一个适合汉语的带有范畴转换的组合范畴语法

王庆江，陈淑娴
华北水利水电大学 信息工程学院，河南 郑州 450046
wangqingjiang@ncwu.edu.cn, 1152825510@qq.com

摘要

为使汉语句子中词或短语的范畴对应其句法功能，在组合范畴语法中添加范畴转换。把词类和短语结构的范畴分别按出现率和是否由结合规则得到分为典型和非典型，建立短语结构中词类和短语结构的范畴转换规则。实词或短语结构通过范畴转换与虚词搭配，让虚词的句法功能趋于明确。树库显示，35%的短语结构形成需要范畴转换，使用范畴转换的短语直接成分中99.67%是实词或短语结构，范畴转换使组合范畴语法适合缺乏屈折的汉语。

关键词： 组合范畴语法 ； 范畴转换 ； 范畴类型透明性 ； 树库

A Chinese-Suitable Combinatory Categorial Grammar with Categorial Conversions

Qing-jiang Wang, Shu-xian Chen
School of Information Engineering
North China University of Water Resource and Electric Power
Zhengzhou 450046, China
wangqingjiang@ncwu.edu.cn, 1152825510@qq.com

Abstract

To make categories of words or phrases in Chinese sentences correspond with their syntactic functions, categorial conversions are added into Combinatory Categorial Grammar. Categories of parts of speech and phrasal structures are divided into the classical and the non-classical respectively by occurrence rate and whether obtained via combinatory rules, category conversion rules are established for parts of speech and phrasal structures in phrasal structures. Notional words or phrasal structures collocate with functional words by categorial conversions, making syntactic functions of functional words tend to definite. Treebank shows, 35% of phrasal structure formations require categorial conversions, and 99.67% of phrasal immediate components using categorial conversions are notional words or phrasal structures, and categorial conversions make CCG adapt to inflectional-absent Chinese.

Keywords: Combinatory Categorial Grammar , categorial conversion , categorial type transparency , treebank

1 引言

基金项目：河南省重点研发与推广专项（212102210495）

在深度学习推动自然语言处理全面发展的时代，语言语法的研究仍然重要(冯志伟, 2021)。组合范畴语法[†] (Steedman, 2019)（Combinatory Categorial Grammar，缩写CCG）是范畴语法的一种扩展，用类型提升和函数组合解释宾语提取、状中、中补等短语结构，用斜线类型将句法类型结合的精确控制由规则一侧转到词汇一侧，使规则跨语言通用，且仍具有句法类型结合伴随语义组合的范畴语法亮点 (Jayant and Mitchell, 2014)，这些使CCG具有重要的计算语言学价值 (陈鹏, 2016; 满海霞, 2022)。

CCG给词指派句法类型（即范畴）及其关联的语义解释，例如 (Steedman, 2019)给sees指派及物动词范畴和解释其语义的λ-项，sees := (S\NP$_{3s}$)/NP: $\lambda x\lambda y$.sees' x y，这里句法类型(S\NP$_{3s}$)/NP指右结合NP得句法类型S\NP$_{3s}$，后者左结合NP$_{3s}$得句法类型S，语义式$\lambda x\lambda y$.sees' x y在先后应用于右侧句法类型NP相伴的语义式x'、左侧句法类型NP$_{3s}$相伴的语义式y'后得到谓词-论元结构sees' x' y'。一个句法类型对应一种句法功能，而词类是词按句法功能分的类，故可按词类考虑词的句法类型。但是汉语里"词有定类"和"类有定职"两难 (胡明扬, 1995)，词类问题复杂，厘清词类问题后才能基于词类给词指派范畴。

另一方面，范畴语法的规则通常指两个范畴结合为一个范畴的规则，但其实还可以是一个范畴转为另一个范畴的规则 (Carpenter, 1991)，前者是跨语言通用的范畴结合规则，后者是因语言而异的范畴转换规则。若结合衍生的范畴不对应短语结构充当的句法成分，就需要转换范畴实现对应。汉语里短语结构缺乏屈折，存在大量的这种不对应。

关于词类及其句法功能，语言学界经过漫长争论，逐渐倾向词有定类而类无定职，并可按出现率把词类的句法功能分为典型和非典型，而短语结构的范畴也可按是否由结合规则得到分为典型和非典型，这样就可以对词类和短语结构统一讨论范畴从典型到某个非典型的转换，建立一个带有范畴转换的组合范畴语法（CCG with Category Conversions, 缩写CCC-C^2）：词典里力争词与范畴一一对应，在句子结构里可以通过范畴转换给同一词类或短语结构指派不同范畴。CCC-C^{2}中的范畴转换是句法层面的，与文献 (Carpenter, 1991)"词:=范畴"闭包中范畴转换的不同在于，后者是词法层面的。要使CCG-C^{2}适合汉语的各种句法特征，只能通过建立树库，才能逐渐明确词类和特殊词的典型范畴，形成词类和短语结构的范畴转换规则体系。

本文的创新性工作有两个方面：（1）在句法层面统一考虑汉语词类和短语结构的范畴转换，即词类和短语结构都可以按所在短语结构的需要由典型范畴改用非典型范畴。（2）通过构建树库，形成适合汉语的范畴转换规则体系。

本文第2节给出组合范畴语法的基本定义；第3节阐述范畴转换规则的汉语言学依据；第4节按范畴结合规则建立句法成分到范畴的一一映射，使实词类或短语结构有典型范畴，由虚词短语的典型范畴得到虚词的典型范畴，并论述一些特殊词的范畴设立依据；第5节举例说明短语结构中的各种范畴转换规则；第6节由树库统计评价这个带有范畴转换的CCG；最后总结工作，指出下一步的研究方向。

