从语言工程的角度看“中心扩展条件”和“并列条件”*

詹卫东
北京大学中文系 北京大学中国语言学研究中心

摘要 沈家煊先生近年来提出汉语词类包含模式，主要立论依据是传统的汉语词类体系中关于名、动、形等的处理方式有一个困境：要么违背语言学理论中的“中心扩展条件”“并列条件”，要么违背理论构建时应该遵循的“简约”原则。如果采用词类包含模式去看待汉语词类，就可以摆脱这个困境。本文从树库语料的分析出发来说明语言事实中确实存在违反“中心扩展条件”和“并列条件”的情况，并将原因归结为言语使用中的“简约”（或“经济”）原则使然。换言之，言语交际中的简约，造成语言理论模型的无法简约。

关键词 汉语语法 中心扩展条件 并列条件 句法规则 词类 简约原则

一 问题的提出

汉语词类划分作为构建汉语语法体系的基础工作，却长期以来困扰着汉语语法研究者和中文信息处理研究者。下面通过一个例子说明主要问题之所在。

例 1 a 科学出版社非常迅速地**出版**了这本书。

b 这本书的**出版**标志着我国思想界的进一步解放。

c 这本书的**封面**显然是用中华书局的原版封面翻拍印制的。

例 1a中的“出版”是动词（v），没有争议，例 1c中的“封面”是名词（n），也没有争议，例 1b中的“出版”是什么词，在汉语语法学界则一直争议不断：是动词、名词，还是动名词、名动词？沈家煊先生（2007,2009a，2009b）提出汉语词类的包含模式理论，即汉语里的名词包含动词，动词包含形容词，因此例 1b中的“出版”既是动词，也是名词[1]。

沈先生的理由是：这样处理之后的语法理论既不违反“简约原则”，也不违反“扩展条件”（本文称为“中心扩展条件”）和“并列条件”。

我们对此的质疑是：

（1）怎么知道（或评判）哪个理论更简约？理论的简约是一种主观感觉，还是有客观的评价标准？

（2）“中心扩展条件”“并列条件”是否真的不能违反？沈先生的理论真的做到了既不违反“中心扩展条件”，又不违反“简约原则”吗？

下面就带着这两个疑问，来重新审视沈先生的词类包含模式理论的上述立论依据是否可靠，或者说沈先生所声称的这种理论的好处是否真的存在。

二 关于“简约原则”

“简约原则”更具学术性的表述是“奥卡姆剃刀”（Ockham’s Razor）原则。这是被学术

* 本文的研究工作得到教育部人文社科基地重大项目“大规模中文树库建设及其应用研究”（课题编号：06JJD740001） 和霍英东基金项目“大规模中文树库构建及其在对外汉语教学中的应用”（课题编号：111098）资助。

[1] 中文信息处理学界黄昌宁等（2009a，2009b）主张这里的“出版”在语料标注中应标为名词（即“出版”是动、名兼类词）。而俞士汶等（2003）则主张标为名动词（vn），即“出版”是动词，但在例 1b 中属于“特殊用法”，即动词用作名词。沈先生的理论观点落实到语言工程上，如果直接应用其词类包含模式理论的话，应该是例 1b 中的“出版”标 v 和标 n 两可，但是根据沈家煊（2009b）对“哭没用”中的“哭”应标为 n 的意见，可以推想，对例 1b 中 “出版”的词性标注，沈先生跟黄先生应是持相同意见，即应该标为 n。我们曾从计算机进行句法结构分析需要什么样的词语分布知识角度讨论过这里的“出版”的词性标记问题（詹卫东 2010a），结论是工程上标 n 并不带来系统性的好处，反而增加知识管理的麻烦。“出版”理论上不是兼类词，在语言工程上就应该是单标记词，应标为 v。标 n 或 vn 都不好。

界普遍接受的一种观念，即所谓的“若无必要，勿增实体”。强调在竞争的两个理论中，如果理论的结果相同，那么简单的那个理论胜出。这个观念很好理解，但是在具体操作层面，要比较两个理论哪个更简单，却绝对不是一件简单的事情。有时候直觉简单的理论，其实并不简单。下面举一个例子说明。

“开飞机容易”“打是疼，骂是爱”是汉语中合法的句子，句子主语由“开飞机”“打”“骂”等“动词性成分”充任。如何描述这样的语言现象呢？理论甲的描述方法是提出两个假设：（1）汉语中动词做谓语不做主宾语；（2）动词做主宾语的时候变成了名词。理论乙的描述方法是提出一个假设：（1）汉语中动词本来能做谓语也能做主宾语。理论乙是朱德熙先生的看法，理论甲是传统的说法，是朱先生反对的说法。是不是能根据理论甲有两个假设，而理论乙只有一个假设，得出理论乙比理论甲简约的结论呢？

答案是否定的。用自然语言描述的理论有时候会给人错觉，比如上面在对比理论甲和理论乙的时候，以为理论甲需要用到两条假设，而理论乙只需要用到一条假设。其实，要解释同样多的事实，理论乙需要的假设（规则）一点也不比理论甲少。如果用S代表句子，VP代表动词性成分（包括动词和动词短语），NP代表名词性成分（包括名词和名词短语），那么，理论甲的两条假设对应的形式规则可以表达如下[2]（规则中的箭头可以解作“推导出”或“变为”）：

（R1） S → NP VP （说明：S 变换为 NP+VP，NP 作主语，VP 作谓语）

（R2） NP → VP （说明：NP 变为 VP，这样 VP 也就可以作主语了）

同样的，理论乙的一条假设对应的形式规则其实也是两条：

（R’1） S → NP VP （说明：同 R1）

（R’2） S → VP VP （说明：S 变换为 VP+VP，第一个 VP 作主语，第二个 VP 作谓语）

把自然语言表达的理论假设“翻译”成严格的形式文法后，就会发现，其实两种理论的规则条数是一样的，谁也不比谁“简约”。理论甲和理论乙都可以“解释”两种组合模式：“名+动”主谓结构（如“他+开飞机”）和“动+动”主谓结构（如“打+是疼”），但都无法解释汉语中的另一种组合模式“名+名”主谓结构（如“他+黄头发”）。要解释这种模式，按照理论甲的思路，仍然可以用两条规则来做到，同时还能解释原来可以解释的事实（前两种组合模式）。理论甲修改后得到如下新版本：

