

几个汉语典型语言现象的范畴语法分析*

姚旭晨 马建强

格罗宁根大学语言与认知研究中心 格罗宁根, 荷兰

E-mail: {yaoxuchen, majianchiang}@gmail.com

摘要: 范畴语法 (Categorial Grammar) 的词汇化的本质和其语法、语义紧密相连的特点使之成为计算语言学的重要基础之一。本文尝试用范畴语法的理论分析了一些典型汉语语言现象, 如“把”字句、“被”字句、动态助词“了”、结构助词“的”“地”“得”和形容词动化。作者提出, 在分析汉语典型语法现象的范畴语法时, 应采用“左结合”(逆向结合), 而非像英语中采用“右结合”(顺向结合)的偏好。在做上述分析时, 作者利用了一个基于推理规则的范畴语法学习算法帮助和验证分析并保证分析结果的一致性。

关键词: 范畴语法, 推理规则, 汉语语法

A Categorial Grammar Analysis of Some Typical Chinese Language Phenomena

Xuchen Yao Jianqiang Ma

Center for Language and Cognition Groningen, University of Groningen, Groningen, The Netherlands

E-mail: {yaoxuchen, majianchiang}@gmail.com

Abstract: Categorial Grammar (CG) is a lexicalized grammar formalism with a high level of transparency between syntax and semantics, which makes it an attractive formalism for computational linguistics. This paper proposes a categorial grammar analysis of some typical Chinese language phenomena, such as BA construction, BEI construction, aspect marker LE, structural particles and the verbalization of adjectives. Rather than a preference of the right-combining rule, the left-combining rule is proposed in the analysis of these typical phenomena. This is a difference between Chinese and English categorial grammar. The authors adopt an inference-rule based categorial grammar learning algorithm to assist and validate the analysis and the result shows this algorithm is capable of analyzing the typical Chinese grammar phenomena mentioned above.

Keywords: categorial grammar, inference rule, Chinese grammar

1 引文

范畴语法 (Categorial Grammar^{[1] [2] [5]}) 作为一种语法词汇主义的典型, 其词汇化的本质和语法、语义紧密相连的特点使之在国际计算语言学界受到重视。虽然早在 1975 年我国就出现了对范畴语法的介绍^[9], 但未引起广泛关注, 而近年来词汇主义的盛行使我国一些计算语言学学者又开始关注范畴语法, 比如孙红举^[10] 讨论了其在计算语言学上的应用前景以及在汉语中应用的相关问题。遗憾的是, 相关文献中的介绍多数是基于英语的, 很少强调范畴语法在汉语与英语中的不同, 或提出一个对汉语范畴语法的系统性描述。

本文在借鉴英语范畴语法和传统汉语语法的基础上, 尝试用范畴语法的理论分析汉语的一些典型语言现象, 如“把”字句、“被”字句、动态助词“了”、结构助词“的”“地”“得”和形容词动化。通过这些分析, 作者试图在简单、初步描述汉语范畴语法上做出尝试, 并提出, 尽管一般而言汉语的范畴语法分析与英语类似, 采用“右结合”(顺向结合), 但是对某些汉语语言

*两位作者对文章有同等重要贡献。他们均受到EU Life Long Learning Program, scholarships for the European Masters Program in Language and Communication Technologies (LCT) 和CLCG, University of Groningen的资助。

现象分析中，更宜采用“左结合”（逆向结合）。本文采用的范畴语法是未经扩充的经典形式，对汉语语言现象的分析也仅限于语法层面上的探索。在分析过程中，本文利用了一个基于推理规则的范畴语法机器学习算法辅助和验证分析，使得全部分析均可由此算法前后一致地推理获得。