2 组合范畴语法的基本定义

英语的基本范畴一般有n、np、pp和s，分别对应名词、名词短语、介宾短语和句子，若考虑数、格、屈折等特征，可有更多基本范畴。汉语里数、格、屈折不明显，本着能归于其他基本范畴就不单设的想法，基本范畴可只有np和s。范畴（Category）用巴科斯范式（BNF）定义如下，其中斜线（Slash）类型'.'、'$\diamond$'、'$\times$'、'$\star$'实现词对规则选择的控制 (Steedman, 2019)，含有斜线的范畴是衍生范畴。若斜线左侧或右侧是衍生范畴，要用圆括号括起来，以保持二分性，例如(s\.np)/.np。衍生范畴可看作函数，斜线左侧为函数结果，右侧为函数参数，方向反映范畴序列中参数位于函数的哪一侧。

SyntaxType ::= np | s | SyntaxType Slash SyntaxType

Slash ::= /. | \. | /$_\diamond$ | \$_\diamond$ | /$_\times$ | \$_\times$ | /$_\star$ | \$_\star$

范畴语法的规则有关联的语义运算，且有类型透明性 (Steedman, 2019)，即句法类型决定语义类型。高阶组合规则（>B^n、<B^n、>B$^n_\times$、<B$^n_\times$）可以无限枚举，但目前也只发现二阶组合的作用，如前向二阶组合规则（>B^2）允许副词或能愿动词与双宾语动词的结合。提升规则

[†]早期对范畴语法（CG）的扩展，即对相邻范畴的包裹（Wrap）、组合、提升、替换等函数运算，本质上都是结合的 (Steedman, 2011)，都源于Moses Schönfinkel的结合子（Combinator），故Combinatory Categorial Grammar应该是这一类范畴语法扩展的总称或者是基于各种结合子的范畴语法，并译为"结合范畴语法"。Mark Steedman的CCG主要是新增了结合子Z（Haskell Brooks Curry称之为B）对应的函数组合，这也许是国内将CCG译为"组合范畴语法"的原因。

（<T、>T）总是和组合规则连用。用下列规则形成汉语各种短语结构，‘⇒’左边匹配短语内部结构，右边得到短语整体功能。在范畴关联的语义部分，a是常量，z、w是变量，f、g是函数，函数也可表示为λ-抽象。

$$\mathrm{X}/_{\star}\mathrm{Y}{:}f \quad \mathrm{Y{:}a} \Rightarrow \mathrm{X}{:}f\ \mathrm{a} \qquad (>)$$
$$\mathrm{Y{:}a} \quad \mathrm{X}\backslash_{\star}\mathrm{Y}{:}f \Rightarrow \mathrm{X}{:}f\ \mathrm{a} \qquad (<)$$
$$\mathrm{X}/_{\diamond}\mathrm{Y}{:}f \quad \mathrm{Y}/_{\diamond}\mathrm{Z}{:}g \Rightarrow \mathrm{X}/_{\diamond}\mathrm{Z}{:}\lambda z.\, f\,(g\,z) \qquad (>\mathrm{B})$$
$$\mathrm{Y}\backslash_{\diamond}\mathrm{Z}{:}g \quad \mathrm{X}\backslash_{\diamond}\mathrm{Y}{:}f \Rightarrow \mathrm{X}\backslash_{\diamond}\mathrm{Z}{:}\lambda z.\, f\,(g\,z) \qquad (<\mathrm{B})$$
$$\mathrm{X}/_{\diamond}\mathrm{Y}{:}f \quad (\mathrm{Y}/_{\diamond}\mathrm{W})/_{\diamond}\mathrm{Z{:g}} \Rightarrow (\mathrm{X}/_{\diamond}\mathrm{W})/_{\diamond}\mathrm{Z}{:}\lambda z\lambda w.\, f\,((g\,z)\,w) \qquad (>\mathrm{B}^2)$$
$$(\mathrm{Y}\backslash_{\diamond}\mathrm{W})/_{\diamond}\mathrm{Z{:g}} \quad \mathrm{X}\backslash_{\diamond}\mathrm{Y{:f}} \Rightarrow (\mathrm{X}\backslash_{\diamond}\mathrm{W})/_{\diamond}\mathrm{Z}{:}\lambda z\lambda w.\, f\,((g\,z)\,w) \qquad (<\mathrm{B}^2)$$
$$\mathrm{X}/_{\times}\mathrm{Y}{:}f \quad \mathrm{Y}\backslash_{\times}\mathrm{Z}{:}g \Rightarrow \mathrm{X}\backslash_{\times}\mathrm{Z}{:}\lambda z.f\,(g\,z) \qquad (>\mathrm{B}_{\times})$$
$$\mathrm{Y}/_{\times}\mathrm{Z}{:}g \quad \mathrm{X}\backslash_{\times}\mathrm{Y}{:}f \Rightarrow \mathrm{X}/_{\times}\mathrm{Z}{:}\lambda z.f\,(g\,z) \qquad (<\mathrm{B}_{\times})$$
$$\mathrm{X{:a}} \Rightarrow \mathrm{T}/_{i}(\mathrm{T}\backslash_{i}\mathrm{X}){:}\lambda f.f\,\mathrm{a} \qquad (>\mathrm{T})$$
$$\mathrm{X{:a}} \Rightarrow \mathrm{T}\backslash_{i}(\mathrm{T}/_{i}\mathrm{X}){:}\lambda f.f\,\mathrm{a} \qquad (<\mathrm{T})$$

不同语言的语法区别仅在词法层面，按词法形成的词汇通过“结合”规则[‡]映射到语言的句子 (Steedman, 2019)，即确定句子里每个词的范畴，按规则一步步结合相邻范畴，就可以得到句子结构。由“结合”规则得到的范畴再与相邻范畴结合，有是否引入范畴转换的两种做法 (王庆江, 张琳, 2020)。词库里可记录每个词的典型和非典型范畴，但词选择哪个非典型范畴即范畴转换，只发生于范畴转换规则表示的上下文，这避免了范畴转换对语法表达力的过度增强。范畴转换其实是把句子的词范畴歧义转化为句法歧义，是词或短语句法功能转换的客观反映，仍属于词法层面。