（R1） S → NP NP （说明：S 变换为 NP+NP，第一个 NP 作主语，第二个 NP 作谓语）

（R2） NP → VP （说明：NP 变为 VP，这样 VP 就既可以作主语也可以作谓语）

但按照理论乙的思路[3]，要“解释”上述三种组合模式，就需要三条规则：

（R’1） S → NP VP

（R’2） S → VP VP

（R’3） S → NP NP

那么，是不是由此就可以得出结论理论甲比理论乙更简约呢？

答案仍然是否定的。理由是理论甲虽然规则少，但它描述的事实也跟着就出问题了，它在能够解释三种主谓结构组合模式的同时，会产生一种新的错误的组合模式：从 R1 和 R2 可以推导出 S → VP NP 这样的组合。推导过程是：R1 中的第一个 NP 变换为 VP，第二个 NP 保持不变。这样形成的组合会把“开+飞机”这样的述宾结构看成是主谓结构，而这显然是不对的。

通过上面这个简单的示例，不难看到，要评判两个理论谁更简单，并不能诉诸直觉。事实上，在比较两个理论的复杂程度之前，我们得有办法评估一个理论能解释多少事实。只有

[2] 为节省篇幅，这里规则简化了，能说明问题即可。比如只有 NP 作主语的规则，不涉及 NP 作宾语的规则。

[3] 理论乙的思路跟理论甲的思路的本质区别是：理论乙不主张无标记转类（即兼类）：VP 不能（或不需要）变成 NP（即动词没有名词化）。理论甲主张可以无标记转类：VP 可以变成 NP（动词可以名词化）。

满足了两个条件，才能客观评价两个理论的复杂程度：（1）一个理论所能解释的事实多少是可以量化的；（2）一个理论模型本身是可以计量的（可比较大小）。上面例子中理论甲和理论乙用自然语言表述的时候，是无法计量其大小的。

直到上个世纪六七十年代，数学家和信息科学家才把诞生于十四世纪的“奥卡姆剃刀原则”从一个抽象的观念变成了操作上可计算的算法（algorithm）。数学上用柯尔莫哥洛夫复杂度（Kolmogorov complexity）[4]这个概念来表达一个对象（比如负载了一定信息的字符串）的复杂程度。这个概念有一个别名是描述复杂度（descriptive complexity）。具体怎么定义这个复杂度呢？数学家想出的办法是用打印程序来模拟，即一个字符串的复杂程度，可以定义为打印出这个字符串所写的打印程序的长度。对同一个字符串，人们可以写出不同的打印程序来打印出这个字符串。这些打印程序就是解释该字符串的不同理论。这样，理论复杂度的比较问题，就转化为打印程序的长短比较问题。短的打印程序简单，长的打印程序复杂。那么，为什么同一个字符串，会有长短不同的打印程序呢？或者说，打印程序（理论）跟打印对象（字符串）之间的本质关系到底是什么呢？答案是，打印程序的本质是发现打印对象的规律，如果发现了字符串的规律，就可以把长的字符串压缩成更短的字符串（比如“010101010101”这个字符串可以表述为“6 个 01”，后者比前者缩短了）。这样一来，好的理论其实就是好的压缩程序，它可以把所描述对象压缩成更短的代码。而发现好的理论的前提是，我们能发现对象的内在规律。信息科学中用最短描述长度原则（minimum description length，MDL）来表达为一个对象发现最好的打印程序（或压缩程序）的算法[5]。至此，简约原则（或奥卡姆剃刀原则）才算是有了一个可操作的版本。那么，是不是什么理论都可以用上述办法来度量其复杂度呢？答案再次令人遗憾。有些描述对象太复杂了，比如自然语言，以至于要发现一个符合MDL的理论模型是不可能的。自然语言的句子理论上是无限多的，而且语言学家至今也没有实现用“有限规则生成无限个句子”的理想[6]。在自然语言信息处理领域，只能针对某个具体的语料库（即数量有限的句子集合），发现一个解释该语料库的MDL理论模型，但这个“好”的理论模型在处理其他的语料库（句子集合变了）时，很可能就成了一个表现糟糕的模型。

以上是我们关于简约原则的看法。我们的目的是说明，关于理论简约性的客观度量需要比较复杂的数学和计算技术。如果仅仅停留在哲学层面或直觉层面谈论简单与复杂，很容易陷入公说公有理婆说婆有理的无效争论。对此，本文不再展开讨论。下面回到具体的容易判别的问题上来，看看“中心扩展条件”“并列条件”在汉语中的实际表现情况如何。当论述中需要用到“简约”这一概念时，我们再给出关于简约与否的明确判别标准。

三 “中心扩展条件”“并列条件”的基本含义

在到树库（treebank）语料中考察中心扩展条件和并列条件的表现情况之前，我们需要先明确一下这两个概念的具体内容。这里引沈家煊（2007）对这两个概念的说明。中心扩展条件（Head Expansion Condition，记作 HEC）指的是“**以一个成分为中心加以扩展，扩展后的结构的语法性质跟中心成分的语法性质一致**”。并列条件（Coordination Condition，记作 CC）指的是“**在非临时活用的场合，并列的两个成分应该属于同一词类或同一语类**（Radford 1988:76）”。

沈家煊（2007）在说明“中心扩展条件”和“并列条件”的性质时，还引用了 Lyons（1968: 331）的论述并补充了自己的意见：

[4] 可参考维基百科对柯尔莫哥洛夫复杂度的解释：http://en.wikipedia.org/wiki/Kolmogorov_complexity。

[5] 可参考维基百科对 MDL 的解释：http://en.wikipedia.org/wiki/Minimum_description_length。

[6] Chomsky 提出的产生式规则可以做到用有限规则生成无限多的字符串，但不能做到生成一种自然语言中的全部句子且只生成这种语言中的句子。现有的自然语言形式规则总是在生成正确句子的同时也生成一堆不正确的句子。

> “N 和 NP 之间，V 和 VP 之间都存在一种必不可少的（essential）的联系，对哪种语言都一样。..... NP 和 VP 不仅仅是帮助记忆的符号，而是分别表示句法成分 NP 必定是名词性的，VP 必定是动词性的，因为两者分别以 N 和 V 作为其必需的主要成分。”他（指 Lyons）接着说，如果有哪位语言学家提出诸如“ NP→V+VP，NP→V，VP→T（冠词）+N ”的规则，“ 那不仅是有悖常情的，在理论上也是站不住的。”这些话是就“ 扩展条件 ”而言的，但是也适用于“ 并列条件 ”，提出有“ NP 和 VP ”这样的并列结构也是有悖常情的，理论上站不住的。

