现代汉语广义虚词知识库的建设*

俞士汶　朱学锋　刘云
北京大学计算语言学研究所
yusw@pku.edu.cn

摘要

在北京大学计算语言学研究所在已有的语言资源的基础上，笔者提出了建设面向信息处理的“广义虚词知识库”计划。本文具体探讨了“广义虚词”的所指，“广义虚词知识库”的主要内容及其建设的技术路线。

关键词

现代汉语　广义虚词　知识库　信息处理

1. 背景

北京大学计算语言学研究所自 1986 年成立以来，在语言信息处理基础资源建设方面已经取得了一些成果：

（1）现代汉语语法信息词典；
（2）大规模标注语料库；
（3）面向机器翻译的语义词典；
（4）面向信息检索与信息提取的中文概念词典；
（5）英汉对照的双语语料库；
（6）信息科学技术领域术语库；
（7）语言知识库建设的系列工具软件（汉语切分与词性标注软件、汉语文本自动注音软件、双语对齐软件、科技术语自动提取软件等）。

这些成果特别是（1）、（2）两项知识库已在语言信息处理学术界和产业界产生广泛影响。考虑到语言信息处理的发展，现在有必要在这些成果的基础上，开发“综合型语言知识库”。

为了高效率地建成一个能够为语言信息处理提供全方位、多层次支持的综合型语言知识库，首先要开发一个支撑软件，它将集成现有成果，形成一个整体：各个独立

*本文相关研究得到中国国家自然科学基金项目 69973005、973 项目G1998030507-4、863 项目 2001AA114040 的支持。

的知识库（如：语法词典、语义词典、语料库等）之间可以相互参照，不断消除瑕疵，提高质量；提供支持数据挖掘和知识发现的工具软件，促使现有的知识库从初级产品形式向深加工产品形式不断发展；提供多种形式的知识传播和信息服务机制，让综合型语言知识库在语言信息处理研究和传统语言学研究中发挥实际的作用。开发基于综合型语言知识库的应用系统，既可以使基础研究成果在社会生活中真正发挥作用，又可以检验知识库是否切合实际需求，检验质量到底如何。

综合型语言知识库当然应该是可以不断扩充、动态更新的，要能够永葆活力。为了向自然语言理解研究前进一步，需要在句法和词汇语义研究已经取得一些成果的基础上，开展句法语义的研究。本文提出了建设“广义虚词知识库”的计划。笔者相信建设中的“广义虚词知识库”是句法语义研究的基础之一，也是综合型语言知识库大家庭中的新成员。

2. “广义虚词”的所指

本计划所指的“广义虚词”的范围还没有最后圈定。目前的考虑如下。

“广义虚词”当然包括《现代汉语语法信息词典》（以下有时简称为《语法信息词典》）中的全部虚词：介词、连词、助词、语气词。

在《语法信息词典》中，副词划归实词，但朱德熙先生认为副词是虚词。本计划将副词包括在“广义虚词”中。

方位词（“上”、“下”、“里”、“中”等），有时单独起英语中的介词（prepositional）的作用，或者包含了日语中的格助词（postpositional）的功能。汉语中，有时介词要求与一个方位词配合形成一个框架，才能完成英语介词的功能。方位词也算作“广义虚词”。方位词和时间词、处所词一样，还有可以直接作状语的功能。这种功能相当于副词的功能。但是否据此将时间词、处所词（和方位词）也划归“广义虚词”，还没有最后决定。

量词：个。除夕之夜，喝了个痛快。

代词：他。吃他两大碗。

每。由“每”构成的“每年”、“每天”等对时态有明确的提示。

谁。需要进一步区分是“特指”（如：谁是王永和？王永和是谁？）和“泛指”（房间里好像有谁。）

本来代词同其他词类不一样，不是按照句法层面的功能划分出来的，有必要仔细研究，至少应该完成基于语法功能的细分类的工作。

动词：形式动词（像“进行”、“给予”、“加以”等），朱德熙先生曾将它们命名为“虚化动词”）、助动词（如“会”、“应该”、“可以”等，有些学者认为它们是副词）颇有资格进入“广义虚词”的行列。其他像补语动词（例如：“了 liao3”、“着 zhao2”、“看 kan4”即“试试看”中的“看”）、趋向动词（例如：“吃上两顿饱饭”中的“上”）

