关于德语背景汉语学习者语料库的构建

张宝林
北京语言大学语言科学院
baolin08@126.com

提　要

* 1．汉语学习者语料库的作用与价值
* 2．外国人汉语习得研究现状
* 3．德语背景学习者语料库总体设计
* 4．潜在的问题与对策
* 5．结语

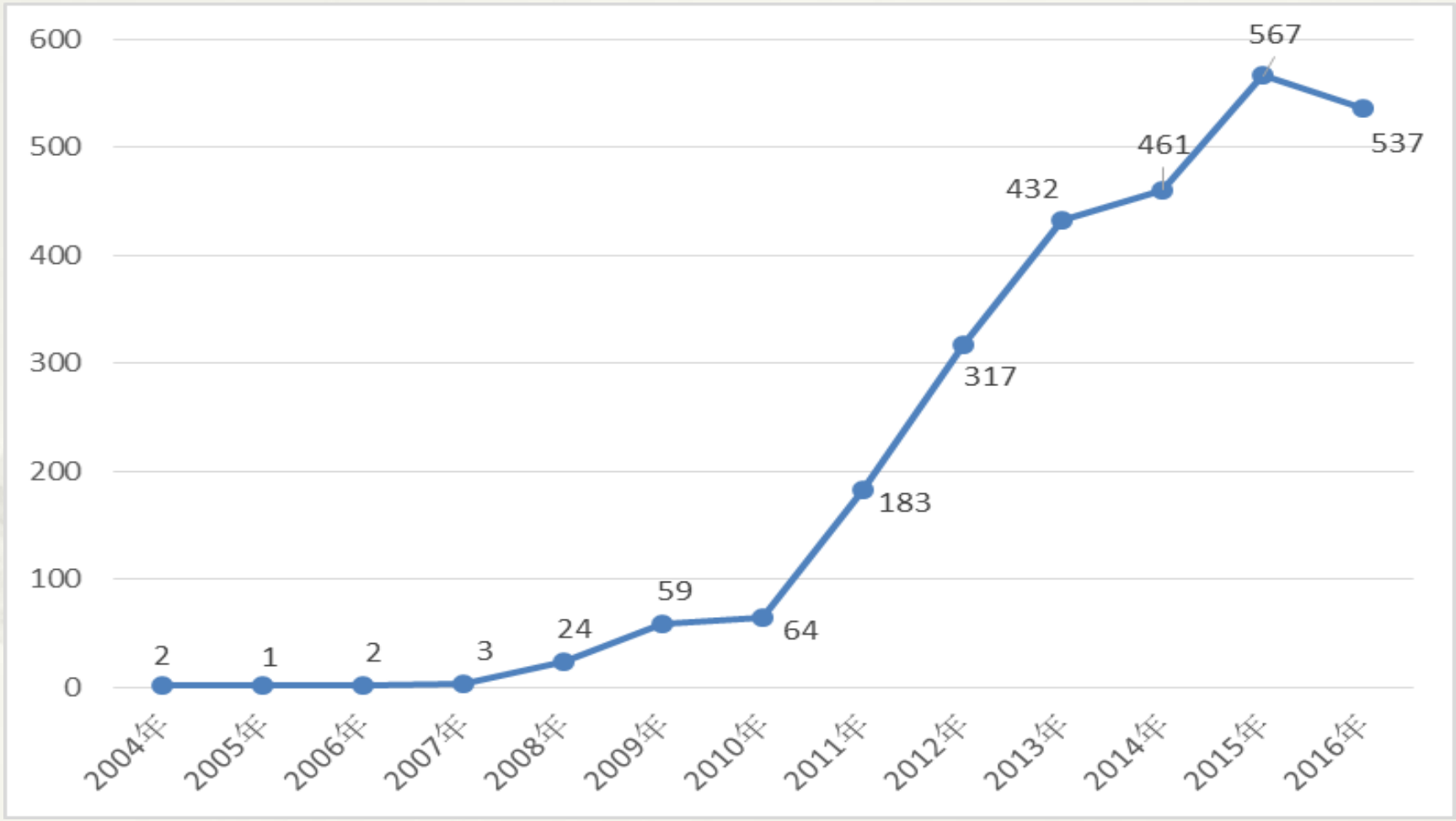
1．语料库的作用与价值

* 自1995年北京语言学院建成第一个汉语中介语语料库，20余年来，汉语中介语语料库从无到有，从小到大，从少到多，其建设与应用研究得到迅速发展，取得了众多且非常重要的研究成果。