根据以上关于“中心扩展条件”和“并列条件”的说明，可以把以下例子区分为两类：

		I	II
例 2	a	这本书的封面	这本书的出版
	b	封面和封底	图书和出版

例 2 中 I 组的例子显然是符合“中心扩展条件”和“并列条件”的。例 Ia 中的中心成分是“封面”，扩展后的结构“这本书的封面”的语法性质跟“封面”一致。例 Ib 中“封面”和“封底”并列，属于同一词类。

如果按照严格的“中心扩展条件”和“并列条件”的定义，例 2 中 II 组的例子显然不符合“中心扩展条件”和“并列条件”。汉语语法的研究者们也正是因为很自然地感觉到了上面 II 组例子跟 I 组例子的显著区别，同时又想把两组例子的句法结构“统一”起来，使得统一后的结构满足“中心扩展条件”和“并列条件”的要求，才会费尽心力地去发明语法理论模型，来解释如何才能让 II 组例子满足这两个条件。沈家煊（2007）在提出自己的理论主张之前，批评了程工（1999）、司富珍（2002、2004）、陆俭明（2003）、熊仲儒（2005）等学者提出的 DP、DeP（“的”字作结构中心）等理论模型，指出这些模型为了满足“中心扩展条件”，付出了违背理论的简约性原则的代价。在分析了已有的理论模型的“不足”之后，沈家煊（2007、2009a、2009b）提出了汉语实词包含模型，在这个理论模型中，动词可以无标记地“构成”为名词，因而“出版”作为“这本书的出版”的中心语，既不违反中心扩展条件，也不必付出违背理论简约性的代价。

对相关理论模型及争议情况做了简要梳理后，不难概括出两个要点：（1）尽管沈先生跟之前的 DP、DeP 理论模型的主张不同，但大家有一个共识，就是普遍接受**“中心扩展条件”（以及“并列条件”）不能违反**这个前提。（2）沈先生跟其他学者的不同在于：他在坚持这个前提的同时，还要兼顾“理论的简约性”这一原则（其他学者应该也主张理论应该简约，但在分析这个具体问题时，没有显式地强调这一原则）。

上文第二节我们对“简约原则”的讨论中已经指出，评价一个理论简约与否，如果不做形式化和定量分析，是无法得到客观结论的。仅仅在理论模型中声称“动词可以无须加名化标记就构成为指称语（名词）”[7]（类推到短语就是动词性短语vp可以无标记地“构成”为名词性短语np），并不能证明这样的理论模型就是简约的。当然，同样也不能证明这种理论模型更复杂。如果要问上述这些理论模型哪个简约哪个复杂，我们认为，实事求是的回答是无法评价。

能够进行评价的是相对客观的“中心扩展条件”和“并列条件”。这两个条件的定义相对而言是清楚的。本文打算调查实际语料中到底有多大比例的例子是跟例 2 中 II 组例子类似的情况。在此基础上反思这两个条件是不是一定不能违反？语法理论模型坚持遵循这两个条件的理由是什么？

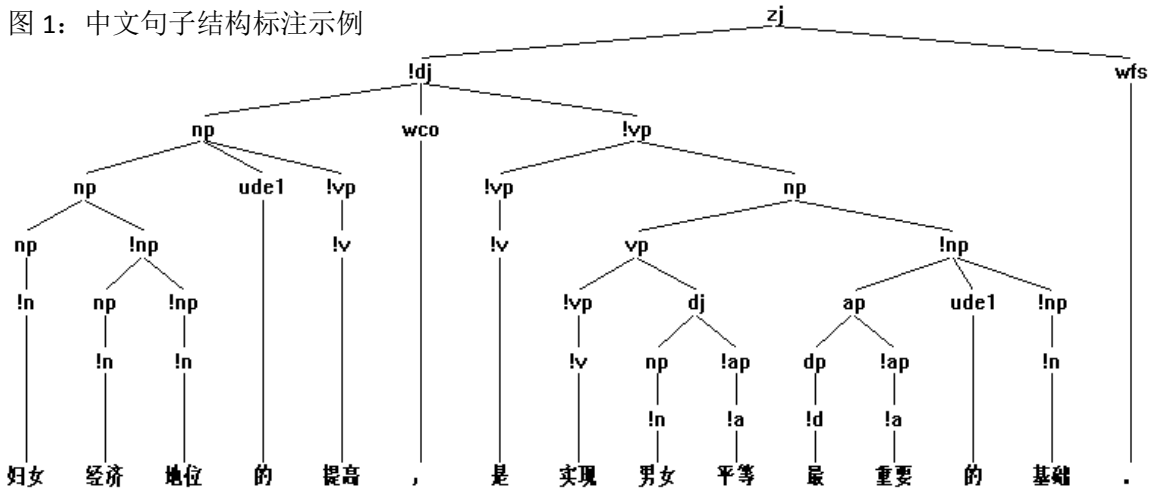
[7] 沈先生提出的“动词可以无须加名化标记就构成为指称语”这样的主张，如果翻译成形式文法，就是“NP → V”（或 N→V）这样的规则。而这正是 Lyons 所谓的“那不仅是有悖常情的，在理论上也是站不住的。”从这个意义上讲，即便承认沈先生的理论模型是简约的，那这个理论模型也是以违背“中心扩展条件”为代价的。

四 树库中的 NHE 结构和 NCC 结构

基于上一节中对“中心扩展条件”和“并列条件”的认识，我们可以把符合这两个条件的结构分别记作 HE 结构（中心扩展结构）和 CC 结构（同类成分并列结构），相应地，把不符合这两个条件的结构记作 NHE 结构（违反中心扩展条件的结构）和 NCC 结构（违反并列结构条件的结构，也即非同类成分并列结构）。

树库是对句子进行了句法结构标注的语料库（周强，2004、陈锋等 2008、Abeillé,2003、Xue et al.，2005）。有了树库语料，就可以方便地调查 HE、NHE、CC、NCC 结构在真实文本中的分布情况。下面是北京大学树库（詹卫东，2008）中一个句法结构标注实例，例句来自中国政府白皮书（1994）《中国妇女的状况》。其中涉及的标记的含义为：zj 整句，dj 单句，np 名词性短语，vp，动词性短语，ap 形容词性短语，dp 副词性短语，ude1 的，wfs 句号，n 名词，v 动词，a 形容词，d 副词。标记之前的叹号！表示该成分是整个结构的中心语。

