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

詹 卫 东

提要　本文从计算机自动句法分析的角度来审视汉语词类问题,主要观点是:(一) 过分强调“分布”与“分类”的严格对应关系,并不是正确的词类观,其负面作用是造成汉语词类的“不可承受之重”。(二) 计算机自动句法分析要求对词语的分布特征进行非常细致的刻画。确定“词类”是为了描写词语的分布特点,但词语的分布特点并不都是靠“词类”来反映的。依靠“属性特征”描述手段,可以灵活且更细致地描述词语的分布特点。(三) 现有的词语语法信息知识库主要是在两两组合的结构框架中描写词语的分布特点,而计算机自动句法分析需要在更复杂的“树”结构框架中描写词语的分布特点,即计算机自动句法分析需要颗粒度更细的词语分布知识。此外,关于词语的分布知识,还需要拓展到每个词语对其组合对象的选择限制的描述。

关键词　自动句法分析　汉语词类问题　分布　特征结构描述　树结构

近年来,汉语语法学界对现有的汉语词类体系做了很多有益的思考和讨论,形成了一些关于汉语词类的新的认识,如沈家煊(2007,2009) 。在操作层面也提出了一些判定词语词类归属的新的方法,如袁毓林等(2009) 。中文信息处理界结合对大规模真实语料进行词性标注的语言工程实践,对目前的主流汉语词类体系进行了深入反思,如宋柔(2003,2009) ,宋柔、邢富坤(2009) ,黄昌宁等(2009) ,黄昌宁、李玉梅(2009) 。本文对这些代表性观点和做法以及这些观点和做法背后的词类观念进行分析(见第 1、2 节) ,并尝试从中文信息处理中自动句法结构分析的需要出发,探讨在挖掘汉语的词类知识时应该加强研究哪些方面的问题(见第 3 节) 。

1. 当前学界关于汉语词类问题讨论中的若干代表性意见

本节概述当前学界关于汉语词类问题的讨论中有代表性的意见。虽然对汉语词类问题进行研究的文献以及相关的研究成果远不止这里提及的,但限于篇幅,本文只能有选择地加以介绍和分析。下面的述评涉及四个有代表性的研究。其中前两个从中文信息处理的角度讨论汉语的词类问题,后两个是语言学界在汉语词类问题上新的代表性研究成果。

---

*　本文部分内容曾在“语言信息处理与汉语语言知识研讨会”(2010. 5. 29 – 5. 30,北京语言大学) 和“第 16 次现代汉语语法学术研讨会”(2010. 6. 7 – 6. 9,香港城市大学) 上报告过。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。

1.1 宋柔(2003,2009)及宋柔、邢富坤(2009)的看法

宋柔对北京大学词类体系(下称"北大词类体系")进行了深入的分析,指出了其中有关兼类处理上的两个重大问题:

问题一:兼类中有逻辑错误。具体是指,北大词类体系对区别词和副词的定义造成不可能有一个词兼属区别词和副词。区别词的定义是"只能在名词或助词'的'前边出现的粘着词"。副词的定义是"只能充任状语的虚词"①。按照这个定义,区别词跟副词是互斥的,不可能有交集,也就不可能存在一个词,既是区别词,又是副词。说"自动、长期"这样的词,兼属区别词和副词②,逻辑上是不成立的。

问题二:兼类中有概念混乱。具体是指,一个词 c 的词类是根据 c 的所有分布得到的结果。如果 c 兼属 A 类和 B 类,那么,对于 c 在实际使用时的任何一次出现,即 c 在语料库中的每一个词例(token),c 的词性(词类属性)仍然是兼属 A 类和 B 类,不能根据 c 的上下文环境来定标记,即 c 在某种环境下是 A 类,在另一种环境下是 B 类。如果这样的话,A 和 B 就成了句法角色标记而不是词类标记了,词性标注也就变成了句法角色标注了。

我们的看法是:宋文指出的上述两个问题对深入认识汉语词类问题很有启发,但所指出的问题在语言工程实践中并没有那么严重,甚至在深入分析后可能会发现它们并不成为真正的问题(参看詹卫东,2009)。下面分别展开说明我们对这两个问题的分析。

我们对于问题一的思考:

1) 人们对兼类词的认识有一个"时间先后"的观念在里面。一个词 c 要先判断它属于区别词,再判断它属于副词,然后才说它兼属区别词和副词。这种看法当然不错。不过,也可以不持这种有"时间先后"的"兼类"观念,而把"区别词跟副词兼类"看作是跟区别词和副词并立的第三种词类。这样,一个词兼属区别词和副词,就不能单独说它属于区别词,也不能单独说它属于副词。比如"自动"就是兼区别词和副词,不能只说"自动"是区别词,或者只说"自动"是副词(参看张化瑞,2009)。

2) 上面这种说法给人的感觉有点像文字游戏或者诡辩。不过,这个诡辩放在汉语词类划分的语境中却有其内在道理。因为这样处理的实质是清楚地说明了"自动"的分布特点:既可以做定语,也可以做状语——而这正是我们描述一个词的词类归属的初衷。从定义的逻辑严密性角度讲,朱德熙(1982)有关区别词和副词的定义确实有逻辑错误,但从划分词类的目的角度讲,这个"逻辑错误"并不带来应用上的任何问题,它"止步于"定义的表述层面③。

我们对于问题二的思考:

1) 词典中标记一个词(词型/type) c 的词性,是概括说明 c 的所有可能的分布。在语料库中标记一个词(词例/token) c 的词性,是标记 c 在当前环境中实现了它所有可能的分布中哪个具体的分布。

2) 在上面这个表述中,"分布"也可以换成"功能"或"句法角色",在这个语境中,这三个

① 参看朱德熙(1982)《语法讲义》第 4.14.1 和第 14.1.1 小节。

② 参看朱德熙(1982)《语法讲义》第 4.14.3 小节。

③ 郭锐(2002)把区别词和副词处理为"饰词"的下位词类。以饰词为基础,可以对区别词和副词的定义进行重新表述:区别词是做定语的饰词,副词是做状语的饰词,这是避免定义表述上的逻辑问题的一种可行的方案。

