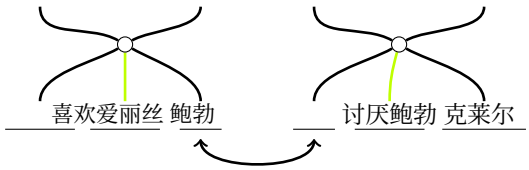


我们通过示例展示这些重写如何将代词链接解释为文本图的复合。

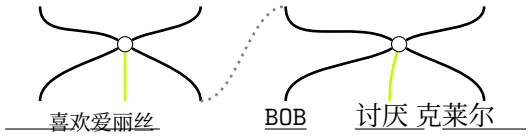
例 3.4（重新审视主语关系代词）。 我们从以下文本开始：



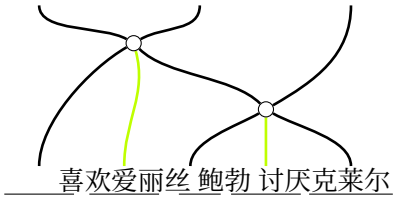
对于这些内容，我们可以将每棵树替换为一个文本图示：



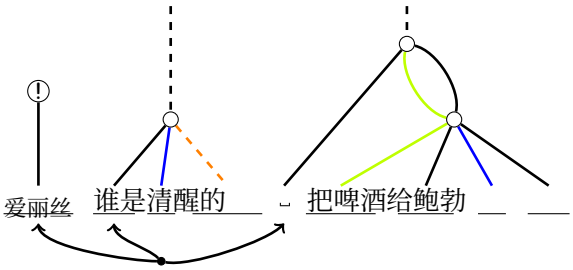
现在，我们可以用文本图中的对应形式替换代词性连接。现在，鲍勃（这是一个专有名词，而不是代词WHO）标注了唯一可用的输出线：



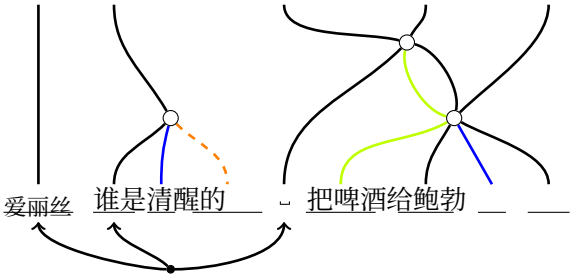
现在可以将代词性连接重写掉：



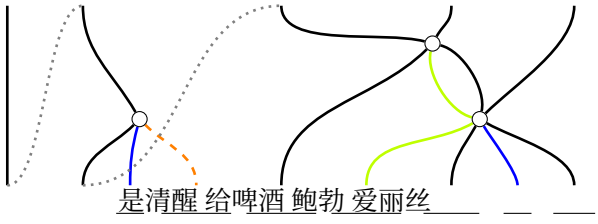
例3.5（主语关系代词再探）。 在文本图的情境中，代词链接“解释”了混合语法中为容纳关系代词而必需的伪影。回想一下，对于主语关系代词的特殊规则，我们会得到一个带有！和伪影的结构：



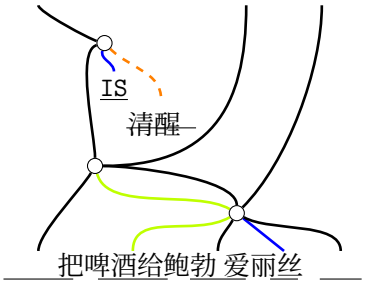
转到文本图时，！伪影就消失了：



现在，我们用文本图对应项替换代词链接，从而得到 g:



将代词链接重写掉后，我们得到以下文本图：



3.5 将短语作用域表示为短语气泡。

我们必须以不同的方式处理复合句，因为在从带语法的文本过渡到文本图的过程中，我们用依赖于句子的名词线集合取代了句子类型 S。这就需要使用带区域的字符串图。[8] 我们将介绍带区域的文本图生成元，用来标明划定句子子短语边界的界线。对于带句子补语且右侧接短语的动词，我们有一些图示片段，能够让一条名词线从右侧进入有界短语，并从左侧离开：



对于同样在左侧接短语的连词，我们有一些重写规则，能够让一条名词线从左侧进入有界短语，并从右侧离开：



在这两种情况下，我们都还需要一条重写规则来“封顶”短语区域。严格来说，我们必须区分左侧短语和右侧短语，因为它们的封顶方式不同；但在图示中，我们省略了这一区分，因为视觉上很容易辨认：

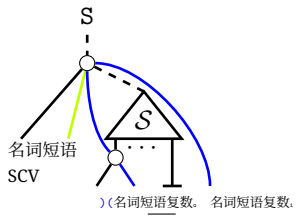
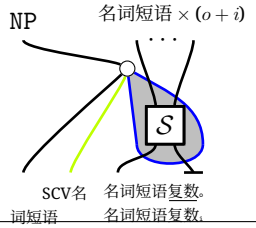
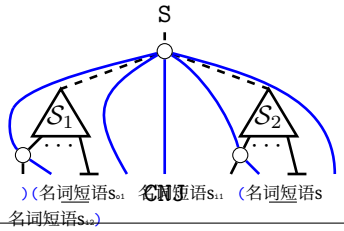
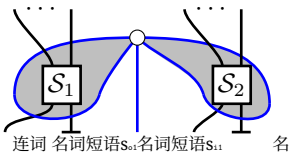


约定 3.6 （短语作用域规则）。 我们在此说明短语作用域和短语区域的图示惯例。首先，短语作用域区域在文本图中表现为平面障碍，但名词短语线和代词链接除外。这会带来两个结果：出现在短语作用域内的节点和导线会被“困在其中”；而且多个短语作用域结构可以逐层嵌套，但彼此不会重叠。其次，名词短语标签只能出现在短语作用域之外。

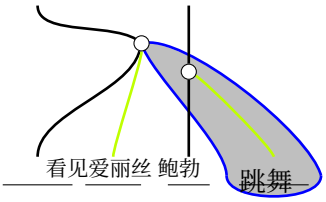
对于语法中的短语作用域结构——带句子补语的动词和连词——为了确保所得文本图的输入名词线与输出名词线数量相等，我们的对应关系必须考虑短语作用域内子句中带标签的名词线（用“脚”绘制）的数量。我们在名词标签下标中加入索引 i 和 o，分别表示“作用域内”和“作用域外”。对于带句子补语的动词和

[8] 此处我们不作阐述，但这里展示的所有内容都可以在结合的 n-范畴 [23], 框架下表达，事实上也可以在证明助手中进行原型设计和实现 [65]。

连词，我们的图示对应如下：

语法	Rule	图示
	句子补语动词	
	连词	

例3.7. 对于爱丽丝看见鲍勃跳舞，我们现在得到：



在这个例子中，句子子短语“鲍勃 跳舞（句子）”被“包含”在一个灰色的“短语泡泡”中。回顾约定 3.6，尽管文本图通常不受平面性限制，但我们规定，除了名词线（它们有特定的图示，允许其穿过边界）之外，任何线都不得穿透“泡泡”的边界；这样，短语区域便捕捉到了短语作用域作为子树，或更一般地说，在依存语法中的行为。

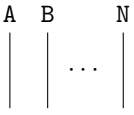
4 文本电路

4.1 定义

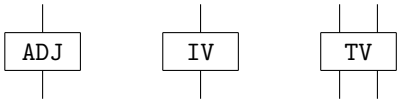
我们的文本电路由三种要素构成：

- 线
- 方框，或门
- 带孔方框，可容纳一个方框，或二阶门

首先，名词用导线表示，每个“不同的”名词都有自己的导线：

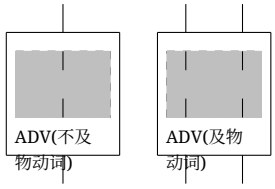


我们表示通过作用于名词线的门来表示形容词、不及物动词和及物动词 es:

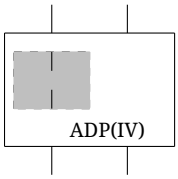


由于及物动词同时有主语和宾语名词，因此会有两条名词线，而形容词和不及物动词只有一条。

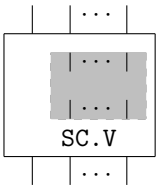
修饰动词的副词用带孔的方框表示，孔中垂下的若干根线表示所需门的形状；这些线应与整个方框的输入线和输出线相匹配：



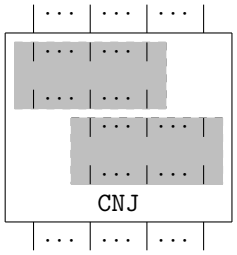
类似地，介词也会修饰动词，此外还会在右侧添加另一根名词线：



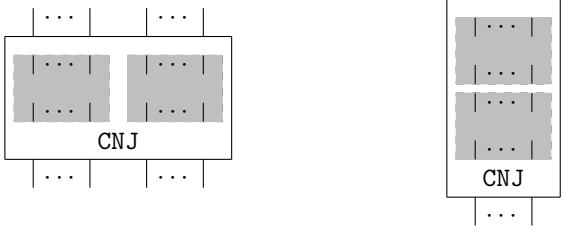
对于带句子补语的动词和连词，我们有一系列方框，以适配任意大小的输入电路。它们会在电路左侧添加另一根名词线：[9]



而连词是方框，它们接收两个电路，这两个电路可能在线上共享某些标签：

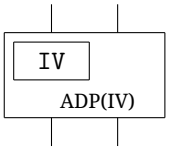


作为特殊情况，当两个电路的名词线互不相交（左图）或重合（右图）时，连词的表示如下：

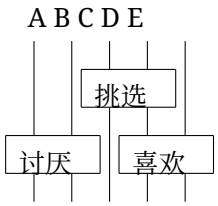


[9]方框内的“点点点”记法在图形上是形式化的 [84]。此类方框的解释此前已在 [52, 63, 86] 中形式化。

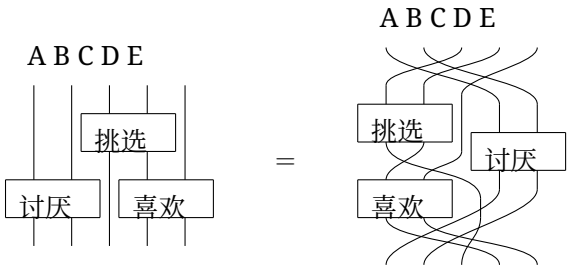
当然，填满的方框就是门



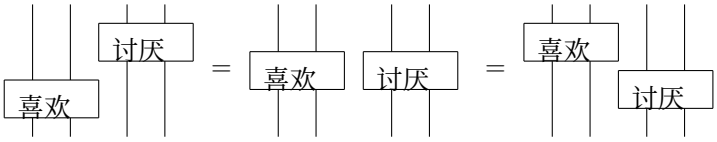
现在我们将讨论它们如何组合。门可以通过匹配部分名词线上的标签进行串联组合；如果它们没有共享名词线，则可以并联组合，从而形成文本电路。按照惯例，我们从上到下阅读文本电路：



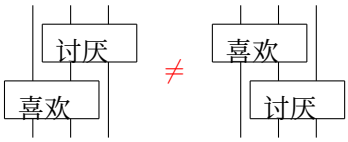
约定 4.1. 有时我们允许导线彼此交叉，并且如果两个电路的门的连接关系相同，就认为它们相同：



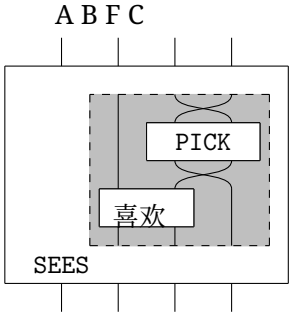
由于只有门的连接关系才重要，因此，只要电路之间的差异仅在于并联门的水平位置，我们就认为它们相同：