图 1：中文句子结构标注示例



上例中“妇女经济地位的提高”是一个类似“这本书的出版”的短语结构，在北大树库标注规范中，为了区别“这本书的出版”和“这本书的封面”，前者被标记为“np ude1 !vp”结构，而后者则标记为“np ude1 !np”结构。我们的想法是，通过这种标记方式，可以很方便地检索出树库中哪些“的”字结构 np 的中心语是由 np 充当的（即 HE 结构），哪些“的”字结构 np 的中心语是由非 np（如 vp，ap 等）充当的（即 NHE 结构）。上例中，“妇女经济地位的提高”就是一个 NHE 结构，“最重要的基础”则是一个 HE 结构。二者的共同之处是整个结构都被标记为 np，句法上属名词性短语，语用功能为指称。但前者的中心语“提高”是 vp，跟整个短语的功能类（np）不一致。后者的中心语“基础”是 np，跟整个短语的功能类（np）一致。

从标注好的树库中可以抽取短语结构组合规则。根据本文调查的需要，我们抽取的规则包括（1）全部短语结构规则；（2）符合中心扩展条件的规则；（3）不符合中心扩展条件的规则；（4）符合并列结构条件的规则；（5）不符合并列结构条件的规则。北大树库的规模情况及抽取出的各类具体规则数量如下面表 1——表 5 所示。

表 1：北大树库语料类型及规模

语料类型	字数	百分比
语文课本	744,563	56.85%
句型语料	174,123	13.29%
新闻语料	170,226	13.00%
科技语料	123,500	9.43%
政府白皮书	97,308	7.43%
合计	1,309,720	100%

表 2：树库句数、词数、短语规则数[8]

总句数	55,742
总词数	899,365
总规则 Type 数	1,930
总规则 Token 数	1,318,488

吾规则的数量（类似统计汉字个数时所有的字种个数，pe 数）。规则 Token（例）数指树库中实际出现过的短、数，每个汉字的一次出现都计一次）。

表 3：树库中HE规则、NHE规则、CC规则、NCC规则的数量统计及所占比例总表 [9]

规则类别	规则种数（type）	规则例数（token）	结构示例	示例
全部规则	1,930	1,318,488	dj → np !ap	人 多
HE 规则	1682(87.15%)	1,289,068 (97.77%)	ap → dp !ap	最 冷
NHE 规则	248(12.85%)	**29,420 (2.23%)**	np → qp !vp	这次 失事
CC 规则	76(65.52%)	26,261(96.52%)	ap → !ap c ap	光荣 而 艰巨
NCC 规则	40(34.48%)	**946 (3.48%)**	np → !np wsc ap	政治、经济、安全

表 4：NHE 规则分类别统计表

序号	短语	规则种数（type）	规则例数（token）
1	np	150	24132
2	tp	30	284
3	dp	24	3910
4	ap	21	506
5	vp	15	414
6	sp	8	174
合计		248	29420

表 5：NCC 规则分类别统计表

序号	短语	规则种数（type）	规则例数（token）
1	vp	17	540
2	np	17	265
3	ap	6	141
合计		40	946

说明：上面表中 NHE 及 NCC 规则的统计数据是程序根据规则形式自动判别的。判别方式是：对于形如 XP→ α !XP β 这样的规则（其中 α、β 可以空字符串），就判定为 HE 规则，否则就判定为 NHE 规则。并列结构规则的一般形式为 XP→!XP c XP 或 XP → !XP wsc XP 其中 c 为连词、wsc 为顿号。XP 同类并列，得到的就是 CC 规则。如果抽取到的并列结构规则形如 XP → !XP c YP 或 XP → !XP wsc YP，则判定为非同类成分并列。因为是程序自动判别，所以没有考察无并列标记的并列结构，只考察了中间有连接标记（连词或顿号）的并列结构。通过程序自动判别得到的数据可能有一定误差。不过，我们的目的并不是统计出精确的数据做量化分析，而仅仅是指出实际语料中存在 NHE 规则和 NCC 规则的实例。很显然，从实例频次的对比来说，NHE 规则和 NCC 规则相对于 HE 规则和 CC 规则都是绝对少数（显然可以认为后二者是常规情况）。换言之，真实语料中的大部分短语组合都是符合“中心扩展条件”和“并列结构条件”的，但是我们想强调的是，也确实存在不符合这两个条件的实例，尽管比例不高，但违反中心扩展条件和并列结构条件的实例也并非特例。下面即展示和分析 NHE

[9] 表 3——表 5 中涉及的短语类标记含义为：c 连词，wsc 顿号，tp 时间词性短语，sp 处所词性短语，qp 数量短语。

结构和 NCC 结构的具体实例。

表 6：NHE 规则示例

序号	内部构成/中心成分	NHE 规则	示例
1.	跟“的”相关的 NHE	np → np 的 !vp np → pp 的 !vp np → sp 的 !qp	时间 的 **推移** 在电子产品 可靠性方面 的 **应用** 他们中 的 **三个**
2.	跟其他助词（似的、地）相关的 NHE	ap → !np 似的 ap → !dj 似的 dp → !qp 地 dp → !vp 地	**雪片** 似的 **他是这个地方的主人** 似的 **一寸一寸** 地 **有秩序** 地
3.	ap 扩展为 np	np → vp !ap	（联系） 教学 **实际**
4.	dj 扩展为 np	np → qp !dj	这种 **再狭窄发生率降低**
5.	vp 扩展为 np	np → qp !vp	这次 **失事**
6.	qp 扩展为 np	np → np !qp	这 **三 本**
7.	ap 扩展为 vp	vp → !a p	**远** 在 数百米甚至 数公里之外 **粗心** 到 这个地步
8.	ap 扩展为 sp	sp → vp !ap	过桥 **不远** （有一座寺庙）

上表的例 1、例 2 都是通过结构助词，系统地改变结构的性质，比如“的”“地”“似的”等结构助词，可以系统地使得结构整体的功能不同于其中中心成分的功能[10]。此外，汉语中也存在结构功能不需要标记成分的帮助，直接发生功能转换的情况，比如例 3、4、5 都是这类情况。陈述性成分、修饰性成分都直接转为指称性成分。最有意思的转类情况是像表中例 8 这样的组合，“过桥”是vp，“不远”是ap，二者组合后，表示一个处所，因而整个短语成了sp（处所词性短语）。

