

<<基于动态流通语料库的汉语熟语单>>

图书基本信息

书名：<<基于动态流通语料库的汉语熟语单位研究>>

13位ISBN编号：9787561923108

10位ISBN编号：7561923104

出版时间：2009-6

出版时间：北京语言大学出版社

作者：杨建国 著

页数：270

字数：309000

版权说明：本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问：<http://www.tushu007.com>

## 前言

《基于动态流通语料库的汉语熟语单位研究》的初稿是杨建国博士的学位论文，完成于2005年6月。2008年底，书稿经过校外匿名审稿专家的评审和北京语言大学学术委员会评定，获得了“北京语言大学青年学者文库出版基金”的资助，将于2009年6月出版。

校外匿名审稿专家认为“基于大型语料库针对汉语熟语的研究较为少见”，而本书“紧跟国际语言学界研究热点，基于动态流通语料库，结合计算语言学方法和认知语言学理论探讨汉语中的熟语问题，使得研究具有较大的理论意义和实践意义”，并且文章“纵横捭阖，涉及流行语、字母词、简称及x字格等多种语言言语现象，其中各论述多有精彩之笔”。不乏美言，但非溢美之词。

建国本人在《后记》中说：“这是我的第一本专著，由于稚嫩，又偏重于应用，‘专著’之前不敢冠以‘学术’二字，唯恐玷污了‘学术’的尊严和高雅。

因此私下里将自己的处女作勉强定性为‘研究’专著。

”他的自我评价，体现了传统的严谨和谦逊，尽管一些措辞如“稚嫩、不敢、唯恐、勉强”等等，又似乎过于谦恭或惶恐。

他是我的博士研究生，他要我为此书做一篇序。

作为他的导师，我的这篇序该怎么措辞呢？

## &lt;&lt;基于动态流通语料库的汉语熟语单&gt;&gt;

## 内容概要

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 $7 \pm 2$ ）；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、语义以及音节约束关系。

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、

## <<基于动态流通语料库的汉语熟语单>>

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）。

#### 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。

## 书籍目录

摘要Abstract第一章 引论 1.1 本研究提出的背景 1.2 本研究的目标 1.3 本研究的意义 1.4 本研究的创新点和难点 1.5 小结第二章 汉语熟语单位 2.1 熟语单位的界定 2.2 熟语单位的判定原则 2.3 熟语单位的范围 2.4 熟语单位的判定方法 2.5 小结第三章 基于2001~2003年《人民日报》的汉语熟语单位提取研究 3.1 语料的选取 3.2 语料库及语料库语言学 3.3 词语自动提取研究的历史和现状 3.4 我们对中文信息处理及汉语的认识 3.5 提取熟语单位的方法和技术路线 3.6 辅助提取手段分析 3.7 部分结果验证及相关分析 3.8 小结第四章 三字格熟语单位研究 4.1 已有的研究 4.2 三字格概况 4.3 音节为“1+2”式、结构为“V+N/NP”式的三字格 4.4 音节为“2+1”式、结构为“V/VP+N”式的三字格 4.5 小结第五章 四字格熟语单位研究 5.1 已有的研究 5.2 四字格概况 5.3 “N+V”式的四字格 5.4 “V+N”/“V+V”式的四字格 5.5 小结 5.6 附论五字格第六章 流行语研究 6.1 引言 6.2 关于“流行”的界定 6.3 关于流行语的语言学研究 6.4 流行语的科学认定 6.5 余论第七章 字母词语研究 7.1 引言 7.2 基于词典的字母词语的分类及相关分析 7.3 基于报纸语料库的字母词语的使用情况举隅 7.4 关于字母词语规范的两点思考 7.5 附论插入语第八章 简称考察研究 8.1 引言 8.2 简称的界定 8.3 简称的分类 8.4 固定简称 8.5 临时简称 8.6 通过形式标记提取的简称例示 8.7 小结第九章 结语——兼论熟语单位的应用价值 9.1 本书的研究方法 9.2 熟语单位的应用价值 9.3 存在的问题与下一步工作附录1 两本新词语词典所收的部分新词语比较附录2 基于大学生的词语语感调查表附录3 从2001~2003年《人民日报》中切出的部分2字串附录4 从2001~2003年《人民日报》中切出的部分3字串附录5 从2001~2003年《人民日报》中切出的部分4字串附录6 从2001~2003年《人民日报》中切出的部分5字串附录7 从2001~2003年《人民日报》中切出的部分6字串附录8 从2001~2003年《人民日报》中切出的部分7字串附录9 从2001~2003年《人民日报》中切出的部分8字串附录10 2001~2003年《人民日报》的部分三字格熟语单位附录11 2001~2003年《人民日报》的部分四字格熟语单位附录12 2001~2003年《人民日报》的部分五字格熟语单位附录13 2001~2003年《人民日报》中相同的部分熟语单位附录14 2001年《人民日报》的部分引号抽取串附录15 2002年《人民日报》的部分引号抽取串附录16 2003年《人民日报》的部分引号抽取串附录17 2001~2003年《人民日报》中相同的部分引号抽取串附录18 1998年1月《人民日报》的部分“V+V”实例附录19 《现代汉语词典》(2002年增补本)收录的简称词条附录20 2002年《人民日报》中的部分简称附录21 本书所使用的标记集参考文献后记

