文章编号：1000-8608(2012)06-0902-06

数据库语义学在古汉语自动分析上的应用

冯秋香*， 汪榕培

（大连理工大学 计算机科学与技术学院，辽宁 大连 116024）

摘要： 在以认知为基础的面向计算机和人工智能领域的语言学框架下，对古汉语的自动句法语义分析进行研究，希望能对古汉语教学与研究，以及现代汉语的分析和处理起到一定的推动和促进作用. 运用以左结合语法为基础的数据库语义学方法对古汉语的两个基本结构(函词-论元结构和并列结构)进行自动句法和语义分析，以可接续性为前提，遵循自然语言的时间线性顺序，采用规则和模式匹配的方法，过程简便，计算效率高，且符合古汉语本身的特点. 分析结束时自动生成的语义关系图清晰、准确，更彰显了数据库语义学方法的独特性、适应性以及分析能力.

关键词： 古汉语；数据库语义学；左结合语法；句法分析；语义分析

中图分类号： H030；TP391.43 **文献标志码：** A

0 引 言

现代汉语一直是自然语言处理的重点研究对象之一，而古汉语却少有人问津. 和现代汉语一样，古汉语也是中国历史和文化的载体，一些经典的古代著作已经成为不可多得的各个学科研究相关古代课题的重要资料. 因此，古汉语文本的自动处理对于相关学科的教学与研究有一定的重要意义. 鉴于古汉语和现代汉语之间的继承性和连贯性，古汉语自动分析也能对现代汉语的句法分析有一定的促进作用.

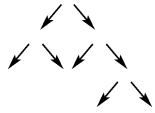
虽然古汉语典籍众多，但是作为一种语言，古汉语不存在继续发展变化的空间. 应用基于规则的方法从句法和语义角度对古汉语进行自动分析具有较强的可操作性.

数据库语义学(database semantics，DBS)[1]是以认知为基础的面向计算机和人工智能领域的语言学框架. 其提出的目的是建立一个功能完备、数据完全、计算复杂度小而计算机操作性强的关于自然语言交流的理论[2]. 本文应用该方法对古汉语的函词-论元结构和并列结构进行句法和语义的综合分析.

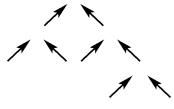
1 左结合语法

左结合语法是数据库语义学的核心算法. 最早由 Hausser[3]在斯坦福大学的报告中提出，其全称是 left-associative grammar(LAG). 该语法以可接续性(possible continuation)为前提，旨在按照时间线性顺序进行句法语义的综合分析. 该语法在概念上区别于传统的以可替代性(possible substitution)为前提的短语结构语法[4]和范畴语法[5]. 这 3 种语法在推导顺序上的区别如下.

短语结构语法(自顶向下扩展)：



范畴语法(自底向上合并)：



左结合语法(自底向上左结合)：

收稿日期： 2011-03-10； **修回日期：** 2012-08-25.

作者简介： 冯秋香*(1977-)，女，博士，E-mail：feng.qiuxiang@gmail.com；汪榕培(1942-)，男，教授，博士生导师.

左结合语法与这些语法框架的最根本区别在于左结合语法遵循的是语言的时间线性顺序[6]. 而时间线性顺序是自然语言的最基本特征之一[7].

在数据库语义学的最新发展[8]当中，Hausser提出了图示法，如语义关系图(semantic relations graph，SRG)和词性关系图(part of speech signature)等. DBS语义关系图以树形图的结构更为直接地展示句子的语法和语义关系. 词性关系图则是同一内容的抽象展示. 这两个图都可以由左结合语法的句法语义分析结果直接生成. 下文将借助这两个图来进一步说明例句的分析结果.