我们确实关心输出到输入的连接关系，因此特别是，不认为电路在顺序组合的门彼此交换位置后仍然相等：



例4.2。 句子“爱丽丝 看见 鲍勃 喜欢 花 THAT 克莱尔 挑选”可以直观地表示为以下文本电路：



例4.3. 图2通过提供一个示例，说明如何直接生成文本电路；该示例类似于图1中使用混合语法生成的文本。

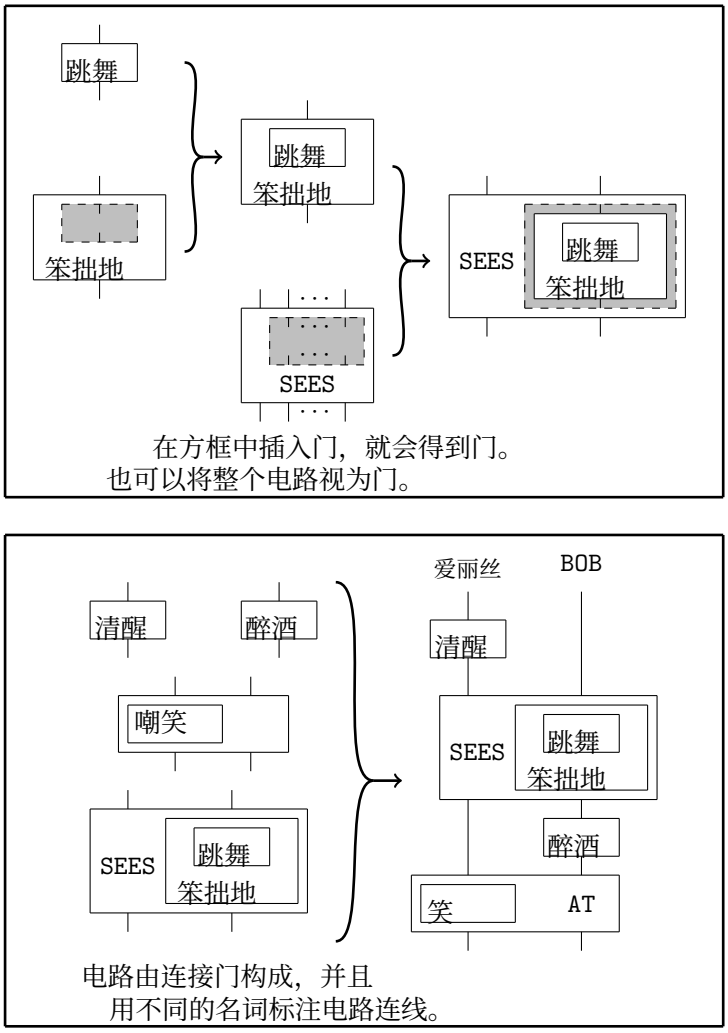


图2：直接生成文本电路。

5 数学结果

5.1 MainText 电路定理

我们假设有一个有限的词汇表，其中包含迄今为止考虑过的语法类型的词。令 $\mathfrak{T}$表示由我们的语法生成的所有原始文本的集合。令 $\mathfrak{C}$ 表示所有文本电路的集合。我们证明以下结果：

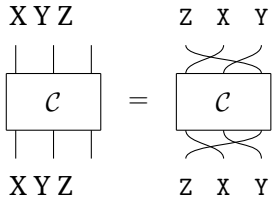
1. 我们证明，第5.3节的翻译规则可以通过算法将任何带有语法的文本与一个文本电路关联起来。2. 所有文本电路都可以通过这种方式获得。

定理5.1。 第5.3节的翻译规则定义了一个满射 $\mathfrak{T} \twoheadrightarrow \mathfrak{C}$。

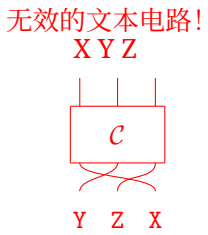
5.2 修订、扩展与惯例

在证明该定理之前，还需作出一些澄清。该定理所论断的是所有 文本电路，以及所有 由混合语法生成的原始文本。约定 5.2 澄清了就我们的目的而言，文本电路究竟如何界定。我们在细化 5.3 中处理一种边缘情况电路——其中一条导线出现在一个方框内，却没有连接到任何门。对于引理 5.15 的证明策略（该策略依赖于一个将多个门并行视为单个句子的步骤），我们通过通用连词 [&] 来容纳逗号等文本元素，以及 AND ALSO 这类“无内容”的连词。我们在细化 5.5 中指出，混合语法中的短语辖域区别可以通过通用的无内容补语标记 [THAT] 在原始文本中恰当地表达。最后，我们扩展文本电路的定义，将“反身代词框”纳入其结构组成部分。

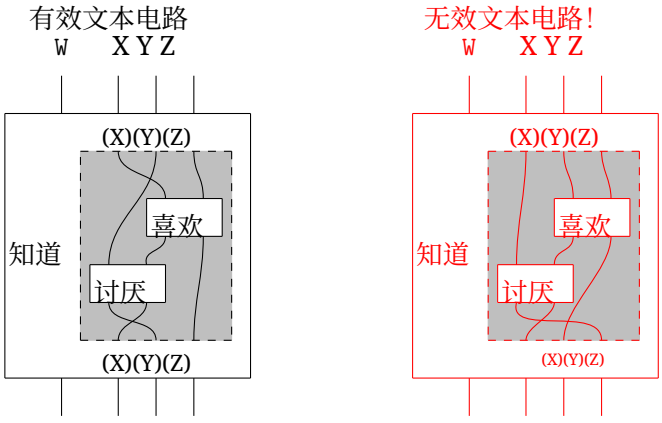
约定 5.2 （导线排序）。 我们对用于描绘电路的约定 4.1 作进一步阐释。电路的每根导线都标有一个名词；在图示中，这些名词标签被强制按从左到右的顺序排列。按照惯例，通过在必要处引入导线扭转，使电路的输入和输出保持相同的顺序。请注意，连接关系相同但名词标签顺序不同的两个电路，仍被视为相等：



作为一个反例，下面的图示 **违反了** 视觉惯例，因为输入导线和输出导线的名词顺序不同：

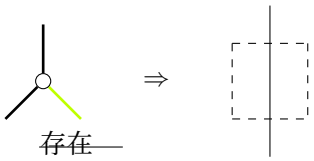


方框中的内容也属于电路，必须遵守相同的输入输出一致性规则：



细化 5.3 （动词 存在）。 我们引入一条特殊的结构规则，用于处理不参与任何门的恒等线。在这种情况下，我们将原始文本层面的恒等线解释为特殊的及物动词“存在”。从文本图到电路的“存在”重写规则如下

如下：



精化 5.4（使用 [&]进行句子复合）。 到目前为止，我们主要考虑将句子列为文本来组合它们，例如：

ALICE RUNS. BOB DRINKS.

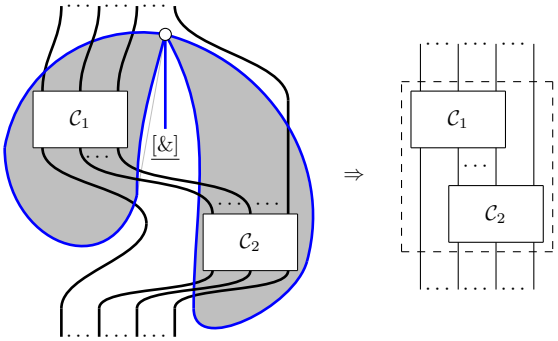
为了证明引理 5.15，我们在文本层面定义了一种特殊连词 [&]，它使我们能够以明确无歧义的方式，将多个门视为由单个句子并行产生。在英语中，有多种不带具体内容的方式可以将两个句子组合成一个句子。例如，使用与号：

ALICE RUNS [&] BOB DRINKS.

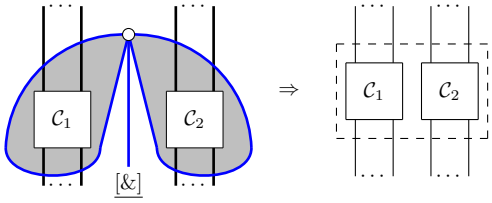
或者使用诸如“而且”之类不带具体内容的并列短语，例如：

ALICE RUNS AND ALSO BOB DRINKS.

特殊连词 [&] 旨在概括这类不带具体内容的连词。在电路层面，[&] 被解释为门组合，其从文本图到电路的重写规则如下：



请注意，当名词线彼此不相交时，[&] 的重写规则只会得到电路的并行组合：



细化 5.5（短语作用域内的句子复合，以及补语标记 [THAT]）。 考虑以下句子：

克莱尔 看见 爱丽丝 RUNS [&] 鲍勃 DRINKS。

对我们而言，当原始文本配有混合语法结构时，克莱尔究竟是看见了爱丽丝跑步和鲍勃喝酒这两件事，还是只看见前者，并不存在歧义。回想一下精化 5.4，将 [&] 读作逗号或 AND ALSO，都无法消除这种歧义。由于我们的总体主张是，我们处理的是自然语言的一种限制，因此必须证明，混合语法结构所作的这种区分，在原始文本层面确实有对应形式。为了像先前的精化那样消除歧义，我们引入补语标记 [THAT],，其作用方式

与第2.4节中引入的短语作用域形式类型（ 和 ）非常相似。补语标记是 HOW 或 WHAT 这样的词，它们位于句子补语之前，例如：

CLAIRE SEES HOW ALICE RUNS.

CLAIRE SEES WHAT BOB DRINKS.

我们在此不讨论一般意义上的补语标记词。相反，我们只使用一个“无内容”的补语标记[THAT]，，当它隐含时可以省略。回到最初的例子，补语标记的存在使我们能够在原始文本层面区分这样一种情况：克莱尔**只**看见爱丽丝在跑：

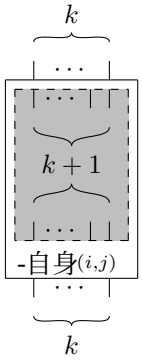
CLAIRE SEES [THAT] ALICE RUNS [&] BOB DRINKS.

与克莱尔看见**既**爱丽丝在跑、又鲍勃在喝酒的情况：

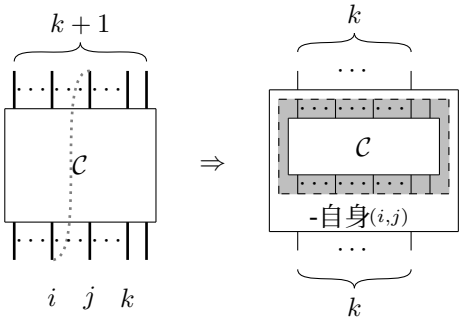
CLAIRE SEES [THAT] ALICE RUNS [&] [THAT] BOB DRINKS.