1．语料库的作用与价值

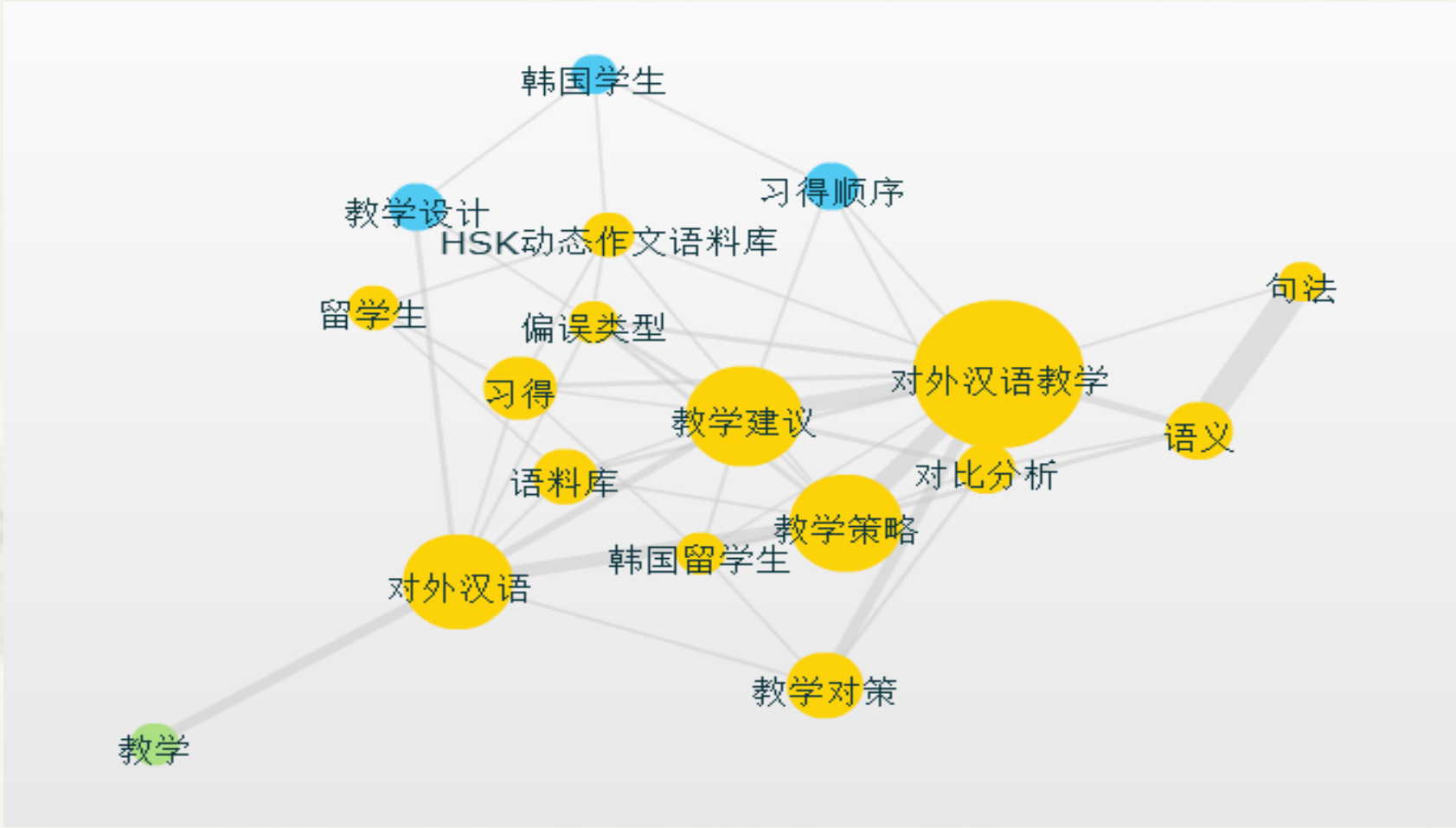
* HSK动态作文语料库：
* 2006年12月24日建成上线。
* 在CNKI中查询各类论文：2696篇（截至2017年6月30日）。