2　古汉语分析实例

函词-论元结构和并列结构是古汉语的两个基本结构. 函词指句子中的谓语核心词，在汉语当中可能是动词，也可能是形容词或者名词. 论元指与函词搭配构成重要语法关系的句子成分，一般为名词性结构，如谓语动词的主语和宾语等. 以“晋侯梦大厉”为例，动词“梦”是函词，名词“侯”和“厉”分别是它的主语论元和宾语论元. “晋”和“大”是分别修饰两个论元的定语成分，由此形成的修饰性结构不构成句子的核心结构. 并列结构指在句子中不同的词扮演相同的句法角色而存在并列关系的结构. 以“弃甲而复”为例，动词“弃”和“复”并列作句子的谓语，本句为并列结构. 下面以这两个句子为例具体介绍古汉语的函词-论元结构和并列谓语结构的分析过程和结果，其中嵌入修饰性结构.

2.1　函词-论元结构分析

例句“晋侯梦大厉”中的各个词在词典中的存储情况如表1所示(由于篇幅所限，只列出关键特征).

每个词条以命题粒(proplet)的形式存储在词典里. 命题粒是一种非递归的特征结构，是有限个属性/值对的集合. 第一个词条“晋”，其核心属性为[noun]，即名词，其值为“晋”，二者一起构成一个特征. [cat](category，范畴)和[sem](semantic，语义)是语法属性，其值分别为[pn](proper noun，专有名词)[nms](name of a state，国家).

左结合语法分析过程中其他常用的属性还有9个. [verb](verb，动词)和[adj](adjective，形容词)，是除[noun]之外的另两个核心属性. 也就是说，古汉语传统语法中的词类在本词典中分别划归名词、动词和形容词三大词类之下. 除核心属性之外，还有[fnc](functor，函词)、[arg](argument，论元)、[mdd](modified，被修饰语)、[mdr](modifier，修饰语)、[nc](next conjunct，下一个并列项)和[ic](initial conjunct，首个并列项)等属性. 这些是用于表现句法语义关系的接续属性. 还有一个属性[prn](proposition，命题编号)是簿记属性. 同属一个命题的命题粒的“prn”值相同. 由于篇幅所限，本文不引用任何命题粒的“prn”特征.

表1　“晋侯梦大厉”各词在词典中的存储情况

Tab. 1　Lexical lookup of words in “jin hou meng da li” in dictionary

词	关键特征
晋	[noun：晋] [cat：　pn] [sem：nms]
侯	[noun：侯] [cat：　nr] [sem：nmt]
梦	[verb：　梦] [cat：s′p′v] [sem：　+nr]
大	[adj：大] [cat：adj] [sem：　]
厉	[noun：　厉] [cat：　cn] [sem：object]

应用左结合语法分析“晋侯梦大厉”的具体步骤如下：

1　[noun：晋]
[cat：　pn]
[sem：nms]
[fnc：　]

2　[noun：晋]　[noun：侯]
[cat：　pn]　[cat：　nr]
[sem：nms]　[sem：nmt]
[fnc：　]　[mdr：　]
[fnc：　]

3　[noun：晋]　[noun：侯]　[verb：梦]
[cat：　pn]　[cat：　nr]　[cat：s′p′v]
[sem：nms]　[sem：nmt]　[sem：+nr]
[mdd：侯]　[mdr：晋]　[arg：　]
[fnc：　]　[fnc：　]　[mdr：　]

4	[noun：晋]	[noun：侯]	[verb：梦]	[adj：大]	
	[cat：pn]	[cat：nr]	[cat：p′v]	[cat：adj]	
	[sem：nms]	[sem：nmt]	[sem：+nr]	[sem：]	
	[mdd：侯]	[mdr：晋]	[arg：侯]	[mdd：]	
	[fnc：]	[fnc：梦]	[mdr：]		
5	[noun：晋]	[noun：侯]	[verb：梦]	[adj：大]	[noun：厉]
	[cat：pn]	[cat：nr]	[cat：p′v]	[cat：adj]	[cat：cn]
	[sem：nms]	[sem：nmt]	[sem：+nr]	[sem：]	[sem:object]
	[mdd：侯]	[mdr：晋]	[arg：侯]	[mdd：]	[mdr：]
	[fnc：]	[fnc：梦]	[mdr：]		[fnc：]
6	[noun：晋]	[noun：侯]	[verb：梦]	[adj：大]	[noun：厉]
	[cat：pn]	[cat：nr]	[cat：v]	[cat：adj]	[cat：cn]
	[sem：nms]	[sem：nmt]	[sem：+nr]	[sem：]	[sem:object]
	[mdd：侯]	[mdr：晋]	[arg：侯厉]	[mdd：厉]	[mdr：大]
	[fnc：]	[fnc：梦]	[mdr：]		[fnc：梦]