所以，当我们在短语作用域内遇到连词 [&] 时，我们在文本层面将其解释为“并且也.....”。更一般地说，我们用 [THAT] 表示原始文本中标志作用域短语起始的左括号。

约定5.6（反身代词框）。我们扩展文本电路的定义，以容纳反身代词，例如 HERSELF、ITSELF。在其他所有情况下，我们都将代词链接视为门组合；但根据定义，当电路中的一个输出名词线必须与其相对过去位置上的一个输入名词线复合时，就会出现反身代词，而这种情况无法画成电路。因此，在电路层面，我们设定一族“反身代词框”。对于所有 $k > 1 \in \mathbb{N}$，以及所有成对的索引$1 \leq i < j \leq k$，我们都有一个方框：

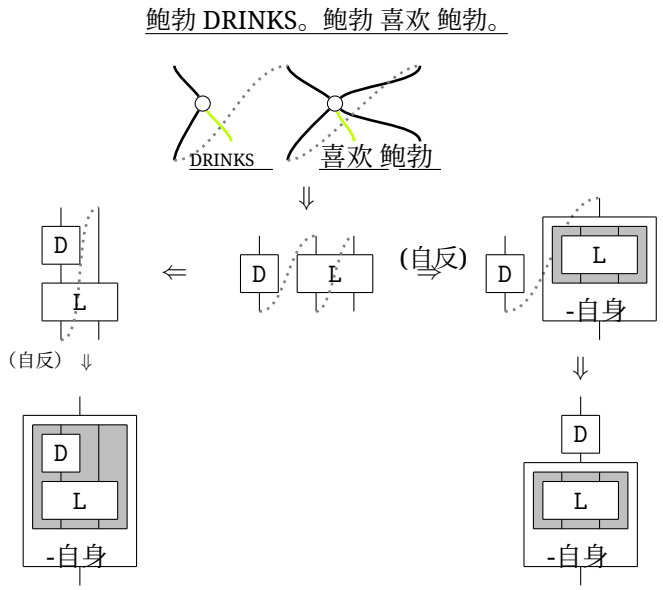


这些方框是文本图重写的目标：当消去代词性连接会将电路的一个输出连接到一个输入，从而使名词线“折返回去”时：

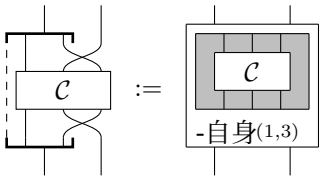


细化5.7 (反身代词框的一致性)。 重写系统中的临界对 是指这样一种情况：两条规则均适用于某个项，并产生不同的最终结果。因此，临界对是命题5.14所述主张的唯一障碍。所有临界

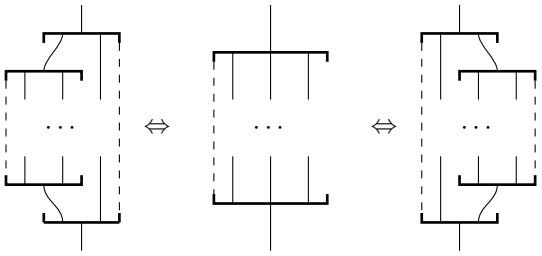
在将文本图重写为电路时出现的配对。考虑以下示例，其中的文本图对应于这段文本：



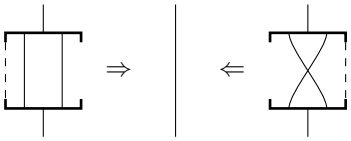
因此，代词链接消除的顺序以及（作为其特殊情况的）反身代词框引入的顺序十分重要。正如我们在推论 5.9 中所示，通常可以通过约定来处理这些临界对：所有反身代词框引入都先于其他代词链接的消除。回想第 2.3 节的脚注，文本图中的代词链接可以有两个方向。对于反身代词而言，这两个方向都不完全合适，因为代词链接应当识别两条导线。因此，我们在此为反身代词框引入语法糖，并引入重写规则，以确保导线的识别得到遵守。下面给出反身代词链接的简写形式：



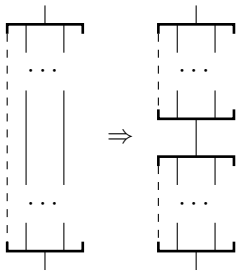
I标识导线具有结合律，我们通过以下双向重写来强制执行这一点 :



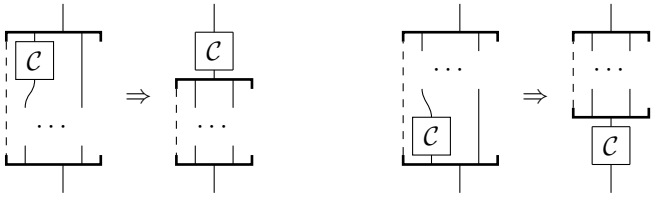
自反 性代词方框内只有恒等导线时，可以通过恒等规则将其消去 :



反过来，我们可以沿着已识别导线拆分现有的反身代词框，从而引入新的反身代词框：



最后，反身代词方框有一个特殊之处：单根导线上的门可以从中滑出。这反映了这样一个事实：只修饰单个名词的句子从不需要反身代词，因为反身代词会指向另一个名词论元：



5.3 从文本图示到文本回路

本节介绍若干引理，它们共同构成了命题5.14证明中的核心工作。该命题声称存在一个函数 $\mathfrak{T} \to \mathfrak{C}$。

首先，引理 5.8 和推论 5.9 以构造性的方式将细化5.7中的重写整理为一个函数，该函数将文本图示映射为不含代词链接的文本图示。其次，引理 5.11 引入了新的重写规则和辅助类型，以构造性方式定义一个函数，将不含代词链接且不含短语作用域的文本图示转化为与单个电路门相对应的范式。第三，引理 5.12 扩展了引理 5.11，使其也涵盖具有短语作用域的文本图示。结合第3节的内容——该节提供了一种对应关系，可将每个混合语法文本映射为一个文本图示——上述三个阶段共同完成了从混合语法文本到文本电路的函数描述。

引理 5.8（缩减反身代词）。 反身代词的结合律、恒等律、拆分和滑动重写规则，足以将文本电路重写为一种形式，使每个反身代词方框恰好包含一个门。

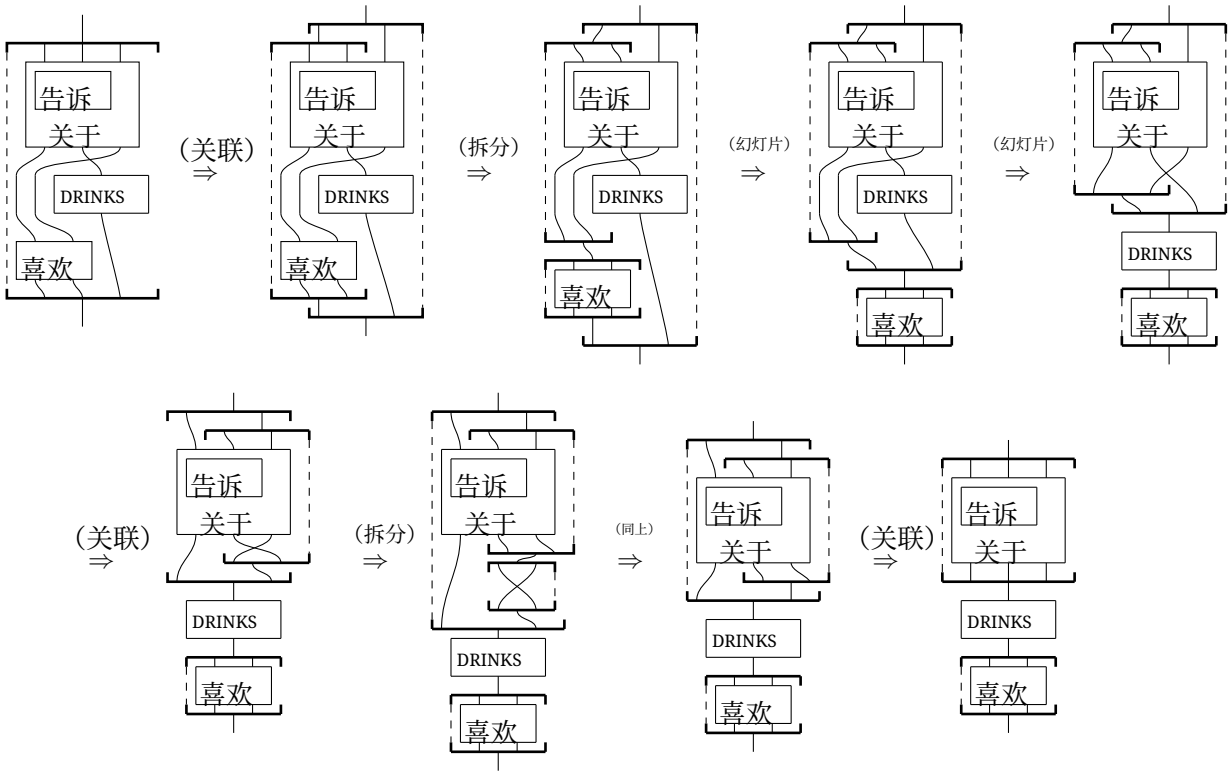
证明。 嵌套的反身代词方框可以根据结合律规则融合并重新排列。任何包含两个或更多门、且这些门沿一条已识别导线依次组合的反身代词方框，都可以根据拆分规则进行拆分。对于仅包含一个门的反身代词方框，可以应用拆分和恒等律规则消去扭结，使输入和输出对齐。 □

推论 5.9。 通过缩减引理得到的文本电路，与从文本图示得到的文本电路相同；在后者中，反身代词方框的缩减先于其他代词链接消除而应用。

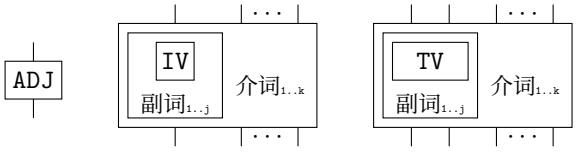
证明。 根据构造，反身代词只出现在句子内部，而句子对应于单个门。引理 5.8 的重写规则保留门连通性，而门连通性由非反身代词链接决定。 □

例 5.10（收缩引理的一个应用）。 考虑一个释义为以下文本的文本电路：

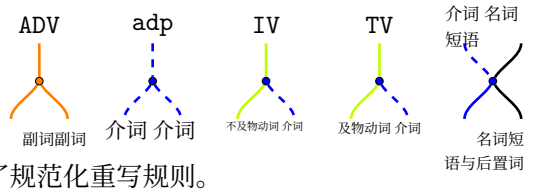
鲍勃告诉鲍勃关于鲍勃。鲍勃 DRINKS。鲍勃喜欢鲍勃。



引理 5.11。 我们在表 3 中给出辅助类型和重写，使得每个不含代词连接且不含短语作用域的文本图都可以视为由以下形式的门构造而成的唯一文本电路：

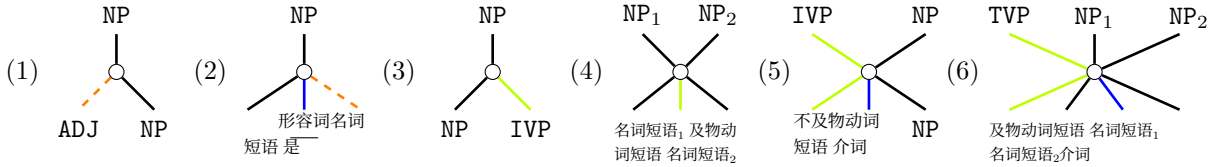


证明。 我们引入一种辅助线类型 adp，以及以下辅助文本图组件。



我们在图3的表格中给出了规范化重写规则。

文本图的名词短语线对应于文本电路中的名词线。对于来自不含代词链接和短语作用域的混合语法结构的文本图，涉及名词短语线的非标签文本图节点有六个；每个这样的节点都对应一个门或方框。