本文第二节简单介绍了范畴语法。第三节给出了一个基于推理规则的学习算法。第四节重点分析了汉语的若干典型语言现象并描述了一些汉语特殊词的范畴语法。

2 范畴语法简介

范畴语法所有的语法信息都包含在“范畴”中，该语法通过建立句子中词语相对应的“范畴”，只使用几条简单的语法规则就能实现对整个句法结构的“演算”。范畴语法中通常只有少数几个基本范畴，比如一般认为英文范畴语法的基本范畴为 S （句子）、 NP （名词短语）和 N （名词）。句子中的其他范畴都可以由特定的函数操作获得。基本的函数操作有两种，如下：

- (1) 前向函数应用： $A/B \ B \rightarrow A$ （记为 $>$ ）
后向函数应用： $B \ A \setminus B \rightarrow A$ （记为 $<$ ）

上式中的斜线（/）和反斜线（\）称为范畴构造符， A/B 和 $A \setminus B$ 称为范畴函数。在范畴函数中，范畴构造符右边的类别表示其接受的输入（如 $A \setminus B$ 中的 B ）；范畴构造符左边的类别表示其产生的输出（如 $A \setminus B$ 中的 A ）。顺向范畴构造符（/）表示其接受的输入在其右方；逆向范畴构造符（\）表示其接受的输入在其左方。图 1 是一个范畴语法解析的简单例子。

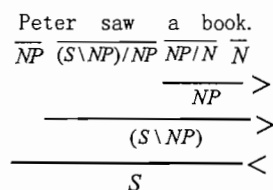


图1 范畴语法函数应用示例

Sentence Type	CG Category	Examples
Sentence	S	Peter saw Marry
Noun Phrase	NP	Peter, the book
Noun	N	book, telescope
Determiner	NP/N	the, a
Intransitive Verb	S\NP	sleep, run
Transitive Verb	(S\NP)/NP	read, saw
Ditransitive Verb	((S\NP)/NP)/NP	give, send
Adjective	N/N	green, red
Adverb	(S\NP) \ (S\NP)	happily, furiously
Proposition	(N\N)/NP, ((S\NP) \ (S\NP))/NP	with, in

表1 常用英语词类的范畴语法

表 1 则显示了通过简单的前向、后向函数应用推理出的常用英语词类的范畴类别。如表所示，范畴语法中的范畴大致相当于传统语法中的词类，但又不是严格的一一对应关系。

上述的范畴语法是上下文无关语法（Context Free Grammar）的弱等价形式，并不能满足描述复杂自然语言的要求。但是扩充过的范畴语法如组合范畴语法（Combinatory Categorical Grammar^{[6] [7]}，简称 CCG）在引入了其他运算后能够用于描述自然语言。

3 基于推理规则的学习算法

计算语言学对大规模文本的处理^{[3] [4]}往往依赖一套统一的语法描述系统。本小节描述一个基于推理规则的范畴语法学习算法^[8]。本文中的几个典型汉语语言现象的范畴语法分析均是利用该算法中的推理规则获得的，分析结果也经过该算法验证，保证了得出的范畴语法具有一致性。

在句子中只有一个未知范畴的情况下，此未知范畴可以由推理规则获得。最简单的推理规则来自于前向、后向函数应用，即：

- (2) 零级推理规则。当 $A \neq S$ 时有：
 $B/A \ ? \rightarrow A \Rightarrow ? = A$ 和 $? \ B \setminus A \rightarrow B \Rightarrow ? = A$
- (3) 一级推理规则。当 $A \neq S$ 时有：
 $A \ ? \rightarrow B \Rightarrow ? = B \setminus A$ 和 $? \ A \rightarrow B \Rightarrow ? = B/A$

我们定义级数为有效的范畴构造符的数目。范畴构造符仅在其接受参数时有效。因此名词短语(NP)的范畴级数为零级。不及物动词的范畴($S \setminus NP$)级数为一级。需要指出的是,英语中的副词的范畴($S \setminus_f NP$)\($S \setminus NP$)级数为二级:尽管它的范畴含有三个范畴构造符,但只有前两个范畴构造符有效(以下标 $_f$ 标出)。第三个范畴构造符只是不及物动词的范畴($S \setminus NP$)的一部分,并不接受参数。