词可以看做是同义语。词典中给出一个词 c 的词性，是综合说明它所有可能的句法角色，在语料库中标记 c 的词性，则是说明它在当前位置具体充当什么句法角色。在语料中标词性的目的就是说明词当前所起的具体句法功能（或扮演的句法角色）。

对于词典中的多标记词（兼类词），在语料中标注其具体语境中的词性显然是有价值的。比如，如果在词典中把“畅销”的词类定为 v 和 a，即认为“畅销”兼属 v 和 a 两类，就意味着我们认为“畅销”可以分布在谓语、述语等位置，同时还可以在“不”、“很”的后面位置上。这些句法位置实际上可以分为三类：（Ⅰ）“很”后面的位置是形容词位置；（Ⅱ）述语位置是动词位置；（Ⅲ）谓语位置和“不”后面的位置，是形容词和动词都可以占据的位置。

因此，在语料库标注中，当“畅销”出现在“很”的后面的时候，就标为 a，出现在述语位置（如“畅销 海内外”），就标为“v”，而出现在“这款车型 畅销”这个环境中，则既可以标 v，也可以标 a。两种标法都是对的，或者说，标 v 和标 a 对自动句法分析的贡献是相同的。当出现在“畅销 车型”这个环境中，标 v 和 a 理论上也都可以，但标 a 可能更好，理由是：基于目前主流的汉语规则体系，标 v 提供的句法角色信息不够清楚，v 这个词类在“v + n”语境中包含了做“述语”和做“定语”两种可能性，而标 a 则意味着这里的“畅销”只有充当定语角色的可能性，因此，标 a 对句法分析的贡献更大④。

对于词典中的单标记词，比如“搏斗/v”。在语料库中标不标它的词性，效果确实都是一样的。在给定规则的条件下，“搏斗”标了 v，也不能为判断“搏斗”在当前位置具体扮演什么句法角色提供更多的信息。但为标注语料库应用上的便利考虑，一般对语料库中所有的词，还是都标上词性比较好，省得在需要用到词性信息的时候，还要再去查词典。

1.2 黄昌宁、李玉梅（2009）与黄昌宁等（2009）的看法

其主要观点是，站在语言工程的立场上，朱德熙（1985）关于名词和动词兼类问题的四种处理策略里面，第一种策略（即“同质策略”）更好。而朱德熙先生本人以及随朱先生之后对汉语词类理论做了深入研究的郭锐，都采用的是朱先生提出的第四种策略，即“优先同型策略”。黄先生举“n + x”、“a + x”、“v + x”定中结构为例，认为其中的 x 位置上的词语的词性应该标为 n，而不是 v 或其他谓词性词类。理由是定中结构中心语应属名词的典型语法位置，处于该位置的词语只能是名词。

尽管黄先生跟朱、郭的选择不同，但从双方的论述来看，对两种策略的特点的认识并无异议。两种策略的主要区别是，同质策略要求词类与词类性质一一对应，这样就会带来大量的狭义兼类词。优先同形策略要求词类数目少，尽量不兼类，但造成词类与词类性质不完全对应。

我们的看法是：定中结构的中心语位置的词语可以统一标为 n，也可以不统一标为 n，即这个词语原来是动词，就还标 v，原来是形容词，就还标 a。对于句法分析来说，这两种处理方式没有明显的优劣之分。下面通过具体的例子分析来说明我们的看法。

---

④ “标 a 更好”的结论是有条件的。条件就是规则体系中“v + n”有歧义，而“a + n”无歧义，如果一个规则体系把“v + n”也设计为无歧义的规则，则标 v 和标 a 在这个环境中就是效果一样的。有人可能会说在相似语境下，同一个词有不同的词性标记，是工程中最忌讳的处理不一致问题。但如果认识到 a 和 v 这两个词性标记在有的语境中对于句法分析的贡献是相同的，就应该更客观地评价这种“表面上的不一致性”，其实这种不一致是语法理论体系允许的。对于一个语法知识点（比如词性标记），不能孤立地评价它的好坏，而应该在整个系统的框架中去看。如果采用“词类 + 属性描述”的模式来刻画词语的分布知识（详见下文 2、3 节），就不难理解，有的语境中的“畅销”标 a 或 v 是两可的。

引黄昌宁等(2009)的一个具体例子(原文第32例):

(1) 周若冰/nh 经过/p 认真/a 调查/v,/w 为/p 刘志明/nh 洗刷/v 了/u 罪名/n。/w

按照黄昌宁等(2009)的分析,"认真/a 调查/v"是定中词组,"调查"应该标为n。

我们的分析是 "认真/a 调查/v"并不一定需要定性为定中词组。"经过/p"的后面,可以出现大量的单个动词。例如:

(2) 经过/p 调查/v 经过/p 修改/v 经过/p 换算/v 经过/p 激战/v

也可以出现一些"d + v"或"v + v"的词组。例如:

(3) 经过/p 反复/d 调试/v 经过/p 持续/v 观测/v

还可以出现"动作"感比较强的"a + v"词组。例如:

(4) 经过/p 激烈/a 搏斗/v

以上这些词组,跟"认真/a 调查/v"的分布都是一样的,在这个位置上,作为中心成分的"调查"如果标为n(也就是把"认真/a 调查/v"定性为定中结构),那么以此类推,"修改、换算、搏斗"等等,就都要标为n。尽管黄文通过对语料统计的数据来说明真实语料中动名兼类的实际词数占比不高(大致在30%以下),但我们认为,30%的比例其实已经相当高了,另外,语料中的统计数据并不能反映理论上的潜在可能性。采用同质策略的思路,理论上可以造成所有的动词都兼名词,这也正是沈家煊先生在经过理论分析之后得出的结论(见下文1.3)。黄先生认为同质策略"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,反而可以彰显汉语名词和动词这两大实词词类各自的词类特性。"我们认为难以做到。同质策略必然导致名动形三大类实词的全面兼类,绝非少数⑤。实质上导致"词无定类"的局面。

其实,问题并不局限于这里讨论的定中结构的中心语如何标记词性,而是汉语中有一些基本的谓词性结构可以无标记地起指称表达功能。或许把上面的例(1)稍作变换,可以更便于阐释这个现象的理论意义:

(5) 经过/p 周若冰/nh 认真/a 调查/v,……

上例"经过/p"后面的词组"周若冰 认真 调查"显然可以分析为主谓结构式小句,这样,其中的"调查"就无法再分析为n,如果要把"调查"标记为n,则整个词组就得分析为一个定中结构,该结构的中心语"认真 调查"也是定中结构。可见,如果要坚持同质策略,"硬性"规定定中结构的中心语必须为n,那么后果就是,"周若冰 认真调查"将成为一个歧义(兼类)结构,它既可以分析为小句(dj),也可以分析为定中结构(np)。显然,同质策略在将"词类"跟"词类性质"一一对应起来的同时,不仅带来词的全面兼类,还会带来词组的全面兼类。

我们认为,同质策略并不能为句法分析带来实质好处,而优先同型策略可以通过在词类信息之外增加特征结构描述手段,来弥补词类与词类性质无法做到一一对应的不足(详见下文第2、3节的讨论)。黄昌宁先生根据语言工程实践(语料库词性标注)中动词占据名词位置的数据比例分析,得出在处理汉语动名兼类问题上"同质策略"更优的结论,其实可能是对语言工程的认识局限造成的。语言工程并不只是语料库中的分词和词性标注。语料库中除了标注词性,还可以标注丰富的次范畴(subcategory)信息。如果不标注次范畴信息,也可以把语料库

⑤ 采用同质策略造成名、动、形等实词兼类的具体词数有多少,要看采用什么样的标准来界定"词类的质",如果用作主语这条标准界定一个词具有名词性,按照同质策略将该词归入名词,那么,可以说绝大多数(从理论上说甚至是全部)动词都能做主语,因而动词都兼属名词。从这个角度看,采用同质策略,肯定会导致兼类词的数量太多。参看郭锐(2002:159)第7.3节的论述。

跟标注了丰富次范畴信息的词典(比如北大的《现代汉语语法信息词典》)配合起来使用。如果把语言工程简单地理解为只是标注词性的语料库,就会夸大词类标记在语言工程中的作用,或者更准确地说,是让词类标记承担本不该它承担的重负。

1.3 沈家煊(2007,2009)的看法

沈家煊先生的主要观点是:相对于英语词类的分立模式,汉语词类属于包含模式,即动词是名词的一个次类,形容词是动词的一个次类。

我们的看法是:

1) 沈先生的分析准确地揭示了汉语中谓词性成分可以比较自由地(即不加形式标记)起指称表达功能的特点。

2) 如何构造语言知识模型来表达这种特点,存在多种可能性。既可以采用词类包含模式假说(其在语言工程中的实质就是用兼类的办法来表示词的多功能性)来强调动词可以无标记地实现为名词;也可以像朱先生所主张的那样,采用优先同型策略,尽可能避免兼类,在大的格局上保持词类之间的区分相对清晰。对于具体词语的多功能性,则用特征描述的方式来补充说明(俞士汶等,1998/2003)。

3) 从语言工程的角度说,词类包容模式的后果必然是名动形三大类实词的全面兼类,这对计算机自动句法分析并无实质的好处。缺点则是增加知识表示的复杂度。因为仅靠 n,v,a 这个层面的词性标记,是很难真实全面地反映一个词的分布性质的(详见下文第 2、3 节的讨论),为了句法分析的目的,n,v,a 必然需要进行更细的下位分类(次范畴)。既然次范畴是不可避免的,那么在 n,v,a 这个层次上再做兼类的处理就没有必要了。

下面通过对两个简单例子的分析来展开说明我们的观点。

(6) "语文/n 学习/"中的"学习"是 v 还是 n?

因为这是一个定中结构(指称性结构),黄先生和沈先生的处理办法都是标 n。但对于句法分析来说,"学习"标成 n 并不能说明下面的分布差异,因而也无法满足句法分析的实际需要。

表 1

分布环境	语文 学习	语文 课本	语文 问题
(a) 两个 ____	-	-	+
(b) 有 ____	-	+	+
(c) 加强 ____	+	-	-

"语文/n 学习/n"中的"学习"如果标 n,则在词性标记层次上,这个短语就跟"语文/n 课本/n"和"语文/n 问题/n"没有区别了,三者都是"n n"组合。但实际上,通过上表我们构造的三个分布环境,可以显示这三个短语的分布是有差异的(只要想找分布差异,总是能找到不同短语之间的分布差异)。尤其是"语文学习"跟其他两个短语的差异更为明显。在我们看来,把"语文 学习"中的"学习"标为 n,是过度强调了这里"学习"的名词性,而忽视了"学习"固有的动词性。跟更为典型的名词"课本"、"问题"相比,"学习"的非 n 属性是不应忽视的。要说明"学习"的分布特点,并不一定要在 n 和 v 的对立中二选一,我们认为,并没有特别充分的理由支持必须把这里的"学习"标为 v,或必须标为 n,如果离开了对"学习"的分布情况的更为精细的特征描述,标 n 还是标 v 对句法分析的帮助都十分有限。而从知识表达的整体一致性和知识库易于管理等角度考虑,避免过多兼类是更好的策略。

(7) "哭/? 没用/a"中的"哭"应该标 v,还是 n?

沈家煊(2009)的看法是标 n,认为标 n 后可以排除"哭"作为陈述语的可能性。但问题并不是这么简单的。在回答上面的问题时,还应考虑下面两例中的"哭"该如何标词性:

(8) 光/d 哭/? 没用/a　　(9) 你/r 哭/? 没用/a

以上三个"哭"没有意义上的明显不同,应看作是同一个"哭",其词性标记也应该统一才对。如果都标为 n,则"光/d 哭/n"、"你/r 哭/n"结构分析就发生困难。把这三例放在一起不难看出,问题仍然出在汉语的陈述性成分(谓词性结构)可以比较自由地(无标记地)用作指称。这个问题并不仅仅针对词这个层次,也涉及词组这个层次的语言单位。"光/d 哭/v"是典型的状中结构,即状中式 vp。"你/r 哭/v"是典型的主谓式小句 dj。这两种谓词性结构,都可以用在主语位置上,起到一般 np 所起的作用。如果要把主语位置的 v 都标记为 n,那就同样地要把主语位置的 vp,dj,都标记为 np。所以我们说,如果遵循词类包含模式的理念,或者采用同质策略来处理汉语的词类问题,必然导致全面的兼类,即同时涉及词层次和词组层次的兼类。我们认为,这样处理并不能为句法分析带来实质好处,但会为语言知识的管理带来麻烦。

