

第七章 结语

§ 7.1 对本文研究工作的总结

本文研究工作可以看作是一个更为宏大的目标——“编写一部给计算机用的现代汉语语法”——的一部分。虽然距离语法大厦的最终建成还有许多路要走，但我们希望，已经迈出的这一步能够或多或少、或正面或反面地昭示未来的方向。如果本文的研究工作能够成为将来真正完整意义上的“计算机用汉语语法”的一个组成部分，那么毫无疑问走这一步是值得的，因为它是通向成功的足迹中的一个；如果将来的“计算机用汉语语法”全然是另一幅图景，那么这一步也是值得的，因为它虽然没有留下一个成功的印迹，但至少竖起了一个“此路不通”的标牌。

是成功的足迹也好，是失败的警示也好，也许以最精练最准确的话来概括本文研究工作的内容，对今后的研究更有实际的参考意义。这既包括本文研究工作所产生的有形而具体的结果，也包括在研究过程中体味揣摩到的无形而抽象的观念。前者对中文信息处理和汉语语法研究固然会有直接的支持作用，后者则可能跟面向中文信息处理的汉语语法研究的宏观面貌和发展走势更为相关。

就具体的成果而言，本课题研究工作的主要贡献可以归纳为以下三方面：

(一) 明确提出了计算机分析现代汉语短语结构所需要的句法语义范畴体系。

句法范畴方面，我们所做的工作实际上是在对“词组本位语法体系”的核心思想作清晰定位的基础上，进一步深化和拓展了已有的研究成果。具体包括两部分：

(1) 提出了一个比较完整的短语结构功能分类体系；

(2) 提出了一个初步的，具有一定规模的关于汉语词和短语句法功能的属性描述体系。

上述短语句法功能范畴体系的直接目标是描述汉语短语结构的句法性质。此外我们也可以按照这个框架，对《现代汉语语法信息词典》中的属性项目做系统地整理工作，使词典对每个词语的用法性质的记录更系统、更合理、更准确。

语义范畴方面，我们所做的工作主要是结合对已有的语义理论的理解，从实际的中文信息处理实践出发，提出了一个描述汉语实词和短语语义性质的框架。具体包括两部分：

(1) 一个简明的汉语实词语义分类体系；

(2) 描述动、形、名三类实词配价信息的广义配价模式；

按照上述语义描述模式，我们已经在—个汉英机器翻译系统的开发工作中初步完成了对四万多汉语实词语义信息的描述。

(二) 初步完成了对现代汉语短语结构的句法语义知识的形式化描述工作，给出了一套关于现代汉语短语结构的形式化分析规则。尽管其中具体的规则还有待改进提高，但已有的框架为今后的研究奠定了进行理论探讨和开展更广泛的实际工作的基础。

(三) 全面考察了现代汉语短语结构歧义的不同类型，并对计算机分析现代汉语短语结构可能碰到的歧义格式进行了统计分析。在此基础上，探讨了消解歧义的策略。我们把歧义问题的能否解决归结为能否准确地描述相关的两项成分的组合条件。如果可以在已有的句法语义范畴上描述相关两项成分的组合条件，就可能对包含这两项成分的歧义格式进行正确地分析，否则不可能解决歧义问题。

就抽象的观念而言，本文在研究过程中，形成了以下认识：

(一) 所谓语法研究，特别是面向计算机进行的语法研究，如果不是全部，至少也是在

很大程度上,可以看作是关于语言成分与语言成分之间搭配(组合)的研究。无论是范畴知识也好,规则知识也好,最终的目标只有一个,那就是回答“X跟Y能不能搭配”这样的问题(X跟Y是任意的语言成分,能否搭配也包括搭配条件在内)。什么时候计算机能轻松而正确地回答所有这样的问题了,什么时候计算机“看上去”就是理解自然语言了。

