现代汉语短语结构知识库规格说明书*

俞士汶

北京大学计算语言学研究所，北京 100871，中国

yusw@pku.edu.cn

Submitted on 17 May, 2003, Revised and Accepted on 28 May, 2003

摘 要

笔者于 1993-1994 年间开发了一个“现代汉语短语结构知识库”。十年来，这项语言资源在语言信息处理领域得到了应用。《现代汉语短语结构知识库规格说明书》介绍短语的定义、分类、短语结构形式化描述方法以及短语结构数据库的构成。

关键词

现代汉语短语，短语分类，短语结构的形式描述，短语结构知识

1.前言

《语言文字应用》1993 年第 3 期曾发表拙文《关于计算语言学的若干应用》。其中简要地介绍了笔者建立的汉语短语结构体系及其形式化描述方法，并报告笔者实际写出了 600 余条汉语短语规则。“现代汉语短语结构知识库” 随后便有了雏形。尽管这个语言资源数据库在本单位的应用系统的开发中发挥了作用，也转让给了一些外单位使用，但一直未正式发表它的规格说明书或更详细的文章。

这次发表的“现代汉语短语结构知识库规格说明书”基本上保持了 1994 年版的原貌。只是因为“现代汉语短语结构知识库”是在《现代汉语语法信息词典》的基础上开发的，而《现代汉语语法信息词典》已经有了很大的发展（俞士汶等，2003），因此做

* 本文有关研究得到 973 项目（G1998030504-01，G1998030507-4）、863 项目（2002AA117010-08，2001AA114040 ）和国家自然科学基金（69973005）的支持。

了一些适应性的修改。

拙文在介绍汉语短语结构体系的同时，曾论述：“在面向计算机时，则需要更明确地指出哪个子类的或具有什么属性的动词和哪个子类的或具有什么属性的名词能构成什么样的短语，这个短语的特性如何，它继承了构成成分的哪些属性，丢失了哪些属性，又派生了哪些新的属性”。

笔者一直在思考这样一些问题：面向语言信息处理，汉语的语言单位应该如何划分？语言单位的界线是清晰的还是模糊的？如何利用这种清晰性或者如何驾驭这种模糊度？汉语的语言单位和英语、日语等外国语的语言单位到底有怎样的对应关系？汉语的各个语言单位之间的属性有什么样的关联？较大的语言单位的属性同其构成成分（较小的语言单位）的属性之间有什么样的继承、变异、衍生关系？

上面引用的论述只是这些想法的萌芽而已。10 多年过去了，虽然时时有所领悟，可惜还未能形成完整的思路，更没能找到一个统一的描述框架。“路漫漫其修远兮，吾将上下而求索”。这次将尘封了 10 年的、现在看来显然粗糙的阶段性成果发表出来，目的乃是抛砖引玉，期盼与学界同仁一起探索、一起切磋。也希望有更多的人了解北大计算语言所还有这样一个语言资源库，能有更多的研究者应用它，扶植它，这样它就会成长得更快一些。

2.短语的定义

现代汉语的短语(phrase，又叫词组)是由两个以上的词或短语按照一定的规则（公式）构成的、能在句法结构中承担某种句法成分的语法单位。

对这个定义的具体解释包含以下几层意思。

（1）短语是一个语法单位。现代汉语短语结构知识库的构建以词组本位语法体系为指导（朱德熙 1982，1985；陆俭明，1992），将语法单位划分为“语素”、“词”、“短语(词组)”和“句子”。

（2）短语能在句法结构中承担某种句法成分。句法结构指的是主谓结构、述宾结构、述补结构、定中结构、状中结构、联合结构等，汉语的短语和句子都可以由这样的结构实现。句法成分就是指短语能作这些结构中的主语、谓语、述语、宾语、补语、定语、状语、中心语等。

（3）短语是由两个以上的词或短语组成的。这里的“词”指的是《现代汉语语法信息词典》数据库中“词语”字段的登录项。关于“词语”以及词典数据库各个字段的说明，请参见（俞士汶等，2002）。有的短语是由“词”和“词”组成的，例如

<述宾短语>::=va＋n

这里，::= 即“ 定义为 ”。这个公式的意思是“ 述宾短语由体宾动词（代码是 va）后接名词（代码是 n)构成 ”，这种短语的实例有 “ 踢足球，学英语，吃面包，整理图书,…… ”。有的短语是由 “ 词 ” 与 “ 短语 ” 或 “ 短语 ” 与 “ 短语 ” 组成的，又如

