基于偏误反馈的对韩汉语词汇教学信息库的建设及应用*

焉德才 胡晓清
鲁东大学国际教育学院，山东烟台 264025
yandecai599@sina.cn

摘要：汉语中介语语料库建设完成以后，词汇教学自主应用平台的建设应该提上议事日程。本文从服务于课堂词汇教学的角度对“基于偏误反馈的对韩汉语词汇教学信息库”的建设和应用进行了阐述。文章论述了信息库的建库原则，展示了信息库的内容构成和主体框架，讨论了信息库内容的编写流程, 并认为汉语中介语语料库的词汇偏误生态描写应该逐步由过去的“多词一面的群体研究”向“一词一面的个体研究”转向，基于偏误反馈的国别化汉语词汇教学信息库的建设很可能是未来汉语语料库建设的方向之一。

关键词：偏误 词汇教学 信息库 应用

Construction and Application of Chinese Vocabulary Teaching Information Corpus for Korean Students Based on Errors Feedback

Yan Decai Hu Xiaoqing
International Exchange College of Ludong University，Shandong，Yantai 264025

Abstract: After the completion of the construction of Chinese Interlanguage Corpus, Chinese vocabulary teaching platforms should be put on the agenda. This article expounded the scheme about "Chinese Vocabulary Teaching Information Corpus based on errors feedback for Korean students". It expounded the principle of building corpus, content and the main framework, at the same time, discussed the factors that affect the vocabulary cognitive coding degrees. The ideas about construction of information corpus will provide a new perspective for the application of Chinese Interlanguage Corpora.

Keywords: errors; vocabulary teaching; information corpus; application

0 引言

语料库的建设和语料库语言学的崛起，是语言学战略目标转移的一个重要标志。（冯志伟，2011：13）语料库语言学主要包含两方面的内容：一是对自然语料进行加工、标注，二是用已经标注好的语料进行语言研究和应用开发。（黄昌宁、李涓子，2007：3）一般来说，中介语语料库的应用开发价值主要体现在“语言研究”、“教材编写”和“词典编纂”

* 本文得到教育部人文社科规划项目（11YJA740107）和国家社科规划项目（11BYY050）的资助，谨致谢忱！

三个大的方面。不过，在 CSL 研究领域，汉语中介语语料库的价值主要表现为“语言研究”领域的偏误分析和偏误描写，而在“教材编写”和“词典编纂”领域对偏误信息的反馈和应用则显得相对迟缓和滞后。以词典编纂为例，目前对外汉语教学领域的一些有代表性的参考词典，除了吕叔湘主编的《现代汉语八百词》、杨庆蕙主编的《现代汉语正误辞典》、杨继洲、贾永芬编著的《1700 对近义词语用法对比》、赵新、李英编写的《商务馆学汉语近义词词典》、朱丽云主编的《实用对外汉语重点难点词语教学词典》和施光亨、王绍新主编的《汉语教与学词典》等包含少量的偏误反馈信息以外，其他词典如《汉语 8000 词词典》（刘镰力主编）、《当代汉语学习词典（初级本）》（徐玉敏主编）、《汉语动词用法词典》（孟琮、郑怀德、孟庆海、蔡文兰编）、《汉语形容词用法词典》（郑怀德、孟庆海）、《北大版学习词典：汉语副词词典》（岑玉珍）等大多是基于汉语本体研究的相关信息而编写的，对中介语偏误信息的反馈和吸收重视得不够。探究造成这种状况的原因，可能不外乎两点：一是主观原因，即对外汉语学界对“偏误入词典” 理念重视得不够。二是客观原因，即纸质词典的容量限制使“偏误入词典”理念不易实施。为了突破这一瓶颈，研究者的目光就很自然地投向了建设一系列充分吸收偏误反馈信息的容量更大、检索更便捷的汉语词汇自主教学应用平台上。这样的词汇自主教学应用平台实际上就是基于偏误反馈的汉语词汇教学信息库。

对外汉语词汇教学信息库的建设需要紧密服务于课堂教学，因此其内容就必须包含词汇在汉语中介语语料库中的偏误反馈信息。这些信息起码应该包括：每个 HSK 大纲词的语音、汉字、语法、广义语义的编码难度等级的标注信息、认知难度的评价信息、常见迁移情况的描写信息（包括易混淆信息）、典型偏误的展示信息等。目前，这些来自于偏误反馈的、服务于课堂教学的中介语偏误信息尚未完全纳入到词典编纂和词汇信息库建设的视野。这一空白，亟待填补。

“基于语料库的语言描述的应用是语料库进化中最具有创新性的一项活动。”（黄昌宁、李涓子，2007：20）下面，本文将对“基于偏误反馈的对韩汉语词汇教学信息库”的建设谈一下初步的设想，并对其内容的编写应用进行初步的展示。