3 范畴转换规则的汉语言学依据

能不能绕过词类，直接考虑词或短语的范畴转换，本质上是语法中能否去掉词类这一概念的问题，汉语言学里该问题的争论止于上世纪五十年代，之后就是争论如何划分词类 (吕叔湘, 1954)。划分词类的目的是为了进行句法分析，划分词类的标准必须考虑句法分析的需要 (胡明扬, 1995)。上世纪五十年代前，词类划分和给词定类都采用意义标准；五十至八十年代，词类划分渐趋句法标准，而给词定类仍坚持意义标准。九十年代及以后，给词定类趋于句法标准，即只根据词参与构建短语和充当句子成分的能力 (沈家煊, 2009)。

给词定类必须先于句法分析，否则词类对句法分析起不到任何作用 (胡明扬, 1995)。然而，“离句无品”，词性通过词所在的例句体现，先确定词在句中充当的句子成分再确定词性的反向逻辑长期存在。《现代汉语词典》第五版才对词做词类标注 (徐枢, 谭景春, 2006)，可见语言学界对词语是否有固有词性是多么纠结。“词有定类”指每个单词力求归于一类，“类有定职”指词类与句子成分一一对应。汉语里，“类有定职”则“词无定类”，词类失去存在意义，故只能是“词有定类”，让词类与句子成分的关系错综复杂 (朱德熙, 1985)。

“词有定类”与词兼类、词同形可能是相容的。兼类词在作不同词类时，词的意义虽然相关但已经不同，而同形词的意义更是毫无联系。词类有意义基础 (张斌, 2005)，“词有定类”可能蕴涵词在一种意义下只属于一个词类，这与范畴转换发生在词类内部是一致的，即发生范畴转换时词的意义并未改变。若该假设成立，兼类词或同形词在每个义项下的词类就是唯一的。

与印欧语词类分立不同，汉语词类句法功能可以重叠甚至包含。在汉语实词类包含模式 (沈家煊, 2009)中，凡动词皆名词，即动词有名词的所有功能，“这本书的出版”中“出版”是动词也是名词，从动词角度讲符合简约原则，即不增加不必要的步骤和名目，从名词角度讲符合中心扩展，中心扩展指以一个成分为中心加以扩展，扩展后结构的语法性质跟中心成分的语法性质一致。

基于词类充当不同句子成分的出现率，可以把词类功能限制为充当出现率较高的句子成分 (胡明扬, 1995)，也可以把出现率最高的句子成分作为词类的典型功能，把其它出现率的句子成分作为词类的非典型功能 (沈家煊, 1997; 沈家煊, 1999)。如果按词有定类、类对应典型功能建立“词:=范畴”词典，则词在短语结构里需要使用非典型功能时，就需要转用非典型功能对应的范畴。

汉语缺乏系统的形态标记，不仅导致词类多功能，也造成语法的词组本位特征 (张伯江, 2011)。汉语句子的构造原则跟词组的构造原则基本一致，可以把各类词组作为抽象的句法格式

[‡]这里指实现范畴结合的所有规则，而非特指Schönfinkel结合子对应的那些规则。

来描写它们的内部结构以及每一类词组作为一个整体在更大的词组里的分布状况，把各类词组的结构和功能描写清楚了，句子的结构也就描写清楚了 (朱德熙, 1985)。词组本位思想表现为结构包孕，即短语结构的基本类型虽然很有限，但每一种结构都可以包孕与它自身同类型或不同类型的结构 (朱德熙, 1982)，这种结构包孕也被称做结构套叠 (陆俭明, 1990)。结构套叠对应到CCG里，就是短语结构的范畴由其直接成分的范畴结合而来，这样得来的范畴如果不是短语结构作为整体在更大短语里应该采用的范畴，就转换范畴然后再范畴结合，依此递归下去。

4 词类和特殊词的典型范畴

首先为句法成分指派范畴，使基本句法结构能按范畴结合规则形成，然后由句法成分的范畴确定词类的典型范畴。给定句子范畴 s 和名词短语范畴 np，由后向应用（<）、前向应用（>）、前向组合（>B）、后向组合（<B）可得谓语范畴s\.np、述语范畴(s\.np)/.np、定语范畴np/.np、状语范畴$(s\backslash.np)/_{\diamond}(s\backslash.np)$和$(np/.np)/_{\star}(np/.np)$、补语范畴$np\backslash_{\star}np$、$(s\backslash.np)\backslash_{\times}(s\backslash.np)$和$(np/.np)\backslash_{\star}(np/.np)$，使主谓（Subject-Predicate，SP）、定中（Attribute-Headword，AHn）、状中（aDverbial-Headword，DHv或DHa）、中补（Headword-Complement，HnC、HvC或HaC）、述宾（Verb-Object，VO）等结构的范畴是其成分的范畴按范畴结合规则得到的结果，这里中心成分H的语法性质可以是名词（n）、动词（v）或形容词（a）。

实词类按短语结构需要选择范畴，例如主语位置上的形容词选用范畴np，定语位置上的名词选用范畴np/.np。出现率最高的范畴是词类的典型范畴，如名词、形容词的典型范畴分别是np和np/.np。与一级词类相比，二级词类的典型范畴更明显，如谓语动词、单宾语动词、双宾语动词的典型范畴分别是s\.np、(s\.np)/.np和((s\.np)/.np)/.np。用范畴中斜线类型精确反映词类的句法功能，如数词、量词的典型范畴分别是$np/_{\star}np$和$(np/_{\star}np)\backslash_{\star}(np/_{\star}np)$，因类型'$_{\star}$'只匹配前向应用（>）、后向应用（<）中的线性类型，故数词可单独做定语，也可与量词结合后再做定语，而量词除与数词结合没有其它句法功能。