<主谓短语>::=n(p$子类)+<述宾短语>

这里，“ p$子类 ” 的含义是字母 p 包含在名词库的 ‘ 子类 ’ 这个字段中，也就是限定了这个名词是指人的。这个公式的意思是 “ 主谓短语由指人名词后接上面的述宾短语构成 ”。

（4）短语是按照一定的公式组成的。“ va+n ” 构成的是述宾短语，“ n+va ” 所构成的就不是述宾短语。现代汉语短语结构知识库中包含了数以百计的这样的公式，用它们生成的短语在语法上是正确的，并尽可能多地覆盖现代汉语的语言事实。

（5）由于在上述的短语定义中把比词更大的语法单位即短语也看作短语的构成成分，当《现代汉语语法信息词典》中包含一些比词较大的单位时，如成语、习用语等，在逻辑上就不会发生困难。然而《现代汉语语法信息词典》中实际上也还包含了一些比词要小的语法单位，如前接成分、后接成分等。由这些成分构成的实际上是 “ 词 ” 而不是 “ 短语 ”，例如

h(n$后接词性)+n(音节数=1)

所代表的就是词(如 “ 阿叔，阿妹，老赵，小李 ”)。对于这样的合成词，现代汉语短语结构知识库也给出了它们的构成公式。

（6）“ 坐着，研究了，谈过，闲着，胖了 ” 这一类由动词或形容词加助词 “ 着、了、过 ”(有的语法书叫做“ 后缀 ”)等构成的结构介于短语与合成词之间，可以归入短语，这里称做 “ 准短语 ”。

3.短语的分类

可以从不同的角度对汉语的短语进行分类。

3.1 固定短语与临时短语

固定短语是指内部词语相对固定而不能随意替换的短语,例如 “ 固若金汤、胸有成竹、大逆不道 ” 之类的成语，又如 “ 大风大浪、总而言之、由此可见 ” 之类的习用语。这类固定短语不是现代汉语短语结构知识库所考虑的对象，可以作为 “ 词语 ” 的登录项

收入《语法信息词典》。

临时短语是指按照一定的结构与规律可以临时组合成的短语，如“很大、非常大、特别大、大房间、大操场、困难大、脾气大、克服困难”等等都是临时短语。现代汉语短语结构知识库以临时短语为研究对象。

