

# 预本衔接的科技汉语语料库建设与应用

杜修平, 周子晗, 尹晓静  
(天津大学国际教育学院, 天津 300072)

**摘要:** 针对不少来华留学理工类学生在进入本科学习阶段都会陷入因专业汉语词汇量不足导致的学习效果低下的困境问题, 本研究建设了基于理工类教材的教材语料库和基于理工类课程教学视频的口语语料库(教学语料库)。首先, 使用语料分析工具对语料进行分词、字频和词频统计, 进而在词频阈值确定理论指导下, 使用词频分析工具筛选大学科技汉语词汇, 再经过专家干预, 最终构建了 2000 词的科技汉语词表和科技汉语常用表达句式。在研究成果应用方面, 本研究开发了科技汉语词汇在线查询平台, 以词表和科技汉语常用表达句式为主要参考依据, 编写《理工中文》教材以及与教材配套的语料素材, 后者可供科技汉语教师作为布置课前知识导入、课后阅读作业或者课堂测试的参考。本研究通过研制大学科技汉语, 衔接了预科科技汉语, 串起来华留学预本衔接科技汉语的全链条服务, 体现了提供预科后专业汉语服务和建设科技汉语资源基地的理念。

**关键词:** 科技汉语; 语料库; 词表; 预本衔接; 预科后  
**中图分类号:** H195.3

## Construction and Application of Scientific Chinese Corpus Which is Aimed at Connecting pre-college and College Chinese Teaching

DU Xiuping, ZHOU Zihan, YIN Xiaojing

(School of International Education in Tianjin University, Tianjin 300072)

**Abstract:** In view of the dilemma that many foreign students majoring in science and technology will fall into the low learning effect caused by the lack of professional vocabulary at the undergraduate stage. This paper constructs a textbook corpus based on science and technology textbooks and a spoken corpus based on science and technology course teaching videos through word segmentation, word frequency and word frequency statistics. Firstly, the corpus analysis tool is used to conduct word segmentation, word frequency and word frequency statistics of the corpus. Then, under the guidance of the word frequency threshold determination theory, the word frequency analysis tool is used to winnow the scientific Chinese words at the undergraduate stage. Finally, after expert intervention, the scientific Chinese word list of 2000 words and the common expression sentence patterns of scientific Chinese are constructed. In application of research fruits to teaching, this study developed an online search platform for scientific Chinese vocabulary. With vocabulary and expression sentence patterns of scientific and technological Chinese as the main reference, and compiled the textbook Chinese for Science and technology as well as the corpus material are compiled, the latter can be used as a reference for scientific Chinese teachers to assign pre-class knowledge introduction, after-class reading assignments or classroom tests. Through the development of college science and technology Chinese, this study connects the preparatory science and technology Chinese, and connects the whole service chain of studying science and technology Chinese in China, which reflects the concept of providing post-preparatory professional Chinese teaching service and building the scientific Chinese resource base.

**Key words:** Science and technology Chinese; Corpus; Glossary; Advance book connection; post-preparatory

**基金项目:** 教育部中外语言交流合作中心 2022 年国际中文教育研究课题重点项目“中文+职业教育”融合发展机理研究(22YH30B)

**作者简介:** 杜修平(1974-), 男, 天津大学国际教育学院副院长、教授、博士生导师, 主要研究方向: 国际中文教育、学习科学和跨文化教育研究

**通信联系人:** 尹晓静(1979-), 女, 天津大学国际教育学院汉语进修系主任, 专职对外汉语教师, 主要研究方向: 语言学和应用语言学. E-mail: yin\_jing99@126.com

## 0 引言

预科教育的教学任务是为来华学习各种专业的学生打下必要的汉语基础及专业课知识基础<sup>[1]</sup>。在政府奖学金预科教育中,理工类预科生是规模较大的群体。科技汉语是来华留学生预科(理工类)阶段的必修课,是预科普通汉语课和专业基础知识课衔接的桥梁。科技汉语教学质量与留学生预科(理工类)教学质量紧密相关。