虚词不单独充当句法成分，虚词、虚词附着的实词（短语）、附着形成的虚词短语三者之间存在范畴结合规则的约束。实词类有典型和非典型范畴，不妨让虚词附着的实词（短语）使用虚词典型搭配词类的范畴，使虚词的范畴按虚词短语的范畴确定。虚词数量少，句法功能差异大，需要对每个虚词专门考虑其句法范畴。例如'的'、'地'、'得'字短语的典型功能分别是做定、状、补语，让'的'附着的实词（短语）无论是否名词性都使用范畴np，'地'和'得'附着的实词（短语）无论是否形容词性都使用范畴np/.np，'的'、'地'、'得'的典型范畴就可分别明确为$(np/_{\star}np)\backslash_{\star}np$、$((s\backslash.np)/_{\diamond}(s\backslash.np))\backslash_{\star}(np/.np)$和$((s\backslash.np)/_{\times}(s\backslash.np))/_{\star}(np/.np)$。

个别词有特殊的句法功能。助词'所'接述语形成'所'字短语，表述转指称，如"所说"指说的话，'所'的范畴可令为$np/_{\star}((s\backslash.np)/.np)$。介词'把'接宾语，再结合述语，形成谓语，如"我把馍吃了"中的"把馍吃"，故'把'的范畴可令为$((s\backslash.np)/_{\star}((s\backslash.np)/.np))/_{\star}np$，其用法如图1，其中"Desig"表示词按词典"词:=范畴"取得指派的范畴，'X'匹配任意范畴，'α'匹配任意λ-项，语义式中的函数都写为λ-抽象。范畴结合规则具有范畴类型透明性，即结合前后都有范畴类型决定语义类型。按λ-应用的左结合优先，图1句子的语义式也可写为"了' (((把' 馍) '吃') 我')"。

```
--Desig ---------------------------Desig ----------Desig ----------Desig ----------Desig
np      ((s\.np)/⋆((s\.np)/.np))/⋆np    np              (s\.np)/.np     X\.X
:我,    :λxλfλy.把' x f y               :馍,            :λxλy.吃' x y   :λα.了' α
        ------------------------------------------->
        (s\.np)/⋆((s\.np)/.np)
        :λfλy.把' 馍' f y
        ------------------------------------------------------------->
        s\.np
        :λy.把' 馍' 吃' y
-------------------------------------------------------------------------<
s
:把' 馍' 吃' 我'
-------------------------------------------------------------------------------<
s
:了' (把' 馍' 吃' 我')
```

图 1: 介词'把'的范畴及其用例

介词‘被’是被动标记，被动句可分为无施事的短被动句和有施事的长被动句 (姚从军, 俎孟晨, 2022)。在缺少出现率依据情况下，不妨设介词‘被’大多用在长被动句中，典型功能是接引宾语提取，形成句子谓语，如“饭被我吃了”中‘被’字短语是“被我吃”而非“被我”。‘被’与介宾结构中介词的功能不同，宾语提取的范畴为s/.np，故‘被’的范畴是(s\.np)/$_\star$(s/.np)，其用法如图2。若‘被’用在短被动句中，如“饭被吃了”，就让及物动词‘吃’转用其非典型范畴s/.np。

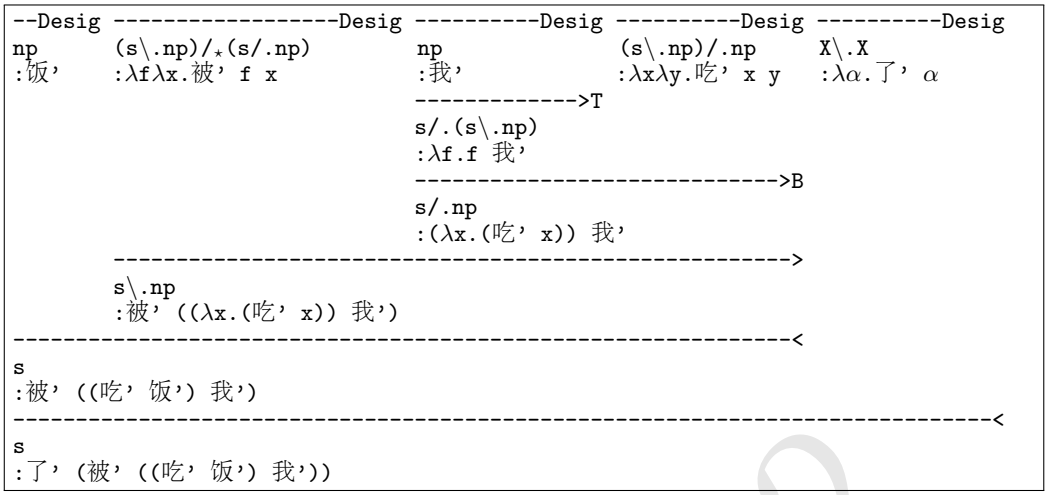


图 2: 介词‘被’的范畴及其用例

图2中，‘饭’的范畴与‘被’字短语的范畴结合，形成“饭被我吃”的范畴，而从伴随的语义结合看，语义项“吃’”是与语义项“饭’”结合，产生这一现象的根源是使用了组合规则。组合规则允许相邻的两个函数范畴先组合，预留的参数空位最终被相距较远的参数范畴填补 (满海霞, 2022)。这种参数占位可保持语义项二元结合的顺序不变，如在规则“>B”关联的语义层面，函数f和g组合为函数λx.f(g x)，没改变函数g先应用到参数x、函数f再应用到参数(g x)的顺序。

5 词类和短语结构的范畴转换规则

对于实词类或向心结构，转换标记“句法成分/词类”表示词类或以该词类为中心成分的短语结构使用句法成分对应的范畴。在构建树库过程中收集词类和短语结构的范畴转换规则，典型范畴为np的名词（短语）的范畴转换规则如表1，其中一些是时间、方位等二级名词特有的，示例下标是词或短语的范畴，‘⇒’下标是规则标记。动词、形容词、副词等实词类或以这些词类为中心的短语结构也有相应的范畴转换规则。