3.2 自由短语与粘着短语

自由短语是可以独立成句的。绝大部分短语都是自由的，粘着短语不能独立成句。设 p， n 分别是介词、名词代码，按公式

p＋n

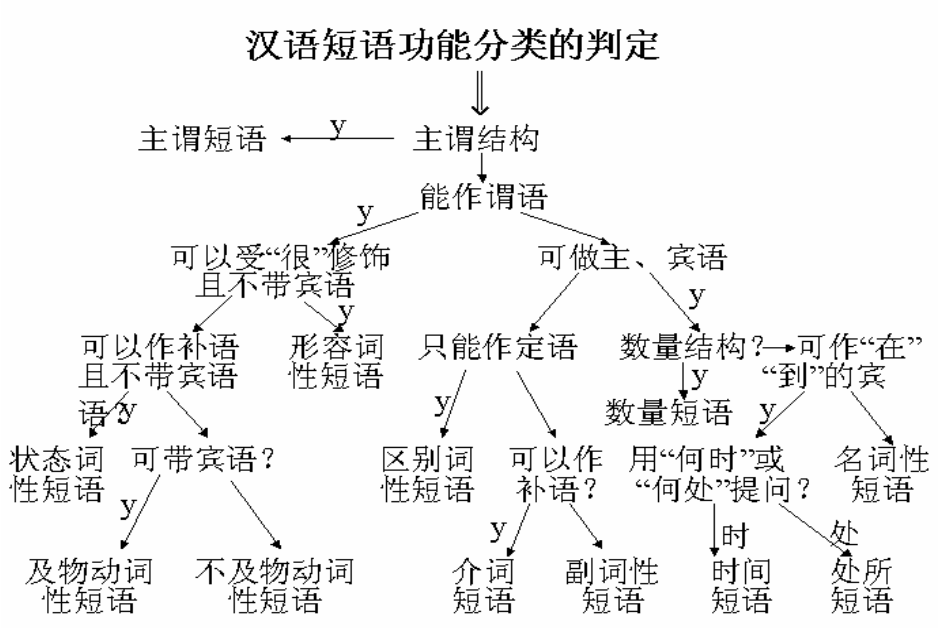
构成的介词短语就是粘着的，例如:“对爱情”、“关于交通”、“为人民”等。

在判断一个短语是“粘着的”或“自由的”的时候，会受到个人语感的影响，需要仔细斟酌。

3.3 依据语法功用的分类

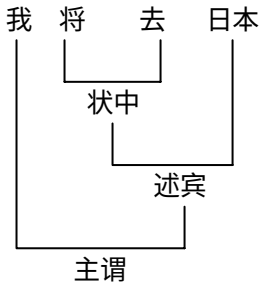
朱德熙先生主张的词组本位语法体系的要点之一是词类体系应当依据词的语法功能优势分布建立。笔者将这个理论贯彻到各个语法单位（语素、词、短语）的分类中，建立了一个短语的语法功能分类体系。可以根据短语在更大的句法结构中承担的句法功能将短语分为名词性短语（例：理论基础、有机化学、北京大学、告别宴会、经济建设、科学技术），时间短语（例：今天上午、会议之后、 上星期三），处所短语（例：北京中关村、黄河以北）， 数量短语（例：三本、五公斤、一杯），动词性短语（例：继续发展、联系过、吃苹果），形容词性短语（例：聪明伶俐、不平凡、很平静），状态词性短语（例：冰凉冰凉的、满满当当、鼓鼓囊囊），区别词性短语（例：超大规模、大中小型、金银），副词性短语（例：不经常、飞快地、历史地、讽刺地），介词短语（例：关于专家系统、在学校、被老师）。名词性短语、时间短语、处所短语、数量短语可以合并为更大的类，即体词性短语；动词性短语、形容词性短语、状态词性短语也可以合并为更大的类，即谓词性短语。各类短语的功能同相应词类的功能大致相同。介词短语可以承担补语、状语、定语等。主谓短语（例：树叶绿了、 学业进步）通常也被看成谓词性的，但它是完整单句型短语，或者说汉语多数句子是主谓短语， 因此这里将主谓短语自成一类，不归属于谓词性短语。

一个具体的短语大致上可通过如下判别步骤进行归类。



3.4 从内部结构上分类

可以划分为主谓、述宾、述补、定中、状中、连动、复谓、方位等多个类别。由于短语可以由词也可以由短语构成，如果在短语的构成公式中只有词类代码或具体的词语(即第 4 章要介绍的终极符)，那么称这样构成的短语为线性的，如果包孕了短语,那么称这样构成的短语为分层次的。对于分层次的短语，其结构以最后构成的短语结构归类，而不考虑作为其构成成分的短语的结构类别。例如：



只称短语“我将去日本”为主谓短语。

某些短语有明显的中心语，如“状中”、“定中”等，这样的短语将继承中心语的某些属性，因此指出这类短语的中心语是有意义的。

3.5 粘合式短语与组合式短语

还可以将短语分成粘合式与组合式的两类（陆俭明，1992）。以述宾结构为例。若述语是单独的动词(不带补语与“着、了、过”等助词)、宾语是单独的名词(不带定语，不是代词)的，则这样的述宾短语是粘合式的。此外，凡不符合上述条件的述宾结构都是组合式的述宾结构。其它如述补结构、定中结构都可区分为粘合式的与组合式的。

粘合式与组合式的划分虽然是从组成成分着眼的，但两者的功能也是有明显区别的。仍以述宾短语为例，粘合式的述宾短语可以直接修饰名词(即不带“的”字)，而组合式的述宾短语则不能直接修饰名词。例如

绣花	绣花鞋	绣了花	绣了花的鞋	*绣了花鞋
讲课	讲课老师	讲大课	讲大课的老师	*讲大课老师
观察事物	观察事物方法	观察新事物	观察新事物的方法	*观察新事物方法