本研究通过对于已经进入本科阶段的往届预科生的访谈发现:他们普遍认为读懂专业教材、听懂专业课讲授等是非常困难的,即便是在预科阶段非常出色的学生,也需要一段相当长的时间、付出相当大的努力才能够跟上本科学习的步伐。学生们反映自己对于专业知识特别是专业词汇量的掌握有所欠缺,导致听课效果不佳,甚至影响其顺利完成学业。由此可见,理工类留学生在进入本科的专业学习后,在读懂专业教材、完成书面作业等方面存在障碍,在听懂专业课教师讲授、就专业相关话题做口头交流等方面困难重重。

为了使预科生顺利衔接专业学习,提升毕业率,本研究提出了预科后服务理念,由预科院校继续提供专业汉语的进阶延伸服务。基于此,本研究建设了基于大学公共基础课的教材语料库,基于慕课、视频课的教学语料库,经过语料分析和专家人工干预等方式筛选词汇,构建了科技汉语 2000 词词表,提供在线查询服务,可以查询词汇释义,汉语拼音,词汇的英语、法语、阿拉伯语翻译等。并且依据词表编写了《理工中文》教材,与以前出版的《科技汉语读写教程》一起,拍摄微课,配套习题,提供在线学习支持,整体上打造从预科到本科的科技汉语学习全链条服务。

## 1 科技汉语语料库建设

2009 年,天津大学国际教育学院以中学数理化教材为语料对数理化科技汉语词汇的词频作定量研究,共分析了上百万字的语料,并对科技汉语词汇的统计结果进行人工干预,结合对汉语教师、数理化教师以及留学生的调查结果,制定出适用于留学生的科技汉语词汇大纲。2015 年,《预科教育理工类大纲及词汇精选》由北京语言大学出版社出版<sup>[2]</sup>。

本研究吸纳天津大学预科系过去针对预科科技汉语的研究经验,对大学本科的数理化教材以及相关课程视频进行研究,将语料库建设划分为本科科技汉语教材语料库、本科科技汉语教学语料库,并逐步开发建设基于在线语料的语料库。

### 1.1 科技汉语教材语料库

#### 1.1.1 语料来源

考虑到教材的权威性与可采集性,科技汉语书面语语料库语料的具体来源是理工类本科一年级学生的数理化教材:数学选用的是由高等教育出版社出版,同济大学数学系教研室编著的《高等数学》;物理选用的是《普通物理学(第七版)》,由程守洵、江之永等人主编;化学选用《大学化学》,由天津大学出版,天大无机化学教研室编著。由于计算机专业的知

识随着科技的发展更新迭代速度极快,因此,计算机教材采集电子科技大学出版社、清华大学出版社和上海交通大学出版社出版的三本同名教材《大学计算机基础》。

80 上述四门课程教材均为国内现有的、常用的且具有权威性的理工类本科一年级学生专用的公共课教材,可以较好地呈现基础、真实的科技汉语语料。

### 1.1.2 语料处理

85 在语料整理与入库过程中需要经过相应的预处理,包括文本的整理、语料元信息标准、分词、词性标注等<sup>[3]</sup>。在电子文本成为纯文字语料前,通过网络下载、人工录入和扫描识别等方式获得的文本存在各种不符合语料库建设规范的字符信息和格式,因此需要人工进行核验,从这些教材文本语料中除去图片、专业符号以及其他特殊标点等与纯文本不相关的内容,并对文本转化时出现标注内容不当或错误的部分进行人工核对。避免无关字符信息影响后续分词的正确率与词频统计的数值信息,出现词性标注错误或搭配统计不准确的情况。

### 1.1.3 语料规模

90 经过人工整理后的语料文本达到 150 万左右的规模,表 1 为各主题语料的文本规模。

表 1 语料文本规模

Tab. 1 The textual scale of corpus

书名	作者	出版社	出版年份	语料规模 (字符)
《大学计算机基础》	主编:马大勇,王洪艳	清华大学出版社	2020	149563
《大学计算机基础》	主编:韦鹏程,罗军,吴海霞	电子科技大学出版社	2016	169568
《大学计算机基础》	主编:王伍柒,周飞	上海交通大学出版社	2014	197483
《高等数学》	同济大学数学系	高等教育出版社	2014	329740
《普通物理学》 (第七版)	程守洙、江之永主编, 胡盘新、汤毓骏、钟季康、 胡过图、钟宏木修订	高等教育出版社	2016	408429
《大学化学》	天津大学无机化学教研室 编,杨秋华主编	高等教育出版社	2014	245052