第 1 步 输入第一个词“晋”；

第 2 步 句首“晋”与下一词“侯”组合成新的句首，依据规则“AN＋N(adnominal noun ＋ noun，指名词性定语＋名词)”的一条子规则，见表 2.

表 2 规则 AN＋N

Tab.2 Rule AN＋N

AN＋N{S＋V}	说明
[noun:_, cat:(pn), sem:(N)]	句首模型
[noun:_, cat:(nr), sem:(nmt), mdr:_]	下一词模型
acopy(SS.noun nw.mdr)	将句首属性 noun 的值复制给下一词的属性 mdr
ecopy(nw.noun SS.mdd)	把下一词属性 noun 的值复制给句首属性 mdd，因该词条在词典中不具备属性 mdd，系统先为其自动生成该属性
copy(nw)	复制下一词到输出当中

当句首和下一词模型都完全匹配时，规则被调用，规则中的操作，如“acopy”、“ecopy”和“copy”等依次执行. 例句前两个词之间的修饰与被修饰的关系表示为两个命题粒的“mdr”和“mdd”的值. 操作执行完毕的输出结果成为新的句首等待下一个输入. 表 2 中规则名“AN＋N”后面的部分，即“{S＋V}”，称作规则包(rule package)，是任一条左结合语法规则的重要组成部分，包中的规则是接下来的分析可能调用的规则，代表句子延续的可能性.

第 3 步 句首“晋侯”与下一个输入“梦”组合，调用规则“S＋V(subject ＋ verb，即主语＋谓语动词)”. 这一步的句法语义关系主要发生在“侯”和“梦”之间. 二者之间的函词-论元关系得到再现之后，动词“梦”的第一个“cat”值，即“s′”，被删除，表示该词的第一个价得到满足. 被赋予新“arg”值的动词和句首一起构成新的句首.

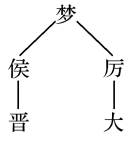
第 4 步 句首“晋侯梦”与下一个词“大”组合，调用规则“V＋ADJ(verb ＋ adjective，即动词＋形容词)”的一条子规则. 因为动词“梦”和形容词“大”之间的句法语义关系并不明确，“大”被直接复制输出，等待下一个输入.

第 5 步 句首“晋侯梦大”与下一词“厉”组合. 句首模型和下一词模型匹配规则“ADJ＋N(adjective ＋ noun，即形容词性修饰语＋名词)”的一个定义. 这一步要表现两个语义关系：“梦”和“厉”之间的函词-论元关系；“大”和“厉”之间的修饰与被修饰的关系.

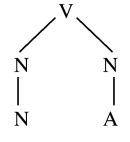
最后输入句号，调用规则 S＋IP，“mark”被赋予函词命题粒的“cat”属性. 全句分析结束，最终的分析结果如下：

[noun：晋]	[noun：侯]	[verb：梦]	[adj：大]	[noun：厉]
[cat：pn]	[cat：nr]	[cat:v,mark]	[cat：adj]	[cat：cn]
[sem：nms]	[sem：nmt]	[sem：+nr]	[sem：]	[sem:object]
[mdd：侯]	[mdr：晋]	[arg：侯厉]	[mdd：厉]	[mdr：大]
[fnc：]	[fnc：梦]	[mdr：]		[fnc：梦]

该分析结果的语义关系图为



其词性关系图为



DBS 的树状图和短语结构语法的树状图有本质的区别，但与依存语法[9]有相近之处. DBS 的树状图中共有 4 种线，各自代表一种语义关系. 左斜线“/”代表主语论元-函词关系；右斜线“\”代表宾语论元-函词关系；垂直竖线“|”代表修饰与被修饰的关系；水平线“—”代表并列关系. 词与词之间的位置关系仅仅出于图示化的需要，不代表重要性上的区别.