也可以划到“广义虚词”加以深入研究。

名词：日语中有形式体言的概念，是重要的语法现象。不过，词汇形式只有很少几种：“もの”、“こと”、“の”、“ところ”等。汉语中某些语境种的“东西 dong1xi5”颇像“もの”，“事情”、“行为”颇像“こと”，“地方”、“场所”颇像“ところ”。其他像“问题”、“现象”、“状态”等抽象名词也许具有“形式名词”的功能。例如：“研究语法问题”大致等同于“研究语法”，“问题”似乎可有可无，也可以算作“广义虚词”。

“广义虚词”应当是《现代汉语语法信息词典》所含词语的一个有限子集（在现代汉语中很可能是一个封闭的子集）。

“广义虚词”在《人民日报》标注语料库中的频度应该是很高的。其频度对时间应该是稳定的，其分布对题材和体裁应该是均匀的。这些认识都可以通过对《人民日报》标注语料库进行统计加以验证。

以下有时就用“虚词”指称“广义虚词”。

3. “广义虚词知识库”的主要内容

3.1 面向语言信息处理

虚词研究历来是汉语语法研究的重点，论著可谓浩如烟海。之所以重新选择这个课题，是为了满足语言信息处理技术发展的需要。目前，面向计算机分析和生成汉语的虚词研究几乎是空白，更没有一个可供语言信息处理系统使用的电子版的虚词知识库。《现代汉语语法信息词典》是汉语信息处理学界最有影响的一部囊括汉语词汇句法知识的电子词典，但它提供的虚词的句法知识（反映在各个虚词数据库的属性字段数目上），同实词相比，特别是同动词相比，要贫乏得多，基本上也没有涉及语义。北大建设中的《现代汉语语义词典》和《中文概念词典》基本上只考虑实词。其他的通用型汉语电子词典也都是研究实词的。研制“广义虚词知识库”显然是一项填补空白的工作。

既然一直无人问津，是否在语言信息处理中虚词并不重要。显然不是。吕叔湘先生在《现代汉语八百词》中指出了汉语语法的 4 个特点，其中第一条是“没有严格意义的形态变化”，那么只好使用虚词和语序来表达各种不同的语法关系；紧接着第二条又是“常常省略虚词”。可见，虚词在汉语语法研究中占有特殊重要的地位。不过，对于某些文本信息处理系统，像文献检索、文本分类、信息提取等，虚词确实常常被忽略掉（被看作 stop word），其原因还是由这些系统所采用的技术决定的，也许这是实用系统的合理抉择。当某些应用必须采用基于内容理解的分析技术或者希望生成自然通顺的汉语句子乃至篇章，虚词的作用就提到显要的位置上来了。下面举两个实例加以说明。

（1）虚词以及语序的作用

例句 1：这么一个工程，三年才完成。

是说该工程用的时间多。

例句 2：这么一个工程，才三年完成。

是说该工程用的时间少。

不用副词“才”，整个句子的意思不明确。“才”的位置不同，表达的意思又不一样。如何表示这样的知识，能够让计算机“记住”并会运用它们呢？

（2）虚词的省略

例句 3：分配部下一个任务。

例句 4：接受上司一个任务。

从两个句子的表层看，它们的词性序列完全一样；从词汇语义角度看，“分配”和“接受”同属一个上位语义范畴（授受关系），“部下”和“上司”也是如此。两个句子的相似度应该很大。不论采用基于规则的方法还是采用基于实例的方法，如果按照相同的句型把这两句话翻译成英语或日语，肯定有一句是错的。如果利用《现代汉语语法信息词典》，在动词库中，可以查到“分配”是双宾语动词，例句 3 可以变换为