1 基于偏误反馈的对韩汉语词汇教学信息库建设

1. 1 对韩汉语词汇教学信息库的建库原则

1. 1. 1 服务教学

不同的语料库，其主要功能也不尽相同。就二语中介语语料库来说，有的偏重“研究导向”（Research guidance），有的偏重“教学导向”（Teaching guidance）。总体来讲，国内已知的汉语中介语语料库大部分偏重“研究导向”。“基于偏误反馈的对韩汉语词汇教学信息库”则是一种偏重“教学导向”的信息库。它主要服务于从事对韩汉语教学的国内外广大教师，向他们提供词汇教学上的参考。我们的设想是通过这个开放型的信息库，让每个从事对韩汉语教学的教师都能够对韩国学生习得汉语 HSK 大纲词的认知难度、常见迁移情况、易混淆信息以及典型的偏误形态有一个比较全面和充分的了解，可以随查随用。

1. 1. 2 聚焦偏误

本信息库的关注焦点是中介语中的偏误因素。偏误语料主要是以汉语 HSK 大纲词为搜索项，从已经建成的“韩国留学生汉语中介语语料库”中提取偏误句，建成子库，为汉语

词汇教学信息库的建设提供数据和信息支持。

1. 1. 3 语料真实

基于偏误反馈的对韩汉语词汇教学信息库所搜集的偏误语料必须是真实自然的，所有偏误形式都必须是在学生的书面作业或者口头话语中出现的真实句子。这一点无需赘言。

1. 1. 4 开放共享

信息库的“开放”包含两个层面的内容：一是信息库的建设是一个长期的开放过程，二是信息库会不断吸收学界最新的研究成果，随时修正和完善相关内容；信息库的“共享”是指所有从事对韩汉语教学的国内外教师，均可通过固定网址凭密码登录这个信息库免费查询所需要的信息和语料。语料库的“开放”和“共享”是未来的大趋势，崔希亮和张宝林先生（2011）所倡导的“全球汉语学习者语料库”就预示着这一趋势即将到来。

1. 2 对韩汉语词汇教学信息库的内容构成和主体框架

1. 2. 1 对韩汉语词汇教学信息库的内容构成

对韩汉语词汇教学信息库直接面向查询者的内容分五大部分，即：基础附码、编码度标注、认知难度评价、迁移情况描写、典型偏误展示。在这五部分中，“典型偏误展示”是最关键的内容，因为其他内容的撰写大多来源于对这些偏误信息的分析和归纳。

具体来说，“基础附码”分“词性附码”、“词法附码”和“词调附码”三种。比如“半天”一词的词性附码是“n”（名词），词法附码是“pz”（偏正结构），词调附码是“41”（四声+一声）；“编码度标注”是对每个词从语音、汉字、语法和广义语义四个维度标注认知难度系数；“认知难度评价”是对词汇的“认知难度”做出解释和评价；“迁移情况描写”是在对偏误语料综合分析的基础上对词的正负迁移情况做出描写和说明，同时对该词的易混淆信息进行分析和解释；“典型偏误展示”呈现的是搜集到的典型偏误例句。面向查询者的信息库简化界面如图 1：

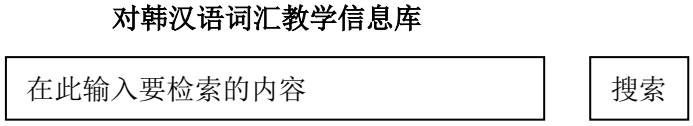

图 1 信息库面向查询者的简化界面

查询者进入此界面，只要输入查询的单词，然后点击搜索键，就会进入该词的页面浏览需要的信息。同样，输入不同的基础附码，也可以检索出具有相同属性的某一类词。比如，输入词性附码“lhc”，就可以将大纲词中所有离合词检索出来，输入词法附码“pz”，就可以将所有偏正结构的单词检索出来，输入词调附码“32”，就可以将所有声调是“三声+二声”的单词检索出来。单击检索出来的词，就可以直接进入该词的信息页面。所有检索出来的信息，既可以用于课堂教学，也可以用于大规模的集合研究。

1. 2. 2 对韩汉语词汇教学信息库的主体框架

对韩汉语词汇教学信息库的主体框架包括“语料处理系统”、“数据库”和“用户检索系统”三个部分。“语料处理系统”中存储的是从“韩国留学生汉语中介语语料库”中提取的所有偏误语料以及韩国留学生汉语音频语料的偏误信息；“数据库”中包括 HSK 大纲词、大纲词基础附码集、大纲词词频统计、大纲词汉字偏误统计、大纲词认知编码度集等各种