(二)一个成分跟另一个成分能否搭配(这实际上包括以何种关系搭配),是由多种因素决定的。通常,我们把这些因素称为句法因素、语音因素、语义因素、语用因素、常识因素、……等等。具体到两个实际使用的语言成分(比如某两个具体的词)能否搭配,可能是这些诸多因素中的一个起作用,也可能是若干个在起作用。实际上,能与不能的情况是不平衡的。两个成分要能够搭配,需要这两个成分满足所有制约因素的要求,而只要有一个制约因素的要求它们不满足,就足以造成它们不能搭配了。

(三)要让目前的计算机像人一样掌握全套关于语言的知识(即在各个层次上面描述成分间的所有搭配制约因素),难度太大。合理的做法应该是,尽量依托那些比较死板的,看上去没有多少道理好讲的句法知识(比如“数词能跟量词搭配这样的知识”),同时在适度的范围内引入一些语义知识(这是跟计算机讲人的“道理”,比如只能“吃面包”,不能“*吃车辆”)。这样做的理由其实也很简单,就是相对而言句法知识最容易范畴化,可操作性强。干一件复杂的事情,从简单开始,可以避免犯一些不必犯的错误。事实上,没有必要去强调汉语研究跟其他语言的研究有多么大的差异,或者强调中文信息处理跟对其他自然语言的信息处理有多么大的不同。如果有不同的话,也不过是:关于成分间的搭配,汉语能够发掘的句法制约因素少一些,相应地,语义制约因素更复杂一些。而像英语这样的屈折型语言,句法制约因素比较丰富,语义制约的负担可能稍轻一些罢了¹。当然,我们只是说不用过去去强调汉语研究应该多么有“特色”。这并不意味着要走向另一个极端,变成有意忽视汉语的特点。上述两种态度实际上都偏离了研究的初衷,对于开展研究工作都没有什么好处。

(四)原有的汉语语法理论的确有许多问题解决不了,这个令人尴尬的局面常常被归结为是受印欧语研究思路的影响造成的,于是有人主张:汉语应该走有自己特色的,语义语法的路线,应该全部抛弃句法概念(范畴),等等²。我们则不这么看,相反提出这样的反思:在句法层面描述语言成分间的搭配,真的就山穷水尽,没有潜力可挖了吗?屏弃任何句法概念,仅靠一套语义范畴体系来描述成分间的搭配制约,真的可能吗?说汉语的语法是句法为主也好,是语义为主也好,甚至冠以语用语法也不错,有了这些帽子,并不意味着我们就可以忘记,语法本来是干什么用的。面对人来谈语法的用处或许多少有些见仁见智,面对计算机来看语法的用处就容易一目了然。说得直白一些,还是回到我们上面提到的认识,即一部面向计算机的语法,就是要告诉计算机,任意两个成分X跟Y,它们能不能搭配(这涵盖了在什么场合下搭配,以及以什么方式搭配等等诸多内容)。人可以选择靠语义范畴来告诉计算机,也可以选择用句法范畴来告诉计算机,也可以选择以语用范畴来告诉计算机。在目前的研究阶段,最好不要忙着下结论,说其中只有一种选择是对的。我们的态度是,只要有办法组织起一套明确的范畴体系,可以尽可能广泛而准确地描述语言成分的搭配知识,就是好的选择。至于所选的范畴是句法的,还是语义的,语用的,或者干脆就是杂糅的,都可以有意无意地淡化。因为我们并不是为了区分句法和语义而区分句法和语义。区分这些层次,仅仅是因为可以使我们在选择描述用的范畴时显得有层次,有条理,而不是一锅粥搅在一起(如果有人能够做到一锅粥搅在一起而不慌乱,那他就绝对有理由选择不分层次)。

(五)在哲学意义上,语法研究无所谓面向什么,最高目标是搞清楚语言的秘密。在现实层面,面向人的语法研究跟面向计算机的语法研究,多少还是有别的。面向人的研究只要描述一个成分跟另一个成分的差别(特别是对那些看起来很像的两个成分)就大致可以,而且用自然语言描述就足够了,并不要求以形式化的方式来描述。比如“很多”跟“很少”,这两个成分看上去很像,但语言学家有义务指出它们的差别。再比如可以说“上菜”,不说

