**第十二章读书报告**

09118223 吴亦珂

读书进度：第十二章读完

1. **问题列表**

（我提出）P15页最后一段说的traces具体是什么意思吗？以及后面的例子能否解释一下具体是什么意思？

讨论结果：在一般的情况下应该出现的没出现就标记为none，然后解析器解析到none是就可以知道按照正常规则，后面部分是缺失的。因此可以成功完成对句子的解析。在书本的例子中，一般说话的内容接在say之后，但是这个例子中却是在say之前，因此在say后标注none，方便解析器解析。

（我提出）P17页12.4.2说的“flat”应该是什么意思？为什么会导致规则很长？

讨论结果：因为语法规则是直接根据语料库中的句子直接总结出来的，由于句子结构的冗余，可能导致规则也非常冗余。

（别人提出）比起原始算法，CCG的优势是什么？  
讨论结果：书上在P21页有提及，传统方法的解决方法会产生冗余的语法，很难管理，并且很脆弱。而新的这些方法能克服这些缺点。

1. **读书收获**

**12.3.5 Coordination**

可以使用连词如and，or，but之类来连接两个相同性质的词：

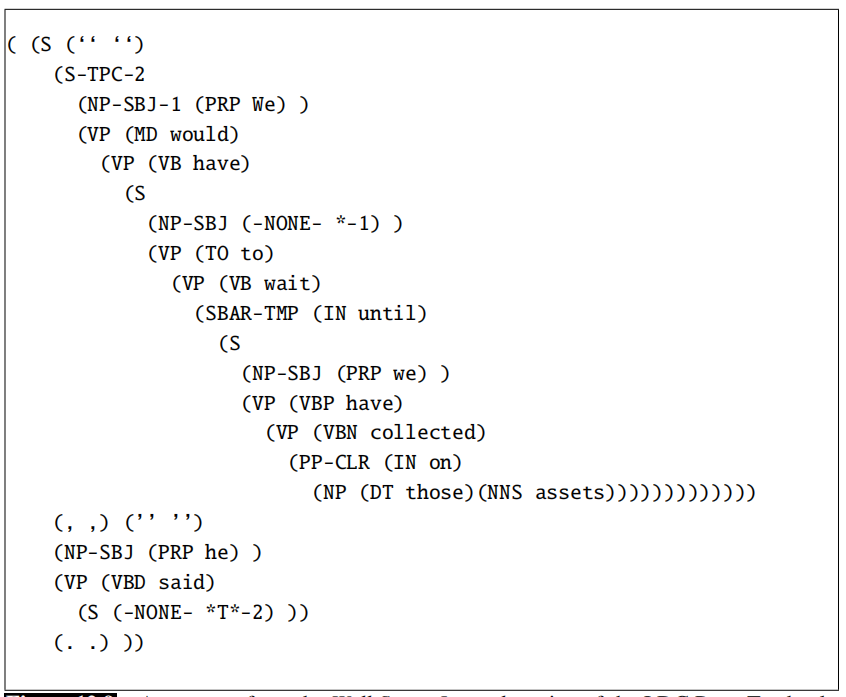


**12.4 Treebanks**

树库是一种深加工语料库，其可被用于对句子进行分词、 词性标注 和句法 结构 关系的标注。

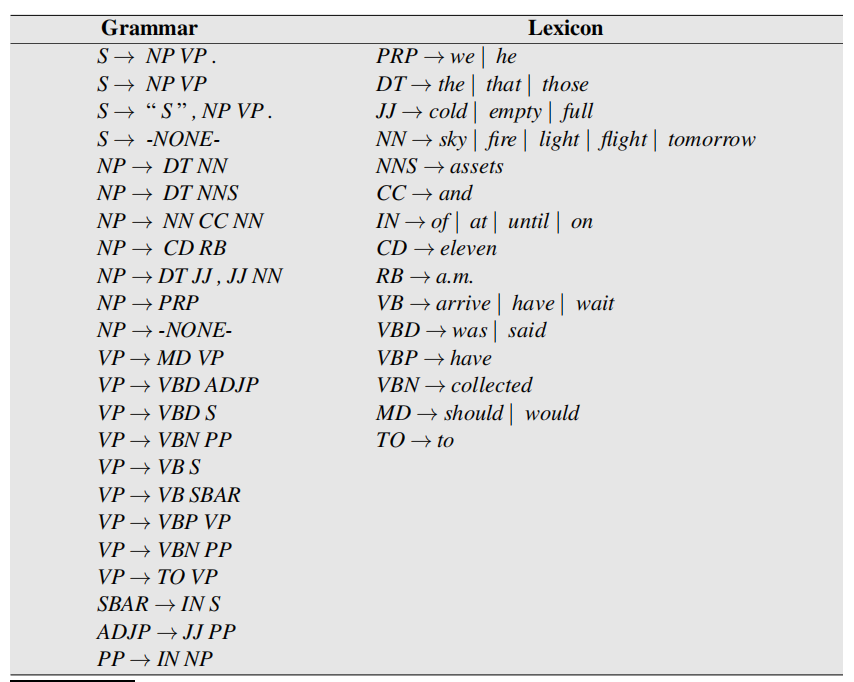
**12.4.1 Example: The Penn Treebank Project**