1.4 袁毓林等(2009)的看法

袁文的主要观点是:词属于某个词类有一个程度问题,借鉴模糊数学中的隶属度概念,可以对一个词的各项分布能力进行打分,计算一个词属于某个词类的隶属度。

我们的看法是:详细考察每个词的具体用法(各项分布能力)是很有价值的,但把分析结果用隶属度来描述,实用价值则比较有限。对学习者来说,隶属度并不是一个特别直观的概念。对计算机来说,隶属度信息则不如直接从带标语料库中统计一个词语不同分布的频次更可靠。打个比方,隶属度就像是加法的结果,而计算机在句法分析时需要使用的是词的具体分布特征信息,即需要的是具体的加数。比如 5 +4 =9,如果仅仅知道结果为 9,是无法还原出加数为 5 +4、还是 3 +6、2 +7、或者 1 +2 +6 的。而如果对词的各种具体用法做了详细的描述(知道了各个具体的加数),则最终的和数(隶属度)是可以用规则直接导出的,不需要一一标注。

下面是袁文中一些词语隶属度计算的实例。这里只给出基于隶属度计算的词类归属结果,隐去了计算过程。

媲美:动词(隶属度 0.6,不太典型)

免费:动词(隶属度 0.6,不太典型);区别词(隶属度 0.4,无法归入);副词(隶属度 0.1,无法归入);修饰词(0.4,无法归入)

临时:状态词(隶属度 0.6,不太典型);区别词(隶属度 0.8,比较典型);副词(隶属度 0.7,不太典型);修饰词(隶属度 1,典型)

结巴$_1$:动词(隶属度 0.3,无法归入);形容词(隶属度 0.9,比较典型)

结巴$_2$:名词(隶属度 0.7,不太典型)

口吃:形容词(隶属度 0.6,不太典型);动词(隶属度 0.7,不太典型)结论:兼类

很显然,无论对于人还是计算机,如果仅有词类信息和隶属度信息,对于了解这些词语的分布(功能)特点,帮助是有限的,只有知道了具体的隶属度计算过程,也就是对每个词语的每个分布特征的评分,才是真正了解了该词语的分布情况,才能为计算机句法分析所用。

2. 为什么学界对汉语词类划分问题有这么多争议

我们认为,学界对汉语词类问题的主要争议,其实质都是源于一个相同的假定:词类可以而且应该等同于词的分布。这一假定也可以表述为"严格同分布的词类观"。

上述这种词类观是语言学家在以结构主义为理论指导进行词类划分时追求的目标。但人们(特别是做中文信息处理的学者)在深入学习和实践后却会发现,已知的词类划分结果都并没有能够真正贯彻这种"严格同分布的词类观"。

为了句法分析,"词的分布差异"当然是描述得越细越好,但这并不等同于"词类分得越细越好",也不等于让一个词兼更多的类,就能更好地描述"词的分布差异"。为说清楚这个道理,必须联系划分词类的目的,或者更大一些,联系语法研究的目标来讨论。从计算的角度讲,语法研究的目标有两个:

1) 判断一个表达形式 S 能说不能说(合语法或不合语法⑥)?

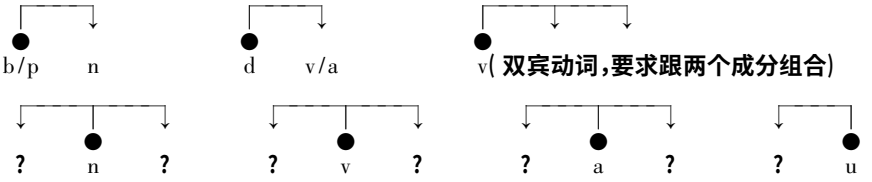
2) 如果一个表达形式 S 合语法,它的结构是什么?

在工程上,人们常常通过回答(2),来回答(1),也就是说,当计算机能分析出 S 的结构,计算机就认为 S 合语法,反之,不合语法⑦。为了分析出 S 的结构,人们给 S 中所有的词标上词性(分词类)。人们希望,一个词 w 的词性标记 t 最好能准确传递这样的分布信息⑧,或者说是分布信息至少要能回答下面三个问题⑨:

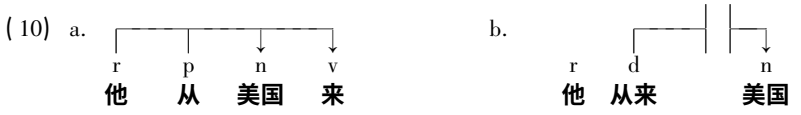
w 的组合方向:(1) w 在参与序列组合时朝哪个方向组合? (分布特征 1);

w 的组合对象:(2) w 要求跟几个成分组合? (分布特征 2)

(3) w 要求跟什么类型的语言成分组合? (分布特征 3)



在上面的图示中,词性标记 b(区别词)、p(介词)、d(副词)在多数情况下都能很好地传递出分布信息。因为这三个标记都可以清楚地回答上面三个问题。标记为 b、p、d 的词都是向后组合,且只跟一个成分组合,其中 b、p 只跟 n 组合,d 只跟 a 和 v 组合。像 b、p、d 这样的词性标记,可以为句法分析做出很大贡献。不妨看下面两个简单示例。



例(10a)是合语法的句子,例(10b)是不合语法的,仅根据两个词性标记序列,"r p n v"和"r d n",计算机就可以对此作出正确的判断。因为前一个序列中,p 要求

⑥ 影响一个表达形式能说还是不能说的因素很多,既有语法(句法)层面的因素,也有语义、语用(篇章)层面的因素。这里的讨论,仅限于考虑句法层面因素的影响。

⑦ 当然这只是一种做法。也有别的做法,不用分析结构,直接回答问题(1)。还有的情况是对于 S 的结构有不同的表示方法,比如 HPSG,链语法,依存语法,等等,都是不同的结构表示方法。甚至可以有所谓的基于记忆的方法,就是把所有的合法的句子死记硬背下来,不用管内部结构。这样也能判断一个句子是"合法"的,如果碰巧待判断的句子是背过的。但由于句子的无限性,这种方法的缺点是显而易见的。