“*下菜”，可以说“下锅”，不说“*上锅”，而既可以说“上馆子”，也可以说“下馆子”等等，都要求语言学家显性地描述“上”跟“下”的同与异。通常语言学家是通过分类，分次类，分次次类，……这样的方式来进行描述的。面向计算机的语法研究总体来讲也是如此，所不同者，对人而言很明显的同与异，要以形式化的方式告诉计算机，却非常困难。而相当多的困难实际上是由于烦琐造成的（即描述的负担太重造成的困难）。比如上文曾经举到的例子，对计算机要描述“90年代”跟“*95年代”的差异。类似的例子还有，描述“1978年以前”跟“3500年以前”的差异，等等。人对“1978年以前”的理解和对“3500年以前”的理解不同，显得那么“轻松自如”。而要以形式化的方式告诉给计算机掌握，描述起来却绝非易事。这样的例子举不胜举。除了描述信息量的负担重以外，还存在着的确不那么好描述的差异。比如上文提到的“散步去”跟“走路去”的差异。值得注意的是，对具体的短语结构的差异问题，如果以系统的整体眼光来看待，就是如何描述“vp+vp”结构的问题。一个语法理论可以不用vp这样的范畴来称说“散步+去”和“走路+去”这样的组合，但它无论如何都要回答这类问题：“散步去”跟“走路去”到底有什么不同，以及这个不同是如何造成的。

（六）需要强调的是，回答上述问题时，要有清醒的“已知——未知”观念，即只有从确定的已知条件开始，一步步推导，最终得到问题的答案。这个过程才是合乎逻辑的。如果这个过程能够以形式化的方式加以描述，计算机就可以重复这个过程。否则，就是没有真正解决问题。比如，人可以区分“走路去”跟“散步去”，而所谓区分，通常也就是指出：在“走路去”中，“走路”是“去”的方式，而在“散步去”中，“散步”是“去”的目的，这样看起来的确是把这两个短语结构区分开了。对面向人的研究来说，这样基本上能够说明问题的。但对计算机来说，问题并没有解决，因为我们还是没有告诉计算机，从什么起点开始，可以推导出“走路”是“去”的方式，“散步”是“去”的目的。实际上，对计算机分析而言，“方式”、“目的”这样的抽象范畴应该是推导的结果，而不是起点。对人来说，要得到这个结果只要稍加体会就可以了（人的已知条件是隐性的），但计算机却无法靠“体会”得出这个结果，计算机只能依靠显性的已知条件。面向计算机进行的汉语语法研究，就是要以范畴和规则的方式尽可能多地来描述已知条件，帮助计算机在掌握了这些已知条件的基础上，理解或生成实际的汉语句子。

§ 7.2 本文研究工作的意义

以上我们大致归纳了本文研究工作的具体成果和在研究过程中形成的一些观念性的认识。我们希望这些对中文信息处理的实际应用系统开发和现代汉语语法理论的建设产生积极作用。下面我们大致勾勒本项研究在这两个方面可能提供的支持。

（一）首先，短语结构规则库可以为自然语言处理应用系统的研究和开发提供直接支持，特别是那些需要对汉语进行深层分析的中文信息处理系统。比如，汉外和外汉机器翻译。事实上，本文概括的短语结构规则正是一个汉英机器翻译系统中的语言知识库的组成部分之一。从分析的实际效果来看，这个系统对汉语短语结构，以及单个句子范围内的语言现象的处理能力还是不错的。当然，要扩大处理能力，提高分析质量，还需要对短语结构分析规则做进一步的深入研究，同时，也需要增加对**句子结构规则**的研究。

（二）本文的研究结果可以对汉语大规模语料库加工提供理论指导。比如汉语树库（treebank）的构建³，就可以在本文提出的短语结构功能分类及句法属性范畴的框架基础上展开。当然，实际语料的情况远远不是短语结构可以覆盖的。但汉语的短语结构是句法结构的核心。以本文提出的短语结构功能分类体系为基本框架，可以循序渐进，向句子结构标记和句群结构标记扩展。最终就能够形成一套相对完整的，为汉语大规模语料库标注加工服