这里，“绣花”、“讲课”、“观察事物”是粘合式短语，“绣了花”、“讲大课”、“观察新事物”是组合式短语。前面有“*”号的短语表示该短语有问题或站不住。

4.短语的歧义消解及其描述方法

分析短语结构的歧义现象是短语结构知识库研究的重要目标之一。这里不研究词汇范畴内的歧义，例如“吃饭”这个短语中的“饭”有两个意思，一是指“饭食”，二是指“米饭”。“吃饭”当然也就有两个意思。“吃饭”还可以引申为“生活”的意思。这些就不予考虑了。

“v＋n”可以构成述宾短语，也可以构成定中短语。这种语法上的同构歧义现象及其消解策略是短语结构知识库的描述重点之一。策略是在描述短语结构的过程中逐步消解这类歧义结构。对基本词类作进一步的子类划分及详细描述各个基本词类的语法属性（俞士汶等，2002）是消解歧义、分化歧义结构的两个基本手段。仍以“v＋n”为例，设vi是不及物动词，va是体宾动词。由此可以判定“vi＋n”只能是定中结构(如：“工作技巧、作战方案、睡觉时间”)，而“va＋n”就可能组成两种不同结构的短语。若在动词属性库中有关于“动词是否可直接修饰名词”的属性描述（<>是“不等号”），则

va(后名<>“可”)＋n

只能构成述宾结构(如“骑自行车、骑马、采取措施”)，而

va(后名＝“可”)＋n

还是仍有可能形成两种不同的结构(“驾驶汽车”是述宾结构，“驾驶技巧”是定中

结构)。这就需要进一步研究分化的方法。当然，有些短语的歧义在语法知识的范围内是没法消解的。例如“学习文件”可以理解为述宾结构（学习了文件)，也可以理解为定中结构(用于学习的文件)。这样的歧义在短语结构知识库的范围内消解不了。

注：《语法信息词典》中有相当多的词类划分了子类，并规定了代码，如：名词、量词、形容词等，也有的词类，其数据库未定义“子类”字段，像动词。而短语结构知识库为了描述简洁，应用了动词的子类。这里动词的子类是根据动词库中的“体谓准”字段的值划分的。

vi：表示不及物动词，即不能带宾语的动词
va：表示只可带体词性宾语的动词
vb：表示只可带谓词性宾语的动词
vc：表示只可带准谓词性宾语的动词
vd：表示可带体词性和谓词性宾语的动词
ve：表示可带体词性和准谓词性宾语的动词
vf：表示可带谓词性和准谓词性宾语的动词
vg：表示可带体词性、谓词性和准谓词性宾语的动词

5.关于短语结构描述方法的解说

5.1 汉语短语的命名方法

可用汉字,也可用代码。汉字直示其义,如“整数”、“小数”、“概数量”等。代码一般分为三部分，前 2 个或 3 个字母代表短语的结构，如“zw”代表“主谓”,“sb”代表“述宾”，“sbu”代表“述补”,“dz”代表“定中”,“zz”代表“状中”等等。 最后一个字母可能是“p”或“w”。“p”代表短语(phrase),“w”代表词(word)。既不是 p 又不是 w 者，则表示是准短语，如“vzhe”,“ale”。中间的若干个字母代表构成成分的词类。例如“dznnp”表示“由名词与名词构成的定中短语”,“sbvnp”表示“由动词与名词构成的述宾短语”,“dzmqnp”表示“由数词、量词、名词构成的定中短语”,“dzhnw”表示“由前接成分与名词构成的定中结构的合成词”。若构成成分超过 4 个，则取前 3 个词类与最后 1 个词类用于命名。

5.2 短语结构的形式描述

5.2.1 一般公式

短语结构的一般公式为:

<<短语名称代码>>::=<<表达式>>{;<<属性>>=<<值>>}[/*<<例>>*/] [/*<<注>>*/]
这里,“::=”代表“定义为”,由于一个短语往往需要用若干个公式才能完全定义,这里的“定义为”不妨理解为“表达式之一”。

<<表达式>> 就是指“v+n”,“d+a”,“vi+n”,“va(后名<>“可”)+n”这一类的式子,详见 5.2.2 的解说。

花括号表示其中的内容可以不出现或重复多次,方括号表示其中的内容可有可无。

<<属性>> 指所构成的短语的特性,目前列出的属性有:

① 自粘(自由与粘着),其<<值>>为自或粘;
② 功用,其<<值>>为体、谓、区、副;
③ 粘组(粘合与组合),其<<值>>为粘或组;
④ 线层(线性与层次),其<<值>>为线或层;
⑤ 结构,其<<值>>为主谓、述宾、述补、定中、状中等。
⑥ 中心,其<<值>>为 n,v,vi,va,aa,……