IS消去规则将所有情形（2）节点的实例重新表述为情形（1）节点。根据约定 3.3，文本图中只有连接关系才重要，因此我们可以收缩与每个

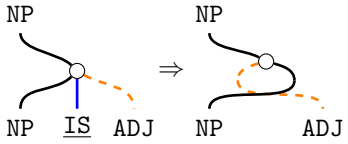
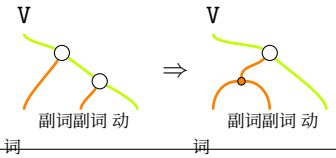
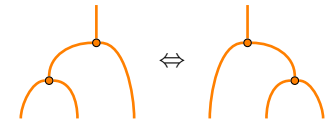
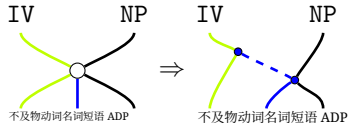
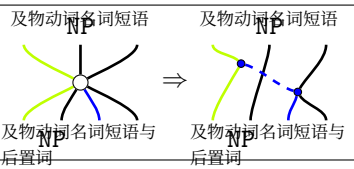
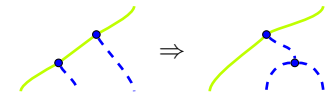
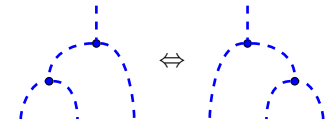
规则名称	文本图重写
IS 消去	
ADV 聚合	
副词关联	
不及物动词 ADP 辅助项	
及物动词 ADP 辅助项	
adp-gather	
后置词结合律	
后置词-副词顺序	

图3：门归一化重写

情形（1）实例相关联的形容词标签，并将其解释为文本电路中的一个门，如下所示。



情况 (5) 和 (6) 是介词，分别只能在存在 IV 或 TV 线时出现。为简便起见，我们将它们统称为 V 线。这类线只能分别通过情况 (3) 和 (4) 出现，而它们只能由 V 标签终止。

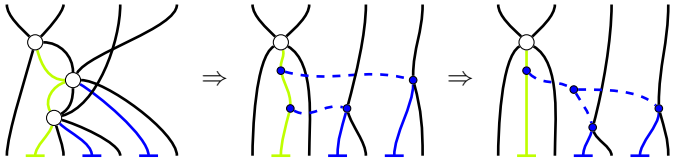
我们无法像处理 ADJ 标签那样，直接对 V 标签重复使用相同的收缩技巧，因为在情况 (3) 或 (4) 的节点与其对应的 V 标签之间，可能存在 ADV 节点，或像 (5) 和 (6) 中那样存在 ADP 节点。不过，所有这类节点都有一条输入 V 线和一条输出 V 线，因此所有 ADV 和 ADP 节点及其标签，都可以明确地对应到一个 V 标签。

辅助类型和重写旨在为收缩到其父 V 节点的 ADV 和 ADP 节点及标签求得范式；这些父 V 节点属于情况 (3) 或 (4)。对于给定的 V 线，我们按以下顺序分析三种可能的情况，以说明这些规则的作用：仅出现 ADV 节点；仅出现 ADP 节点；两种节点都出现。

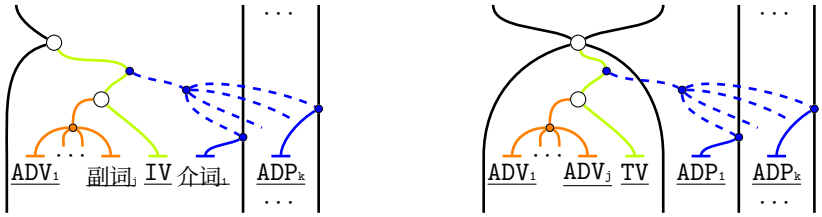
在第一种情况下，如果同一条 V 线上相邻出现两个 ADV 节点，我们就应用 ADV-gather 规则。对于同一个动词有多个副词的情况，ADV-assoc 双向重写规则使 ADV-gather 的应用顺序无关紧要。收缩标签后，这些规则允许将同一条 V 线上的多个 ADV 节点重新表示如下：



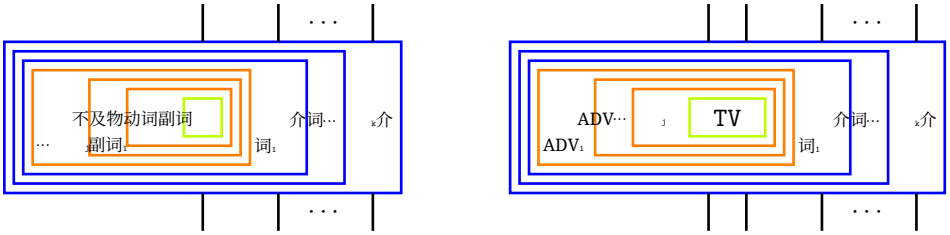
在第二种情况下——这也涉及类型 (5) 和 (6) 的节点——如果存在多个 ADP 节点，我们首先应用 ADP-ancilla 规则，引入 adp 辅助类型。然后，我们可以应用 adp-gather，以与处理 ADV 线类似的方式进行：对于同一个动词上有多个介词的情况，adp-assoc 规则同样使 adp-gather 重写的顺序无关紧要。这些规则允许将同一条线上的多个 ADP 节点重新表示如下；这里以两个 ADP(TV) 为例说明。



在 V 线上同时存在 ADV 和 ADP 节点的第三种也是最后一种情况中，我们首先通过 ADP-辅助节点重写，消去所有类型为 (5) 或 (6) 的 ADP 节点。然后应用后置词-副词顺序规则，可以将 V 线上所有 ADP 节点排列在所有 ADV 节点上方，接着按照前两种仅含 ADV 节点或仅含 ADP 节点的情况进行处理。将 ADV 和 ADP 标签收缩后，这一过程分别为每个类型为 (3) 和 (4) 的节点得出以下范式。



Whi分别可以读作以下形式的门，并以不同颜色标示以便辨认。



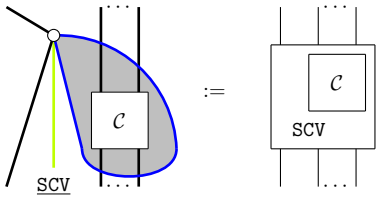
这些与论断中的最后两个门仅在语法糖方面有所不同。所得电路的唯一性成立，因为这些规则都不会改变类型为 (1) 至 (4) 的节点之间的相对连接关系。

□

引理 5.1 **2.** 每个不含代词链接的文本图都可以视为一个唯一的文本电路 it.

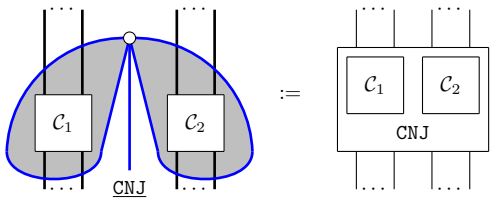
证明。 我们只需扩展引理 5.11 的论断，使其涵盖短语作用域。回顾约定 3.6：短语作用域只允许名词短语线和代词链接穿过；名词短语标签只能出现在短语作用域之外；并且图示中任意位置的短语作用域嵌套都没有歧义。我们通过（强）归纳法证明该论断，归纳假设为：所有嵌套深度至多为 k层的短语作用域文本图，都可以视为唯一的文本电路。对于没有短语作用域的基础情形，由引理 5.11 可知结论成立。进行归纳时，假设我们有一个短语作用域嵌套深度为 $k+1$ 的文本图。最顶层的短语作用域节点要么是 SCV 节点，要么是 CNJ 节点。

带句子补语的动词 由于进入短语作用域的名词短语线数量必须与离开的数量相同，根据归纳假设，我们可以将短语作用域的内容表示为文本电路 $\mathcal{C}$，并可明确地读出如下结果（忽略副词以及此前讨论的 SCV 标签所用的介词）。

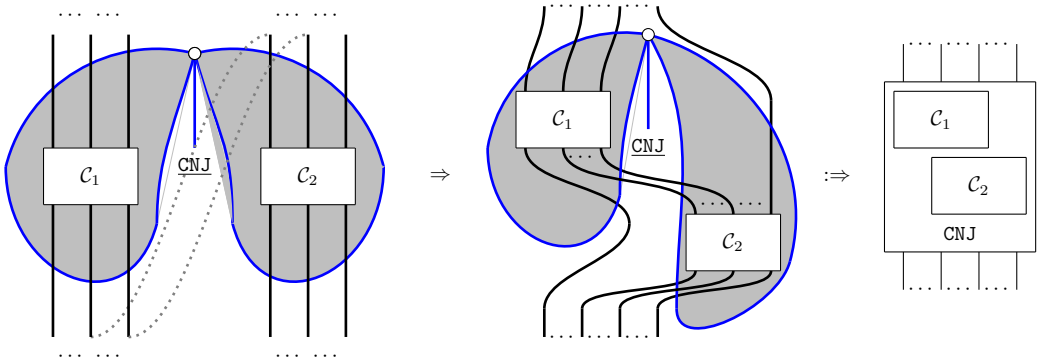


文本图的其余部分可由归纳假设处理。

连词 与之前类似，我们可以将左侧和右侧合取项的内容分别表示为文本电路 $\mathcal{C}_1$ 和 $\mathcal{C}_2$。我们依次考虑两种情形：合取项不共享名词短语线，以及合取项共享名词短语线。在前一种情形下，我们从一个 CNJ 节点得到以下门。



在后一种情形中，两个合取项共享名词线——或者，预见后续内容，合取项之间存在代词链接——我们可以引入一个参数重叠的共轭框，作为反身代词框内“常规”合取框的语法糖：



□

示例 5.13。 图 4 展示了如何从文本图示中获得文本电路。

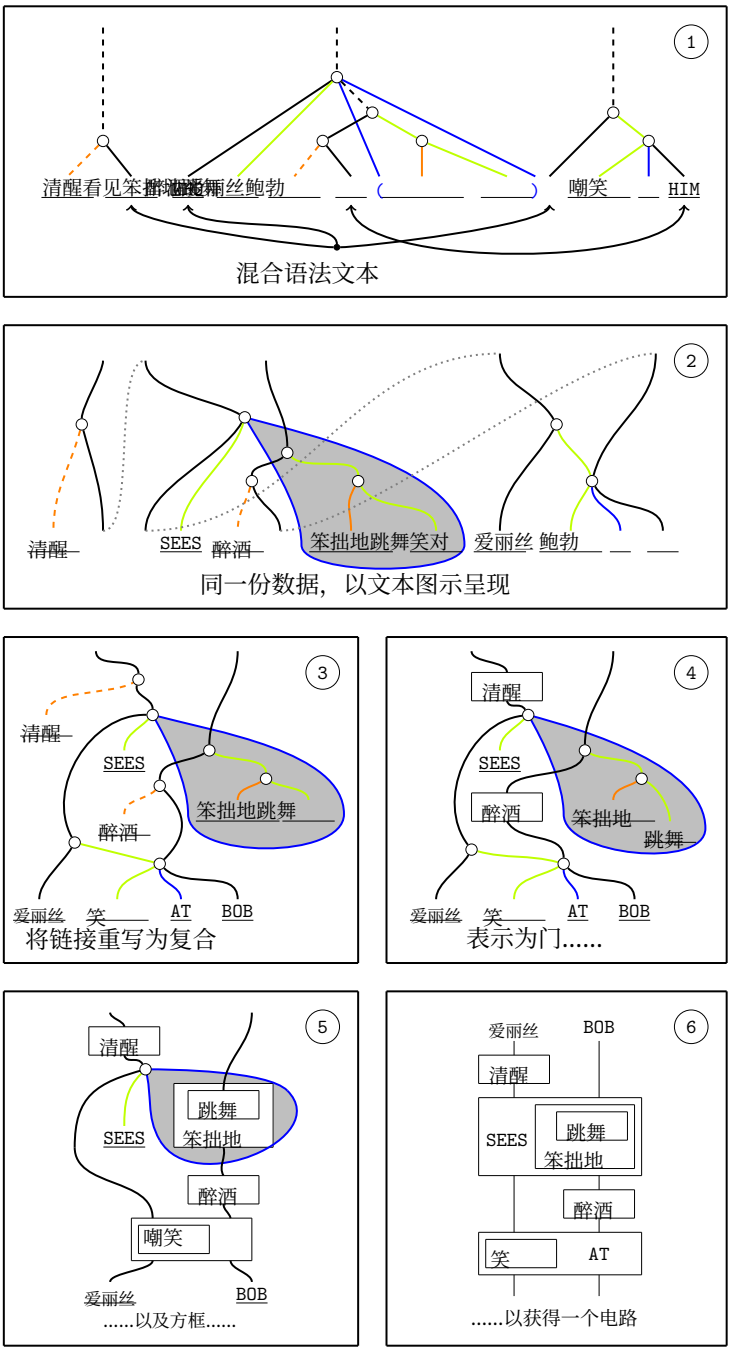


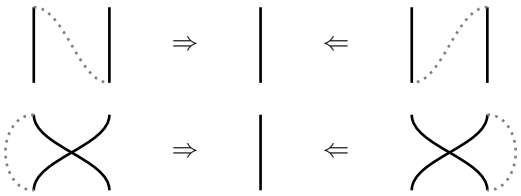
图 4：从文本图中获取文本电路。

5.4 定理的证明

只需证　　明以下内容。首先，从文本到电路的翻译过程给出　　n
在第5.3节中产 生一个函数 $\mathfrak{T} \to \mathfrak{C}$。其次，该函数是满射。