使用-NONE-nod来标记长距离依赖或句法移动。

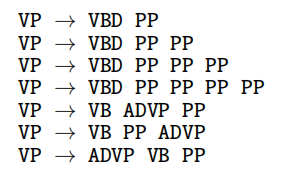


**12.4.2 Treebanks as Grammars**

树库中的句子隐含地构成了被注释的语料库所代表的语言的语法。



由于会直接根据语料库中的句子直接进行解析，导致出现很多很长的规则。



**12.4.3 Heads and Head Finding**

模型一：每个解析树中的非终结符用一个单词注释，构成它的词法头。

缺点：复杂的大多数短语都有争议。

模型二：一个句子被解析后，生成的结果树被遍历并找到合适的头部。

**12.5 Grammar Equivalence and Normal Form**

一个正式的语言可以定义为一串单词（可能无限）。

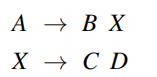
Grammar equivalence：

Strong equivalence：生成了相同的一组字符串并且每个句子具有相同的短语结构。

Weak equivalence：生成了相同的一组字符串并且每个句子不具有相同的短语结构。

任何上下文无关的语法都可以转换成弱等价的Chomsky normal form grammar。





使用二进制分支实际上可以生成更小的语法。

Chomsky-adjunction：一个符号a与一个可能无限的符号序列B的生成a→B形式的规则。

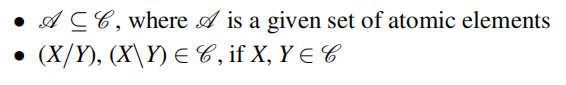
**12.6 Lexicalized Grammars**

**12.6.1 Combinatory Categorial Grammar**

分类法三个主要要素：

一组类别，一种将单词与类别联系起来的词典，以及一组规则来管理类别如何在上下文中组合。

Categories：可以是原子元素，也可以是单参数函数，当提供所需的category作为参数时，它们返回一个category作为值。



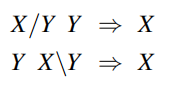
斜杠符号用于定义语法中的函数。它指定所需参数的类型、预期找到该参数的方向以及结果的类型。

原子类别的集合通常非常小，包括句子和名词短语等熟悉的元素。功能范畴包括动词短语和复合名词短语等。

Categorial方法中的词汇是由对词的类别分配组成的。这些赋值可以是原子类的，也可以是函数类的。由于词汇的歧义性，单词可以被分配到多个类别。

Rules:

the basis for all categorial grammars:



第一个规则forward function application：将函数应用于其右侧的参数。

第二个规则backward function application：将函数应用于其右侧的参数。

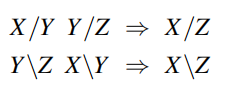
英语允许同一类型的两个成分进行协调，因此添加一个规则：



这条规则规定，当同一类别的两个成分被一个CONJ型成分分开时，它们可以合并成同一类型的一个较大的成分。

问题：categorial approach和传统的CFG规则比较起来，没有更多的表达能力。它只是将信息从语法转移到词典。

为了超越这些限制，CCG包括对功能进行操作的操作。



第一个规则forward composition：第一个是一个在其右边寻找Y类型参数的函数，第二个是作为结果提供Y的函数。这个规则允许我们将这两个函数组合成一个单一的函数，其中是第一个成分的类型和第二个是成分的参数。这两种成分在CCG图中都用B表示，并伴有一个<或>指示方向。

type raising：可以将简单类别提升到函数的状态。

下面显示了两个版本的类型提升：一个用于右侧的参数，另一个用于左侧的参数。