举一个典型的实例:

```
dzmqnp::= <整数>+qa+na;自粘=自;功用=体;粘组=组;线层=层;结构=定中;
          中心=na /*三匹马*/
```

这个公式的意思是:数量名定中短语可以由<整数>后接个体量词 qa 再后接个体名词 na 组成,这种短语是自由的,体词性的,组合式的,层次的,定中结构,中心词是 na。<<例>> 和<<注>> 都是注解,可以随便写什么。这里,在<<例>>的位置上写“三匹马”,是该公式的实例。

5.2.2 表达式的组成

表达式是将构成成分按其在短语中出现的次序用加号“+”连接起来的式子。表达式只有一个构成成分则无需用加号。少量表达式只有一个构成成分,是为了递归地描述短语的结构(见 5.2.3 和 5.2.4),和短语的定义没有矛盾。

表达式的构成成分为 3 类:终极符,非终极符,函数。

终极符包括词类、子类及属性描述,也包括常用的具体的汉语词语(如“的”、“了”、“着”、“一”等等)。

词类与子类各用一个小写字母表示 (俞士汶等,2003)。词类与子类代码 n,na,nf,p,m,ma,r,a,qd 等等)实际上表示的是《语法信息词典》中属于该词类或子类的词语的集合,属性描述就是对这个集合的限制。若有属性描述,则用圆括号括起来,紧接着置于词类代码或词类子类代码之后,像下面例子中用到的 n(“p”$子类)。属性描述又可分为两个部分,第一部分可以是词语分类库中的字段名与值的对。每一个字段名与它的值之间可用=(等号)、<>(不等号)及$(被包含)连接,也可以是其它可利用的属性,如音节数等。有些属性前面可以加上逻辑算子.NOT.(即“否定”),如

da+a(.NOT.(很=“否”))。 不同的属性之间用逗号连接，表示这些属性之间是“逻辑与”的关系。第二部分为排除(用“cex”或“wex”表示)，即从已限定的集合中排除一些元素，“cex：” 之后排除的是某些子类，“wex：”之后排除的是某些具体词语。

表达式中属性描述的“属性”同5.2.1置于表达式之后的“<<属性>>=<<值>>”中的“属性”分别指向不同对象：前者指构成成分的属性，后者指所构成的短语的属性。请见以下实例及简要说明

nf＋n(“p”＄子类)

这里，第一个成分是专有名词nf，第二个成分是名词n，但对名词的属性有限制，其子类应包含“p”，即指人名词。

ma＋n(数名＝“数”)

这里，第一个是系数词，第二个是名词库中可直接受数词修饰的那些名词。

r(后名<>“否”)＋n

这里r是代词，在代词库中，“后名”的意思是该代词若不能直接修饰名词，则此字段填“否”，后名<>“否”则表示可直接修饰名词。

ma＋“百”＋ma(wex:“两”)

三百八、两百三都是正确的百位整数，但一般中国人不说“三百两”，故从“百”后的系数词ma中排除具体的词“两”。

表达式的第2类构成成分为非终极符，非终极符即是短语名称代码，如:dznnp，sbvnp，整数等。凡非终极符在表达式中出现时，两端加上尖括号，如

<整数>＋qd＋na

n+<zzdvp>

表达式的第3类构成成分为函数。在某些词类数据库中列入了词语的NN、VV、AA、AABB、ABAB、A里AB之类的形态变化。可以采用函数形式表达这些属性。例如有些形容词(如“孤单”，“亮堂”)的ABB形态构成的是状态词，可用如下公式表示

aazw:=ABB(AB:a，ABB=“ABB”)

其中ABB是函数名，括号内说明，AB应是形容词，且有ABB的变化形态，则此形态构成状态词。像“高兴高兴”、“会不会”之类的结构，前后两个成分是同一的，则用如下方式表达

a>(ABAB="ABAB")+a<

v>(助动词=“助”)+“不”+v<

表达式的前一个形容词a或动词v之后有‘>”，表示后面会重复出现该词，即‘a>”与“a<”配对，“v>”与“v<”配对。

5.2.3 公式的分立与合并

使用同一个名称代码的短语往往要用若干个不同的表达式才能完全表达，这可以用分

立的公式表示。例如:

数字序列::=<数字元素>

数字序列::=<数字元素>+<数字序列>

这种分立的公式之间是“ 逻辑或 ”的关系，它们也可以写在一个公式中，不同的表达式之间用“ | ”隔开，如

数字序列::=<数字元素>|<数字元素>+<数字序列>

使用分立的公式便于同数据库中的纪录对应。

5.2.4 递归表示法

从上述关于<数字序列>的公式中可以看出，短语结构的递归表示（<数字序列>的定义中又用到“ <数字序列> ”本身)不仅是必要的，也是很简洁的。

5.3 短语结构数据库

为了便于管理，建立了一个现代汉语短语结构知识库，实际上采用关系数据库形式。其字段为：名称代码、表达式、自粘、功用、粘组、线层、结构、中心、例、注。数据库的每个记录可以很方便地映射到 5.2 所描述的短语结构公式。

在数据库基本建成之后，可进行各种排序、统计、分析、对比，不仅可以改进现代汉语短语描述体系，而且可以发现很多有价值的研究问题。若按“ 表达式 ”排序，则可发现大量同构歧义现象，从而为消解歧义结构提供启示。

从现代汉语短语结构知识库中摘取若干纪录列在下面（只取 1 个例，实际库中例子更多）：

名称代码	表达式	自粘	功用	粘组	线层	结构	中心	例
zzaap	a(状=“ 可 ”)+a	自	谓	粘	线	状中	后 a	绝对可靠
zwaap	a+a	自			线	主谓		谦虚好
aaccp	a	自	谓		线			多
aaccp	a+<aaccp>	自	谓	粘	层	词串		多快好省
dzccp	aa+n	自	体	粘	线	定中	n	新衣服

这 5 个记录可以很方便地映射为

zzaap::=a(状=“ 可 ”)+a；自粘=自；功用=谓；粘组=粘；线层=线；结构=状中；

中心=后 a /*绝对可靠*/

说明：可以直接作状语的形容词（即属性状=“ 可 ”）后接形容词可以构成状中结构，中心语是后面的形容词。

zwaap∷＝a+a；自粘=自；线层=线；结构=主谓 /*谦虚好*/

说明：形容词和形容词可以构成主谓结构，“ 功用 ”、“ 粘组 ”、“ 中心 ” 这 3 个字段未指定特别的值或取默认值，就略去。

aaccp∷＝a；自粘=自；功用=谓；线层=线 /*多*/

说明：这是为下一个递归公式作准备的。类似于两个形容词构成的联合结构，本知识库定义多个形容词构成的词串 aaccp。这里先假定词串 aaccp 只有 1 个形容词构成。

aaccp∷＝a+<aaccp>；自粘=自；功用=谓；粘组=粘；线层=层；结构=词串 /*多快好省*/

说明：这是构成形容词词串的递归公式。因为公式中有非终极符<aaccp>，这个公式是分层次的，所以有 “ 线层=层 ”。

dzccp∷＝aa+n；自粘=自；功用=体；粘组=粘；线层=线；结构=定中；中心=n /*新衣服*/

说明：aa 子类的形容词（即可以直接修饰名词作定语）和名词构成定中结构，中心语是 n。

参考文献

[1]陆俭明. 1992. 80 年代现代汉语语法研究理论上的建树. *世界汉语教学*. No.4, pp.193－203

[2]俞士汶等. 2003.《现代汉语语发信息词典详解》(第 2 版) . 北京: 清华大学出版社

[3]朱德熙. 1982.《语法讲义》. 北京：商务印书馆

[4]朱德熙. 1985.《语法答问》. 北京：商务印书馆

1994 年第一版
2003 年 4 月修订
2003 年 5 月 17 日定稿

Specification for the Phrase Structure Knowledge-base of Contemporary Chinese

Shiwen Yu

Institute of Computational Linguistics, Peking University, Beijing, 100871, China

yusw@pku.edu.cn

Abstract: *From 1993 to 1994, the author developed a Phrase Structure Knowledge-base of Contemporary Chinese, which has been used in language information processing during the past decade. This paper introduces the definition and categorization of phrases, the formalized description of phrase structure, and the schema of the phrase structure database.*

Keyword: *phrase of contemporary Chinese, phrase categorization, phrase structure formalization, phrase structure knowledge-base*