命题5.14。 我们有一个函数 $\mathfrak{T} \to \mathfrak{C}$。

证明。 引理 5.12 处理了所有由不含代词链接的文本得到的文本图。接下来需要证明，在文本图中，代词链接可以被唯一地消解。以下变换会消除代词链接。

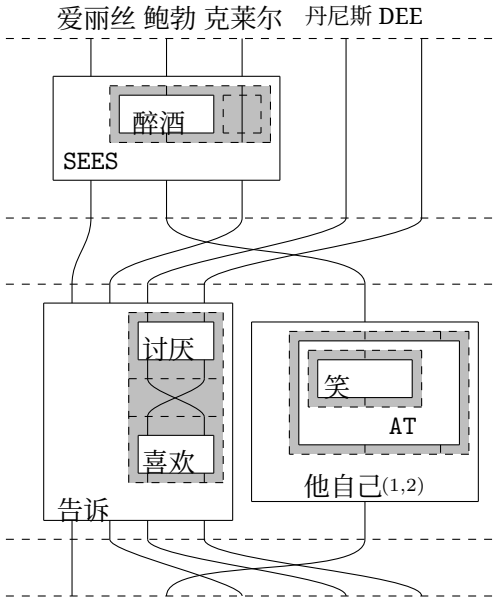


要将这些变换直接解释为文本电路的复合，唯一的障碍由约定5.6中的反身代词框解决。由反身代词框产生的临界对由细化5.7解决，而引理 5.8 则以构造性的方式给出了反身代词框的范式。因此，初始混合文本决定了相应的电路。

□

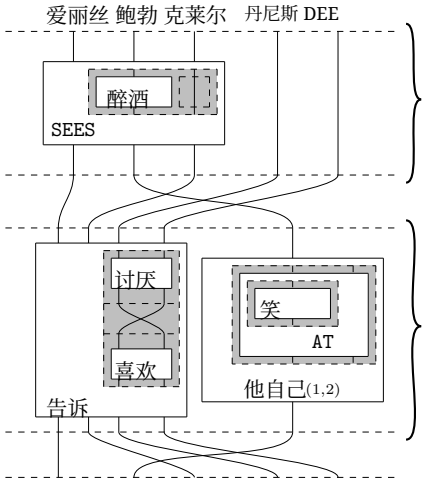
命题5.15。 $\mathfrak{T} \to \mathfrak{C}$ 是满射的。

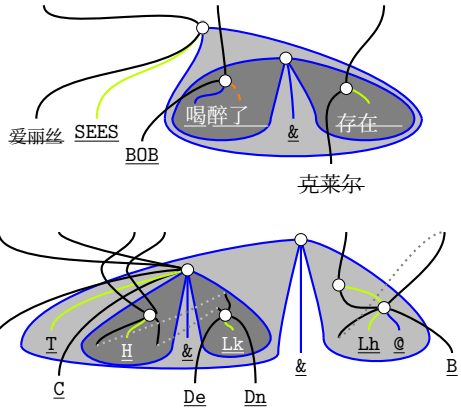
证明。 满射性意味着要证明，任意电路都对应一个能够转换成该电路的文本；换句话说，每个文本电路都可以文本化。首先，我们将电路划分为水平切片，使得每个切片要么至少包含一个门，要么只包含扭转线。这里我们利用细化 5.3：当方框上下文中的一条导线没有连接到任何门时，我们假设它连接到一个存在门。我们在方框的孔洞中重复进行这种切片：



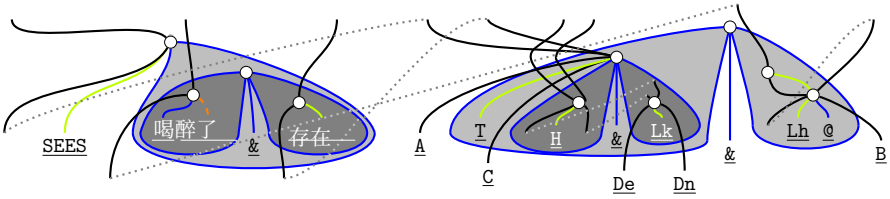
现在，我们运用精化 5.4；由此得到的每个带有门的水平切片，都是由若干门并联组成的集合，其中可能包含恒等导线。对于方框孔洞内未连接到门的恒等导线，我们将其视为连接到一个存在门，例如示例中 SEES 的孔洞。其他恒等导线无需在文本中提及。我们通过使用系词 IS 结构来处理形容词门，例如示例中的 DRUNK。我们通过引入

适当的代词性连接，并重述名词，来处理反身代词，例如示例中的 HIMSELF。我们为每个带有门的切片生成一个文本图：

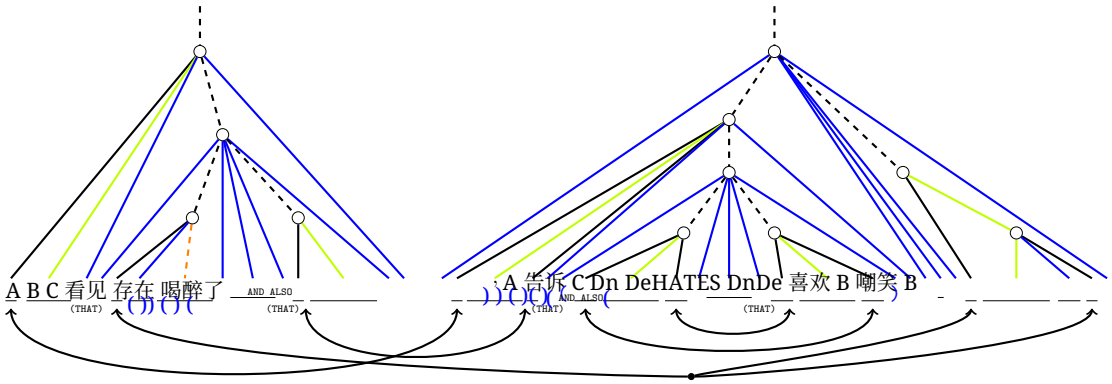




然后，我们用代词链接连接每个切片的图示，这些链接映照了线扭结切片的连接关系：



文本中代词链接的结构与电路的组合结构相吻合。根据第3节中将文本图与混合语法联系起来的规则，以及精化 5.4 的解释规则，我们得到如下混合语法文本：



ALICE SEES THAT BOB IS DRUNK AND ALSO THAT CLAIRE EXISTS.
ALICE TELLS CLAIRE THAT DENNIS HATES DEE AND ALSO THAT DEE LIKES DENNIS, BOB LAUGHS AT BOB.

按构造，我们的重写规则会将文本还原为起始电路。 □

示例 5.16. 图 5 展示了一个文本电路的文本化示例，该电路曾出现在我们之前所有类似的插图中。

5.5 借助方程扩展语法

文本电路定理的一个直接推论是，解析文本上存在一种非平凡的等价关系。用文字来说：如果两个文本生成相同的文本电路，我们就认为它们等价。

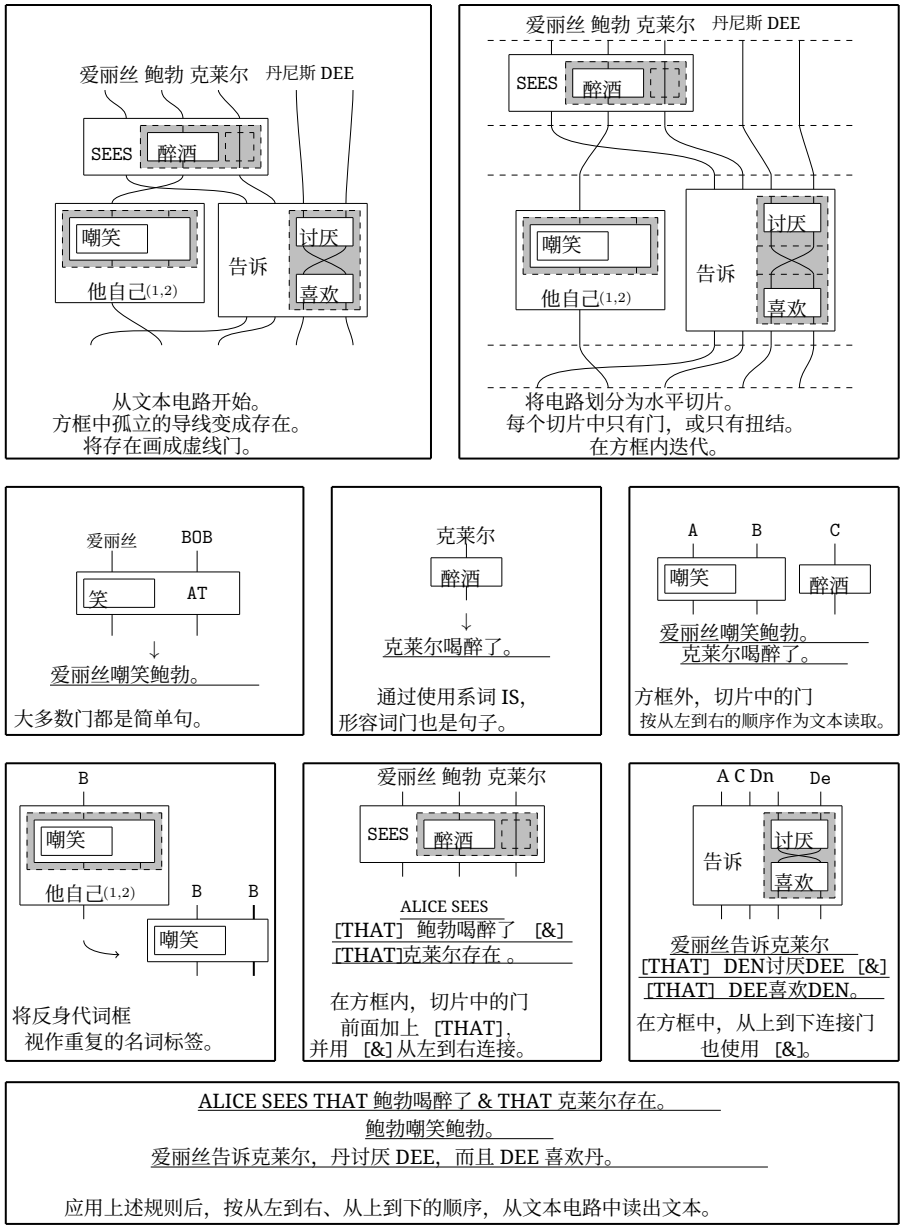


图5：电路文本化示例。

所以，我们可以如下定义文本等价 $\mathcal{T}_1 \equiv \mathcal{T}_2$：

$$\left.\begin{array}{l}\text{exists rewrite } \ \mathcal{T}_1 \Rightarrow \mathcal{C}_1 \\ \text{exists rewrite } \ \mathcal{T}_2 \Rightarrow \mathcal{C}_2\end{array}\right\} \ \text{such that } \ \mathcal{C}_1 =_{\mathfrak{C}} \mathcal{C}_2$$