如上面的语义关系图所示，“梦”是函词，通过

左斜线相连的“侯”是主语论元，通过右斜线相连的“厉”是宾语论元. 通过垂直线与这两个论元相连的分别是“晋”和“大”两个定语修饰语. 该句不存在并列关系，图中也就没有出现水平线. 把语义关系图中的节点替换为代表其词性的字母，具体的语义关系图就转变为抽象的词性关系图.

2.2　并列结构分析

“弃甲而复”虽然是一个只有 4 个字的句子，但是包含多种语法现象，如函词-论元结构、并列关系，以及主语省略等. 其句法语义分析的过程涉及主语的表示、动宾和并列关系再现，以及虚词吸收等.

句中 4 个词在词典中的存储情况如表 3 所示.

表 3　“弃甲而复”各词在词典中的存储情况

Tab. 3　Lexical lookup of words in “qi jia er fu” in dictionary

词	关键特征
弃	[verb：　弃] [cat：　s′p′v] [sem：　+nr]
甲	[noun：　甲] [cat：　cn] [sem：object]
而	[adj：　而] [cat：　conj] [sem：　]
复	[verb：　复] [cat：　s′v] [sem：　]

应用左结合语法对该句进行的句法语义分析可以分解为如下几个步骤：

1　[verb：　弃]
[cat：　s′p′v]
[sem：　+nr]
[arg：　]

2　[verb：　弃]　[noun：　甲]
[cat：　s′p′v]　[cat：　cn]
[sem：　+nr]　[sem：object]
[arg：　]　[fnc：　]

3　[verb：　弃]　[noun：　甲]　[adj：　而]
[cat：　v]　[cat：　cn]　[cat：　conj]
[sem：　+nr]　[sem：object]　[sem：　]
[arg：　#甲]　[fnc：　弃]

4　[verb：　弃]　[noun：　甲]　[verb：　复]
[cat：　v]　[cat：　cn]　[cat：　s′v]
[sem：　+nr]　[sem：object]　[sem：　]
[arg：　#甲]　[fnc：　弃]　[arg：　]
[nc：　#]

5　[verb：　弃]　[noun：　甲]　[verb：　复]
[cat：　v]　[cat：　cn]　[cat：　v]
[sem：　+nr]　[sem：object]　[sem：　]
[arg：　#甲]　[fnc：　弃]　[arg：　#]
[nc：　复]　[ic：　弃]

第 1 步　动词“弃”作为句首输入.

第 2 步　句首“弃”与下一个词“甲”组合成新的句首，调用的是规则“V＋O(verb ＋ object，指动词＋宾语)”的一条子规则. 无论是现代汉语还是古汉语，通用的语法规则是主语在前，谓语动词在后. 所以这一步不但要再现动宾关系，还要表示出主语缺失的情况. 左结合语法分析过程中，当某个值不确定时，用符号“#”来占位. 所以，这一步分析完成之后“弃”的属性“arg”有两个值，分别是“#”和“甲”. 其属性“cat”的前两个值“s′”和“p′”同步删除.

第 3 步　句首“弃甲”与下一个词“而”进行组合. 左结合语法的分析过程中，实词基本保留，而虚词则被实词吸收. 连词是虚词，因此，当“而”输入时，系统自动为句首部分的函词命题粒，即“弃”，添加属性“nc”并赋值“#”. 除此之外，不执行其他任何操作，也就是说，虚词“而”被读入，但在分析结束时不输出，也就不出现在下一步的句首部分. 如第 4 步所示.

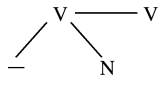
第 4 步　动词“复”被读入. 根据规则“V＋V(verb ＋ verb，指主动词＋主动词)”，系统自动为“复”添加新属性“ic”，并复制命题粒“弃”的核心属性值作为这个新属性的值. 同时“弃”的原“nc”属性值被替换. “复”的属性“arg”被赋值“#”.