表 7：NCC 规则示例

序号	非同类并列	NCC 规则	示例
1.	ap – vp	ap → !ap c vp	对朋友诚实 和 帮助老人
2.	vp – dj	vp → !vp c dj	地震 与 火山喷发
3.	ap – dp	ap → !ap c dp	（失恋以后，会是）颓废 或 奋力
4.	np – dj	np → !np wco dj	电视机的改进和电视的普及，广播频道增多
5.	np – vp	np → !np wco vp	一间红瓦灰墙的小屋，一排白漆的大栅栏，或许还有三五个人影（，眨眼就消失了。）
6.	dj – tp	dj → !dj c tp	我应该今天开始 还是 明天

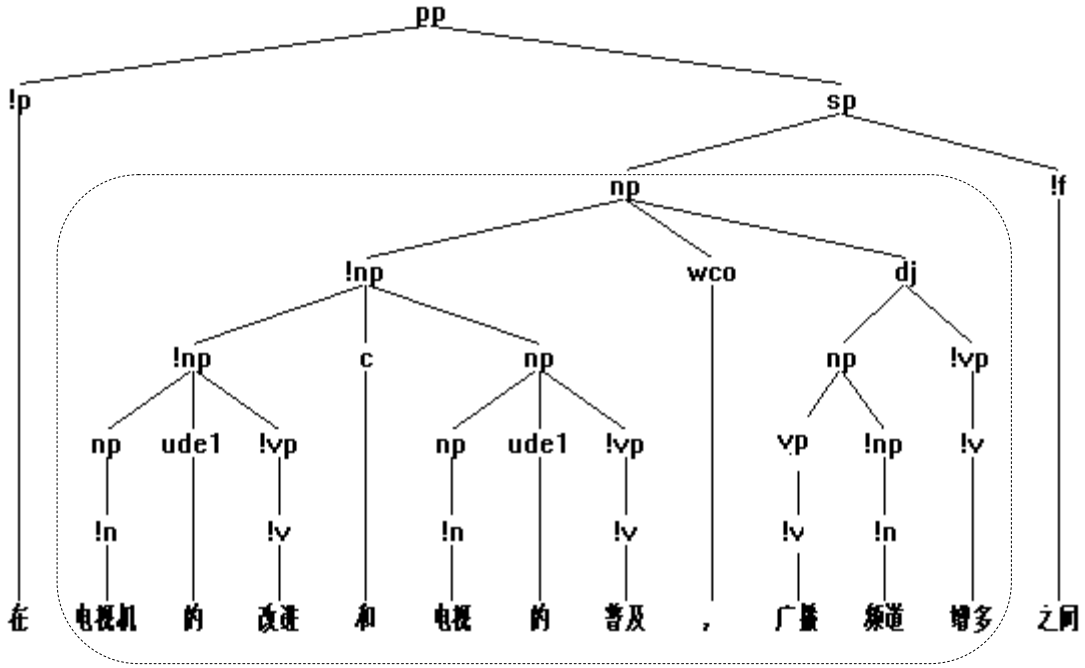
从规则形式看，上表中的规则都符合 NCC 的判定标准，但内部情况有区别。如果从宽泛的标准看，例 1 至例 3 也可以属于 CC 结构，因为例 1 和例 2 中 ap 跟 vp 并列，ap 跟 dj 并列，可以理解为都是谓词性短语，例 3 中 ap 跟 dp 并列，可以理解为都是修饰性短语，因而也可算是同类并列。尤其例 2 中“地震”因为词典收录为动词，在参与短语组合时就成为 vp。“火山喷发”是主谓式 dj，这样就形成了 vp 跟 dj 并列的 NCC 结构。但显然二者的内部结构是平行的。不过，从严格的同类并列标准看，这三种情况又应属于 NCC 结构。这些例子的启示在于：并列条件是相对的。短语类（词类）是同一范畴还是不同范畴，有一定的相对性。例 5 中并列前项是“一间红瓦灰墙的小屋，一排白漆的大栅栏”，属 np。并列后项是“还有三五个人影”，属 vp。从语义上说，后项 vp 中的“还有”几乎没有语义功能，这

[10] 有的语法理论模型主张这类结构中“的”“地”“似的”是中心成分，我们不持这种观点，可参见本文第五节的分析。关于树库语料中“X 的 Y”结构的不同类型和分布数据，参见詹卫东（2010b）。

里参与并列结构语义组合的其实是 vp 中的“三五个人影”。但在句法形式上，例 5 仍然构成 NCC 结构。例 6 的并列前项是“今天开始”（dj），后项是“明天”（tp），属典型的因为成分省略造成的 NCC 结构（传统语法中一般称为“骈合结构”）。

下面重点分析一下例 4 中 np 跟 dj 构成并列结构的情形。通过分析这个例子可以看出，汉语中陈述性成分确实可以比较容易地转为指称性成分。这个性质造成的后果是：中心扩展条件、并列条件、理论的简约性之间的关系其实鱼跟熊掌，无法兼得。

图 2：一个 NCC 规则的结构树图



例 4 的句法结构分析引出的一个核心问题是：汉语中主谓结构是陈述性成分还是指称性成分？如果是陈述性成分，则主谓结构跟 np 并列的时候，就违反了“并列结构条件”，如上面图 2 所示，np 跟 dj 并列了。如果把“广播频道增多”分析为指称性成分，则不违反并列结构条件。但是，当它不违反“并列结构条件”的时候，还要进一步追问，主谓结构的中心语又是什么呢？如果主谓结构的中心语是谓语 vp，那么，例中作为指称性成分的主谓结构（跟 np 并列的 dj），其功能就又跟 vp（“增多”）不一致了，前者是起指称作用，而后者是起陈述作用。这样，就又违反了“中心扩展条件”。这个例子显示，在短语结构的组合过程中，如果在一个层次上要遵守“并列结构条件”，就可能在另一个层次上违反“中心扩展条件”，二者难以兼顾。如果要兼顾，就必然要迫使“增多”从动词类转为名词类。这样才能使得“广播频道增多”符合中心扩展条件的要求。这样处理后，上面的结构就同时满足中心扩展条件和并列条件了。但带来的问题是显然的，就是理论不再简约。不仅“增多”现在要兼属动词和名词两类，而且“广播频道增多”要兼属主谓式单句（dj）和名词性短语（np）两类。更为严重的是，这种情况并非个例，而是普遍存在的。下面引沈家煊（2009b）举过下面的例子：

> 美国的介入是肯定的。无非是硬介入还是软介入，以及介入力度大小的问题。……所以美国介入是有条件的，这些条件也是我们可以利用的，要让美国感觉到它的介入将付出他所不能承受的代价，这样它就会选择不介入或少介入。