其中，$\mathcal{T} \Rightarrow \mathcal{C}$ 表示使用语法 $\mathcal{T}$将文本重写为文本电路 $\mathcal{C}$，而 $\mathcal{C}_1 =_{\mathfrak{C}} \mathcal{C}_2$ 表示文本电路相等——参见约定 4.1。

到目前为止，我们已经看到，文本电路似乎能够捕捉自然语言繁复表象之下意义的基本连接关系。基于这一点，我们所定义的文本等价关系看起来很像“意义等价”。现在，我们将把这一观察转化为一个指导性假设。

论题 5.17 (文本电路论题)。 相同的电路代表相同的文本含义。

这篇论文为扩展我们的语法提供了一套方法，以涵盖我们在此尚未讨论的语法现象。下面我们将使用 [ING] 明确演示如何纳入被动语态、物主代词和动词形容词化。

当人们普遍认同 $\mathcal{T}_1$ 和 $\mathcal{T}_2$ “意思相同”时，我们就假定情况确实如此，并据此反向推导文本图中应包含哪些部分，以及相应语法的重写规则应是什么。如果 $\mathcal{T}_1$ 使用了某种语法现象 $\mathcal{GP}$，超出了我们迄今所考虑的范围，而 $\mathcal{T}_2$ 则属于我们的图示/语法范围，那么我们就通过添加 $\mathcal{GP}$的文本图部分及相应的重写规则，使 $\mathcal{T}_1$和 $\mathcal{T}_2$ 的电路相等。以下示例最能说明这一过程，前三个示例详见 [21]。[10]

例5.18 （被动语态）。 在英语中，及物动词采用被动语态时，主语和宾语的顺序会颠倒。例如，下面这个被动语态句子：

鲍勃 被 喜欢 爱丽丝

表达的事实与下面这个主动语态句子相同：

爱丽丝喜欢 鲍勃

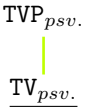
为了处理被动语态，我们引入一种新的线类型：及物动词短语（TVP）$_{psv.}$。在下图中，左侧我们引入一个字符串图，使 TVP 能在短语作用域内变为 TVP$_{psv.}$；右侧则引入一个字符串图，通过标注 IS 和 BY 来闭合被动语态短语作用域：



我们还要求每个标签：



都有一个对应的被动语态形式：



例如，喜欢变为 LIKED。我们还规定，除上方左侧的图示外，TVP$_{psv.}$对于所有其他图示都模仿及物动词短语。这样，我们就可以进一步生成被动语态辖域内的文本语法/图示，从而得到例如：

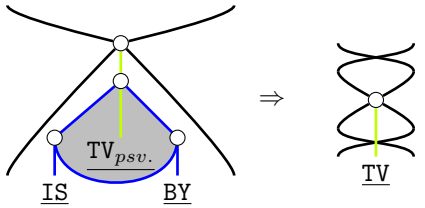
鲍勃深受爱丽丝喜爱

作为一点说明，左图的例外以及单独设置及物动词短语$_{psv.}$ 是必要的，这样才能避免被动结构被反复应用；例如，否则就会过度生成**不合乎语法的**：

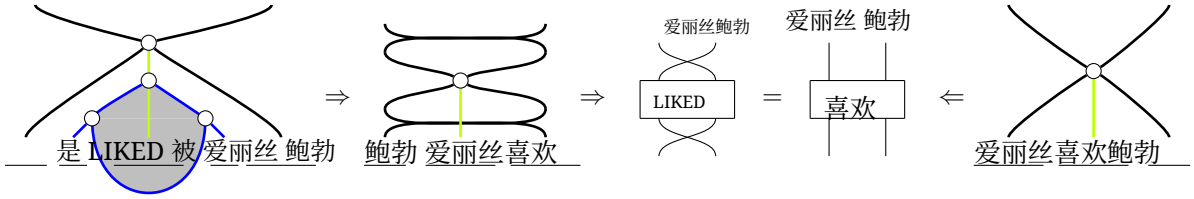
BOB IS IS LIKED BY BY ALICE

[10]在 [21] 中，我们没有像这里这样通过假设方程来扩展图示/语法，而是使用合乎语法的结构和所谓的“内部连线” [67, 68, 28, 39, 17, 13, 18] ，来推导句子之间的方程。事实上，我们正是沿着这条路径得出了本文呈现的结果，因此我们也将相应的功劳归于这些内部连线。即将发表的一篇论文将专门探讨“内部连线”这一广泛主题及其与本文所呈现结果的关系 [77], ，我们还会在第6节中简要讨论它们。

回到这个例子，我们根据以下重写规则，将被动语态改写为主动语态：



所以我们得到以下文本等价证明：



例5.19 （物主代词）。 请看短语中的物主代词“他的”：

BOB DRINKS HIS BEER

我们可以将名词短语“他的啤酒”改写为名词短语“他拥有的啤酒”，而不改变其核心意义。应用这一转换，我们得到：

BOB DRINKS BEER THAT HE OWNS

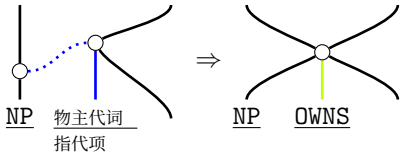
我们知道，此处的关系代词 THAT 会产生与以下文本相同的电路：

BOB DRINKS BEER. BOB OWNS BEER.

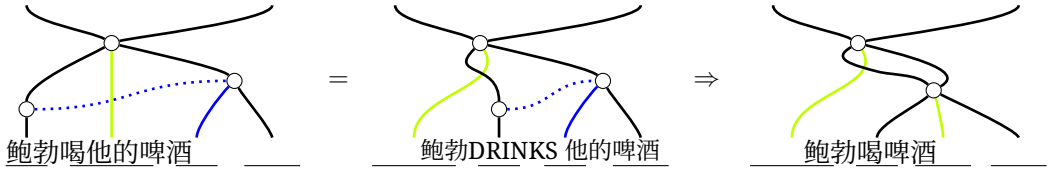
我们可以通过为物主代词标签引入以下生成规则和重写规则，使这段文本等同于原句“鲍勃喝他的啤酒”。我们引入一种点状蓝色的物主代词指代链接类型，其中包含两个生成元，使领属者由物主代词生成器以代词形式指代：

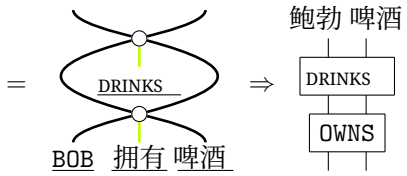


重写 消除物主代词和物主代词指代项的重写规则如下 s:



So, 回到我们的句子，我们有一系列从文本图到文本电路的重写 t:





请注意，虽然 文本生成的文本图是平面的，但在重写过程中，我们可以自由扭转导线 s.

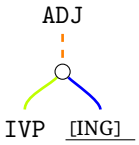
例5.20 （动词通过动名词 -ING 形容词化）。 在动词后添加 -ING，可以使该动词像形容词一样使用。例如，名词短语：

DANC-ING ALICE

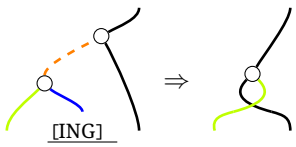
可以改写为名词短语：

ALICE WHO DANCES

我们将 [ING] 建模为一个生成元，它把一个形容词转换为一个不及物动词短语。



化简规则很简单：

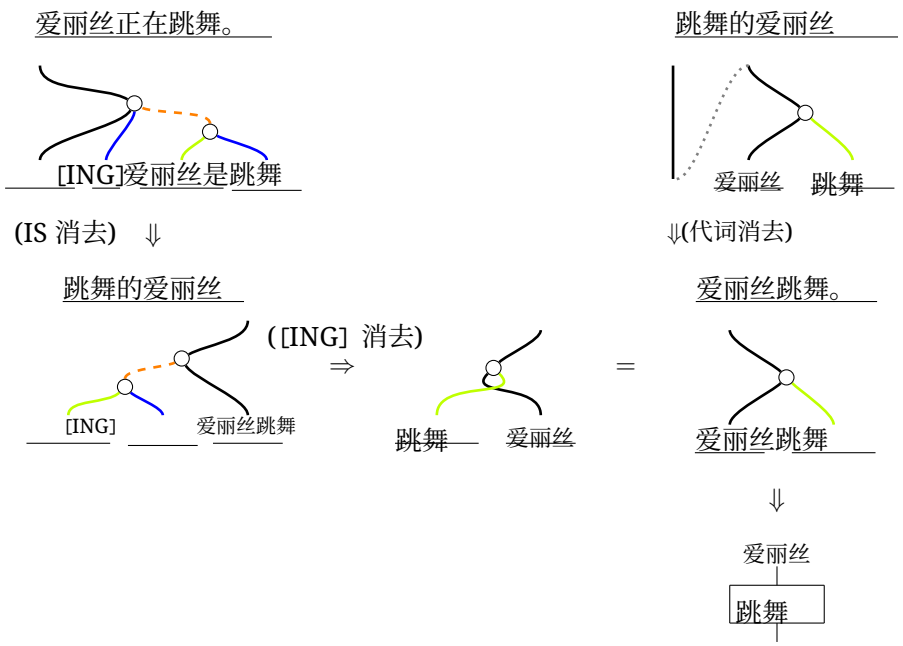


由此，我们得到这些句子之间的等价关系：

- 爱丽丝正在跳舞。
- 爱丽丝跳舞。

以及名词短语：

- 跳舞的爱丽丝
- 跳舞的爱丽丝



6 讨论

6.1 文本解析

正如我们在引言中提到的，我们还希望机器能够从真实文本中提炼出电路。因此，一般而言，提炼过程更可能如下所示:

文本图 文本电路 对应于重写为文本+语法 解析

其中，句法分析是自动为句子和文本赋予合乎语法的及其他结构的过程。如 [49], 所述，常见的依存句法分析器（例如 SpaCy） [30] 在处理简单句时表现足够好——能够处理短语结构和短语作用域这两个组成部分——从而提供可系统地解释为构成文本电路的各个门的输出。就我们的目的而言，要分析真实文本，还需有一种识别代词链接的程序——这一问题称为回指消解，相关算法早已问世 [45]。事实上，现代机器学习语言模型在这项任务上的表现已接近人类水平 [8]。含有关系代词的句子可以通过模式匹配，系统地分解为多个由代词链接的简单句。由此可得到一系列通过代词指代相连的简单句。

6.2 任意文本

请回想一下，我们在这里讨论的只是语言的一个片段，但我们已经证明，可以通过隐式方程扩展这一片段，以涵盖语言的更多方面。要分析任意文本，可能需要先依据这些隐式方程对文本进行预处理，以得到一系列简单句。为了从质的层面将我们的片段扩展到隐式方程所能实现的范围之外，在未来的工作中，我们还将纳入需要更高阶表示的语法，而不仅仅是作用于门上的运算，从而捕捉例如量词和限定词等现象。

6.3 与其他句法和语义形式体系的关系 S

在本节中，我们开始将混合语法、文本图和文本电路与文献中的一些主流形式体系进行比较和对照。我们的分析主要集中在文本电路上。本节也有助于说明，尽管我们是从混合语法出发得到文本电路的，但也可以

