《基于动态流通语料库的汉语熟语单位研究》

书籍信息

版 次：1
页 数：
字 数：
印刷时间：2009年06月01日
开 本：16开
纸 张：胶版纸
包 装：平装
是否套装：否
国际标准书号ISBN：9787561923108

内容简介

本书基于汉语词典学界和中文信息处理界重“ 词 ”轻“ 语 ”的现象，对词组研究、熟语研究进行了反思，提出“ 熟语单位 ”（Idiom Unit，IU）的概念。所谓“ 熟语单位 ”，就是“ 结合紧密，使用稳定 ”、功能相当于熟语的结构单位，是符合人的认知规律并被人们经常当做一个词来使用的定型化了的固定短语或凝固表达式。我们判别IU的三条原则为：是否“ 结合紧密，使用稳定 ”；是否符合人们的认知规律（IU的长度一般为7±2）；流通度是否达到一定的阈值。IU理论上包括一切具有熟语性的词语组合单位。本书讨论的IU范围包括三字格中的惯用语和像“ 差不多、靠不住、来不及 ”这样介于词和短语之间的结构串，四字格中的成语和新固定短语，简称略语、插入语和字母词语等。
本研究选用的是《人民日报》2001-2003年三年的文本，约8000万字。文章以动态语言知识更新理论为指导，以流通度理论为基础，以年平均流通度阈值（0.5）作为主要筛选依据，运用规则和统计相结合的方法对“ 熟语单位 ”（IU）进行了初步的提取研究，并对部分提取结果的噪声环境作了定量与定性分析。
对于IU的提取，我们采取的策略和基本步骤是：
（1）利用点号和“ 的、是、在、和、了、有 ”等高频词（字）将文本化短；自动提取时牺牲包含切分点的字符串，该部分字符串另行补救。

（2）数
据格式转换。将
切分得到的形式上“ 完整 ”的2-
8字串转为数据库格式。
本书基于汉语词典学界和中文信息处理界重“ 词 ”轻“ 语 ”的现象，对词组研究、熟语研究进行了反思，提出“ 熟语单位 ”（Idiom Unit，IU）的概念。所谓“ 熟语单位 ”，就是“ 结合紧密，使用稳定 ”、功能相当于熟语的结构单位，是符合人的认知规律并被人们经常当做一个词来使用的定型化了的固定短语或凝固表达式。我们判别IU的三条原则为：是否“ 结合紧密，使用稳定 ”；是否符合人们的认知规律（IU的长度一般为7±2）；流通度是否达到一定的阈值。IU理论上包括一切具有熟语性的词语组合单位。本书讨论的IU范围包括三字格中的惯用语和像“ 差不多、靠不住、来不及 ”这样介于词和短语之间的结构串，四字格中的成语和新固定短语，简称略语、插入语和字母词语等。
本研究选用的是《人民日报》2001-2003年三年的文本，约8000万字。文章以动态语言知识更新理论为指导，以流通度理论为基础，以年平均流通度阈值（0.5）作为主要筛选依据，运用规则和统计相结合的方法对“ 熟语单位 ”（IU）进行了初步的提取研究，并对部分提取结果的噪声环境作了定量与定性分析。
对于IU的提取，我们采取的策略和基本步骤是： （1）利用点号和“ 的、是、在、和、了、有 ”等高频词（字）将文本化短；自动提取时牺牲包含切分点的字符串，该部分字符串另行补救。
（2）数据格式转换。将切分得到的形式上“ 完整 ”的2-8字串转为数据库格式。
（3）统计3-5字串的频度、散布度和流通度。
（4）用字符串全年的平均流通度阈值进行筛选。 （5）对五音节（含）以上字串进行分词并加以词性标注，对其中的3字串、4字串和符合“ N+N ”、“ N+V ”、“ V+N ”、

“ V+V ”等语法组合规则的相邻字符串（二元组）进行抽取；再对抽取的字符串重复上面的第（3）和第（4）步。 （6）对筛选得到的字符串进行噪声剔除，全部进行重新切分并加以词性标注，然后运用静态规则模板（共30条规则）再次过滤。
（7）借助辅助手段对熟语单位进行直接抽取。
（8）得到三至五字格熟语单位表（约13500条）。 本书还对提取出来的2001年的5500个三字格、2002年的6500个四字格作了简单的分类和例示性的分析说明，重点考察了具有熟语性的短语。 三字格中我们重点探讨了音节为“ 1+2 ”式、结构为“ V+N/NP ”式和音节为“ 2+1 ”式、结构为“ V/VP+N ”式的两类，验证了冯胜利有关三音节组合的论断：音节为“ 1+2 ”式的是短语，音节为“ 2+1 ”式的是韵律词。 四字格中我们重点探讨了“ N+V ”式和“ V+N ”式。N和V之间存在复杂的语法、语义以及音节制约关系。关于“ N+V ”式，通过考察，我们发现：定中关系的“ N+V ”式四字格熟语性最强，数量也最多；状中关系次之，主谓关系的四字格熟语性最弱，且N与V之间存在离散性。关于“ V+N ”式，我们发现：第一，“ V+N ”式四字格如果表示通名，它往往是或者容易成为一个NP习惯性搭配。第二，“ V+N ”式四字格中的N如果是比较抽象的双音节名词，则这类四字格构成的NP其熟语性相对较强。第三，“ V+N ”式四字格中的V如果是双音节述宾式动词，那么这种“ 双音节述宾式动词+宾语 ”形成的NP熟语性很强。 本书还从应用的角度对流行语、字母词语和插入语进行了考察研究，对流行语的科学评定和字母词语的规范发表了意见。 本书对简称略语的研究主要以《现代汉语词典》（2002年增补本）所收的134个简称和报纸语料中的约350个简称为考察对象，将简称分为固定简称和临时简称两种，少数临时简称随着使用次数的增加、使用范围的扩宽，可以成为固定简称。我们对两种简称的构成及固定简称的成因进行了初步探讨，重点考察了简称在真实文本中的使用情况。 本书主要有以下三方面的创新： （1）依据熟语性定义了“ 熟语单位 ”（IU）。IU是基于大众语感的认知结构单位，它使得固定短语的范围适当扩大，更加有利于中文信息处理、语言教学和汉外翻译等。 （2）第一次基于动态流通语料库（DCC），从大规模真实文本中提取通用的报纸固定短语，而且是采用相似于公众共同语感的流通度来由计算机自动提取。 （3）提出按照文体集合对应语体原则构建报纸分类语料库的短语提取策略，减少系统处理开销，提高短语识别的召回率（recall rate）和准确率（precision rate）。
显示全部信息

作者简介

杨建国，北京语言大学首都国际文化研究基地副研究员，硕士研究生导师。主要研究方向为语言学及应用语言学、汉语文化教育等，已发表语言、文化及教育类论文30余篇。曾参与编写《四库大辞典》《中国传统文化》等工具书及教材。

目 录