1．语料库的作用与价值



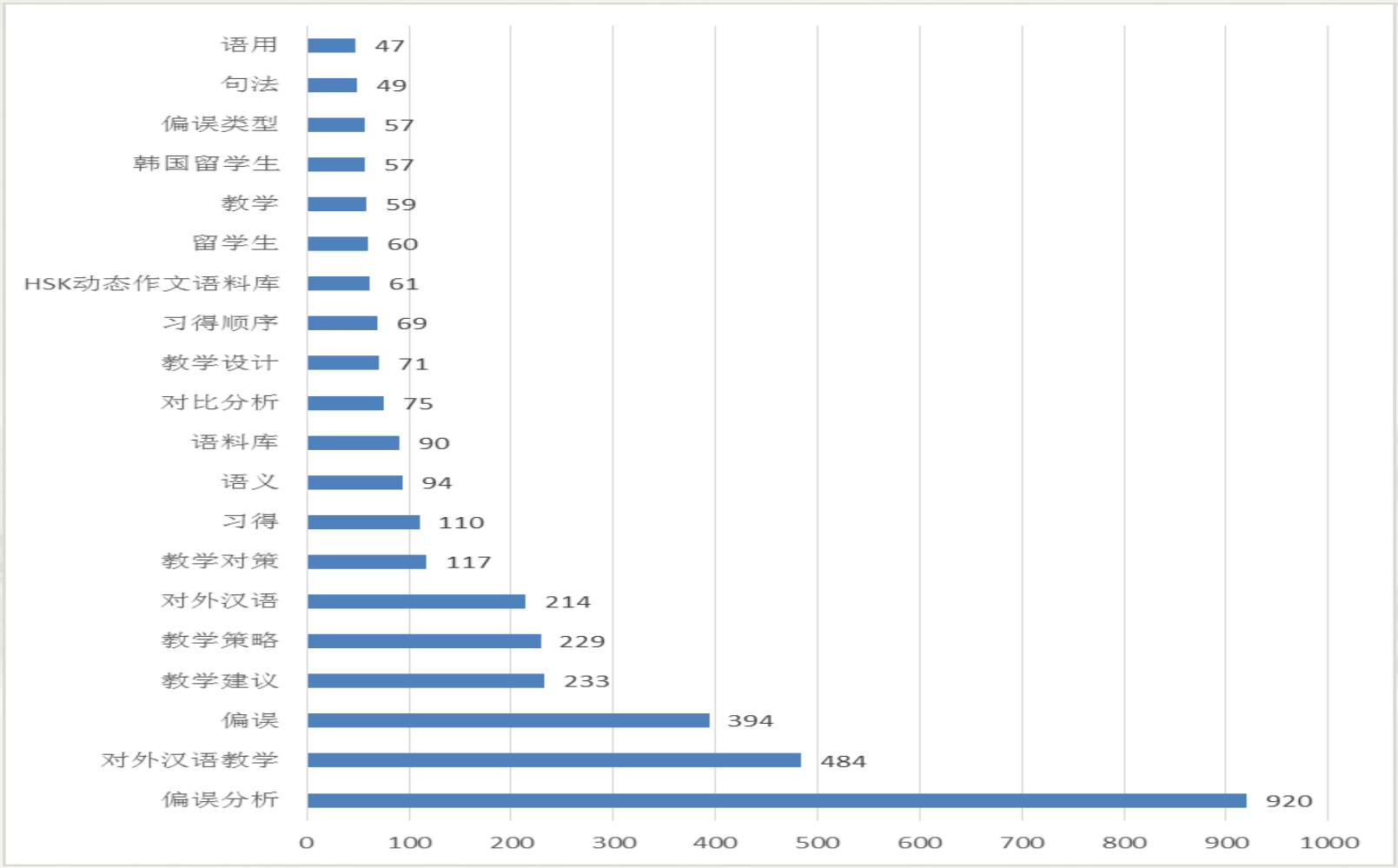
* 图 年度发文数量统计分析图（单位：篇）

1．语料库的作用与价值



* 图2 关键词共现网络

1．语料库的作用与价值



* 图3 关键词分布频次统计

1．语料库的作用与价值

* “目前，汉语中介语语料库(数据库)建设主要着眼于两大方面：一是服务于对外汉语教学学科建设和理论研究；二是适应和满足教学实践的实际需要。”“它已成为汉语教师和汉语教学工作者开展教学、科研的基本方法和手段。”（郑艳群，2013）