使用其他成熟的形式体系（依存语法、类型逻辑语法、话语表征理论等）作为获得文本电路的起点。事实上，对于希望生成文本电路的实际应用而言，这很可能是最简单的途径。

6.3.1 我们的文本混合语法

从单个句子过渡到文本，关键要素是代词链接数据，我们假设这些数据是通过其他方式获取的——第6.1节。要从形式上将混合语法与文献中现有的形式体系联系起来，还需要开展更多研究；不过，我们在此提出，混合语法属于线性上下文无关重写系统所涵盖的语法类别 [79], ，该类别例如包括线性索引语法 [75]、组合范畴语法 [80], 和中心语语法 [81]。

我们有两个主要理由怀疑这一点。首先，线性上下文无关重写系统中所谓的组合函数与文本电路的门十分相似，因为两者对输入论元都是线性的且规则的。

其次，如果我们假设确定代词链接的算法在与“多项式时间复杂度”（即科巴姆论题）相关的意义上是“可有效计算的”，那么用混合语法对文本进行句法分析显然是多项式时间的。温和上下文敏感性的一种刻画[33] 就是多项式时间句法分析，而线性上下文无关重写系统的发展，也受到以这些术语对温和上下文敏感性作出更精确刻画这一目标的推动。

我们将这些对应关系更深入的形式化探讨留待未来工作。

6.3.2 文本图

与依存语法的关系。 文本图在几何形态和概念上都类似于依存语法 [73], ，更广泛地说，也类似于链接语法 [71]。[49],对这种关系作了详细阐述，并以此为理论基础，提出了一种“手写”的文本到电路解析算法。

本文中使用的文本图比依存语法精简得多，因为它们没有规定那么多语法关系。回想引言中关于“二维性”的讨论，这在一定程度上是因为图示的几何结构“吸收”了某些语法角色，例如主语、宾语和关系代词的指称对象。重要的是，这种极简主义并未以苦行式的简约为代价。正如我们在例5.19中所见，我们可以根据需要添加新的线类型和重写规则，以捕捉语法关系。

与类型逻辑语法的关系。 在正在撰写的工作中，我们将展示单个句子的文本图如何从预群语法中产生 [44], ；众所周知，预群语法与上下文无关文法等价（参见上文） [9], ，并且在应用中与组合范畴语法密切相关 [85]。

与转换语法的关系。 文本图的重写规则是转换语法的一种类型受限类比。众所周知，不受限制的转换语法是图灵完备的 [59] ，因此其能力过强，无法刻画自然语言。我们将文本图的形式化刻画及其在乔姆斯基层级中的定位留待未来工作。

6.3.3 文本电路。

我们概述一些论据，说明文本电路与理论语言学和计算语言学的不同研究路径之间的联系，并在每段开头注明这种联系的性质。

话语表征的精炼。 文本电路所体现的文本（或话语）意义建模方法，与动态语义学视角十分契合。二者共同的理念是：将文本片段视为以新信息更新既有上下文，其结果便是更新后的上下文。在动态语义学的诸多理论中，文本电路与话语表征理论 [35] 尤为相似。粗略地说，文本电路的名词短语线对应于 DRT 的话语指称项，而作用于名词短语线的门

则对应于 DRT 中出现在指称对象上的谓词及其他条件。值得注意的是，话语指称项这一概念可以追溯到经典论文 [40], ，该论文在介绍这一概念时，设想了一种假设算法：解析文本，并据此记录所获知的关于每个指称对象的所有信息。在自然语言处理领域，文本电路可被视为对这一目标的直接实现。

一个关键区别在于，话语层面理论通常将句子的句法和语义视为既定内容，并以此为基础建立语法文本理论。从这个意义上说，文本电路是将话语理论向下扩展，以容纳句子语法。因此，文本电路优于话语表征理论的一个方面，是将副词处理为方框，因为副词很自然地可以解释为高阶算子。相比之下，我们尚未发现有哪种话语表征理论变体会用高阶谓词的数据来装饰话语方框，也未发现逻辑中处理高阶谓述的标准方法。

我们用一个简单的例子展示这种对应关系，它使文本电路可以被视为对DRT的话语表征结构进行“细化”。在文本电路中，简单句被描绘为（可能嵌套的）门，这与DRT中由嵌套DRS框表示的复杂条件基本对应。区别在于，DRT中唯一的复合形式是框嵌套，而文本电路还可以使用幺半范畴的图示组合，按照文本推进以图形方式展开话语方框的内容，如图6所示。



图6：左：句子“如果一位农夫拥有一头驴，他就会打它”的话语表征结构。右：同一句子的文本电路。

鉴于 DRT 与 DisCoCirc 之间的紧密类比，我们认为进一步探究两者之间的关系将大有裨益。尤其是，我们认为这样的研究将有助于扩展文本电路的能力，使其能够对更大范围的语言片段进行建模，并处理更复杂的语言现象。我们作出这一判断，是因为使用文本电路对文本建模时遇到的许多挑战，恰恰也是 DRT 曾面临的挑战。

图表示语义学的细化。 图是知识表示的一种常见计算机实现。文本电路可以看作图模型，并通过体现文本顺序的非交换性加以细化。也就是说，总可以将文本电路的门解释为可交换的，从而得到等价的图模型，如图7所示。

非交换性有两方面的益处：一方面是理论上的，另一方面是实践上的。如上所述，理论上的益处在于，我们总可以忽略文本中事实的顺序，从而得到图表示，但反过来却不行：“扁平”的文本意义图表示反映了这样一种情况：在经典逻辑中，对于完美的逻辑学家而言，事实的顺序不会影响推理结果。实践上的益处与这一点有关；我们知道，人类推理者会受到顺序的影响，就像我们知道笑话如果先讲包袱就不好笑一样。要捕捉阅读和理解文本的过程 ，而不只是其结果，非交换性在实践中必不可少。

蒙塔古语法的近亲。 蒙塔古语义学 [54]可（概括地）视为从句法到语义学 [32]的代数同态。代数指的是用于复合的数学框架，而

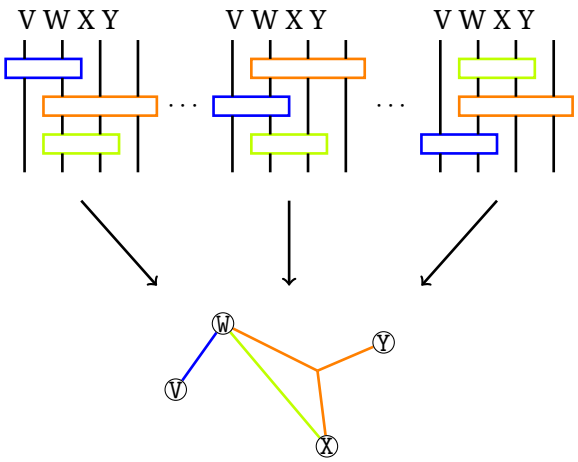


图7：上：门相同但顺序不同的文本电路。下：通过规定门的顺序无关紧要而得到的图表示。

同态则是保留该代数组合结构的映射。文本电路与蒙塔古语义学之间的差异，源于对复合代数应当是什么这一问题有着不同看法。蒙塔古代数是克隆，它们等价于 [41] Lawvere 理论 [46]，而 Lawvere 理论又是 PROPs [56]的特殊情形。Lawvere 理论与 PROPs 的根本区别在于，前者是多输入、单输出，后者则是多输入、多输出。文本电路完全符合蒙塔古关于从句法到语义学的「保结构映射」这一基本思想。事实上，近期量子自然语言处理实验 [50] 的成功，或许正是因为在量子计算机上以这种「保结构」的方式实现了文本电路。

7 致谢

我们感谢 Dimitri Kartsaklis、Konstantinos Meichanetzidis、Anna Pearson、Robin Lorentz、Alexis Toumi、Giovanni de Felice、Ian Fan、Richie Yeung、Stephen Clark 以及 SEMSPACE 的审稿人，他们对本论文（截然不同的）早期版本提出了有益的反馈。我们感谢 Vid Kocijan、Richie Yeung、Amar Hadzihasanovic、Stephen Clark 和 Michael Moortgat 对当前版本提出的意见。

参考文献

[1] S. Abramsky. 追溯过程代数中的若干路径。载于*CONCUR'96*：并发理论，收录于《计算机科学讲义》第1119卷，第1–17页。施普林格，1996年。[2] K. Ajdukiewicz. Die syntaktische konnexit¨ at. *Studia Philosophica*，1：1–27，1937年。[3] D. Bankova、B. Coecke、M. Lewis 和 D. Marsden。组合分布语义学中的分级上下义关系。《语言建模杂志》，6(2)：225–260，2019年。[4] M. Baroni 和 R. Zamparelli。名词是向量，形容词是矩阵：在语义空间中表示形容词-名词结构。载于*2010*年自然语言处理实证方法会议论文集，第1183–1193页，2010年。[5] J. Benabou. Categories avec multiplication. *Comptes Rendus des S´eances de l'Acad´emie des Sci-ences. Paris*，256：1887–1890，1963年。[6] J. Bolt、B. Coecke、F. Genovese、M. Lewis、D. Marsden 和 R. Piedeleu。交互式概念空间 I：概念的语法组合。载于 M. Kaipainen、A. Hautam¨ aki、P. G¨ardenfors,

和 F. Zenker 编，Concepts and their Applications，Synthese Library，Studies in Epistemology, Logic, Methodology, and Philosophy of Science。施普林格，2018年。即将出版。[7] J. Bresnan. Bresnan词汇功能句法。John Wiley &Sons，马萨诸塞州马尔登，2000年9月。[8] T. Brown、B. Mann、N. Ryder、M. Subbiah、J. Kaplan、P. Dhariwal、A. Neelakantan、P. Shyam、G. Sastry、A. Askell、S. Agarwal、A. Herbert-Voss、G. Krueger、T. Henighan、R. Child、A. Ramesh、D. Ziegler、J. Wu、C. Winter、C. Hesse、M. Chen、E. Sigler、M. Litwin、S. Gray、B. Chess、J. Clark、C. Berner、S. McCandlish、A. Radford、I. Sutskever 和 D. Amodei。语言模型是少样本学习者。arXiv:2005.*14165 [*cs*]*,2020年7月。arXiv：2005.14165。[9] W. Buszkowski 和 K. Moroz。预群语法与上下文无关文法。2008年1月。[10] N. Chomsky. 句法理论的若干方面。1965年。[11] S. Clark 和 S. Pulman。结合符号式与分布式意义模型。载于AAAI量子交互春季研讨会论文集。AAAI出版社，2007年。[12] B. Coecke。组合性就在我们身边，无处不在。arXiv预印本 arXiv:2110.05327，2021年。[13] B. Coecke. 文本结构的数学，第181–217页。施普林格国际出版社，2021年。arXiv:1904.03478。[14] B. Coecke、G. de Felice、K. Meichanetzidis 和 A. Toumi。近期量子自然语言处理的基础，2020年。[15] B. Coecke、G. de Felice、K. Meichanetzidis 和 A. Toumi。如何让量子比特说话。载于 E. R. Miranda 编，《艺术与人文学科中的量子计算》，页码待定。施普林格，2021年。[16] B. Coecke 和 A. Kissinger. Picturing Quantum Processes. A First Course in Quantum Theory and Diagrammatic Reasoning. 剑桥大学出版社，2017年。[17] B. Coecke、M. Lewis 和 D. Marsden。笛卡尔动词和介词的内部连线。载于 M. Lewis、B. Coecke、J. Hedges、D. Kartsaklis 和 D. Marsden 编，《2018年物理学、自然语言处理与社会科学组合方法研讨会论文集》，收录于《理论计算机科学电子论文集》第283卷，第75–88页，2018年。[18] B. Coecke 和 K. Meichanetzidis。密度矩阵的意义更新。arXiv:2001.00862，2020年。[19] B. Coecke 和 ´E. O. Paquette. 物理学家的范畴论。载于 B. Coecke 编，《物理学新结构》，《物理学讲义》，第167–271页。施普林格，2011年。arXiv:0905.3010。[20] B. Coecke、M. Sadrzadeh 和 S. Clark。组合分布语义模型的数学基础。载于 J. van Benthem、M. Moortgat 和 W. Buszkowski 编，《献给 Jim Lambek 的纪念文集》，收录于 Linguistic Analysis 第36卷，第345–384页。2010年。arxiv:1003.4394。[21] B. Coecke 和 V. Wang。语法方程。arXiv预印本 arXiv:2106.07485，2021年。[22] G. De las Cuevas、A. Klingler、M. Lewis 和 T. Netzer。猫会爬蕴含哺乳动物会移动：在组合分布语义学中保持上下义关系。arXiv:2005.*14134 [*quant-ph*]*, 2020.[23] C. Dorn. 结合式n-范畴。arXiv:1812.*10586 [*math*]*, 2018。arXiv：1812.10586。[24] T. Duneau. 在 DisCoCirc 中求解逻辑谜题，2020年。[25] T. Duneau. 在 DisCoCirc 中分析连词。SemSpace 2021，第66页，2021年。[26] G. Frege. 致 Jourdain 的信。哲学与数学通信集，第78–80页，1914年。