⑧ 其实,句法分析需要的分布特征信息远远不止这些(见第 3 节的讨论)。

⑨ 为满足句法分析的需要,实际中关于分布还要回答更多问题,如 w 跟其他词的组合类型(分布特征 4)。比如都是 v 向后跟 n 组合,还可以形成述宾(批评 学员)和定中(游泳 教程)两种组合类型。

向后跟 n 组合，组合之后形成的 pp 要求向后跟 v 组合，组合之后形成的 vp 要求向前跟 r 组合，最终成为一个句子。后一个序列中，d 要求向后跟 v 或 a 组合，但它的后面是个 n，二者无法连通，形成一个鸿沟，使整个序列无法成为一个合法句子。

相比之下，标记 n(名词) ，v(动词) ，a(形容词) ，u(助词) 则都无法准确回答上面三个问题，即这些词性标记不能很好地传递出分布信息。

把“词 w 的分布”这个概念分解为关于 w 的组合方向和组合对象两个层次上的问题之后，就很容易看出，词类问题的实质是，基于 n、v、a、b、d、p、……等等这些词类标记，是否能清楚、全面、精确地描述一个词 w 的组合能力。如果 n、v、a 等标记无法做到这一点，那无论是兼类的策略，还是不兼类的策略，就都无法满足句法分析的需要。

事实上，并没有必要去强调“同词类的词应严格同分布”这个观念。把词类看作是根据分布标准对全体词语所做的一个严格的子集划分，这是不现实的。在现实中，我们通常只了解一个词的“部分分布”，而非“全部分布”。词的分布信息是一个潜在无穷的数据集。这个潜在无穷并不针对通常的所谓新词新语或词语的新用法而言，而是针对我们用于刻画分布的尺度而言。分布可以定义为组合环境，组合环境可大可小，可粗可细。如果组合环境的测试尺度不断加细，就意味着词语的分布可以不断细化下去。下面通过两个例子来说明这个看法。

(11) 如何通过给数字分类，来分析(判断) 下列数字形式的合语法性?

a. 二(* 两) 十　　二(两) 百　　二(两) 千　　二(两) 万　　二(两) 亿

b. 三十二(* 两)　　三百二(* 两)　　三千二(* 两)　　三万二(* 两)　　? 三亿二(* 两)

c. * 五百二 万　　——　　五百二十 万

d. ? 五百 万 三千　　——　　五百二十三 万 三千

上面例中 a、b 反映同属数词的“二”跟“两”的分布差异，c 反映了“五百二”跟“五百二十”的分布差异，d 反映了“五百”跟“五百二十三”的分布差异。这些词语都属于数词，为了说明数词的内部构造，数词还可以再分系数和位数等子类(朱德熙，1982) 。但是，要解释例(11) 中的正例(合语法的形式) 和反例(不合语法的形式) ，如果坚持“严格同分布的词类观”，就必须给出关于数词的更细的划分，把“五百二”跟“五百二十”分到不同的子类中去，否则，无法说明二者的分布差异，也就无法满足句法(组合) 分析的需要。

(12) 如何说明下面四个词的分布差异?

表 2

	耐心	热心	信心	点心
有___	+	-	+	+
很___	+	+	-	-
很有___	+	-	+	-

按照一般的划分词类标准，“耐心”和“热心”属形容词，“信心”和“点心”属名词，但在上面更细的分布鉴别框架中，这四个词仍然可以显示出组合能力的差异，如果坚持“严格同分布的词类观”，这四个词就得分属四个词类，否则无法解释它们的分布差异。

上面的例子说明，现实中分出的词类只能提供关于一个词的分布的部分信息，不可能“精确”地提供关于一个词的分布的全部信息。为了描述词的部分分布信息，宋柔(2009) 提出的建议是“合理的做法是，彻底贯彻语法功能分布原则，把具有基本相同词语功能分布的词归做一类，标在词典中，对词例不再标词性”。这种做法可以去尝试，不过，我们预期结果就是可

能得到比现有词类数量增加三到四个数量级的词类个数(大致在几千类的水平)。这会给词类知识的管理带来很多麻烦。

面向计算的当代形式语法理论普遍采用的知识描述模式是"类+特征"的次范畴表达机制(Borsley,1996,Sag & Wasow,1999),对于词语的分布知识,同样也可以采用这种灵活且方便管理的模式。北大计算语言所与中文系合作开发的《现代汉语语法信息词典》采用的正是这样一种知识表达模式(详见俞士汶等,1998/2003)。这里不妨用《现代汉语语法信息词典》(1999 版)动词库中的数据为例,说明如果把用"特征"表达的分布差异,转为用"词类"来表达,会得到多少不同的词类。1999 版动词库中字段共 46 个,其中语法属性 40 个:黏着、系词、助动词、趋向动词、形式动词、准谓宾、有宾、前名、后名、介宾的后、外内、体谓准、双宾、兼语句、后接、后动量词、后时量词、存现、动介、动趋、不、没、很、着了过、在正在、重叠、aabb、V 一 V、V 了 V、V 了一 V、离合、VVO、复数主语、单做主语、单做谓语、单做宾语、单做状语、单做补语、情态词、谓词性主语。动词库中共收 14479 个动词。如果按照 40 个特征中的"单做主语、单做谓语、单做宾语、单做状语、单做补语"5 项特征的取值聚类,可将 14479 个动词分入 32 类。如果按照"动介、动结、动趋、不、没、很、着了过、在正在"8 项特征的取值聚类,可分入 399 类。如果按照全部 40 个语法属性取值的差异来聚类,可分出 7897 类(平均一个类里只有1.84 个词)。其中 6105 类是"一词一类",903 类是"两个词一类"。词数最多的一个类中也只有 89 个词。这还只是语法信息词典中动词列出的 40 个语法属性。实际上,在做句法分析时,需要知道的关于词语的分布信息并不是这 40 个语法属性就能满足的。如果考虑其他词类相关的分布特征,语法属性可能成百上千,如果基于这个数量级的分布差异来界定词类,词类数量达到上万的规模也不奇怪。