输入句号，该句的句法语义分析结束. 其结果如下：

[verb：　弃]　[noun：　甲]　[verb：　复]
[cat：v，mark]　[cat：　cn]　[cat：　v]
[sem：　+nr]　[sem：object]　[sem：　]
[arg：　#甲]　[fnc：　弃]　[arg：　#]
[nc：　复]　[ic：　弃]

其语义关系图如下：

图中“弃”和“复”之间用水平线连接，表示二者之间是并列关系. 同一内容的词性关系图如下：



2.3 左结合语法的几个主要规则

左结合语法的每一条规则都包含 4 个主要部分:规则包、句首模型、接续词(下一个词)模型以及模型完全匹配之后进行的一个或者一组操作.在较常用到的左结合语法规则当中,"S+V"和"V+O"主要用于再现函词-论元关系."V+V"主要用于再现动词的并列关系.此外,修饰与被修饰的关系最常调用的是"ADJ+N(adjective+noun,即形容词性修饰语+名词)"和"V+ADJ(verb+adjective,即动词+形容词性修饰语)"两个规则.这几个规则概述如下:

```
S+V{V+O, S+IP}
[noun:_, cat:(N), sem:_, fnc:(), mdr:(_)]
[verb:_, cat:(_ V), sem:_, arg:(), mdr:()]
acopy(ss.noun nw.arg)
cancel(nw.cat.1)
acopy(nw.verb ss.fnc)
copy(nw)

V+O {V+ADJ, S+IP}
[verb:_, cat:(p′ V), sem:_, arg:_, mdr:(_)]
[noun:_, cat:(N), sem:(O), fnc:(), mdr:(_)]
acopy(ss.verb nw.fnc)
acopy(nw.noun ss.arg)
cancel(ss.cat.1)
copy(nw)

V+V {V+ADJ, V+O, S+IP}
[verb:_, cat:(_ V), sem:_, arg:(_), mdr:(_)]
[verb:_, cat:(_ V), sem:_, arg:(), mdr:()]
ecopy(ss.verb nw.ic)
ecopy(nw.verb ss.pc)
acopy(# nw.arg)
cancel(nw.cat.1)
copy(nw)

ADJ+N {S+V}
[adj:_, cat:(adj), sem:(+N), mdr:(_),mdd:()]
[noun:_, cat:(N), sem:_, fnc:(), mdr:(_)]
acopy(ss.adj nw.mdr)
acopy(nw.noun ss.mdd)
copy(nw)

V+ADJ {V+O, S+IP}
[verb:_, cat:(_ V), sem:_, arg:(_), mdr:(_)]
[adj:_, cat:(adv), sem:_, mdd:()]
acopy(nw.adj ss.mdr)
```

(基于功能词吸收的原则,副词补语不出现在输出结果当中,因此这条规则中没有 copy 操作.)

3 讨 论

目前应用数据库语义学方法在 15 000 句的古文语料库中的 3 次测试准确率结果如表 4 所示.

表 4 句法测试结果

Tab.4 Syntactic test results

结构	准确率/%		
	第 1 次	第 2 次	第 3 次
函词-论元	90	94	92
并列	91	93	100

每次实验从语料中任意抽取 1 500 句,其中包含函词-论元结构和并列结构的比例不限,包含分句的数量也不限.从结果可以看出,第 1 次实验的准确率最低.左结合语法是基于规则的方法.虽然古汉语不具备发展性,但规则覆盖率与歧义消解之间的关系不容易平衡.所以,在第 2 次实验之前,统计并分析了第 1 次实验出现的错误类型,并在此基础上扩展了规则的数量和覆盖面,比如用限制集来代替某一特定变量.第 2 次实验的结果与第 1 次相比分别提高了 4%和 2%.但是在覆盖率提高的同时,歧义结果的数量增加的幅度也比较大.这是因为,汉语的词性与句法角色之间并不存在确定的关系,即汉语偏重语义关系,且词类活用以及词的多义现象较为严重.因此,本文首先进一步完善了词典,对实词的句法和语义功能作了进一步标注和说明,对多义词和兼类词的搭配功能作了更为细致的界定.之后对规则集中的相应部分也作了修正.第 3 次实验结果当中,函词-论元结构的准确率略微下降,但是并列结构的准确率得到了显著提高.综合来看,经过 2 次大规模的改进,第 3 次实验最为成功,同时也验证了左结合语法的适用性和有效性.