范畴转换规则	标记	适用短语示例	转换与结合连用（短语类型）
np np⇒np s\.np	P/n	今天$_{\text{np}}$星期一$_{\text{np}}$	$\Rightarrow_{\text{P/n}}$ np s\.np$\Rightarrow_{<}$ s (SP)
np np⇒(s\.np)/.np np	V/n	组织$_{\text{np}}$学生$_{\text{np}}$	$\Rightarrow_{\text{V/n}}$ (s\.np)/.np np$\Rightarrow_{>}$ s\.np (VO)
np np⇒np/.np np	A/n	门$_{\text{np}}$把手$_{\text{np}}$	$\Rightarrow_{\text{A/n}}$ np/.np np$\Rightarrow_{>}$ np (AH_n)
np np⇒np np\$_\star$.np	C_n/n	商城$_{\text{np}}$沃尔玛$_{\text{np}}$	$\Rightarrow_{\text{C}_\text{n}/\text{n}}$ np np\$_\star$np$\Rightarrow_{<}$ np (H_nC)
s\.np np⇒s\.np (s\.np)\$_\star$(s\.np)	C_v/n	走$_{\text{s\.np}}$一天$_{\text{np}}$	$\Rightarrow_{\text{C}_\text{v}/\text{n}}$s\.np (s\.np)\$_\star$(s\.np)$\Rightarrow_{<}$ s\.np (H_vC)
np s\.np⇒(s\.np)/$_\diamond$(s\.np) s\.np	D/n	寒假$_{\text{np}}$返校$_{\text{s\.np}}$	$\Rightarrow_{\text{D/n}}$(s\.np)/$_\diamond$(s\.np) s\.np$\Rightarrow_{>}$ s\.np (DH_v)

表 1: 名词（短语）的范畴转换规则

对于非向心结构，转换标记“句法成分/短语结构”表示短语结构使用句法成分对应的范畴，标记“词类/短语结构”表示短语结构使用词类的典型范畴，目前发现的非向心结构有主谓（s）、宾语提取（oe）和谓语提取（pe），宾语提取即“主述”短语，谓语提取即“主状”短语，如表2，其中删除线部分为短语结构的上下文，“U1P”指‘的’字短语，方位名词的典型范畴是np\$_\star$np，谓语提取的典型范畴是s/$_\diamond$(s\.np)。

范畴转换规则	标记	适用短语示例	转换与结合连用（短语类型）
s s\.np ⇒ np s\.np	S/s	桃花开$_{s}$是在春天$_{s\backslash.np}$	⇒$_{S/s}$ np s\.np ⇒$_{<}$ s (SP)
np s ⇒ np s\.np	P/s	解放军$_{np}$意志坚定$_{s}$	⇒$_{P/s}$ np s\.np ⇒$_{<}$ s (SP)
(s\.np)/.np s ⇒ (s\.np)/.np np	O/s	认为$_{(s\backslash.np)/.np}$你很努力$_{s}$	⇒$_{O/s}$ (s\.np)/.np np ⇒$_{>}$ s\.np (VO)
s np ⇒ np/.np np	A/s	他上班$_{s}$时间$_{np}$~~不长$_{s\backslash.np}$~~	⇒$_{A/s}$ np/.np np ⇒$_{>}$ np (AH_n)
s np\⋆np ⇒ np np\⋆np	H_n/s	课程安排$_{s}$上$_{np\backslash\star np}$	⇒$_{H_n/s}$ np np\⋆np ⇒$_{<}$ np (H_nC)
s (np/⋆np)\⋆np ⇒ np (np/⋆np)\⋆np	N/s	他上班$_{s}$的$_{(np/\star np)\backslash\star np}$	⇒$_{N/s}$ np (np/⋆np)\⋆np ⇒$_{<}$ np/⋆np (U1P)
(s\.np)/.np s/.np ⇒ (s\.np)/.np np	O/oe	让$_{(s\backslash.np)/.np}$你做$_{s/.np}$	⇒$_{O/oe}$ (s\.np)/.np np ⇒$_{>}$ s/.np (VO)
s/.np np\⋆np ⇒ np np\⋆np	H_n/oe	学生学习$_{s/.np}$上$_{np\backslash\star np}$	⇒$_{H_n/oe}$ np np\⋆np ⇒$_{<}$ np (H_nC)
s/.np (np/⋆np)\⋆np ⇒ np (np/⋆np)\⋆np	N/oe	他读$_{s/.np}$的$_{(np/\star np)\backslash\star np}$	⇒$_{N/oe}$ np (np/⋆np)\⋆np ⇒$_{<}$ np/⋆np (U1P)
s/$_{\diamond}$(s\.np) (np/⋆np)\⋆np ⇒ np (np/⋆np)\⋆np	N/pe	学生在班里$_{s/_{\diamond}(s/.np)}$ 的$_{(np/\star np)\backslash\star np}$~~表现~~	⇒$_{N/pe}$ np (np/⋆np)\⋆np ⇒$_{<}$ np/⋆np (U1P)

表 2: 非向心结构的范畴转换规则

虚词短语使用非典型句法功能，可以通过范畴转换，如‘的’字短语的典型范畴是np/⋆np，通过“S/a”可转用非典型范畴np，也可通过虚词转用非典型范畴得到虚词短语的非典型范畴，如助词‘得’典型功能是接形容词做述语补语，如“干得好”，非典型功能是接修饰形容词的副词（简称形容词副词）做形容词补语，如“灵得很”，存在范畴转换规则“U3d/u3”，“u3”指助词‘得’，“U3d”指右接形容词副词时‘得’的范畴((np/.np)\⋆(np/.np))/⋆((np/.np)/⋆(np/.np))。附带地，副词典型功能是修饰动词，“D_a/d”表示副词使用形容词副词的范畴，“U3d/u3”总是和副词的范畴转换规则“D_a/d”连用。