通过上面的分析,我们的结论是,学界之所以对汉语词类问题存在诸多争论,主要原因是过于强调"严格同分布的词类观"。有意思的是,争论中的具体主张却出现了两个看似截然相反的方向,一个是极端主张兼类的方向(以沈家煊先生和黄昌宁先生为代表),一个是极端主张无兼类的方向,即把"严格同分布的词类划分方法"发展到极致的方向(以宋柔先生和陈小荷先生为代表)。但实际上更可取的词类观是"基于分布而又不过分强调分布"。在这种观念下,也还有不同的做法,袁毓林先生基于隶属度的词类划分办法是一种选择,不过隶属度应用到计算上却并不方便。目前最合理的选择仍然是面向计算的形式语法理论中普遍采用的"类+特征描述"的知识表达范式。

我们认为,如果认识到"严格同分布的词类观"是不现实的,就不会在汉语词类问题上出现这么多不必要的争议。除了词类,还可以用"属性:值"这种配对特征描述形式来表示词语的分布特点。这样"词类"就不用承受那么多本不该由它承受的负担了。当然,从计算机做句法分析的实际需要出发,用"分类+特征描述"的表达方式来刻画词语的分布信息,还有进一步细化的必要。下面我们就进一步来说明,目前关于词语分布的"特征描述"还可以从基于线性串的分布框架扩展到基于树结构的分布框架。同时,要加强对词语组合对象的选择限制的描述。这是以往单纯用词类来刻画词语分布特点很少涉及但又是计算机句法分析非常需要的分布知识(参看詹卫东,2000)。

3. 句法分析需要关于词的分布状况的更精细的描述

为了满足句法分析的实际需要,描写词语的分布特征,还应从两个方面下工夫。

(一) 同类词中的不同个体(比如 X1,X2)与另一类词(比如 Y)组合后,可能发生功能分

歧。这也属于词语的分布特征，应该看作是 X1 跟 X2 的分布差异，在词库中应该表达出来。下面是三个示例。

(13) a. 原本 抽烟 的 不怕烟味

b. 也许 抽烟 的 不怕烟味

c. 老 抽烟 的 不怕烟味

上面这个例子对应的句法格式可表达为：adv + vp + 的 + vp。其中(13b)的内部结构为：[也许 [抽烟 的 不怕烟味]]，(13c)的内部结构为：[[老 抽烟 的] 不怕烟味]。(13a)有歧义，既可以跟(13b)一样理解，也可以跟(13c)一样理解。

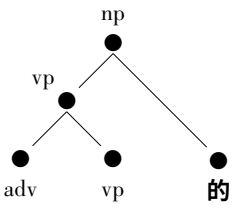
但如果上面这个格式变成：adv + adv + vp + 的 + vp，即两个副词(adv)连用修饰 vp 的格式，其内部结构再发生歧义的可能性就会降低。如例(13a)变成下面例(14)就没有歧义。

(14) 原本 就 抽烟 的 不怕烟味

例(14)的内部结构只能分析为像(13c)那样的结构：[[原本 就 抽烟 的] 不怕烟味]。例(14)之所以没有歧义，是因为第二个副词"就"加 vp 之后，形成的"就抽烟"这个 vp 不能加"的"组成"的"字结构，这样，当前面出现"原本"的时候，结构就只能是"原本"跟"就抽烟的"形成一个结构体 vp。

计算机要对例(13a)和例(14)的句法分析做出正确区分，在词库知识描述中，就需要对"就"这样的副词的分布特点作出更细致的描述，除了描述副词能修饰动词短语(vp)外，还要描述有的副词在跟 vp 结合之后，不能再跟"的"字组合成"的"字结构。关于词语的这种分布知识的考察，我们称之为"基于树结构的分布描写"，以区别于以往主要是基于两两之间成分组合来描写分布能力。后者可以称之为"基于串结构的分布描写"。

上面副词跟 vp 组合后是否再能跟"的"组合，可以用下面的树形图表示：



"老、一直、原本"等副词都可以进入这个树图的 adv 位置，"也许、就"等副词则不能进入 adv 位置。以这个树结构作为考察框架，可以呈现这几个副词的分布差异。

其实传统的语法研究中早就关注过类似的词语分布差异研究。目的都是进一步揭示同类词内部成员之间分布特点上的不同。下面再举两例说明。

关于形容词"好"跟"高"的分布差异：

(15) a. 不 好 —— 不 高

b. 张三 比 李四 更 不 好 —— * 张三 比 李四 更 不 高

作为形容词，"好"跟"高"都可以受"不"修饰。"不好"跟"更"组合后用在比较句中没有问题，"不高"跟"更"组合后，就无法再用在比较句中。这样的分布差异在语法学界研究比较句时是注意到的，但是，还没有人从词语分布知识的角度来看待这种差异，更没有把这种分布知识记入计算机可用的词库中，为计算机句法分析提供帮助。

关于形容词"胖"跟"高"的分布知识：

(16) a. * 有 胖 —— a'. 有 那么 胖

b. * 有 高 —— b'. 有 三米 高

"胖"跟"高"都是形容词，因而不能作"有"的宾语。这是《现代汉语语法信息词典》中可以查到的关于这两个形容词的分布知识。但"胖"跟"那么"组合形成"那么胖"，"高"跟"三

米”组合形成“三米高”之后，就都可以跟“有”组合了。像这样的更复杂的词语分布知识，目前的面向计算机使用的语法知识词典中，都还没有涉及到。这就需要我们今后进一步完善语法信息词典，或者用短语组合规则的办法来刻画这种更为细化的关于词语分布情况的知识。

（二）词语（X）对组合对象的选择限制，对于句法分析是起到实质作用的语法知识。目前已有的词类知识也好，语法信息词典中关于词语分布特征的描述也好，都没有系统地涉及词语对组合对象的选择限制描述。