1．语料库的作用与价值

* 促进习得研究范式的转变：

  * 小规模、经验型、思辨性研究

* 基于大规模真实语料的、定量分析与定性分析相结合的实证性研究

1．语料库的作用与价值

* 基于语料库的偏误分析与习得研究，极大地促进了汉语中介语语料库的建设。

* 1 → 4 → “遍地开花”

2．习得研究现状

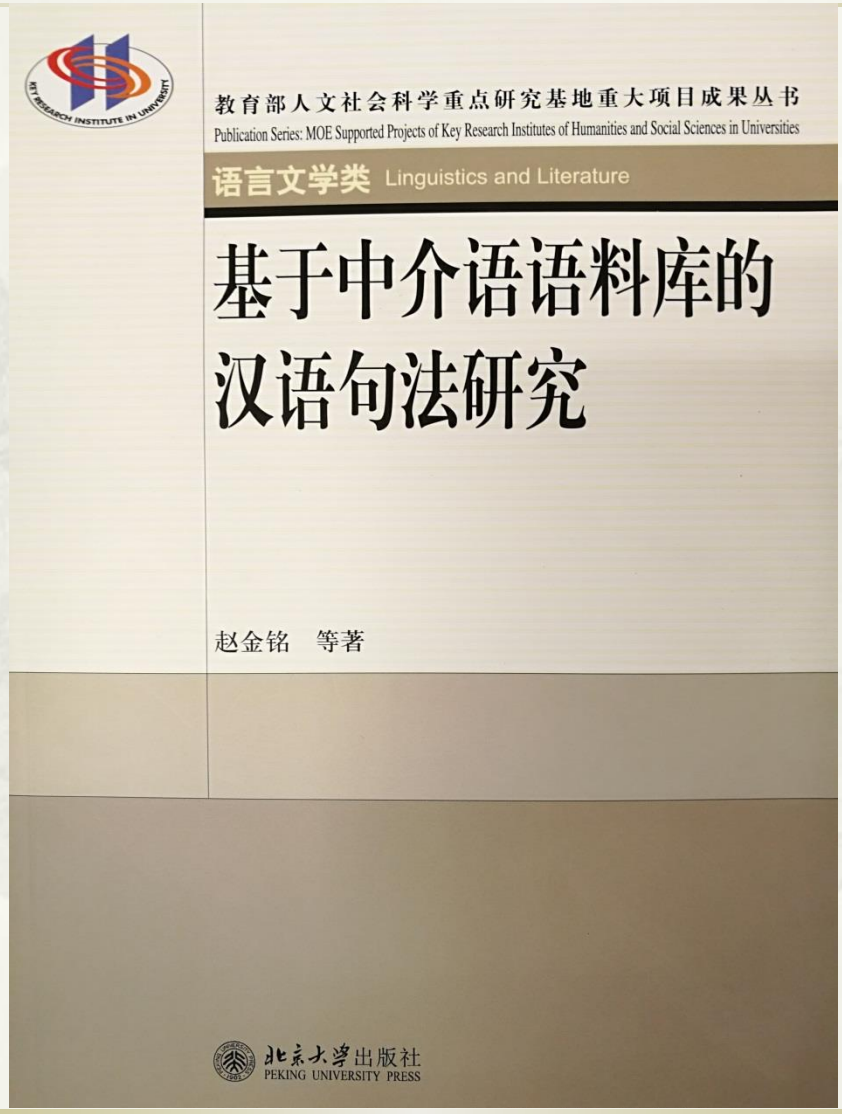
- 2.1研究类型
- 2.1.1基于“HSK动态作文语料库”的研究
- 海内外注册用户：40599人（截至2016年7月27日）；
- 在CNKI中查询各类论文：2311 篇（截至2016年7月28日）。