## 1.2 科技汉语教学语料库

### 1.2.1 语料来源

95 科技汉语教学语料库的语料来源于中国慕课平台上对应科目的课程以及高校线下真实课堂视频,利用语音转录文字技术提取视频中的文字后进行人工校对和整理,构建口语语料库。本研究分别以“高等数学”、“普通物理”和“大学化学”为关键词在中国大学 MOOC(慕课)平台上进行搜索。最终,数学选定同济大学开设的高等数学课程,物理选定南京理工大学开设的普通物理课程,化学选定天津大学开设的大学化学课程。以上课程均为国家精品课。

### 1.2.2 语料处理

100 口语语料相较于书面语语料来说获取难度更高,需借助格式转换及语音自动识别工具,

结合人工干预，完成语料的采集与转写。具体使用的采写工具及采写流程如下：

首先，本研究使用 Format Factory 工具中的视频转 MP3 功能，将中国慕课平台上的视频文件及课堂教学录像逐个转换为音频文件，为后续处理做好准备。继而使用安卓版的“语音转文字”软件，将转换好格式的 MP3 文件导入应用，使用“转文字”功能完成智能转写。

最后，人工对转写后的语料进行精细加工，需要仔细对照视频课程反复核查有无错字别字和漏字，对课程中教师所说的每一句话都如实转写，包括因强调重点或是课堂提问而产生的重复语句，在这个过程中不对语料做无必要的改动。其次，由于高数、物理、化学课堂教学中涉及公式极多，给自动分词带来很大困扰，需要人工进行语料清洗，在 word 里使用“查找替换”功能，将数字、字母、符号全部替换，只留下汉字，提高分析的准确性。

### 1.2.3 语料规模

对校对清洗后的熟语料进行统计之后，共产出 1496093 字口语语料，其中去除高等数学（课堂录像）课程中与讲课无关的内容后，剩余 1363128 字口语语料。其中包括高等数学（慕课）286039 字符、普通物理（慕课）165298 字符、大学化学（慕课）230455 字符、高数数学（课堂录像）814101 字符。

## 2 科技汉语词表研制

科技汉语词表的研制需要对语料库中的文本内容，进行分析、处理、筛选，确定科技汉语词汇的数量和范围。下面对本研究的词汇分析实践工具与词汇阈值确定理论两方面进行简要说明。

### 2.1 词频分析实践工具

词频分析可谓是分析和研究语料库的主要方法的核心，是对所有语料库进行统计和分析的最重要基础。语料库工具根据实际需要按照统一的基准对原始频次进行“标准化”。标准化频次有助于比较不同容量或者不同规模语料库中的词频，从而更易于进行比较分析<sup>[4]</sup>。本研究综合使用了多种词汇统计分析工具，对文本语料进行处理分析，既采用了网络上共享的、可以直接使用的在线或离线工具，也自行开发了词汇统计分析工具。本研究中，使用了 ROSTCM6 大数据计算机工具的文本处理和功能性分析对文本内容进行分词，分析出输入文本的字频和词频，分析结果包含中英文两种形式。

### 2.2 词汇阈值确定方式

本研究采用的词汇阈值确定方式有三种。其一是齐夫第二定律，在对自然语言文本处理的过程中，词出现的频率（frequency）与它在频率表里的序位（rank）的乘积大致是一个常数，即： $f \times r = c$ 。对数据求对数，那么词频与词序之间满足： $\log f = \log c - \log r$ ，两者之间存在线性相关关系。此方法可以总结出的统计文本中低频词的分布态，确定高频词阈值<sup>[5]</sup>。其二是词频 g 指数，它是指某一个研究主题关键词的数量分值为 g，在此研究主题的关键词总量