例句 5：给部下分配一个任务。

“接受”不是双宾语动词，例句 4 应该变换为

例句 6：从上司接受一个任务。

就是说，给例句 3 加上介词“给”（相当于英语的 to，日语的に），给例句 4 加上介词“从”（相当于英语的 from，日语的から），并调整语序，变成例句 5 和例句 6，再让机器去翻译，困难就减少了。找出汉语虚词省略的规律（笔者以为即使汉语语法学界对这些规律的认识也还是模糊的，至少是不系统的），不仅对语言自动处理是极为重要的，对汉语教学（特别是作为第二语言的汉语教学）也十分有意义。笔者曾提出过面向信息处理的“受限汉语”的研究任务，如何规范地使用汉语的虚词，不随意省略应该使用的虚词也是“受限汉语”的重要研究内容之一。

还可以从另外一个角度来看待例句 4 和例句 3 同助词“的”关系。助词“的”是现代汉语的第一高频词。什么时候该用，什么时候可省，什么时候不能用，似乎是永远的话题。如果把例句 4 改为：

例句 7：接受上司的一个任务。

机器也好理解了。即可以认为例句 4 等价于例句 7，只是省略了一个助词“的”。但却不能认为表层同例句 4 相似的例句 3 也省略了一个助词“的”。“分配部下的一个任务”是不通的，至少不是例句 3 的原义。

理解这些问题对说汉语的人都不困难，从机器处理的角度看都成了难题。本计划一定针对机器自动处理的需要，把涉及广义虚词的知识系统化，按计算机可以利用的方式表示出来。

3.2 广义虚词知识库的内容概要

(1) 确认每一个广义虚词的功能语义项 ID 或广义虚词数据库中每个纪录的登录项 (Entries)。

尽管《现代汉语语法信息词典》经常强调其句法功能和义项相结合的收词原则，但其主要依据还是句法功能。经常举的一个例子便是“ 都 ”。《现代汉语语法信息词典》虽然承认副词 “ 都 ” 有 “ 总括全部、甚至、已经 ” 3 个不同的义项，但却粗略地认为 “ 对应于这些不同的义项，‘ 都 ’ 的语法功能并没有什么差别，因此语法词典只收入一个副词 ‘ 都 ’”。在研制《现代汉语语法信息词典》时，这样处理是必要的、恰当的。但在广义虚词知识库中，就需要把他们区分为 3 个不同的 ID 或 Entries 。

至于 ID 的具体表示法，还可以细致地设计。

助词 “ 的 ” 也是这样。《现代汉语语法信息词典》中只有一个 “ 的 ”，尽管助词库中有 “ 子类 ”（值为 “ 结构助词 ”）、“ 组合要求 ”（值为名词、动词、形容词、区别词等等）、“ 结构性质 ”（指 “ 的 ” 字结构为体词性的或谓词性的）、“ 前照应词 ”（值为空）等属性字段，但这些知识对计算机来说，还是太笼统，太空泛，不便应用。《现代汉语八百词》关于 “ 的 ” 的论述要详细得多，其功能语义大约有十几种区分。广义虚词库不仅要更详细，而且要系统化、条例化、规格化，对每一种区分都要赋予一个唯一的 ID。

(2) 对一个广义虚词的每一个 ID，都要建立它的判别条件。

只告诉机器副词 “ 都 ” 有 3 个不同的 ID，分别代表 “ 总括全部 ”、“ 甚至 ”、“ 也 ” 是不够的。要明确给出判断这 3 个 ID 的条件。条件通常就要到它们的使用语境中去找。像适当范围内的前共现词类、后共现词类、前共现义类、后共现义类、甚至具体的前共现词、后共现词都是可以加以描述的条件。在统计意义上，使用频度也是一个条件。仍以 “ 都 ” 为例，表示 “ 总括全部 ” 的 “ 都 ” 前面常有表示复数或泛指的名词或代词，后面的动词常是肯定式；表示 “ 甚至 ” 的 “ 都 ” 后面的动词常是否定式；表示 “ 也 ” 的 “ 都 ” 所在句子的句末常有语气词 “ 了 ”。还有，表示 “ 总括全部 ” 的 “ 都 ” 经常重读，而表示 “ 甚至 ” 的 “ 都 ” 经常轻读，这些信息对语音识别和语音合成也很有用。