[27] G. Gazdar. 自然语言处理中的范式合并。载于 R. Milner 和 I. Wand 编，《计算的未来：计算机科学未来研究方向》，第 88–109 页。剑桥大学出版社，1996年。[28] E. Grefenstette 和 M. Sadrzadeh. 意义的范畴组合分布模型的实验支持。载于*2014*年自然语言处理实证方法会议。第 1394–1404 页，2011年。arXiv:1106.4058。[29]J. Hefford、V. Wang 和 M. Wilson. 语义概念的范畴。*arXiv*预印本 *arXiv:2004.10741*，2020年。[30] M. Honnibal 和 I. Montani. spaCy 2：利用 Bloom嵌入、卷积神经网络和增量句法分析实现自然语言理解。即将发表，2017年。[31] J. E. Hopcroft 和 J. D. Ullman. 自动机理论、语言与计算导论。培生，1979年。[32] T. Janssen 和 T. Zimmermann. 蒙塔古语义学。载于 E. Zalta 编，《斯坦福哲学百科全书》。形而上学研究实验室，斯坦福大学，2021年夏季版，2021年。[33] A. Joshi. 树邻接语法：要提供合理的结构描述，需要多大程度的上下文敏感性? 载于 A. Zwicky、D. Dowty 和 L. Karttunen 编，《自然语言句法分析：心理学、计算与理论视角》，自然语言处理研究，第 206–250 页。剑桥大学出版社，剑桥，1985年。[34] A. Joyal 和 R. Street. 张量演算的几何 I。《数学进展》，88：55–112，1991年。[35] H. Kamp、J. Genabith 和 U. Reyle. 话语表征理论。载于《哲学逻辑手册》，《哲学逻辑手册》第 *15* 卷。*ISBN 978-94-007-0484-8*。施普林格科学+商业媒体 *B.V.，2011*年，第 *125* 页，第 125–394 页。2010年11月。期刊简称：《哲学逻辑手册》，《哲学逻辑手册》第 15 卷。ISBN 978-94-007-0484-8。施普林格科学+商业媒体 B.V.，2011年，第 125 页。[36] D. Kartsaklis. 使用紧闭范畴和弗罗贝尼乌斯代数的组合分布语义学。*arXiv*预印本 *arXiv:1505.00138*，2015年。[37] D. Kartsaklis、I. Fan、R. Yeung、A. Pearson、R. Lorenz、A. Toumi、G. de Felice、K. Meichanetzidis、S. Clark 和 B. Coecke. lambeq：用于量子自然语言处理的高效高级 Python 库。*arXiv*预印本 *arXiv:2110.04236*，2021年。[38] D. Kartsaklis 和 M. Sadrzadeh. 构建句子向量时对词张量进行预先消歧。载于*2013*年自然语言处理实证方法会议。第 1590–1601 页。计算语言学协会，2013年。[39] D. Kartsaklis 和 M. Sadrzadeh. 自然语言范畴框架中的纠缠研究。载于第十一届量子物理与逻辑研讨会论文集（*QPL*）。日本京都，2014年。[40] Lauri Karttunen. 话语指称项。载于《语言学地下笔记》，第 363–385 页。布里尔出版社，1976年。[41] S. Kerkhoff、R. P¨oschel 和 F. Schneider.《克隆简论》。《理论计算机科学电子笔记》，303：107–120，2014年3月。[42] J. Lambek. 句子结构的数学。《美国数学月刊》，65，1958年。[43] J. Lambek. 类型语法再探。《计算语言学的逻辑方面》，1582，1999年。

[44] J. Lambek. 从词到句子。Polimetrica，米兰，2008年。[45] S. Lappin 和 H. J. Leass. 一种代词回指消解算法。《计算语言学》，20(4)：535–561，1994年。[46] W. Lawvere. 函子语义学的代数理论背景下的一些代数问题。载于 S. MacLane 编，《中西部范畴研讨会报告 II》，第 61 卷，第 41–61 页。Springer Berlin Heidelberg，柏林、海德堡，1968年。丛书名：数学讲义。[47] M. Lewis. 利用正算子实现组合式上下义关系。载于《自然语言处理最新进展会议论文集》，保加利亚瓦尔纳，第 638–647 页，2019年。[48] M. Lewis. 朝向组合分布语义学中的逻辑否定。arXiv预印本 arXiv:2005.04929，2020年。[49] J. Liu. 作为电路的语言。理学硕士论文，牛津大学，2021年。[50] R. Lorenz、A. Pearson、K. Meichanetzidis、D. Kartsalkis 和 B. Coecke. 实践中的量子自然语言处理：在量子计算机上运行意义的组合模型。arXiv预印本 arXiv:2102.12846，2021年。[51] K. Meichanetzidis、A. Toumi、G. de Felice 和 B. Coecke. 量子计算机上的语法感知问答。arXiv预印本 arXiv:2012.03756，2020 年。[52] A. Merry. 运用 !- 图进行推理。博士论文，2014 年 3 月。arXiv: 1403.7828。[53] F. Meyer 和 M. Lewis. 利用密度矩阵对词汇歧义建模。载于第24届计算自然语言学习会议论文集，第 276–290 页，2020年。arXiv:2010.05670[cs.CL]。[54] R. Montague. 普遍语法。Theoria，36：373–398，1970年。[55] R. M. Nefdt. 关于机器中组合性的一则难题。Minds & Machines，30(1)，2020年。[56] nLab 作者。nLab 中的 PROP。[57] R. Penrose. 负维张量的应用。载于《组合数学及其应用》，第 221–244 页。Academic Press，1971年。[58] F. Pereira. 形式语法与信息论：再度携手？《英国皇家学会哲学汇刊 A 辑：数学、物理与工程科学》，358(1769)：1239–1253，2000年4月。出版者：英国皇家学会。[59] P. S. Peters 和 R. W. Ritchie. 转换语法的生成能力。《信息科学》，6：49–83，1973年1月。[60] R. Piedeleu. 意义范畴模型中的歧义。硕士论文，牛津大学，2014年。[61] R. Piedeleu、D. Kartsaklis、B. Coecke 和 M. Sadrzadeh. 自然语言处理中的开放系统范畴量子语义学。载于第六届计算机科学代数与余代数会议（CALCO 2015）。Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik，2015年。arXiv:1502.00831。[62] S. Pinker. 语言本能：心智如何创造语言。Penguin，伦敦，新版，1995年3月。[63] D. Quick. !-逻辑：非交换字符串图族的一阶推理。博士论文，2015年。[64] A. Radford. 转换语法：入门课程。剑桥大学出版社，1988年5月。

[65]D. Reutter 和 J. Vicary。《结合式 n-范畴中同伦构造的高阶方法》。arXiv:1902.*03831 [*math */*, 2019. arXiv：1902.03831。[66] B. Rodatz、R. A. Shaikh 和 L. Yeh。《组合式分布语义学中利用现实语境实现会话否定》。arXiv预印本 arXiv:2105.05748，2021年。[67] M. Sadrzadeh、S. Clark 和 B. Coecke。《词义的 Frobenius 结构 I：主语和宾语关系代词》。《逻辑与计算杂志》，23：1293–1317，2013年。arXiv:1404.5278。[68] M. Sadrzadeh、S. Clark 和 B. Coecke。《词义的 Frobenius 结构 II：所有格关系代词》。《逻辑与计算杂志》，26：785–815，2016年。arXiv:1406.4690。[69] P. Selinger。《幺半范畴图形语言综述》。载于 B. Coecke（编），《物理学新结构》，《物理学讲义》，第275–337页。施普林格出版社，2011年。arXiv:0908.3347。[70] R. A. Shaikh、L. Yeh、B. Rodatz 和 B. Coecke。《会话否定的组合》。arXiv预印本 arXiv:2107.06820，2021年。[71] D. Sleator 和 D. Temperley。《用链接语法分析英语》。arXiv:cmp-lg/9508004，1995年8月。arXiv：cmp-lg/9508004。[72] M. Steedman。《组合语法与寄生空位》。《自然语言与语言学理论》，5：403–439，1987年。[73] L. Tesni`ere。《结构句法学纲要》。约翰·本杰明出版公司，阿姆斯特丹，2015年2月。[74] Russell S Tomlin。《基本语序（RLE 语言学 B：语法）：功能原则》。Routledge，2014年。[75] K. Vijayashanker。《树邻接语法研究》。ProQuest 收录的学位论文，第1–185页，1987年1月。[76] V. Wang-Mascianica 和 B. Coecke。《谈论空间：从空间语言意义中进行推理》。arXiv预印本 arXiv:2109.06554，2021年。[77] V. Wang-Ma´scianica、J. Liu 和 B. Coecke。《自然语言中关于蜘蛛的五个故事》，2022年。准备中。[78] M. H. Waseem、J. Liu、V. Wang-Ma´scianica 和 B. Coecke。《DisCoCirc 文本电路的语言无关性：英语和乌尔都语》。载于 M. Moortgat 和 G. Wijnholds（编），《向量语义的端到端组合模型论文集》，《理论计算机科学电子论文集》第366卷，第50–60页。开放出版协会，2022年。[79] D. Weir。《温和上下文敏感语法形式体系的刻画》。ProQuest 收录的学位论文，第1–154页，1988年1月。[80] D. Weir 和 A. Joshi。《组合范畴语法：生成能力及其与线性上下文无关重写系统的关系》。载于第26届计算语言学协会年会论文集，第278–285页，布法罗，纽约，美国，1988年6月。计算语言学协会。[81] D. Weir 和 K. Vijay-Shanker。《树邻接语法与中心语语法之间的关系》。2002年5月。[82] G. Wijnholds 和 M. Sadrzadeh。《动词短语省略句嵌入的组合模型评估》。载于2019年 ACL 会议论文集，第261–271页。计算语言学协会，2019年。

[83] G. Wijnholds、M. Sadrzadeh 和 S. Clark。《类型驱动组合的表示学习》。载于第*24*届计算自然语言学习会议论文集，第313–324页，2020年。[84]P. Wilson、D. Ghica 和 F. Zanasi。《非严格幺半范畴的字符串图》。*arXiv:2201.11738 [math]*,2022年1月。arXiv：2201.11738。[85] R. Yeung 和 D. Kartsaklis。《基于组合范畴语法的 DisCoCat 框架版本》。*arXiv:2105.07720[cs, math]*, 2021。[86] V. Zamdzhiev。《字符串图的上下文无关族重写》。博士论文，2017年5月。arXiv：1705.07520。