零级和一级推理规则可以用于不及物动词的范畴的学习,如英语句子“Peter slept”中的“slept”。注意(2),(3)中的限制条件 $A \neq S$ 旨在防止出现某些不正确的范畴。如果没有这条限制条件,如在英语句子“Peter slept well”中的副词“well”会被学习为 $S \setminus S$ 。正如上文中指出的,英语副词的范畴级数为二级,用二级推理规则才能学得此范畴。

(4) 二级推理规则。

侧位: $A B \rightarrow C \Rightarrow ? = (C/B)/A$ 和 $A B ? \rightarrow C \Rightarrow ? = (C \setminus A) \setminus B$

中位: $A ? B \rightarrow C \Rightarrow ? = (C/B) \setminus A$ (左结合), $? = (C \setminus A) / B$ (右结合)

二级推理规则分为侧位规则和中位规则两种,这取决于未知范畴出现在句子的两端还是中间。算法的二级规则通过递归调用零级和一级规则得到,因此也避免了范畴中包含 $S \setminus S$ 的形式。

同样的,三级推理规则也分为侧位规则和中位规则两种。此时获得的范畴类别已包含三级范畴构造符,略显复杂。但是一门语言的语法复杂度是有限的。如表1所示,多数英语词类的范畴级数均止于三级。下一小节的数据则表明多数汉语范畴的级数也止于三级。需要说明的是,二级和三级中位推理规则对未知范畴的推导中,能得出两种范畴。这两种范畴是通过“左结合”(逆向结合,其过程如下所示)和“右结合”(顺向结合)分别得出的。

(5) 左结合:

$$A ? B \rightarrow C \xrightarrow{\text{左结合}(A?) } (A?) B \rightarrow C \xrightarrow{\text{一级推理规则}} (A?) = C/B \xrightarrow{\text{分拆}(A?) } A ? \rightarrow C/B \xrightarrow{\text{一级推理规则}} ? = (C/B) \setminus A$$

(6) 右结合:

$$A ? B \rightarrow C \xrightarrow{\text{右结合}(?B) } A (?B) \rightarrow C \xrightarrow{\text{一级推理规则}} (?B) = C \setminus A \xrightarrow{\text{分拆}(?B) } ? B \rightarrow C \setminus A \xrightarrow{\text{一级推理规则}} ? = (C \setminus A) / B$$

英语的范畴普遍是采用了右结合得出的,如表1中的全部范畴。多数汉语的范畴在也能从右结合得出。但是值得注意的是,对许多汉语语言现象,如“房子里”中的“里”等后置词,使用右结合不仅有违语言直觉,有时甚至无法推理出范畴。而使用左结合则更为可行和合理,如“里”的范畴向左结合,与“房子”构成后置词短语。这是英语范畴语法和汉语范畴语法不同的一点。

4 汉语的几个典型语言现象的范畴语法分析

汉语中的名词、及物动词、不及物动词、形容词等常见词类在范畴语法中的范畴与其对应的英文词类的范畴并无明显差异。本小节重点分析一些典型的汉语语言现象中的词汇范畴。同时,我们发现在对这些范畴的分析中,系统地利用左结合的策略更为适宜和有效。

4.1 “把”字句

“把”字句是一种常见的汉语语言现象,并存在不同解释,如王力(1946)^[12]从“把”字句式的语法意义的角度提出了“处置说”,吕叔湘(1984)^[11]从动词的意义限制角度,论了“把”字宾语的特性和谓语动词前后成分的要求等问题。本文并不探讨“把”字句的语法和语义本质,而是由常见“把”字句的结构和其中各个词汇的词类出发,分析“把”在“把”字句中的范畴。