这些条件也许可以从广义虚词的使用语境中直接得到，也许要进行一些推理。推理时，可能会用到其他知识库中的知识（例如《现代汉语语法信息词典》，现代汉语语义词典，中文概念词典等等），也可能会遇到知识匮乏知识。反过来，这又促进了其他知识库的发展和完善。

确定 ID 的另一个办法是使用句法分析中的变换分析法。介词 “ 给 ” 有不同的语义。靠用其他介词替换来区分。

例句 8：给妹妹气坏了。

例句 9：给妹妹卖毛衣。

例句 10：给妹妹寄毛衣。

例句 8 中的“ 给 ”可用“ 被 ”替换，例句 9 和例句 10 中的“ 给 ”不可用“ 被 ”替换，可用“ 为 ”替换。因此，例句 8 中的“ 给 ”和例句 9、例句 10 中的“ 给 ”有不同的 ID。至于例句 9 和例句 10 中的“ 给 ”是否一样呢？例句 10 可变换为“ 寄毛衣给妹妹 ”或 “ 把毛衣寄给妹妹 ”。例句 9 不可以这么变换，否则原义变了。因此，例句 9 和例句 10 中的“ 给 ”也是不同的。在把这些例句翻译成英语时，“ 给 ”的用法和意义的不同都会显性地表现出来（要分别使用不同的介词 by，for，to），因此英语译文也是一个重要的参考系。

（3）要研究冗余的虚词

口语中有些虚词是冗余的，不提供任何信息。指出这些冗余的虚词，会减少理解或翻译时的干扰。“ 腰给扭伤了 ”、“ 可不能让囚犯给跑了 ”、“ 猫把鱼给偷吃了 ”、“ 叫雨给淋了 ”中的“ 给 ”就是多余的。广义虚词知识库应当指出哪些语境中的哪些虚词是冗余的。

（4）要研究省略的虚词

《现代汉语八百词》对助词“ 的 ”的省略情况进行了相当详细的描述，但也还只是局限在短语的层次上。到了句子里，特别是在较长的句子里，情况还会有更多的变化。如果不能把握什么地方省略了虚词，对句子的理解也会造成困难。上面所举的例句 4 就属于这种情况。

还有，像“ 我的血 ”中的“ 的 ”，按照《现代汉语八百词》是要用的，但在“ 我以我血荐轩辕 ”中就没有“ 的 ”。有时，用不用一个“ 的 ”字，还会造成语法理论上的歧见。如“ 我的胃疼 ”只是简单的主谓句；而“ 我胃疼 ”就有不同的解释：一种意见认为省略了“ 的 ”，还是简单的主谓句，另一种意见认为是主谓谓语句。

其他助词的省略情况也很多。“ 苹果吃了 ”既可以解释为“ 苹果被吃了 ”，是过去时被动式陈述句，也可以解释为“ 把苹果吃了 ”，是祈使句。

广义虚词知识库也应当指出在哪些语境中可能省略哪些虚词。这个任务或许要比研究虚词的冗余还要困难。

（5）错例分析

对一些虚词，还要给出一些典型的错误的例子。这些错误可能是机器犯的，可能是将汉语作为第二语言学习的人犯的，还可能是母语是汉语的人犯的。要分清错者的类型。具体的错例要逐步积累，也可以利用已有的“ 中介语 ”语料库。

4. “ 广义虚词知识库 ”建设的技术路线

4.1 广度与深度的适当折衷

尽管广义虚词是一个有限集，但其研究内容可能是难以穷尽的。据说，助词“ 的 ”

可以细细分辨出 20 多种不用的用法和语义，论述介词“被”的文章有几百篇。为了在两三年内建成有使用价值的广义虚词知识库，对每一个虚词的研究则要把握好适当的广度和深度，否则一个虚词的内容也许就可以写成一篇论文呢。