强调汉语特点的语法学者在谈到汉语动词的分布性质时，往往只说动词能做主语（或相对的，名词能做谓语），以此作为汉语跟印欧语相比的显著差异。但是，告诉计算机动词能做主语，或者规定动词不能做主语，而是动词变成（实现为）名词后才能做主语，对计算机进行句法分析，并无实质的帮助。对计算机来说，真正的问题不是“哭”这个动词（或名词）能不能做主语，而是它做谁的主语，即它做主语的时候，对其组合对象——谓语提出了什么选择限制要求。“哭”做主语的能力显然没有名词那么强，要受到很多限制。下面把“哭”跟典型名词“计算机”做主语进行对比：

	a		b		c
（17）	哭 解决不了问题	——	* 哭 发现了问题 * 问题 哭 解决了	——	哭 出了问题
（18）	计算机 解决不了问题	——	计算机 发现了问题 问题 计算机 解决了	——	计算机 出了问题

例（17a）和例（18a）中“哭”跟“计算机”表现一样，都是做主语。但在 b 组例子中，“哭”露出了它原本是动词的马脚，两个例子都不合语法。“计算机”则是纯正名词，仍然可以在例中充当主语，包括普通主谓宾句中的主语和主谓谓语句中的小主语。c 组例子中，“哭出了问题”有歧义。“计算机出了问题”则没有歧义。我们可以构造例（19）来展示例（17c）的歧义：

（19） a. 他 眼睛 哭 出 了 问题

b. 他的戏份中，哭 出了 问题，笑 没有问题

对句法分析的任务而言，把“哭”处理为名、动兼类，并没有为准确说明“哭”的分布特征，特别是“哭”对其组合对象有什么选择限制，提供更多的有帮助的信息。所以，当我们为了计算机句法分析的目的更深入地挖掘词语的分布特点时，根本性的问题并不是为“哭”找一个合适的词类，主张“哭”专属动词也好，还是兼属动词和名词也好，都只能传递出比较有限的分布信息。抛开词类问题（主要是兼类问题）上的争议，站在更高的角度去看汉语语法分析问题，不难体会到，句法分析的要求实质上就是要给出词与词（词组与词组）之间的组配约束条件（参看詹卫东，1999）。词类只是表达组配约束条件的一种手段（或者说是一部分信息），并非全部，而且也不可能是全部。

4. 结论

最后，我们把本文分析的各家词类观点概括为表 3（见下页）。

表面上两种截然不同的词类划分做法（表现），怎么会都源自一个相同的理念——都是强调“严格同分布的词类观”呢？乍一看似乎令人费解。其实通过本文的分析，不难体会其中的道理：强调严格同分布的词类观，就会发现当前通行的词类体系其实都不满足严格同分布的要求；为了实现严格同分布，就出现两种不同的策略：1）把严格同分布的理念发展到极致，只要词语分布有差异，就设置新的类，导致词类数目大量增加。2）重新“解释”词语的分布差异，

表 3

关于汉语词类的代表观点	理念	结果(具体做法或表现)
甲: 宋柔(2009),宋柔、邢富坤(2009),陈小荷(1999)	强调(或追求)严格"同类同分布"的词类观	类的数目大量增多⑩
乙: 黄昌宁等(2009),黄昌宁、李玉梅(2009),沈家煊(2007,2009)		类的数目大量减少(或具体词语的兼类增多)
丙: 袁毓林等(2009)	不强调"严格同分布"(认为词类是一种原型范畴)	类的数目无大变化,类的成员带上隶属度
丁: 本文	不强调"严格同分布"(认为词类划分是相对的,分布知识可由特征来表达)	类的数目无大变化,类的成员带上丰富的次范畴特征描述

只要两个词语出现了相同的分布,就"主动忽视"它们的差异。于是当动词出现跟名词一样的分布时,动词就自动地成为名词(形容词跟动词的关系也依此类推)。第一种策略可以概括为"虽千万类,吾往矣",体现了严格按照分布划分词类的一种执着精神。第二种策略可以概括为"你中有我,我中有你",体现了面对汉语主要实词可以不换装扮演多种角色(无标记转指)时的一种无奈态度——既然类分不清楚,干脆不分算了。

本文通过在前人和时贤有关汉语词类问题的深入讨论基础上做对比分析,并结合自己从事计算机句法结构分析的经验,重新审视各家关于汉语词类问题的观点,最终形成了对于汉语词类问题的如下认识:

1) 为了作语法分析,我们需要知道一个词的分布信息。如果能全面掌握任意一个词语的分布知识,就能判断(或预测)这个词能出现在哪儿,不能出现在哪儿(能跟谁组合,不能跟谁组合)。

2) 理想的情况是,根据词的分布不同,把词分到不同类中,同类词具有相同分布,不同类词具有不同的分布——这就是"严格同分布的词类观":词类 = 分布。

3) 上述理想情况是不可能做到的。原因可以分两个方面说,理论上,"分布"本身缺乏严格的定义,建立在一个模糊概念上的词类划分体系,不可能达到严格的同词类同分布目标。实践中,我们对于一个词语的了解,永远都只是局部,关于一个词语的全局分布的知识,只存在于想象(假设)中。

4) 尽管理想目标无法达到,但在现实层面,为了语言信息处理的需要,根据实用的目的描述词语的主要分布知识还是可以做到的。其中最主要的分布知识,可以用"词类"来描述,其他的分布知识,则可以通过"属性:值"这种配对特征结构表达模式来描述。这种表达模式适用于从语素到词到短语到句子等各种不同层级的语言单位的句法语义知识的描写。

5) 在过去二十多年中,北大《现代汉语语法信息词典》采取这种"词类 + 属性描述"的知识表达范式对汉语中超过八万普通词汇进行了分布知识的描写。其中的具体知识对于计算机分析汉语句子的句法结构提供了很大帮助,但同时也要看到,目前的分布知识局限于两两组合的串结构分布,还需要扩展到基于树结构框架的分布知识描写,另外,目前的分布知识中缺少

⑩ 类的数目增多、减少、或无变化,指相对目前比较通行的词类体系中词类个数而言。

对词语组合对象的选择限制条件的归纳描写⑪。而这两方面的知识，正是计算机自动句法分析迫切需要的。从语言工程的角度看，与其过多地把注意力放在“词类”的争论上，还不如花大力气用丰富的属性描述手段来精细地刻画“关于词语的详细分布状况的知识”。