虽然现有研究认为“把”字句并非能够和主动宾(SVO)句任意自由转换,两者在语义上也非完全等价^[13],然而不可否认的是,“把”字句实现了将动词支配或关涉的对象置于动词之前的转换,使汉语在整体呈现SVO结构的同时,还具有了一些SOV结构语言的特征,即:主语(S)+“把”+宾语(O)+谓语(V)。下面是一个简单句与“把”字句的比较:

(7) 他吃了苹果。(SVO结构) → 他把苹果吃了。(SOV结构)

对于是否每个“把”字句都是从某种主动句式转换过来的,以及两种句式语义之间的关系是否对等等问题,我们不予讨论。这里,仅关心“把”字句的特殊形态(S+“把”+O+V)和“把”字在“把”字句里应有的词类范畴。“把”字的作用范围覆盖了主语,宾语和谓语这三大句子主要成分。在“把”字句中,“把”是一个桥梁,它需要和前面的主语,后面的宾语以及宾语之后的动词分别产生作用以组成一个完整的句子。因此,“把”字在范畴语法中的范畴应当包含三个有效的范畴构造符以“吸收”句子中的主语、谓语、宾语,并输出句子范畴S。

图2是一个“把”字句的范畴语法解析例句。这个例子是通过第三级推理规则获得的。注意,在图2中,“吃了”由“他吃了苹果”中的及物动词转换为“他把苹果吃了”中的不及物动词。我们认为“吃了”在正常语句(“他吃了苹果”)中是及物动词,但在“把”字句中其后并没有直接连接宾语,因此“吃了”起不及物动词的作用。这样的分析也使获得的“把”字的范畴更简洁。

4.2 “被”字句

“被”字句是汉语的一种被动句式。这种句式将动词支配或关涉的对象置于主语之前。下面是将一简单句变换为“被”字句前后的比较:

(8) 他吃了苹果(SVO结构) → 苹果被他吃了(OSV结构)

通过“被”字句的转换,汉语在整体呈现SVO结构的同时,还具有了一些OSV结构语言的特征,即:宾语(O)+被+主语(S)+谓语(V)。与“把”字一样,“被”字的作用范围也覆盖了主语,宾语和谓语这三大句子主要成分。因此,类似的,“被”字的范畴语法也需要体现出这种特征。

图3是一个“被”字句的范畴语法解析例句。这个例子是通过第三级推理规则获得的。

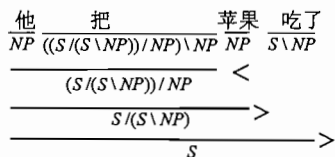


图2 “把”字句范畴语法例句

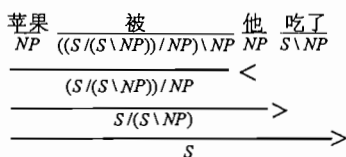


图3 “被”字句范畴语法例句

4.3 动态助词“了”

图2和图3中的“吃了”均被当作单一词(及物动词或不及物动词)处理的。实际上动态助词“了”在汉语中也占有重要的类别,因此有必要进一步细分。

当“吃”为不及物动词时,“了”通过左结合从左边“吸收”掉这个不及物动词,并且又产生出一个不及物动词,从而使“吃了”从整体上观察还是不及物动词的特性,如图4a所示。类似的情况也适用于“吃”为及物动词时,如图4b所示。请注意,在这种情况下,如果采用右结合,则“了”首先会与“苹果”而非“吃”结合,这与其公认的语法功能有悖。

“了”字还能出现在形容词后。唐广厚,车竟^[14]等人认为词后添加动态助词是动词的主要特征,因此后面添加“了”字的形容词本身应该视为动词。吕叔湘^[15],陆俭明^[16]等人则认为词后加“了”是形容词和动词的共有特征。本文在对范畴语法的分析中倾向于后一种观点,即在承认形容词在加了动态助词“了”后与动态助词一起在整体上来看具有动词的特征,在句子中起谓