这个例子中开头的主语是“美国的介入”，后面则用“美国介入”作主语。仅相差一个“的”字，但其中蕴含的汉语句法结构的特点却需要引起足够的重视：汉语中的主谓结构既可以用于陈述表达功能，也可以用于指称表达功能，在表层句法结构上是表现为主谓结构可

以做主语，而且本身不需要添加形式标记。有“的”和没有“的”，都不妨碍“美国介入”起指称作用。区别只是，有“的”之后，“美国的介入”只能用于指称（句法上分析为 np），没有“的”的时候，“美国介入”可以用于指称，也可以用于陈述（句法上分析为 dj）。主谓结构本身的内部中心成分是谓词性的（起陈述表达功能），但主谓结构整体则既可以起到谓词性功能（作谓语），也可以起到体词性功能（作主、宾语）。因为汉语主谓结构存在上述用法特点，构建汉语短语结构规则系统时不违反“中心扩展条件”就很难做到了。当然技术上也并不是没有办法处理这种情况。比如，上面图 5 例子中的 np 并列结构要同时满足同类成分并列的条件和中心扩展条件，可以采取的策略是假设“改进”“普及”和“增多”除动词外，同时也是名词（这就是沈家煊先生汉语词类包含模式的思想，本质上也就是传统的动词名物化的主张），但这样做的后果就是，理论上在词和短语层面都会造成大面积的兼类问题——极端情况就是所有的谓词性短语 dj、vp、ap 等都同属 np。

黄昌宁等（2009a，2009b）主张语言工程中把“n+v”（如“文艺批评”）、“a+v”（如“重大调整”）、“v+v”（如“继续教育”）的中心成分都处理为名词，并认为这样处理后动名兼类词比例并不高，“不会打破‘兼类的词只能是少数’的格局，也不会造成‘词无定类’的恶果”。但实际上，问题远不是黄先生所估计的那么乐观。沈家煊先生在提出汉语实词包含模式时，也主要是以“这本书的出版”这样的结构为分析对象的，认为只要把“出版”处理为既是动词，同时又是名词，就解决了中心扩展条件的问题了。但其实汉语中短语结构分析和词类问题的要害并不在于上面这些结构，而在于像主谓结构和述宾结构这样的谓词性短语（陈述性成分）都可以无标记地转为非陈述性用法。下面是陈述性成分用在非陈述性的句法位置的一些实例：

例 3：dj 直接做主语

a <u>六千名代表汇集北京</u> 标志着 科学的春天的到来。

b <u>我们结婚</u> 已经有三十年了。

c <u>他们能否成功</u> 还不清楚。

例 4：dj 直接做定语

a <u>二氧化硫年均浓度达标</u> 城市

b <u>饮水困难和不安全</u> 问题

c 扣除 <u>物价上涨</u> 因素 后，……

例 5：dj 作“的”字定中结构的定语和中心语

a **呼吸引起** 的 **胸壁及腹腔内压改变**

b **总工程师说** 的 **车身悬浮**

c **官方计算** 的 **人均收入低于 150 元**

例 6：vp 直接做定语

a 妇女经济地位的提高，是<u>实现男女平等</u>最重要的基础。

b 出口 <u>年产 20 万吨纯碱</u> 成套设备

c 在<u>发展日中关系</u>方面做出了很大贡献

例 7：vp 直接作定中结构的中心语

a 国学家们的<u>崇尚国粹</u>自是情理之中的事

b 爱情也会随着天气的<u>变冷</u>而变冷吗

限于篇幅，这里只简要分析一下例 7a 的情况。“崇尚国粹”是一个典型的述宾结构 vp。而“国学家们的崇尚国粹”是一个用作指称表达的短语，整体功能类应属 np，于是，这个结构就成了一个 NHE 结构：vp 作 np 的中心语。如果要避免这种分析，就得想办法把“崇尚国粹”变成 np。一旦这样处理了，问题就接踵而来：这个作为 np 的“崇尚国粹”中的“崇尚”是 v 还是 n 呢？“崇尚国粹”的中心词是“崇尚”还是“国粹”呢？如果一个为了不违

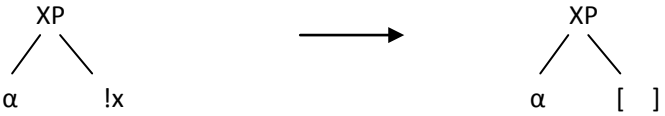
背中心扩展条件而构建的汉语语法理论模型将“崇尚国粹”分析为兼属 vp 和 np 两类，其中的“崇尚”也兼属 v 和 n 两类，我们就没有任何理由相信，这样的理论模型会是简约的，至少，它不会比不主张“崇尚国粹”和“崇尚”兼类的理论模型简约。

五 为什么会有违反“中心扩展条件”和“并列条件”的结构

那么，为什么在常规的 HE 结构和 CC 结构之外，语言系统会有 NHE 和 NCC 结构呢？

我们的假设是：言语交际过程中，说话人（信息发出方）为了编码的简约，会尽量压缩编码长度，省略那些信息量低的成分，而这样的成分在语篇中，往往是上文已经出现的旧信息或背景信息，在句法结构中，这样的成分一般是中心成分。中心成分省略后，新的结构体很可能就违反中心扩展条件了[11]。下面是这个过程的简要示意[12]：

图 3： 甲 常规结构（HE 结构） 乙 非常规结构（NHE 结构）



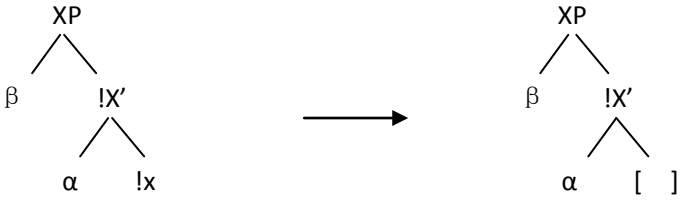
例 8 a 成套的 书 a’ 成套的

b 实际 情况 b’ 实际

例 8a“成套的书”的中心成分“书”省略后，新的结构“成套的”从句法分布功能的角度讲，跟原来的“成套的书”等同，因此仍属 np，但中心成分既不是“成套”（vp），也不是“的”（结构助词）。如果一个语法学理论强调“中心扩展条件”是不能违反的，那么为了解决这个问题，就必然要设计一套办法，或者假设有一个隐形的中心成分，或者让“的”成为中心成分，然后把整个短语分析为 DeP（不同于 np）。