务的，充分体现汉语语法研究成果的句法标注体系。

(三) 本文研究的具体成果以及对于汉语语法研究的观念性认识，可以为受限汉语研究⁴提供支持。受限语言研究的初衷是降低计算机分析自然语言的复杂度和难度，本质上就是把计算机分析自然语言碰到的困难，由人来分担一些。在尽量不丧失基本表达能力的前提下，对自然语言进行限制，尽量以“规范”的语言形式进行表达，这样，提供给计算机的输入句子（计算机的分析对象）就是比较“规整”的，当然分析起来也就相对容易一些了。不难看出，原来的问题并没有减少，只不过现在变成了：**如何对自然语言进行限制，产生一个合适的“受限语言子集”**。就这点而言，本文对现代汉语短语结构的系统研究，显然可以为受限汉语制订合理的规范提供参考依据。特别是对歧义格式的分析，有助于建立短语结构形式跟意义之间尽可能严格的对应关系。比如规定“vp+去”一定表示方式（如“坐飞机去”），而“去+vp”一定表示目的（如“去看电影”）。这样就通过人的“干预”把容易引起计算机分析歧义的短语结构避免了，即如果人们使用事先制订好的受限汉语规范写作，就不会出现“散步去”这样的短语结构，而只能用“去散步”。诸如此类。

(四) 关于短语结构歧义类型的分析和对汉语短语结构歧义格式的全面统计结果，可以用来指导设计针对计算机自然语言处理能力测试的“考题”，从而为中文信息处理应用系统性能的自动评测⁵提供支持。比如，对汉英机器翻译，汉语句法分析等中文信息处理应用系统，可以根据本文分析后得到的短语结构歧义的不同类型，有针对性地设计包含不同类型歧义的测试句子，并可以按照歧义程度的高低来安排具体考题的难度级别，等等。

(五) 本文的研究可以为面向信息处理的大规模电子词典工程建设提供参考。前文已经提到，在描写短语结构的组合规则时，必须依靠一部详细记录了词语语法语义信息的词典为基础。反过来，对短语结构组合规则的描写和探索，又可以指导我们回过头去看已有的词典信息乃至整体框架设计的问题⁶，从而进行相应地调整，最终使整个语言知识库的水平得以提高。同时，本文的研究也可以为汉语词类的理论研究提供一个背景。譬如有关词的理论研究中，词的定义问题，词的兼类问题等等常常纠缠不清，如果从中文信息处理实际应用的角度来看，许多纠缠不清的问题可能会显得简单一些，或者问题背后的实质能看得更清晰一些。

(六) 在本文构建的短语结构框架下，可以系统地现代汉语语法研究提供课题。比如，从结构类型上讲，可以以现有的9种结构类型为基础，向内进一步细分（比如把“连谓结构”分出更多的类型来），向外进一步探索更多的结构模式（比如“每周一三五检查卫生”⁷中“每周一三五”的结构模式）。从成分间搭配组合来讲，可以一个组合模式一个组合模式地进行穷尽性的研究（比如“np+np”组合、“vp+np”组合、……等等）。对此我们在第三章的小结中已经提及。同时，对短语结构层面具体组合模式的研究手段和基本原则，还可以扩展到对句子结构模式，以及句群结构模式的研究中去。最终形成一个比较完整的汉语语法体系。

(七) 本课题研究得到的现代汉语短语结构的具体规则以及关于应该分层次组织语言知识的思想，对汉语的语义研究有参考意义。汉语的语义研究跟句法研究的目标有很大部分实际上是一致的，也是通过范畴制约来描述语言成分间的搭配关系。只不过，语义范畴比句法范畴的抽象度低一些罢了。一方面，尽管通常认为意义是模糊的，难以形式化，但我们想强调，在目标明确的情况下，意义范畴完全可以部分地形式化，并且可以用来帮助对语言成分间的搭配组合进行分析。另一方面，对所谓可以完全不靠句法范畴，仅靠语义范畴来分析汉语（认为这样才是汉语研究的特色）的认识，恐怕要多打几个问号。语义知识跟句法知识，是如何更好地配合的关系，而绝不是一个取代另一个的关系。