语的作用的同时，认为形容词本身的词性并没有发生变化。

如图4c所示，在“苹果青了”中，当形容词“青”与“了”结合时，便从整体上体现了动词的特性。此时，动态助词“了”起到了将形容词+动态助词转换为动词的作用。

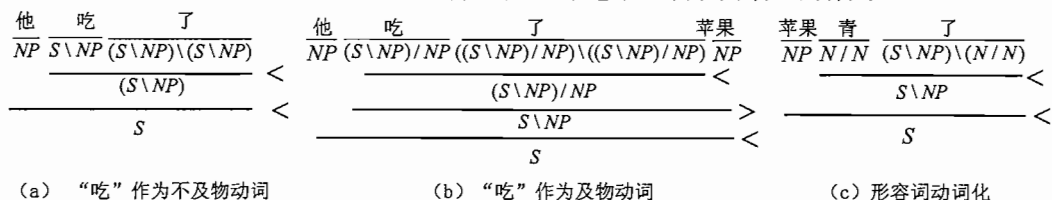


图4 “了”字范畴语法例句

在以上例子中，动态助词“了”通过左结合与其前面的词（或是动词，或是形容词）结合，在整体上体现出动词的特性，并使单独的词不失其基本词类的本意。这种情况下采用左结合，使范畴语法在针对有不同分词颗粒度的句子进行分析时，得出的结果具有内在的一致性，各个范畴的语法功能和范畴间的关系也在句子解析中一目了然。4.4 中对结构助词的分析中采用左结合也是基于相同的原因。

除了“了”字，汉语的动态助词还有“着”、“过”等。其分析过程基本与“了”字相似，图5给出了两个例子。

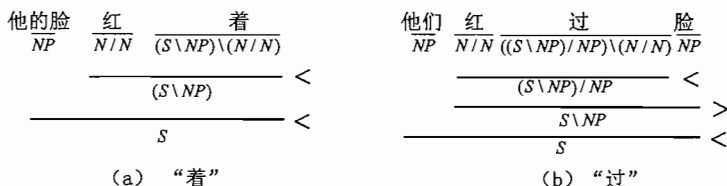


图5 动态助词“着”“了”范畴语法例句

4.4 结构助词“的”“地”“得”

“的”“地”“得”三个词是汉语中特有的连接词。“的”字一般用在介宾词组，连接形容词和名词。“地”字一般用在动宾词组，连接副词和动词。“得”字一般用在动补词组，连接动词和副词。通常将“的”“地”“得”与前面的词组合起来作为一个词不予细分。

但是汉语的多词性特性使得许多词，如“红色”、“跑”和“飞快”，具有名词这一基本词类。因此这提供了对图6中示例进一步细分的可能性。在假设“红色”、“跑”和“飞快”都为名词的情况下，“的”“地”“得”的范畴语法真正体现了其承上启下的连接作用，并使连接后的短语在整体上看不失其原本特征，如图6所示。

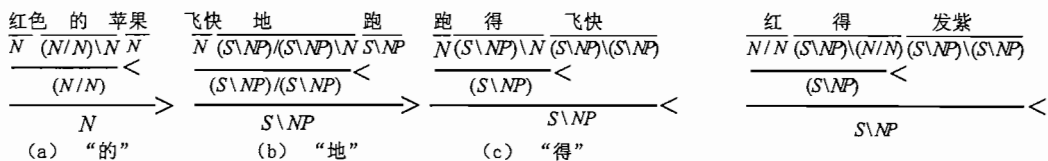


图6 “的”“地”“得”范畴语法例句

图7 “得”字与形容词结合范畴语法例句

由于使用了左结合，“的”“地”“得”三个词首先与其左边词配对，如图6所示。例如，“的”与“红色”结合，组成“红色的”，并具有形容词的词类(N/N)。“地”与“飞快”结合，组成“飞快地”，并具有副词的词类((S\NP)/(S\NP))。“得”与“跑”结合，组成“跑得”，并具有不及物动词的词类(S\NP)。在假设大部分词均为名词的情况下，“的”“地”“得”的

