《英－汉计算语言学术语数据库》——中德合作项目的研制报告摘要

龚彦如　李竹　冯志伟
（国家语委　语用所）

计算语言学是一门新兴的边缘学科，它的兴起和发展推动了计算机科学和语言学的进一步结合。随着计算语言学的发展及其术语研究的不断深入，越来越多的问题出现在我们的面前，急待解决。例如，计算语言学究竟有那些术语？它们处于一个怎样的分类体系之中？它们在各种不同的自然语言中，应该翻译成为怎样的术语？这些译名又如何准确地、科学地反映事物的本身？为此，国家语言文字工作委员会语言文字应用研究所与德国 Trier 大学汉学系、英语系联合研制了《英－汉计算语言学术语数据库》。它是在原先已建成的三个数据库（(1)语用所的《应用语言学术语数据库》；(2)北京大学俞士汶教授等研制的《多语言计算语言学词汇》；(3) Trier 大学的《计算机词汇数据库》）的基础上，以大量的计算语言学科技文献为依据研制的。它使用 dbase Ⅲ 数据库语言为工具，建立在微型机上，它收录术语9471条，库中每条术语是一个记录，每个记录包括八个数据项：英文名、英文结构、汉文名、汉语拼音、汉文结构、注释、分类号和出处。

计算语言学涉及众多学科，它的内涵与外延并不十分清晰，其术语也还没有一个公认的、明确的分类体系。大量的术语无法简单地线性归类，它们往往由于不同的来源和不同的应用而存在于相异的学科领域中。例如，在我们的库中术语“论元匹配(argument matching)”是在机器翻译中用来检验生成的句子成分是否在逻辑上匹配的，但同时，它也是数理逻辑中的常见术语。因此，我们提出了一种分类设想，即用一个三维空间来表示计算语言学的术语分类体系。这个三维空间分别由基础理论领域、应用领域和相关学科领域作为三个坐标轴。在这个空间中“论元匹配”的坐标为(7,2,5)，同时这个坐标也是此术语的分类号，表示出它在分类体系中的位置。

计算语言学术语存在一个英文术语对应几个汉语译名；或多个英文术语只由一个汉语术语表示的情况；有些汉译术语意义比较含糊，歧义现象时有发生。如何使翻译方法更加合理，更具排歧性；如何使汉译术语既能准确表达原义，又符合术语学原则，符合汉语的规律性；如何从汉语的构词特征上着手避免译名的歧义性，从而达到准确性、科学性、简洁性和约定俗成的要求，是我们工作的重点之一。在我们的翻译过程中，常常遇到名动同形词 NVP 形式的术语，这些术语既可能是体词性的术语，也可能是谓词性的术语。要解决这些歧义现象，可以采用下面几种方法：

1、通过添加助词“的”和一些词缀“性”、“型”、“式”等，来避免歧义的产生。例如：augmented dependency grammar 可以翻译为“扩充型从属关系语法”。

2、在词组型术语中，其中的动词可通过选择来避免歧义。如果术语是体词性的，可选择不及物动词，而不及物动词是不能带宾语的；如果术语是谓性的，就可以选择及物动词。

《英－汉计算语言学术语数据库》是采用单标记的二叉树来表示汉译术语的结构的。例如：“论元匹配”的结构表达式为：NVP(N/NV)。而这种结构表达式本身就隐含歧义性，NVP 包含了体词性和谓词性两种可能。当具体的汉语词汇出现时，其意义更加具体化，可能部分歧义被消除；但也可能歧义仍然存在。因为，现代汉语短语是分层次、非线性、多组合结构，不同的组合形式产生不同的语法、语义内涵，从而发生歧义；同样的组合形式也同样可能存在不同的语法、语义解释。要想圆满地解决歧义结构问题，应从逻辑语义关系入手，判断其施受关系的搭配是否得当？各成分在意义上紧密程度如何？只有多种方法综合使用，才可能产生良好的效果。