N 中，有  $g$  个关键词的累计出现频次不少于  $g^2$ ，并且  $g+1$  个关键词的累计出现频次要少于  $(g+1)^2$ <sup>[6]</sup>。此方法可用于研究共词分析，即通过高频出现的词汇来反映这一研究内容或领域的整体结构和知识体系。其三是词频累积占比，这是刘敏娟教授等人提出的共词分析词集范围的确定方法，是一种基于词频、词量、累积词频占比三者变化关系的分析和统计方法<sup>[7]</sup>。此方法主要包括三个环节：①关键指标计算。针对抽取出的内容词，计算其词频、累积词频占比及其对应的词量，为下一步通过三者关系的变化情况确定高、中、低频区域做好准备。

②高、中、低频词区识别。根据词频、词量、累积词频占比三者变化关系，分析不同区域三者相互变化的规律和特点，以此确定高、中、低频词区的阈值及范围。③分析词集，确定范围。取高、中频词区的内容词共同作为共词分析对象，进行后续的共词分析。

2.3 科技汉语词汇筛选

在筛选、整理科技汉语多语种词汇表的过程中，使用了上述三种方式，对语言文本进行计算来确定词汇阈值，具体数据情况如下表所示。

表 2 《高等数学》词频阈值

Tab. 2 Word frequency threshold of <Advanced mathematics>

方法	数值	词序	词频
齐夫第二定律	68.91	69	341
词频 g 指数	/	301	64
词频累积占比	90%	540	26

表 3 《普通物理学》词频阈值

Tab. 3 Word frequency threshold of <General physics>

方法	数值	词序	词频
齐夫第二定律	117.89	118	282
词频 g 指数	/	370	100
词频累积占比	85%	872	32

表 4 《大学化学》词频阈值

Tab. 4 Word frequency threshold of <University chemistry>

方法	数值	词序	词频
齐夫第二定律	118.56	119	185
词频 g 指数	/	277	84
词频累积占比	80%	838	23

通过数据分析，本研究选用词频累积占比的方法对各科词汇情况进行分析，确定中频词阈值，选择高频区和中频区共同作为理想词表的选词范围，保证词表包含更多重要的专业词汇，同时避免低频词给学生带来干扰和过多的词量负担。各科的词频累计占比筛选结果分别为：数学 95%，取到词频 3，筛选后 468 个词汇；物理 85%，取到词频 7，筛选后 646 个词汇；化学 80%，取到词频 6，筛选后 705 个词汇。需要说明的是，最终确定的词汇，需要经过专家干预，并且需要考虑几个科目的特点和词汇分布情况。



### 3 科技汉语语料库研究成果及其应用

#### 3.1 科技汉语多语种词汇在线查询

本研究开发的本科科技汉语多语种词汇在线查询平台,目的是为了辅助用户学习基于理工类科技汉语的高频词表,从而更高效地提取词汇信息,辅助用户学习。

在开发制作科技汉语多语种词汇查询平台之前,需要对前期构建的科技汉语词表进行数据处理,并搜集科技汉语词表相应中文词汇的英语、法语、阿拉伯语翻译,并对词汇标记例句及释义。

其次,将整理好的科技汉语多语种词汇表以“课程”模块的形式导入至 istudy 平台中的科技汉语资源基地“大学科技汉语”模块下。istudy 平台是本研究基于 Moodle 开源软件,经本地化定制和部署,构建的科技汉语资源基地(<http://istudy.tju.edu.cn>)。该平台既是科技汉语的资源平台,也是科技汉语的学习平台、科研平台。利用该平台,不仅可以永久存储从科技汉语书面语语料库中得到的数据成果,使科技汉语多语种词汇对照表得到可视化的展示,更能为汉语教师及汉语学习者提供科技汉语词汇资源,使其更好地发挥相应价值。科技汉语多语种词典最终呈现的结果如图 1 所示。用户可以通过平台查询词汇并获取信息、巩固复习词汇。例如,用户查询“积分”,则可以获取关于积分的英语翻译、法语翻译、阿拉伯语翻译以及中文释义。



图 1 词汇查询示例

Pic 1 example of query for words

#### 3.2 科技汉语语料在线查询

本研究秉承科学性和专业性的原则,依据科技汉语 2000 词表,结合当前科技热点话题,按照学科类别编写与《理工中文》配套的科技汉语语料素材,可供科技汉语教师作为布置课