例 8b“实际情况”省略“情况”后，剩下的“实际”就要担负起原来由“实际情况”担负的功能，也就是“实际”将分布在“实际情况”的句法位置上，比如“教学实际（情况）”。“实际”原本的词类属性为 a（形容词），而“实际情况”的短语属性为 np（名词性短语），为了不违背中心扩展条件，唯一的选择就是让“实际”兼属 a 和 n 两类。而当一个 NHE 结构本身又是某一个结构的中心成分时，就又会造成新的 NHE 结构。这个由 HE 结构通过省略不断滋生出 NHE 结构的过程可图示如下：

图 4： 甲 常规结构（HE 结构） 乙 非常规结构（NHE 结构）



例 9 a 军马的 饲养 方法 a’ 军马的 饲养

b 教学 实际 情况 b’ 教学 实际

[11] 如果碰巧中心成分省略后留下的从属成分的功能类跟短语整体功能类一致，则造成新的组合不违反“中心扩展条件”的假象。比如“电视 节目”省略中心成分“节目”后剩下的“电视”也是 np，跟原来“电视节目”的功能类 np 是一样的，这种情形下，省略后的新结构可以看作没有违反“中心扩展条件”。

[12] 限于篇幅，这里只讨论 HE 结构到 NHE 结构的过程，NCC 结构的形成机制跟 NHE 结构是一样的。从略。图 3 和图 4 中 XP，X’，x 等标记是采用 X-bar 理论中习惯的表达方式，即符合中心扩展条件的结构模式，x 作为中心成分（head），一路向上投射，生成同功能类的 XP 短语。α、β 表示不同的从属成分。

例 9a、b 代表了常规的定中式 np 的两种结构类型：“xp 的 !np”和“xp !np”（xp 代表任意短语类型）。例 9a′、b′则代表了由于省略中心成分造成的定中式 np 的两种新的结构类型：“xp 的 !vp”和“xp !ap”。例 9a′的中心成分不再是 np，变成了动词“饲养”，例 9b′的中心成分也不再是 np，变成了形容词“实际”。如果一定要坚持结构必须满足“中心扩展条件”，那么，就必然要让“饲养”和“实际”都兼属 n。

不过，遗憾的是，仅仅让词兼类，还是解决不了这个问题。根据上一节的分析，要满足中心扩展条件，需要让所有的动词短语 vp、主谓结构 dj，形容词性短语 ap 等谓词性短语类都兼属 np 才行。因为在图 4（例 9）中的 α 位置，不仅可以出现单个谓词，还可以出现谓词性短语。当体词性短语的中心成分（np）省略后，整体的功能必然要由原来的从属性成分（即 α 位置上的谓词性短语）来充任，只有让 α 位置上的成分都归属 np，才能在 α 的父节点所在树结构往上的层次避免违反中心扩展条件的问题。但是，将 α 位置上的成分归属 np 这个操作本身，仍然是违反中心扩展条件的。对此，可以用上面图 4 的树结构演化过程来做进一步说明。例 9a“军马的饲养方法”可以用图 4-甲树结构来解释其构造。完全符合“中心扩展条件”。当其中的中心成分“方法”省略后，造成了“军马的饲养”这样的表层序列，其结构就面临着违反“中心扩展条件”的问题。“方法”省略后的树结构如图 4-乙所示。而怎么处理这个含有“空位”的树结构（把“空洞”删除掉），则有两种方式：

图 5：

图 5-甲所示的方式就是主张兼类的方式，即让原来的从属成分 α 兼属 XP/X′类（这里就是让“饲养”兼属 np 类）。图 5-乙所示的方式则是不主张兼类的方式，即认为汉语词类（短语类）多功能的理论模型。在这个模型下，“饲养”仍然以 vp 身份成为 np 的中心语。显然，无论是哪一种方式，本质上都要违反中心扩展条件，只不过具体发生违反的树结构层次不同罢了。图 5-甲是在 X′→ !α 这个层次违反，图 5-乙是在 XP → β !α 这个层次违反。

上述过程如果用形式文法规则来描述，需要相同数量的规则。为了能同时解释“军马的饲养方法”和“军马的饲养”，主张兼类的理论模型和不主张兼类的理论模型分别对应的形式规则如下表所示：

主张兼类的理论模型	不主张兼类的理论模型
np → np ude1 !np np →!vp	np → np ude1 !np np → np ude1 !vp

显然，为了描述同样的语言现象，两种用自然语言描述的不同的理论模型，在简单程度上并无差异。更重要的是，无论哪一种模型，都必然要包含 NP → (α) !VP 这样的被 Lyons 称为是“有悖常情”的规则（规则中 α 是可选的）。

通过本文的分析可以看到，汉语中最重要的语法现象是：陈述性成分可以无标记地用作指称性成分[13]（沈家煊 2007,2009a，2009b均指出了这一点）。这一“事实”的存在，使我们有机会对主流语法理论中强调的“中心扩展条件”和“并列条件”进行更深入的反思。当代

[13] 实际语料中有的例子甚至同时兼具“陈述性”和“指称性”，比如：“**这一批精品成套的出版**，标志我国蒙古族当代文学的发展已经进入了一个崭新的阶段。”例中“这一批精品成套的出版”，既可分析为 dj，又可分析为 np。

语法理论模型青睐“中心扩展条件”的理由很充分，因为符合条件的组合规则是“简单”的，中心成分与结构整体在功能上的共性使得语法规则的组织更容易系统化，X-bar理论就是这样的理论模型的典型代表。但是，语言“事实”却并不会轻易就范逻辑上“完美”（简约）的理论。言语交际中的实际使用者会不断尝试突破“中心扩展条件”（以及“并列条件”）的限制，因为遵守这个条件的约束，必然意味着更长的编码，而使编码缩短，是言语符号使用者的“天性”。追求言语表达的“简约”，必然造成语言句法系统的“繁化”——即组合规则（或组合模式）的数量（类型）增多。这个“冲突”不可能靠提出更简约的理论模型来回避。