2．习得研究现状

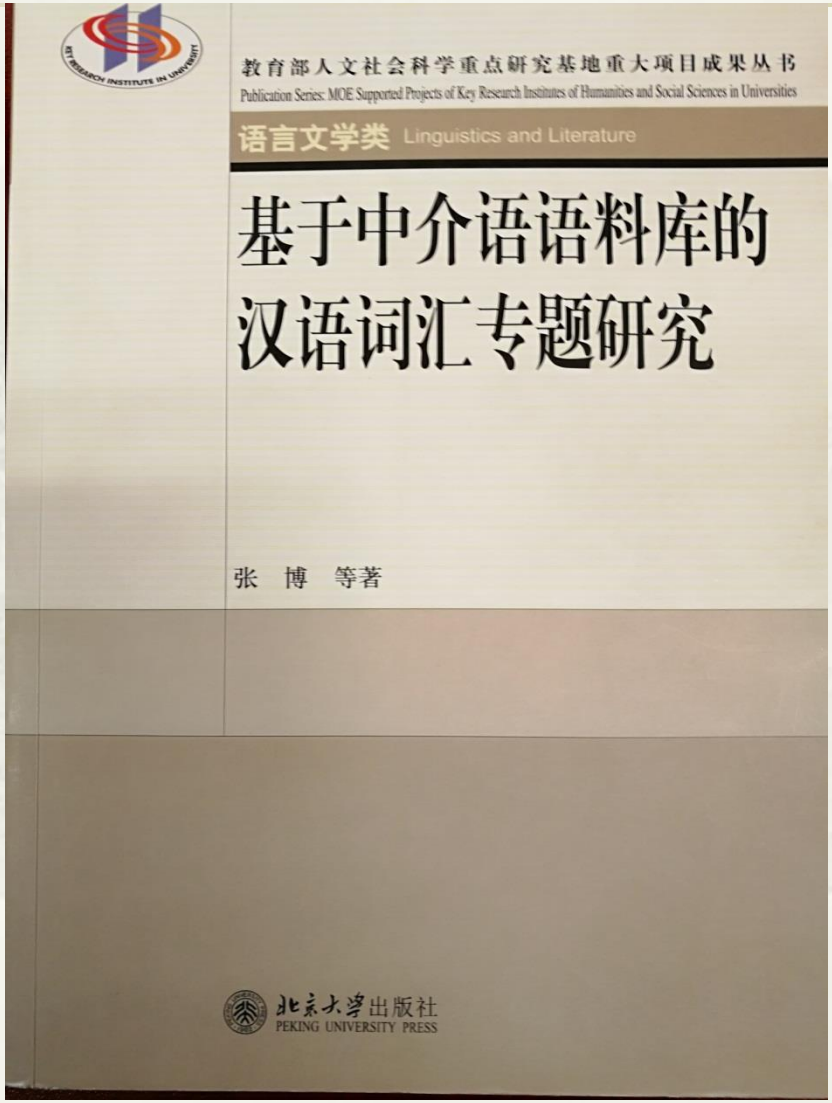
* 2.1.2代表性成果

* 1）专著
* 赵金铭：汉语句法研究（2008）
* 张　博：汉语词汇专题研究（2008）
* 肖奚强：汉语句式学习难度及分级排序研究（2009）
* 张宝林：汉语句式习得研究（2014）