短语的两个直接成分同时使用范畴转换规则，这样的情况如表3。

范畴转换规则	标记	适用短语示例
s np/.np ⇒ np s\.np	S/s-P/a	学生住宿$_{s}$方便$_{np/.np}$
s (s\.np)/.np ⇒ np/.np np	A/s-H_n/v	社会主义现代化$_{s}$建设$_{(s\backslash.np)/.np}$
s np ⇒ np np\⋆np	H_n/s-C_n/n	人懒$_{s}$这个现象$_{np}$
np s\.np ⇒ (s\.np)/.np np	V/n-O/v	组织$_{np}$教学$_{s\backslash.np}$
np s ⇒ np/.np np	A/n-H_n/s	这$_{np}$成绩好$_{s}$~~是必然的~~
np np/.np ⇒ np/.np np	A/n-H_n/a	职务$_{np}$方便$_{np/.np}$
np (s\.np)/.np ⇒ np/.np np	A/n-H_n/v	思政$_{np}$教育$_{(s\backslash.np)/.np}$
np s\.np ⇒ np/.np np	D_a/n-A/v	这样$_{np}$好$_{(s\backslash.np)/.np}$~~的规定~~
s\.np np/.np ⇒ np s\.np	S/v-P/a	出门$_{s\backslash.np}$方便$_{np/.np}$
s\.np (s\.np)/.np ⇒ np/.np np	A/v-H_n/v	毕业$_{s\backslash.np}$设计$_{(s\backslash.np)/.np}$
(s\.np)/.np (s\.np)/$_{\diamond}$(s\.np) ⇒ np/.np np	A/v-H_n/d	~~少数有~~录取$_{(s\backslash.np)/.np}$可能$_{(s\backslash.np)/\diamond(s\backslash.np)}$~~的学生~~
(s\.np)/.np s\.np ⇒ s\.np (s\.np)\×(s\.np)	P/vt-C_v/v	分配$_{(s\backslash.np)/.np}$到基层工作$_{s\backslash.np}$
np/⋆np (s\.np)/.np ⇒ np s\.np	S/a-P/vt	留级的$_{np/\star np}$也算上$_{(s\backslash.np)/.np}$
((s\.np)\×(s\.np))/⋆(np/.np) (s\.np)/$_{\diamond}$(s\.np) ⇒ ((np/.np)\⋆(np/.np))/⋆((np/.np)/⋆(np/.np)) (np/.np)/⋆(np/.np)	U3d/u3-D_a/d	~~灵得~~$_{((s\backslash.np)\times(s\backslash.np))/\star(np/.np)}$ 很$_{(s\backslash.np)/\diamond(s\backslash.np)}$

表 3: 短语两个成分都使用范畴转换规则

范畴转换规则也具有范畴类型透明性，即范畴转换后范畴类型仍然决定语义类型，如图3。

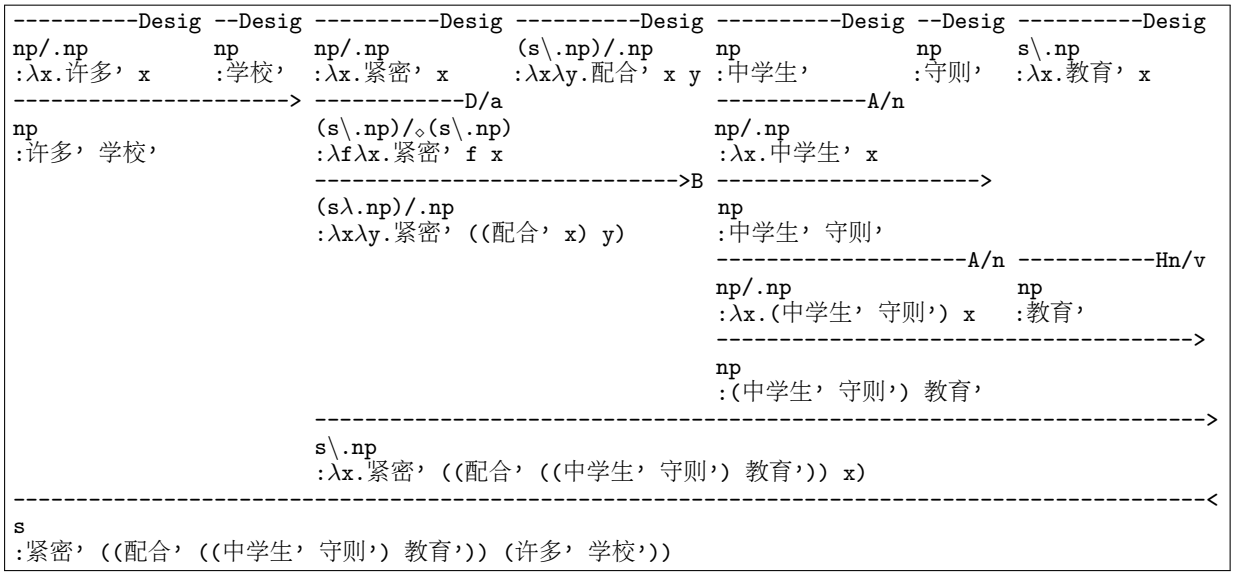


图 3: 范畴转换中的范畴类型透明性

6 语法树库的统计分析

句法分析时，若保留句法歧义，即每次传递时允许使用所有范畴转换，传递后不删除造成歧义的短语，短语集合规模大致会随传递次数指数增长，故句法分析中必须及时消解句法歧义。在构建树库过程中已记录遇到的每一个句法歧义片段，将来可通过挖掘消歧模式，并引入常识和世界知识，探索完全靠机器消歧获得人脑预期分析树的可能性。

句法歧义的主要成因是句法结构层次、句法结构关系的不同 (何洪峰, 2016)，映射到CCG分析里，就是一个范畴既可与左边的范畴结合，又可与右边的范畴结合，毗连的两个范畴有多条结合途径可用，结合途径指范畴结合经过的范畴转换与范畴结合规则。范畴转换使句法歧义严重，是汉语缺乏形态标记、句法成分之间大多可以套叠 (陆俭明, 1990)的直接后果。为限制和及时消解句法歧义，每次计算范畴结合传递时，由用户按需选择范畴转换，然后由机器完成范畴结合，再由用户消解句法歧义，使最后形成的传递闭包只含一棵分析树，这样的句法分析其实是一个计算结合-消歧传递闭包的过程。