目前可以用 2600 多万字的 1998 全年的《人民日报》标注语料库作为考察的范围。《人民日报》标注语料库不仅可以提供每个虚词的各种 ID 的用例，而且可以提供每个虚词的各种 ID 的使用频度。如果《人民日报》没有“吃他两大碗”这样的用法，在这次所建的虚词知识库中就可以不给“他”的这种用法赋予一个 ID。

如果除了词性，在文本中还要对虚词进一步标注，就应该标注这样不同的 ID。如果把研制广义虚词知识库的过程和对《人民日报》语料库中的虚词进行 ID 标注的过程结合起来，那么将能同时得到两个十分有意义的成果。

4.2 现有资源的充分利用

从标注语料库可以派生出一个重要成果，即以所选词语为关键词的相关句列（concordance），由此可以方便地罗列出所选词语的使用语境，进而可以提炼出相关属性信息（如前面列举的前共现词类、后共现词类、前共现义类、后共现义类、前共现词、后共现词以及原始文本潜在的切分歧义等等）。

除了 2600 多万字的标注语料库，北大计算语言所的其他资源，像《现代汉语语法信息词典》、现代汉语语义词典、中文概念词典等，都是可以利用的。广义虚词知识库应当做到与这些知识库的“无缝”对接。《现代汉语语法信息词典》的词语（选中的广义虚词）、拼音、词类、同形、释义、备注等字段的内容都可以拷贝过来。其中“词语”、“词类”、“同形”这 3 个字段联合作为《现代汉语语法信息词典》的主关键字（数据库中每个纪录的唯一标识），它们可以作为确定虚词 ID 的基础。“现代汉语语义词典”也包含广义虚词中的一部分，其“义项编码”和“word1”这两个字段和《现代汉语语法信息词典》主关键字联合又成为确定虚词 ID 的新的基础。“中文概念词典”用同义词集合（synset）表示概念。如果某个广义虚词落在某个同义词集合中，那么这个概念及相关信息（上位-下位、同义-反义、部分-整体等）也是该广义虚词的知识的来源。

这样的广义虚词知识库与《现代汉语语法信息词典》、现代汉语语义词典、中文概念词典、标注语料库是相互补充的，它们都是综合型语言知识库的有机组成部分。

4.3 辨异与求同的研究方法

“求同”与“辨异”是科学研究的基本方法。《现代汉语语法信息词典》的“分类”与“属性描述”相结合的词典框架就是在这样的方法论指导下设计的。广义虚词知识库的建设也要在这样的方法指导下进行。

单纯方位词“上”、“中”、“下”、“里”、“外”同其他合成方位词有很多不同之处，

它们彼此间又有很多共性，这些在《现代汉语语法信息词典》中已得到描述。但《现代汉语语法信息词典》则没有更细致地区分这几个方位词之间的差异。“ 上 ” 可以接在一些抽象名词后面，表示 “ 方面 ” 或 “ 范围 ”，例如 “ 思想上 ”、“ 理论上 ”、“ 感情上 ” 等等，“ 中 ”、“ 下 ”、“ 里 ”、“ 外 ” 就没有这种用法。“ 里 ” 可以接在一些行政单位后面，指称会话双方共知的或上文已交待了的某个具体的行政机关或其领导班子，例如 “ 部里 ”、“ 局里 ”、“ 省里 ”、“ 系里 ” 等等，“ 上 ”、“ 中 ”、“ 下 ”、“ 外 ” 也没有这种用法。

4.4 专家知识的主导作用

北大计算语言所以往在建设各种词典、语料库时充分运用了辅助工具软件并取得了实际成效。很难想象，如果不借助软件技术，一个小小的研究所能建立起如此庞大的语言知识库。不过在人与技术的相互关系中，北大计算语言所更强调人的因素，重视专家在知识库建设中的主导作用和质量保证作用。由于虚词的特性，它所起的句法语义作用既重要，又精微，不充分利用专家多年研究所积累的成果，或者主要寄希望于机器学习，自动发现虚词的各种句法语义特性，可能反而会欲速则不达。