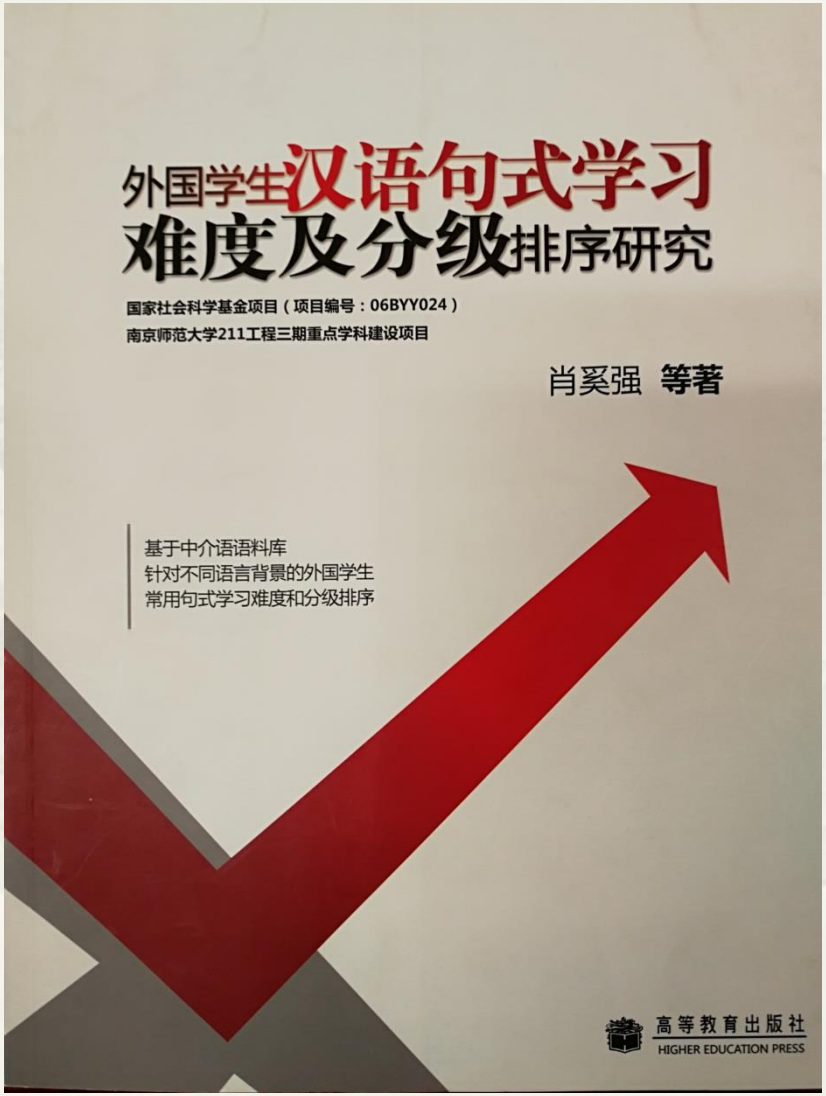
2．习得研究现状


教育部人文社会科学重点研究基地重大项目成果丛书
Publication Series: MOE Supported Projects of Key Research Institutes of Humanities and Social Sciences in Universities
语言文学类 Linguistics and Literature
基于中介语语料库的
汉语句法研究
赵金铭 等著
北京大学出版社
PEKING UNIVERSITY PRESS