§ 7.3 进一步的研究计划

一个研究课题总是针对一个或一些特定问题的。一方面，探索真理的路永远都没有尽头；

另一方面,在一个具体的研究课题范围内,对现有问题的解决通常总是有一定限度的。因此,在一个研究课题暂时告一段落,人们要思量下一步该如何去做的时候,也无非是在这两个方面做更多的努力,即一面结合更多的实践,对现有的框架进行检验并向纵深挖掘;一面在现有的研究成果基础上,探索如何开辟更广阔的研究空间。本课题的研究也不例外。在考虑进一步的研究计划时,主要也是这两个方面:

(一)如何对现有的短语句法语义范畴体系作进一步的精化细化,以使得形式化的短语结构规则系统的表达更有效率,更简洁。为此,我们主要是增加调试例句,通过计算机反馈的分析结果,不断改进现有的短语结构规则系统。

(二)如何以现有的对汉语短语结构组合规则的研究为依托,进一步尝试描述句子的结构规则,乃至篇章的结构规则。这首先要求我们从理论上对句子和短语的差别进行系统的研究。从已有的实践来看,用目前的短语结构规则的组织原则对句子进行分析,碰到的主要两个障碍是复句和由多个短语结构以标点分隔形成的所谓“流水句”。对这两类句子的处理,将是我们拓展研究空间的首选目标。

值得指出的是,对短语、句子、乃至篇章的结构规则进行研究,是个浩大的语言系统工程。工作量是很大的。《现代汉语语法信息词典》的作者曾经大致估算过开发一个5万词左右的现代汉语语法信息词典的工作量,是40个人年。以此为参照,建立一个大规模的实用表现出色的现代汉语短语结构规则库,以及最终建成现代汉语句子、篇章各个层面的分析规则库,总工作量是相当大的。对此要有充分的估计和准备,为此,也特别要求在开始阶段,理论准备和实验性的探索工作要尽量做的扎实。

总而言之,所谓解决一个旧问题,或许更应该看成是打开了一扇通向新问题的窗户。在某种程度上,本课题的研究工作仍然是在提出问题的阶段,离真正全面地解决中文信息处理中遇到的汉语语法分析问题这个最终目标,仍是“路漫漫其修远兮”。随着人们对语言本质的认识不断加深,以及计算机信息处理技术的飞速发展,我们将积极探索,期待能提出更好的解决方案。

附注:

- ¹ 可参见冯志伟(1992)《中文信息处理与汉语研究》,商务印书馆1992年版,P96。
- ² 参见徐通锵(1997)《语言论——语义型语言的结构原理和研究方法》,东北师范大学出版社1997年版。
- ³ 关于汉语树库构建的理论探讨与实践,可参见周强等(1997)《汉语树库的构建》,载《中文信息学报》,1997年第4期。
- ⁴ 关于受限汉语的研究,可参见俞士汶(1995)《关于受限的规则汉语的设想》,载王均主编《语文现代化》,山东教育出版社1995年版。张伟(1998)《受限汉语研究和受限汉语辅助写作系统的设计》,北京大学博士学位论文。
- ⁵ 参见俞士汶等(1994)《机器翻译译文质量评价的实践与分析》,发表在中文电脑国际会议ICCC'94(新加坡)论文集,P26-32。
- ⁶ 前文讨论具体的短语结构规则时,我们已经随文指出了《现代汉语语法信息词典》里一些有待改进之处,既包括属性设置方面的,也包括具体的词语属性信息方面的。
- ⁷ 这个例子参考了朱德熙(1982)《语法讲义》,P156。