180 前知识导入、课后阅读作业或者课堂测试的参考。

同时，本研究开发科技汉语在线语料库检索系统，对自编语料提供在线查询服务。其检索服务的实现主要采用基于 WEB 的 B/S 结构，也就是将 Internet 浏览器作为客户端，MySQL 数据库和 istudy 学习平台作为服务器端。客户端使用网页作为数据检索界面，只要有 Internet 浏览器，不需要任何其它软件就可以随时随地进行检索。

185 首先创建数据库表，用来存放全文语料库样本。语料样本入库前，先按设定的样本容量将大篇幅的文本切分成若干个语料样本，每个语料样本作为一条数据记录存入全文语料库的数据表中。数据表的结构如下所示：

表 5 语料样本数据结构表  
Tab. 5 the data structure of corpus sample

字段名称	数据类型	说明	字段名称	数据类型	说明
ID	数字字段	序号	Key-word1	文本字段	关键词 1
Title	文本字段	标题	Key-word2	文本字段	关键词 2
Editor	文本字段	编辑	Key-word3	文本字段	关键词 3
source	文本字段	来源	Content	文本域字段	正文内容
subject	文本字段	文章主题	Grading	文本字段	适用等级
subtopic	文本字段	副主题			

190 在全文数据导入 istudy 学习平台之前，需要在平台内的 module 中设置与数据库的字段映射关系，再将全文数据表导入，istudy 平台的检索服务支持分页显示，支持语料下载及自定义标注功能。可按篇名、作者、出处、时间和语料正文检索关键词。按语料正文检索时，可键入任何字符串（汉字或英文）。凡符合检索条件的语料样本的全文均可分页显示出来，并统计满足检索条件的数据总数。显示检索结果时，所键入的检索字符串用黄色高亮显示，  
195 以便快速找到所检索的关键词在语料全文中的位置。

找到记录: 1/366 (重设过滤器)

## 自由落体运动

数学 函数 自由落体 位移 瞬时速度

你玩儿过游乐场里的跳楼机么? 感受过蹦极的失重与刺激么? 如果你是亲历者, 那么你一定是一个非常勇敢的人, 即使未曾尝试, 想必也目睹过类似的极限运动场景。它们模拟从高空坠落的自由落体运动, 在下落过程中速度越来越快, 是对身体和心理素质的双重挑战。下面我们就来了解一下自由落体运动中的数学知识。

自由落体运动是一种变速运动, 更为确切地说, 是匀变速直线运动, 这就意味着自由落体运动的加速度恒定不变, 是一个定值, 它恒等于 $g$ , 为了方便计算, 重力加速度 $g$ 通常取近似值 $9.8\text{m/s}^2$ 。

我们试着描述一只小球进行自由落体运动的情况。当小球从距离地面 $h_0$ 的位置自由下落时, 它的初始速度为 $0\text{m/s}$ 。由于小球下落的位移 $h$ 是小球下落时间 $t$ 的二次函数, 它们的函数关系为 $h=1/2gt^2$ , 在不计空气阻力的情况下, 经过 $t_1$ 时间后, 小球经过的位移为 $h_1=1/2gt_1^2$ 。若想得到小球在 $t_1$ 时刻的瞬时速度, 则要先求得速度与时间的函数关系式。由于可以将小球在 $t_1$ 时刻的瞬时速度看作此时的位移变化率, 所以我们可以将 $h$ 与 $t$ 的函数关系式 $h=1/2gt^2$ 进行求导, 导函数关系式就是瞬时速度 $v$ 与时间 $t$ 的函数关系式, 即 $v=gt$ 。那么在 $t_1$ 时刻的瞬时速度为 $gt_1$ 。

接下来我们再看看 $h=1/2gt^2$ 与 $v=gt$ 这两个函数图像分别是什么样的, 以及它们有什么关系。 $h=1/2gt^2$ 是一个二次函数, 其图像是一个开口向上抛物线的右支, 该函数图像有且只有一个零点, 在原点处, 其实际意义在于在自由落体初始时刻, 位移为 $0\text{m}$ 。 $v=gt$ 是一个一次函数, 其图像是一个端点在原点处的射线, 该射线的斜率 (指该射线的倾斜程度, 也是它倾角的正切值) 为 $g$ 。由图可知, 自由落体运动的瞬时速度与时间成正比。