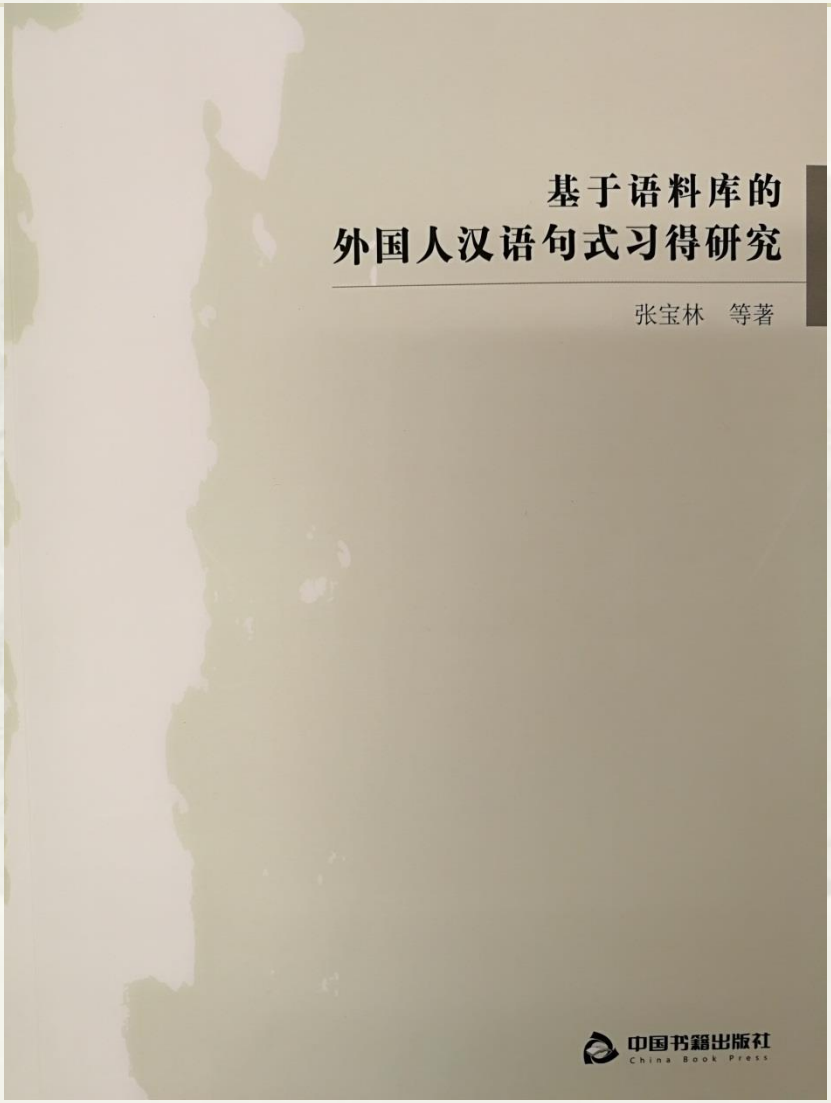
2．习得研究现状


教育部人文社会科学重点研究基地重大项目成果丛书
Publication Series: MOE Supported Projects of Key Research Institutes of Humanities and Social Sciences in Universities
语言文学类 Linguistics and Literature
基于中介语语料库的
汉语词汇专题研究
张 博 等著
北京大学出版社
PEKING UNIVERSITY PRESS

2．习得研究现状


外国学生汉语句式学习
难度及分级排序研究
国家社会科学基金项目（项目编号：06BYY024）
南京师范大学211工程三期重点学科建设项目
肖奚强 等著
基于中介语语料库
针对不同语言背景的外国学生
常用句式学习难度和分级排序
高等教育出版社
HIGHER EDUCATION PRESS

2．习得研究现状



2．习得研究现状

* 2）关于把字句习得情况的新认识
* 学界共识：难点；回避策略。
* （1）学习难度
* “HSK动态作文语料库”（1.0版）：把字句3682句，正确句3221句，87. 48%；偏误句461句，12. 52%。

2．习得研究现状

* （2）回避：
* “证据”：
* （1）外国人把字句使用率不足百分之一；
* （2）1个小时没说一个把字句。

2．习得研究现状

- 依据“HSK动态作文语料库”的发现：
- 外国人对把字句的使用率仅次于“是……的”句、是字句、有字句，而高于其他所有句式。

2．习得研究现状

* HSK动态作文语料库：3682/4000000≈0.0921%
* 母语者：
* 张旺熹2001：1049/650000≈0.16%
* 张　黎2007：46/210000≈0.0219%
* 李　宁、王小珊2001：335万字，≈0.0894%
* 人民日报：
* 俞士汶（98/1-6）：9801/13000000≈0.0754%
* 宋柔（00，12）：1498/1930000≈0.07762%
* 张宝林（2000）：18413/24000000≈0.0767%

2．习得研究现状

* （3）偏误类型
* 偏误句：461句
* “回避”：165句，35.79%
* “泛化”：句有161句，34.92%
* “内部偏误”：135句，29. 28%

2．习得研究现状

* （4）其他偏误
* 是……的句: 偏误句数居首位
* 离合词：共87处，占词汇偏误总数的0.098%
* 词汇量：27000

2．习得研究现状

* 2.1.3 德语国家相关研究较少
* （CNKI，2008.1~2017.6，按主题查询）
* （1）汉语教学
* 对外汉语教学：18933
* 德国汉语教学：20
* 奥地利汉语教学：4
* 韩国国汉语教学：168
* 越南汉语教学：44
* 泰国汉语教学：432
* 印尼汉语教学：42

2．习得研究现状

* （2）偏误分析
* 对外汉语教学+偏误分析/习得研究：
  2859+531= 3390
* 德国汉语教学+偏误分析/习得研究：1+0
* 奥地利汉语教学+偏误分析/习得研究：0+0
* 韩国国汉语教学+偏误分析/习得研究：10+0
* 越南汉语教学+偏误分析/习得研究：7+3
* 泰国汉语教学+偏误分析/习得研究：40+2
* 印尼汉语教学+偏误分析/习得研究：6+1