4 结 论

数据库语义学框架下的左结合语法按照时间线性顺序分析自然语言,将词与词之间的句法语义关系表示为词的特征.遵循自然语言语表的时间线性组合本身有利于提高计算效率.左结合语法操作过程中,基于模型匹配的规则方法也保证了较低的计算复杂度.当出现句法歧义或者词汇歧义时,左结合语法允许不同推导路径并行继续运算.这种并行算法对消歧的要求较高,但同时避

免了大量回溯造成的重复计算.分析结束时自动生成的 DBS 图能够清晰、明确、直观地展示句子的句法语义关系.

应用数据库语义学方法分析古汉语的函词-论元结构和并列结构,实践证明易操作、易理解,且效率较高.

参考文献:

[1] Hausser R. **Foundations of Computational Linguistics, Human-Computer Communication in Natural Language** [M]. New York: Springer-Verlag, 1999.

[2] Hausser R. **A Computational Model of Natural Language Communication: Interpretation, Inference, and Production in Database Semantics** [M]. New York:Springer, 2006.

[3] Hausser R. Left-associative grammar and the parser NEWCAT [R] // Report IN-CSLI-85-5. Stanford: Center for the Study of Language and Information, Stanford University, 1985.

[4] 乔姆斯基 N. 句法结构 [M]. 黄长著,等,译. 北京:中国社会科学出版社, 1979.
Chomsky N. **Syntactic Structures** [M]. HUANG Chang-zhu, *et al*, tran. Beijing: China Social Sciences Press, 1979. (in Chinese)

[5] Bar-Hillel J. **Language and Information — Selected Essays on Their Theory and Application** [M]. Mass: Addison Wesley and Jerusalem Academic Press, 1964.

[6] Hausser R. Comparing the use of feature structures in nativism and in database semantics [C] // **Information Modelling and Knowledge Bases XIX**. Amsterdam:IOS Press-Ohmsha, 2007.

[7] 索绪尔 F D. 普通语言学教程[M]. 高名凯,译. 北京:商务出版社, 1980.
Saussure F D. **Course in General Linguistics** [M]. GAO Ming-kai, tran. Beijing:China Commerce and Trade Press, 1980. (in Chinese)

[8] Hausser R. **Computational Linguistics and Talking Robots-Processing Content in Database Semantics** [M]. New York:Springer, 2011.

[9] 刘海涛. 依存语法的理论与实践[M]. 1 版. 北京:科学出版社, 2009.
LIU Hai-tao. **Dependency Grammar:from Theory to Practice** [M]. 1st ed. Beijing:Science Press, 2009. (in Chinese)

Application of database semantics to automatic analyses of ancient Chinese

FENG Qiu-xiang*, WANG Rong-pei

(School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)

Abstract: The automatic syntactic and semantic analyses of ancient Chinese are discussed in a cognition-based, computer and artificial intelligence-oriented linguistic framework, aiming to support ancient Chinese teaching and research, as well as modern Chinese parsing and processing. The method of database semantics based on left-associative grammar is applied to the automatic syntactic and semantic analyses of two basic sentence structures of ancient Chinese (functor-argument construction and coordination construction). Based on possible continuation, in conformity with the time-linearity of natural language, rules are composed and executed. It is proved to be convenient, computationally efficient, and applicable to ancient Chinese. The graphs, automatically generated as a result of the analyses, are clear and accurate, which highlights the particularity, adaptability and analytical capability of database semantics.

Key words: ancient Chinese; database semantics; left-associative grammar; syntactic analysis; semantic analysis