六 余论

本文从语言工程句法标注实践中观察到的违反“中心扩展条件”和“并列条件”的实例出发，对汉语语法理论模型设计中是否应该以及如何遵循“中心扩展条件”进行了分析和反思。结论是汉语句法理论模型不可避免地需要包含违反“中心扩展条件”和“并列条件”的组合规则（当然这样的规则在实际语料中只占相当少的比例）。沈家煊先生正确地指出了汉语陈述性成分可以无标记地“构成”指称语的事实，但是由此提出的“汉语实词包含模型”并不是一个更简约的理论设计。从语言工程的角度看语法理论模型的设计，一个重要的评价标准是，理论是否能够反映语言事实中存在的区别（而不是掩盖区别）。当我们觉察到“这本书的出版”跟“这本书的封面”，“语言学习”跟“语言系统”是有区别的组合时，语言工程中就应该把区别表达出来，比如“这本书的出版”标记为“np ude1 vp”组合模式，“这本书的封面”标记为“np ude1 np”组合模式；“语言学习”标记为“np vp”组合模式，“语言系统”标记为“np np”组合模式。而不是相反，把这些组合处理成相同的模式，即把“出版”“学习”跟“封面”“系统”一样，都标记为 np。这样的“兼类”做法对揭示区别并无帮助，虽然这样做达到了“表面上规则符合中心扩展条件”的效果。从语言工程和为计算机做句法结构分析的目的来讲，是否把“出版”“学习”处理为动、名兼类，并不重要。重要的是用“类+特征描述”的知识表达方法，尽可能精细地去刻画词语的分布差异（比如“出版这本书”可以合法的变换为“这本书的出版”，而“属于这本书”却不能平行地变换为“这本书的属于”），以及词语在参与组合时对其组合对象的选择约束（詹卫东，2010a）。

针对汉语句法系统中存在违反中心扩展条件和并列条件的组合（虽在少数，但非特例），本文提出了“语用省略”假说来加以解释。对于语用省略的句法后果，有两点需要补充说明：

（1）省略不是无限制的。解码的负担是对编码简约度的一个约束。编码过检，会造成大量歧义，带来解码负担过重，影响交际，社会中的每个人都既是信息发出方，同时又是信息接收方，在编码简约与解码负担加重的博弈过程中，信息发出方跟信息接收方的讨价还价最终达成语言系统的一个相对稳态。

（2）并不是所有的 NHE 结构都能解释为是 HE 结构的省略。比如“这本书的出版”就不容易像上面分析“军马的饲养”那样，做出类似的分析。对此，我们的看法是，一旦一个 NHE 结构（比如 np → xp ude1 !vp）形成，使用时间长了，它就可能跻身常规结构模式，从而不再由“常规结构”（HE 结构）经由省略这个渠道来产生实例，而是以常规的类推的方式，直接由 NHE 规则产生新的实例。比如“这本书的出版”可以解释为由“np → xp ude1 !vp”规则直接生成。

最后，值得一提的是，汉语中“np 的 vp”是典型的书面语结构，而非口语结构。我们猜测，书面语需要“的”来显示标记整个结构的体词性（指称性），尽管不用“的”的“np vp”结构本身也可以兼具陈述性和指称性。而在口语中，上下文语境更清楚，同时，信息发出方更具缩短编码（简约）的动力，人们就会更倾向于使用“np vp”这样的短编码模式（如“美国介入”）来表达指称，而不倾向用“np 的 vp”（如“美国的介入”）这样的长编码模式。有关编码压缩程度与句法结构系统歧义程度大小之间的关系以及达成平衡的机制，“np 的

vp”跟“np vp”在书面语和口语上的语用差异，对于认识汉语陈述语和指称语之间的关系，有重要的价值，这里只是先简要地提出了问题，要把这些问题研究清楚，还有待树库语料扩大规模及更为细致的标注。

参考文献：

陈锋、陈小荷，2008，基于树库的现代汉语短语分布考察，《语言科学》2008 年第 1 期，12-17 页。

程工，1999，《语言共性论》，上海外语教育出版社 1999 年版。

黄昌宁、姜自霞、李玉梅，2009a，形容词直接修饰动词的“a+v”结构歧义，《中国语文》2009 年第 1 期，54-63 页。

黄昌宁、李玉梅，2009b，评动、名兼类词的四种划类策略 —— 来自语言工程的观察，《语言学论丛》第 40 辑。74-92 页。

陆俭明，2003，对“NP+的+VP”结构的重新认识，《中国语文》2003 年第 5 期，387-391 页。

沈家煊，2007，汉语里的名词和动词，《汉藏语学报》2007 年第 1 期。27-47 页。

沈家煊，2009a，我看汉语的词类，《语言科学》2009 年第 1 期。1-12 页。

沈家煊，2009b，我只是接着向前跨了半步——再谈汉语的名词和动词，《语言学论丛》第 40 辑。3-22 页。

司富珍，2002，汉语的标句词“的”及相关的句法问题，《语言教学与研究》2002 年第 2 期，36-42 页。

司富珍，2004，中心语理论和汉语 DeP，《当代语言学》2004 年第 1 期，26-34 页。

吴长安，2006，“这本书的出版”与向心结构理论难题，《当代语言学》2006 年第 3 期，193-204 页。

熊仲儒，2005，以“的”为核心的 DP 结构，《当代语言学》2005 年第 2 期，148-165 页。

俞士汶、段慧明、朱学锋、孙斌、常宝宝，2003，北大语料库加工规范：切分•词性标注•注音，Journal of Chinese Language and Computing, Vol.13, No.2，121-158 页，新加坡。

詹卫东，2008，信息处理用短语功能标记体系的设计及语料库标注实践，两岸四地语言学论坛•澳门，2008 年 12.5-7，澳门理工学院。

詹卫东，2010a，计算机句法结构分析需要什么样的词类知识——兼评近年来汉语词类研究的新进展，第 16 次现代汉语语法学术研讨会，2010.6.7－9，香港•香港城市大学。

詹卫东，2010b，从结构分类到功能分类—— 以“的”字短语和“所”字短语的分析为例，走向当代前沿科学的现代汉语语法研究国际学术研讨会，2010.8.17－19，北京大学。

周国光，2005，对<中心语理论和汉语的 DeP>一文的质疑，《当代语言学》2005 年第 2 期，139-147 页。

周强，2004，汉语句法树库标注体系，《中文信息学报》2004 年第 4 期，1- 8 页。

Abeillé, Anne, ed., 2003, Treebanks: Building and Using Parsed Corpora, Kluwer Academic Publishers.

Lyons, J. 1968. An Introduction to Theoretical Linguistics. Cambridge; Cambridge University Press.

Radford, Andrew 1988. Transformational Grammar: A First Course. Cambridge: Cambridge University Press.

Xue, Nianwen, Fei Xia, Fu-Dong Chiou & Marta Palmer (2005) The Penn Chinese Treebank: Phrase structure annotation of a large corpus, In Natural Language Processing II(2): pp.207-238. Cambridge University Press.