对专家知识的借重将体现在以下几方面：

（1）认真学习吕叔湘、朱德熙等大师的著作，充分利用语言学家在词典、语言学论著中已经发现、整理并表述清楚了的知识。

（2）吸引年轻的语言学家直接投身这项语言工程，并鼓励年轻的语言学家调整自身的知识结构，能够领悟计算机的能力与工作模式，以便与软件专家配合，实现人与机器的恰当分工，充分发挥人与机器的各自优势。

（3）争取陆俭明、王逢鑫、冯志伟等语言学家的现场指导，加强广泛的学术交流。

4.5 数据结构的灵活设计

北大计算语言所已有的多种知识库各自采用的数据结构形式是不一样的，有的采用关系数据库的二维表，有的采用树结构，有的采用 XML 标记的文本，有的采用自定义格式的文本。广义虚词知识库该采用什么样的知识表达形式既要适应虚词本身的特点、便于机器运用，还要继承以往本所开发电子词典的传统，知识表达形式要让语言学家也容易理解，便于表述自己的知识和了解知识库中已有的内容。

5. 结语

广义虚词知识库尚在筹划之中。由于这项工程是基础性的，规模也不会小，寻求支持不容易。不过，北大计算语言所从长期的基础研究实践中体会到，真正有利于应用系统研究和开发的基础研究成果还是能得到业界认同的。目前最需要做的一件事是

向广大专家征询意见。

十分感谢徐杰博士。由于徐杰博士的邀请，俞士汶得以出席“第二届肯特岗国际汉语语言学圆桌会议”。除了聆听各位专家的报告、吸取知识外，俞士汶也得到了一个极好的机会，向与会专家咨询，这件事有没有必要做？会不会做重复劳动？会不会对困难估计不足？在计划启动之前，明确认识这些问题，是十分重要的。

参考文献

陈群秀，现代汉语述语动词机器词典的扩充和槽关系研究，《语言文字应用》，2001年，第4期，98-104

常宝宝，基于汉英双语语料库的翻译等价单位的自动获取研究，《术语标准化与信息技术》，2002年，第2期，24-29

董振东、董强，面向信息处理的词汇语义研究中的若干问题，《语言文字应用》，2001年，第3期，27-33

吕叔湘主编，《现代汉语八百词》，北京：商务印书馆，1984

王惠、詹卫东、刘群，《现代汉语语义词典》的概要及设计，《1998中文信息处理国际会议论文集》，清华大学出版社，1998年

于江生、俞士汶，CCD的结构与设计思想，《中文信息学报》，2002年，第16卷第4期，12-20

俞士汶，关于受限的规则汉语的设想，王均主编《语文现代化论丛》，山东教育出版社，1995年10月，193-205

俞士汶、朱学锋，受限汉语研究的必要性，王均主编《语文现代化论丛第三辑》，语文出版社，1997年10月，150-160

俞士汶、朱学锋、王惠、张芸芸 著，《现代汉语语法信息词典详解》，北京：清华大学出版社，1998年4月

俞士汶、段慧明、朱学锋、孙斌，北京大学现代汉语语料库基本加工规范，《中文信息学报》，2002年第5，6期

朱德熙，《语法讲义》，北京：商务印书馆，1981

The Development of Knowledge-base of Generalized Functional Words of Contemporary Chinese

YU Shiwen ZHU Xuefeng LIU Yun
Institute of Computational Linguistics, Peking University, 100871
yusw@pku.edu.cn

Abstract

Based on the existing language resources of the Institute of Computational Linguistics of Peking University，authors propose a plan of developing the Knowledge-base of the Generalized Functional Words of Contemporary Chinese oriented to information processing. This paper concretely discusses the scope of "Generalized Functional words", and also discusses the main content of this knowledge-base and its technical policy.

Keyword

Generalized Functional Words Knowledge-base Information Processing