由语法的二分性，分析树只有度为0的结点（即词）和度为2的结点（即短语）。把词看作跨度0的短语，使词和短语有统一的数据抽象。短语有外在的位置、句法类型和语义表示，也有内在的结合途径、语法关系，定义为三元组:

$$((Start, Span), (SyntaxType, Tag, Seman, PhraStru, Act), SecStart)$$

其中$Start$是短语在句中的起始位置，$Span$是短语跨度，即所含词数减一，$SyntaxType$是短语的句法类型，Tag是短语的句法类型结合途径的标记，词的Tag为“Desig”，取指定（Designate）之意；$Seman$是短语的语义式，由两个成分的语义式通过函数运算得到，词的语义用词加撇表示；$PhraStru$是短语的语法关系，可以是实词性语法单位间语法关系，如主谓（SP）、动宾（VO），也可以是与虚词有关的语法单位间关系，如宾语抽取（OE）、介宾（PO）、‘的’字结构（U1P），词的$PhraStru$为“DE”，取指定（Designated）之意。已参与形成短语时，Act为False，否则为True。$SecStart$是短语中第二组成成分在句中的起始位置。分析树存储为纯文本，从中可统计句子长度、短语个数、每个短语的外在和内部属性，并可生成树状结构图。

在平衡语料库（http://corpus.zhonghuayuwen.org/CnCindex.aspx）按“学生”检索词类标注语料，对前200句进行小句分割 (邢福义, 1995; 李艳翠, 2013)，共得到727个小句。对每个小句，将词类标注替换为范畴标注，然后进行人机交互的句法分析，得到华水树库1.0（https://github.com/wangqingjiang-ncwu/my-ccg/tree/master/doc）。

华水树库1.0共使用67个不同的范畴转换规则，占转换规则总数的91%，使用各规则的次数如图4，次数为1时不显示。动词（v）、名词（n）、形容词（a）、副词（d）使用范畴转换规则的次数比例分别是33%、32%、14%、7%，若算上连词（c）、量词（q）、非向心结构短语（s、oe和pe），以及短语两个成分都使用范畴转换，则使用次数比例合计99.67%，说明使用范畴转换的几乎都是实词或短语。平均每棵分析树使用范畴转换规则2.46个，形成7.05个短语，故35%的短语形成用到范畴转换。

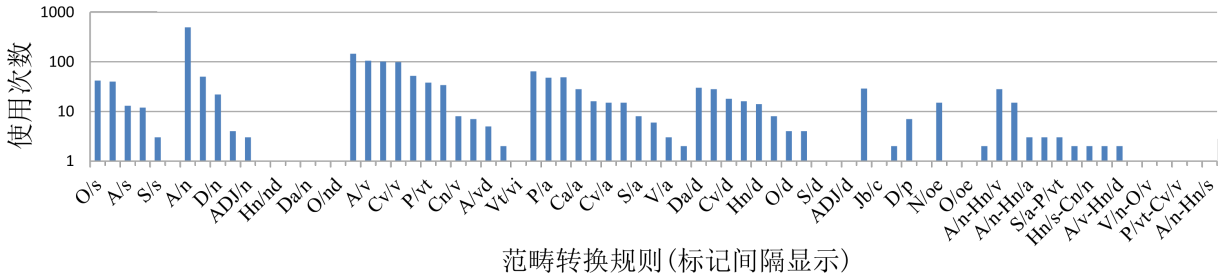
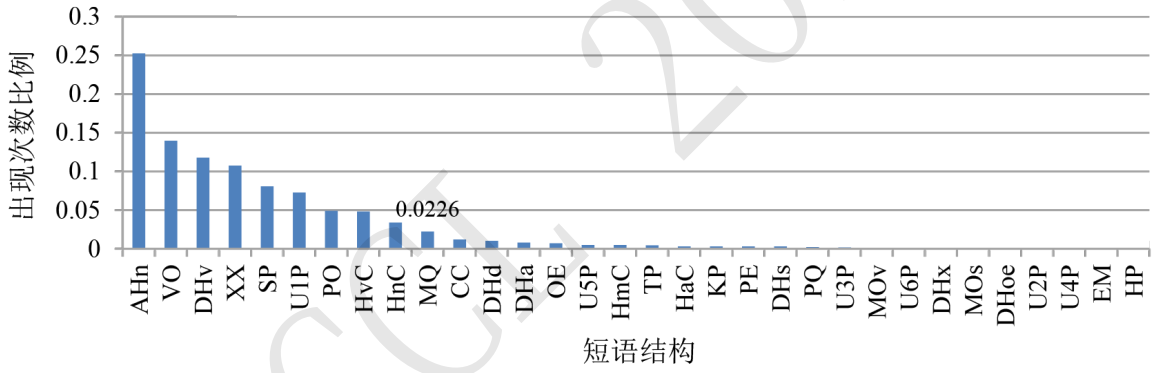


图 4: 华水树库1.0中范畴转换规则的使用次数

华水树库1.0中分析树共包含27种、5127个短语，覆盖全部短语结构类型的84%，短语类型出现次数比例如图5。因结合途径中范畴结合规则相同，连谓并入中补（HvC），兼语归入动宾（VO），复指并入中补（HnC），方位并入中补（HnC），能愿并入状中（DHv），这时定中、动宾、状中、并列、主谓，中补等基本结构可合称广义基本结构，其次数比例合计81%。'的'字短语（U1P）、介宾短语（PO）是广义基本结构外使用比例最高的两种。次数比例最高的前9种短语类型的次数比例总计90%，其它短语类型的次数比例均低于3%。