2．习得研究现状

- 德国汉语教学+偏误分析：1

题名	作者	来源	发表时间	数据库	被引	下载	阅读	热度
德国学生汉语语序偏误分析	谭敏	云南大学	2012-05-01	硕士	8	429		3.5

2．习得研究现状

* 2.1.4 德语国家相关研究少的原因
* 1）学习者人数少，相应的研究少；
* 2）研究材料少，不能满足研究的需求。
* 根据：HSK动态作文语料库中包括的语料

2．习得研究现状

* 全库作文篇数：11569，约424万字
* 德国：43篇，367字*43篇=15781字
* 奥地利：12篇，367字*12篇=4404字
* 德国、奥地利：55篇，20185字
* 韩国： 4171篇，约1530757字
* 越南： 221篇，约81107字
* 泰国： 374篇，约137258字
* 印度尼西亚： 739篇，约271213字

2．习得研究现状

* 2.1.5 可行的改进办法
* 1）在现有语料库中增加德语背景学习者的语料数量
* 2）构建德语背景学习者语料库

3．语料库总体设计

* 3.1 建库的目的与原则
* 1）目的：为面向德语背景的汉语教学与相关研究服务，为相关的统计分析、量化研究奠定基础。
* 2）原则
* （1）全面性：语料类型、学生类别、标注内容
* （2）针对性：突出德语背景学习者的学习特点、重点与难点
* （3）渐进性：语料由少到多，由缺而全；标注由浅而深、由易到难；急用先建、逐步完善

3．语料库总体设计

* 3.2 语料的收集与整理
* 1）真实性：自主成段表达，原稿，初稿
* 2）系统性：各类学生（专业、程度），各年级学生；各次作业，各类语料；可供横向与纵向研究之需
* 3）与教学同步，反映实际教学情况下的自然学习过程
* 4）规模：书面语100万字，口语50万字。

3．语料库总体设计

- 3.3 作者和语料背景信息的收集与整理
- 1）国籍、学校、专业、年级、学习动机与目的，母语、掌握的其他语言，学期课程成绩，是否参加HSK考试、成绩等级
- 2）语料所属课程、文体、性质（考试、平时作业）、完成地点、要求时限与实际用时、成绩
- 3）可采用Excel表的形式

3．语料库总体设计

* 3.4语料标注
* 1）标注原则
* （1）全面标注：满足多种研究需求；
* （2）浅层标注：提供检索方便，不替用户做判断；
* 2）标注模式：偏误标注+基础标注

3．语料库总体设计

* 3）标注内容
* 基本内容：字、词、句、标点符号
* 扩展内容：短语、语篇、语体、语义、语用、修辞
* 口语语料：+语音标注，-字标注
* 视频语料：+语音标注，+体态语标注，-字标注

3．语料库总体设计

* 4）标注方法
* 基本方法：人标机助
* 辅助方法：机器自动标注，繁体字、异体字，分词+词性标注

3．语料库总体设计

* 3.5检索系统与呈现方式
* 1）检索：简单，易用，方便
* 2）呈现：
* 偏误语料/正确语料
* 偏误语料+正确语料
* 语料+背景信息

3．语料库总体设计

* 3.6 语料库的建设与应用方式
* 1）合作共建，方式灵活
* （1）提供语料
* （2）提供语料+相关研究
* （3）提供语料+可单检该项语料的语料库
* 2）免费开放，为全世界的汉语教学与研究服务

4．问题与对策

* 1）背景信息：个人隐私
* 代码，能确保语料和相关信息能对应即可
* 2）语料规模与建设速度
* 边建设边开放，逐步积累，逐步完善
* 得到维大和德语区汉语教学协会的支持
* 3）标注内容的繁与简
* 根据实际需要，由少到多，由浅入深
* 4）标注方式
* 尽量机器标注或机助标注，方向

5. 结语

- （1）汉语中介语/学习者语料库对汉语的教学与相关研究具有重要意义。
- （2）德语背景学习者语料库建设滞后，已影响到相关研究的开展。
- （3）德语背景学习者语料库建设需要海内外学界的通力合作，提别是德语国家汉语学界的努力与支持。

谢谢！