配合使用为组成新词类提供最大的灵活性。例如，“飞快”为名词，与“的”结合便构成了形容词“飞快的”，与“地”结合便构成了副词“飞快地”。在范畴语法的框架下，词类的变换取决于连接词的范畴及其结合方向，整个词组的语法结构更清晰。

与动态助词“了”的词类转换特性类似，结构助词有时也可以将某一词类转换为另一词类。例如，在“红得发紫”中，“红”本是形容词，与“得”结合后便体现了其动词的特征，如图7。

5 结语

范畴语法的词汇化的本质以及语法和语义紧密相连的特点使其在计算语言学方面有广阔的应用前景。本文是一个利用范畴语法的形式化方法描述汉语的初步探索。作者参考了英语范畴语法的范例和传统汉语语法的特点，着重分析了一些典型的汉语语言现象，包括“把”字句、“被”字句、动态助词“了”和结构助词“的”、“得”、“地”以及形容词动词化，并得出了相关词汇的范畴。同时，作者提出，与英语范畴语法不同，汉语特有语言现象中的范畴语法应采取“左结合”（逆向结合）的偏好，并在实际例子中证明这种偏好是合理有效的。

作者分析的基础是一组直觉化的推理规则。这些规则能在句子只有一个未知范畴的情况下，推出未知范畴。最简单的推理规则来自于前向、后向函数应用，更为复杂的规则是基于对简单规则的递归调用。在分析汉语范畴语法时，作者利用了一个基于这些推理规则的范畴语法机器学习算法辅助分析和验证分析结果。这个机器学习算法保证了语法分析的前后一致性，并为以后计算语言学中利用范畴语法处理大规模文本奠定了基础。本文提出了一个对汉语范畴语法的一个简化的、初步的描述，这只是向建立汉语语义范畴的演算系统的目标迈出一小步，而实现这一目标，需要对汉语语言特点更为深入的研究以及解决语法、语义分析的有机结合等进一步问题。

参 考 文 献

- [1] Ajdukiewicz, K., Die syntaktische Konnexität, *Studia Philosophica*, 1935, 1, 1-27.
- [2] Bar Hillel, Y., A Quasi-Arithmetical Notation for Syntactic Description, *Language*, 1953, 1, 47-58.
- [3] Buttery, P. J., Computational models for first language acquisition. PhD Thesis, University of Cambridge, Churchill College., 2006.
- [4] Villavicencio, The acquisition of a unification-based generalised categorial grammar. PhD Thesis, University of Cambridge, 2002.
- [5] Steedman, M., Categorical Grammar. A short encyclopedia entry for MIT Encyclopedia of Cognitive Sciences, 1999.
- [6] Steedman, M., *The syntactic process*, MIT Press, 2000.
- [7] Steedman, M. & Baldridge, J., *Combinatory Categorical Grammar, Non-Transformational Syntax*, 2005.
- [8] Yao, X.; Ma, J.; Duarte, S. & Coltekin, C. (2009) An Inference-rules based Categorical Grammar Learner for Simulating Language Acquisition. *Proceedings of the 18th Annual Belgian-Dutch Conference on Machine Learning*. 2009.29-37.
- [9] 冯志伟.数理语言学简介.计算机应用与应用数学,1975(4).
- [10] 孙红举.“范畴语法”及其在汉语中的应用.现代语文,2008(6).
- [11] 吕叔湘.汉语语法论文集.北京:商务印书馆,1984.
- [12] 王力.中国语法纲要.北京:开明书店,1946.
- [13] 张伯江.论“把”字句的句式语义.语言研究, 2000(1).
- [14] 唐广厚,车竟.形容词接动态助词动词化初探.渤海大学学报,哲学社会科学版,1985(2).
- [15] 吕叔湘.关于汉语词类的一些原则性问题.中国语文,1954(9).
- [16] 陆俭明.关于词的兼类问题.中国语文, 1994(1).