本文继承了朱德熙、陆俭明、郭锐、俞士汶等学者关于汉语词类划分的观点和做法。同时根据我们自己标注语料库以及进行句法结构分析的经验补充了一些新的思考，就教于大家。

参考文献

白　硕　1994　《词类划分的数学理论》，《软件学报》第6期，科学出版社。

陈小荷　1999　《从自动句法分析角度看汉语词类问题》，《语言教学与研究》第3期。

郭　锐　2002　《现代汉语词类研究》，商务印书馆。

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

黄昌宁　李玉梅　2009　《评动、名兼类词的四种划分策略——来自语言工程的观察》，《语言学论丛》第40辑。

陆俭明　2003/2005　《现代汉语语法研究教程》(第三版)，北京大学出版社。

沈家煊　2007　《汉语里的名词和动词》，《汉藏语学报》第1期。

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

宋　柔　2003　《统计和规范中的误区》，孙茂松等编《中文信息处理的若干重要问题》，科学出版社。

宋　柔　2009　《从语言工程看汉语词类》，《语言学论丛》第40辑。

宋　柔　邢富坤　2009　《从语言工程看现行汉语词类体系的本质和困境》，“新视野下的汉语词类问题”系列研讨文稿。

俞士汶　2009　《关于区别词和副词兼类的答辩》，《语言学论丛》第40辑。

俞士汶　朱学锋　段慧明　刘　扬　吴云芳　2008　《汉语词汇语义学研究及词汇知识库建设》，《语言暨语言学》第2期。

俞士汶　朱学锋　王　惠　张芸芸　1998/2003　《现代汉语语法信息词典详解》，清华大学出版社。

袁毓林　2000　《一个汉语词类的准公理系统》，《语言研究》第4期。

袁毓林　2006　《关于等价功能和词类划分的标准》，《语文研究》第3期。

袁毓林　马　辉　周　韧　曹　宏　2009　《汉语词类划分手册》，北京语言大学出版社。

詹卫东　1999　《一个汉语语义表达框架：广义配价模式》，黄昌宁、董振东主编《计算语言学文集》，清华大学出版社。

詹卫东　2000　《面向中文信息处理的现代汉语短语结构规则研究》，清华大学出版社、广西科学技术出版社。

詹卫东　2009　《词类三问——一个汉语词类知识学习者和使用者的反思》，《语言学论丛》第40辑。

张化瑞　2009　《关于区别词和副词兼类的集合论定义》，《语言学论丛》第40辑。

朱德熙　1982　《语法讲义》，商务印书馆。

朱德熙　1985　《语法答问》，商务印书馆。

Borsley, Robert D.　1996　*Modern Phrase Structure Grammar*, Blackwell Publishers Inc..

Sag, Ivan A. & Thomas Wasow　1999　*Syntactic Theory: A Formal Introduction*, CSLI Publications.

(詹卫东　北京　北京大学中文系　zwd@pku.edu.cn)

⑪ HPSG、LFG等形式文法体系中常用Agreement(一致性)特征来表达两个发生组合关系的语言成分之间的相互约束关系。对于汉语这样的所谓“意合”型语言来说，形式上的Agreement特征不显著，因此有必要大力去挖掘语义层面的Agreement特征。从这个角度讲，汉语的词汇语义知识库，基于配价(或论元结构、论旨角色理论)的词语组配知识库，对于句法分析来说，就格外重要了。

Key words: the Yiwu dialects, first person pronoun, *shinong*, *shiwo*

ZHANG Ding, The source of the universal-reference usage of interrogative words in Chinese and the formation of *renhe* (任何)

Interrogative words in Chinese have a universal-reference usage. For example, *shenmo* (什么), when stressed, can mean 'anything'. This paper finds that the usage is traced to the construction 'matrix clause + indirect parametric questions', in which the main verbs of the matrix clause, such as *buguan* (不管, *lit.* not concern), imply a sense of indifference or irrelevance. When the main verb evolves into a concessive conjunction in parametric concessive conditional clauses, it can be omitted in sluicing contexts and the interrogative word becomes a universal indefinite pronouns. The paper also claims that the indefinite determiner *renhe*(任何) is a calqued word in the early years of the 20th century when Chinese widely contacted western languages.

Key words: interrogative words, universal reference, parametric concessive conditional clauses, *renhe*

WANG Lin, LI Wei, Causative markers *jiao* (叫) and *gei* (给) in Ryukyu's Mandarin textbooks

Although *jiao* (叫) and *gei* (给) are both called 'causative' markers in Ryukyu's Mandarin textbooks, they are different in meaning, imperative causative and permissive causative respectively. This distinction is found in parallel with other South dialects in which permissive causative markers usually share the same form with verbs denoting 'giving', while in most North dialects and Mandarin imperative causatives and permissive causatives fall into one form. The comparisons between southern and northern dialects are significant in typological studies.

Key words: Ryukyu's mandarin Chinese textbooks, causative marker, *jiao*, *gei*

CHEN Xiao, The intensifier *suo* (所) in the Beijing dialect of late Qing and early ROC period

The intensifier or degree adverb *suo* (所) is only found in the materials that recorded the Beijing dialect of late Qing dynasty and early ROC period. This paper enumerates exlusively the examples which show that *suo* usually modifies a complicated VP in non-commendatory or unpleasant contexts. Its origin still remains unclear so far.

Key words: Beijing dialect, late Qing and early ROC period, *suo*, intensifier

LIU Yun, Some materials of the early Beijing dialect

This paper introduces some materials of the early Beijing dialect, including books written in both Manchu and Chinese, textbooks for Beijing colloquialism, official archives of the Qing government and novels by native Beijing dialectal speakers. The massive amounts of data record the development of the Beijing dialect over a large time span and deserve more researches.

Key words: Beijing dialect, early documentation

ZHAN Weidong, A review on word classification of Chinese from the perspective of sentence parsing by computer

Based on researches of information processing, this paper endeavors to review some recent opinions on the word classification in Chinese. It argues that although syntactic function or distribution is the criterion of classification, words in the same category can not have the same functions. As in the syntactic theories such as HPSG and LFG, feature structures or so-called 'attribute-value matrices' can serve a basic framework to represent linguistic knowledge. Automatic sentence parsing by computer requires two kinds of knowledge about word distribution: one is the more complicated tree structures and the other is the word-selection restrictions.

Keywords: sentence parsing, word classification, distribution, feature structure, tree structure