## 章节摘录

如果我们不否认熟语也应包括俗语（俗语）、谚语、格言、名联、名句、流行语等，那么，熟语就应该是固定短语的上位概念。

我们知道，短语是词和词的语法组合，它包括自由短语和固定短语两类。

自由短语是出于表达需要形成的词跟词的临时组合（如“看报、撰写论文”等）。

而固定短语是词跟词的固定组合，一般不能任意增减、改换其中词语。

语法学研究关注更多的是自由短语，即一般的短语；词汇学和词典学研究关注的是固定短语。

我们基于中文信息处理、对外汉语词汇教学和汉外翻译等的需要，关注的也是固定短语，是新产生的固定短语。

虽然我们也关注句子层面的熟语，但其使用跟短语层面的熟语比起来，要少得多。

短语层面的熟语，结构上比较固定，功能上相当于一个词；是人们常用的定型化了的固定短语，是一种特殊的词汇单位。

由于它们为人们所常见常用，因此又被称做习用语。

习用语的特点，简言之，就是言简意赅，其中大部分源远流长，富于表现力。

目前汉语词汇学界除了对歇后语和成语的认识比较一致以外，对惯用语、俗语、谚语、缩略语等的认识都存在较大的分歧。

对于歇后语，几乎没有什么不同的意见，都承认它是由近似于谜面、谜底的两部分组成的带有隐语性质的口头用语。

对成语达成的基本认识是，成语以四字格为主，具有书面语色彩浓厚和相沿习用两个特点；争论主要集中在所谓的新成语和四字格以外的成语上，如“多快好省、莫须有、百闻不如一见”等。

对惯用语达成的基本认识是，惯用语以三字格为主，具有意义的整体性、结构的凝固性和明显的口语色彩等特点；争论的焦点主要集中在惯用语是不是具有表意的双层性上，极端的意见认为惯用语应该是“原‘惯用语’中除去三字格的部分”，“是非三字格的一部分俗语”（周荐，1998）。

对于俗语和谚语，一般的教科书不加论述。

因为汉语里的俗语和谚语在内容上有一些共同点，而且都具有明显的口语色彩，所以往往俗谚并称。

崔希亮（1997）从形式上给二者作了区分：俗语都是单体结构，而谚语都是复体结构，其中大部分是双体结构。

由于近两三年有人（王立，2003；曹炜，2004）已经对汉语词汇的研究成果作了较为清晰的梳理，我们在这里再复述一遍，意义不大。

因此，我们除了表明自己不同于前贤观点之处。

编辑推荐

《基于动态流通语料库的汉语熟语单位研究》是作者在其博士论文的基础上修改而成的。

《基于动态流通语料库的汉语熟语单位研究》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，提出了“熟语单位”（IdiomUnit, IU）的概念。基于《人民日报》2001~2003年三年约8000万字的文本，研究以动态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）进行初步的提取；并对部分提取结果的噪声环境作了定量与定性分析。在此基础上，对提取出来的三字格、四字格熟语，流行语，字母词语和插入语，简称略语等分别进行了重点考察。

本研究将汉语的难题研究和应用研究相结合，为熟语单位的分析研究提供了丰富的材料和可资借鉴的方法，具有较大的学术创新。



版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:<http://www.tushu007.com>