AH_n: 定中$_{名}$　CC: 并列小句　DH_a: 状中$_{形}$　DH_d: 状中$_{副}$　DH_{oe}: 状中$_{宾语提取}$　DH_s: 状中$_{句}$
DH_v: 状中$_{动}$　DH_x: 状中$_{趋向动词}$　EM: 语气短语　H_aC: 中$_{形}$补　H_mC: 中$_{数}$补　H_nC: 中$_{名}$补
HP: 前接短语，如“老三”、“第一”　H_vC: 中$_{动}$补　KP: 后接短语，如“工作者”、“中式”
MO_s: 移动宾语到主语前，即‘被’字短语　MO_v: 移动宾语到述语前，即‘把’字短语
MQ: 数量　OE: 宾语提取　PE: 谓语提取　PO: 介宾　PQ: 代量，如“这筐”
SP: 主谓短语　TP: 语调短语　U1P: ‘的’字短语　U2P: ‘地’字短语　U3P: ‘得’字短语
U4P: ‘着’、‘了’、‘过’字短语（已归入 H_vC）　U5P: 以‘等’、‘似的’结尾的比况短语
U6P: 所’字短语　VO: 动宾　XX: 并列短语

图 5: 华水树库1.0中短语类型的出现次数比例

7　结论

与印欧语相比，汉语缺乏屈折，使一种词类或一种短语结构能充当多种句法成分，对应到组合范畴语法中，就是有多个范畴，而这种多功能出现在一个句子里时，必须完成与印欧语屈折对应的范畴转换，范畴语法分析才能继续下去。为此，把词类的范畴按出现率分为典型和非典型，把短语结构作为整体充当句法成分时需要的范畴按是否由范畴结合规则得到分为典型和

非典型，把使用非典型看作源自典型的转换，可建立一套适应短语结构需要的范畴转换规则，而且这样的范畴转换具有范畴类型透明性。

在构建树库过程中，逐渐明确词类和特殊词的典型范畴，形成词类和短语结构的范畴转换规则体系。统计发现，短语直接成分使用非典型范畴的概率是35%，这是迄今已知的现代汉语句法语料中句法功能屈折的首次测量。使用范畴转换的短语直接成分中99.67%是实词或短语结构，说明虚词范畴力争明确，让虚词附着的实词或短语结构通过范畴转换与虚词搭配，这样的范畴转换规则体系是成功的。

范畴转换的上下文是短语结构，这种转换对组合范畴语法的生成能力带来了什么影响，需要做理论分析；分析树刻画的是小句结构，若句子不能正确分割为小句，则一些范畴的转换上下文就会不准确，需要对小句理论特别是小句识别进一步研究。在做好小句识别基础上还要分析更多的句子，纠正词类和特殊词的典型范畴，完善词类和短语结构的范畴转换规则，调整短语类型体系，逐渐让这三方面收敛，使这个带有范畴转换的CCG更简洁、稳定。另外，句子结构与句子语义式的关系，以及基于语义式的理解模型也是一个研究方向。

致谢

感谢匿名审稿人对论文的评审，评审意见在论文补充研究和进一步完善上发挥重要作用。

参考文献

Bob Carpenter. 1991. The generative power of Categorial Grammars and Head-Driven Phrase Structure Grammars with lexical rules. Computational Linguistics, 17(3):301-313.

陈鹏. 2016. 组合范畴语法(CCG)的计算语言学价值. 重庆理工大学学报(社会科学), 30(8):5-11.

冯志伟. 2021. 神经网络, 深度学习与自然语言处理. 上海师范大学学报(哲学社会科学版), (2):110-122.

何洪峰. 2016. 句法结构歧义成因的思考. 语言研究, 23(4):26-31.

胡明扬. 1995. 现代汉语词类问题考察. 中国语文, 1995(5):381-389.

Jayant K and M Mitchell T. 2014. Joint syntactic and semantic parsing with combinatory categorial grammar. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Stroudsburg, Pennsylvania, pages:1188-1198.

李艳翠, 冯文贺, 周国栋等. 2013. 基于逗号的汉语子句识别研究. 北京大学学报(自然科学版), 49(1):7-14.

陆俭明. 1990. 汉语句法成分特有的套叠现象. 中国语文, (2):81-90.

吕叔湘. 1954. 关于汉语词类的一些原则性问题. 中国语文, (9):6-14.

Mark Steedman, Jason Baldridge. 2011. Combinatory Categorial Grammar. Non-Transformational Syntax. Blackwell:181-224.

Mark Steedman. 2019. Combinatory Categorial Grammar. Current Approaches to Syntax – A Comparative Handbook. De Gruyter Mouton:389-420.

满海霞. 2022. 组合范畴语法: 通向人工智能语义理解的一种逻辑经验主义路径. 哲学动态, (1):119-125.

沈家煊. 1997. 形容词句法功能的标记模式. 中国语文, (4):242-250.

沈家煊. 1999. 不对称和标记论. 江西教育出版社, 南昌.

沈家煊. 2009. 我看汉语的词类. 语言科学, 8(1):1-12.

王庆江, 张琳. 2020. 支持中文句法结构套叠的组合范畴语法. 中文信息学报, 34(1):17+22.

邢福义. 1995. 小句中枢说. 中国语文, (6):420-428.

徐枢, 谭景春. 2006. 关于现代汉语词典（第5版）词类标注的说明. 中国语文, (1):74-86.

姚从军, 俎孟晨. 2022. 语言、逻辑与计算互动视角下汉语直接被动句的MMCCG处理. 湖南科技大学学报（社会科学版）, 25(1):42-50.

张斌. 2005. 现代汉语语法十讲. 复旦大学出版社, 上海.

张伯江. 2011. 关于现代汉语词典（第5版）词类标注的说明. 汉语学习, (2):3-12.

朱德熙. 1982. 语法讲义. 商务印书馆, 北京.

朱德熙. 1985. 语法答问. 商务印书馆, 北京.