说明了两个函数图像各自的特征, 我们再看看二者存在什么关系。在位移与时间的函数图像上,  $t_1$ 时刻的瞬时速度就是抛物线在 $t=t_1$ 处的切线斜率 $gt_1$ , 即切点坐标为  $(t_1, 1/2gt_1^2)$ 。

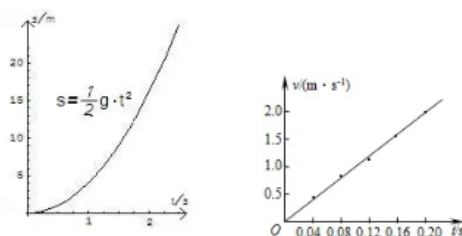


图2 语料查询示例

Pic 2 example of query for corpus

## 200 3.3 大学科技汉语 (理工中文) 教材编撰

本研究以词频分析为主获得的词表和 International 汉语教学专家挑选审核的科技汉语常用表达句式为主要参考依据, 考虑科学性、趣味性、大学课堂相关性等因素, 编写理工中文 (科技汉语) 教材。主要为来华留学本科一年级学生或预科教育高级阶段专业中文学习者服务。

教材划分为数学、计算机、物理和化学四个基础学科版块, 以词表为基础, 凸显了科学性和严谨性, 同时最大程度上复现了大学数理化课堂常用的科技词汇, 为留学生的专业学习提供必要的语言支撑。其次, 教材编写以提高留学生综合语言能力为目标, 重在培养学生独立阅读大学理工类公共基础课教材, 听懂大学理工类公共课程的能力。再次, 教材复现了大量预科科技汉语和预科数学、物理、化学等专业课中的词汇, 较好地衔接了预科教育和本科教育, 配有全部章节的 PPT 课件、习题及微课视频, 并在线提供理工中文的补充语料, 体现了我们提供预科后专业中文服务和建设科技汉语资源基地的意愿。最后, 教材以科普文章为主线, 以科技话题为纲, 选题丰富多样, 遵循了教材编写的时代性和趣味性的原则。部分章节中所包含中国优秀传统文化以及当代科技成果的内容也是将语言、文化和理工知识结合的重要表现, 使留学生在专业语言的学习同时, 不仅可以了解到中国科技发展, 同时透过不同侧面看到发展中的科技中国。



## 215 4 结论

对来华留学学历教育来说, 近些年理工类专业学生逐渐增多, 已成为来华留学的主流, 对其实施同一化教学管理的大方向已基本形成共识。本研究针对来华留学理工类学生在进入本科学习阶段陷入因专业汉语词汇量不足导致的学习效果低下的困境, 建设本科科技汉语语料库, 基于语料分析, 开发本科科技汉语词表, 编写理工中文教材, 建设科技汉语资源基地, 220 提供教材、微课、词汇释义、语料查询等服务, 打造了从预科到本科的科技汉语学习全链条服务, 对来华留学专业汉语学习具有重要作用, 对提高来华留学教育质量具有重要意义。

## [参考文献] (References)

- [1] 钟授. 对外汉语教学初探[M]. 北京: 北京语言大学出版社, 2006.
- [2] 杜修平, 韩志刚. 中国政府奖学金来华留学预科教育研究[M]. 天津: 天津大学出版社, 2015.
- 225 [3] 黄昌宁, 李娟子. 语料库语言学[M]. 北京: 商务印书馆, 2002.
- [4] 杨惠中. 语料库语言学导论[M]. 上海: 上海外语教育出版社, 2002.
- [5] 张忠友. 齐夫定律的理论基础及其实践意义[J]. 情报科学, 1989, (05): 62-66+78.
- [6] 杨爱青, 马秀峰, 张风燕, 薛卫双. g 指数在共词分析主题词选取中的应用研究[J]. 情报杂志, 2012, 31(02): 52-55+74.
- 230 [7] 刘敏娟, 张学福, 颜蕴. 基于词频、词量、累积词频占比的共词分析词集范围选取方法研究[J]. 图书情报工作, 2016, 60(23):135-142.