信息；“用户检索系统”的内容分基础附码、编码度标注、认知难度评价、迁移情况描写、典型偏误展示五大部分。信息库的主体框架，如图 2：

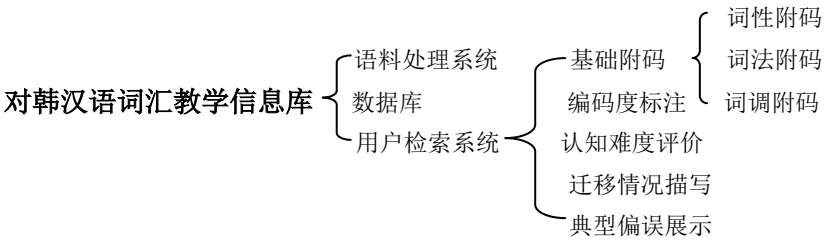

图 2 信息库的主体框架

2 对韩汉语词汇教学信息库的内容编写与应用

2. 1 编码度标注

在对韩汉语词汇教学信息库的建设过程中，“编码度标注”是一个比较重要的内容。这里的“编码度”是一种衡量韩国学生认知汉语词汇难易程度的量化指标。对汉语大纲词进行编码度的标注，有助于对韩汉语教师比较直观地了解韩国学生学习汉语词汇的重点难点，以便采取针对性的教学方法和教学策略，实现教学效果的最大化。词汇编码度的标注可以简单概括为“四维五级”。“四维”是指编码度的标注分“语音”、“汉字”、“语法”和“广义语义”四个维度，“五级”是指学生认知汉语词汇的五个难度等级，认知难度最高的词标为 5，最低的标为 1，其他根据偏误率、词频等指标信息分别标为 2 级、3 级和 4 级，各级都有相应的级差标准。需要指出的是，这里的难度等级主要是指韩国学生学习汉语词汇动态过程中起点时的相对静态的难度评价系数，它既包含预测的信息，比如语音，也包含基于偏误反馈的信息，比如语法。

具体来说，语音编码度的确定，需要跟韩国语的语音进行比照，找出难音的偏误规律，同时要兼顾声调难点、语流音变、音频语料的反馈信息等。比如，发音方面，“ü、ue、f、p、b、zh、ch、sh、r、z、c、s”等是难音；声调方面，汉语的“三声+二声、三声+三声、四声+四声、二声+二声”等四类双音节词的声调以及“一”、“不”的变调对韩国人来说是最难的，这些难音难调会导致词汇的语音认知难度系数高。相反，“不得不、不得已、新郎、不满”等词的韩语发音跟汉语差不多，有正迁移因素，语音的认知难度系数就低。

汉字编码度的确定起码要考虑如下几方面的因素：汉字笔画数（一、繁）、汉字结构的复杂度（旧、翻）、与其他汉字的相似度（爱、受）、简繁体因素（韩、韓）、汉韩字体笔画相似度（吕、모）、汉字偏误的数量等。

语法编码度的确定起码要综合考虑如下几方面的因素：这个词能否纳入“介宾+谓词”框架（对…感兴趣、为…操心）、是否是语法词（了、的、反而、还有、随着）、是否属于高难生成的副词（就、才、都、还）、是否是离合词（见面、结婚）、是否常以高难的特殊句式呈现（“把”字句、“被”字句、“得”字补语句、存现句、主谓谓语句）等等，当然这些最终也要参考语法偏误的数量才能确定。

广义语义编码度的标注起码要考虑如下几方面的因素：是否是汉字词（不满、新郎）、语义实用度（非常、极为）、语义差异（明白、理解、了解、知道、意识到、懂）、语义负迁移（朝、对、向、冲）、语体差异（勤奋、勤勉）、词彩差异（造成、快乐）、文化义差异（白手、黄色）以及语义偏误的数量等。

2. 2 认知难度评价

对汉语词汇认知难度进行评价需要分层次进行。以单词“旅游”的认知难度评价为例，其语音和汉字的认知难度都是 5，语法的认知难度是 3，语义认知难度是 2。这些难度系数的确定是基于以下相关信息：

语音认知：“3 声+2 声”是韩国学生最难掌握的声调形式，若不经过反复强化，这一声调形式极易化石化。

汉字认知：“旅”字在汉语中介语语料库中出现了 11 种偏误书写形式，其中“旅”和“𠂢”两种偏误形式最为常见，而“游”字出现了 6 种偏误书写形式，其中以“遊”和“斿”最为常见。根据以上偏误信息，我们可以将单词“旅游”的汉字认知难度系数标为 5。

语法认知：确定“旅游”的语法认知难度为 3 的偏误信息如下：

1） 趁着放假，我要旅游不少地方。

2） 三周前，我旅游釜山。

3） 我北京旅游的时候，觉得很多地方真美丽。

4） 我最感兴趣的是中国旅游。

5） 在朋友的帮助下，我的香港旅游也成功了。

6） 通过哈尔滨旅游，我学到了很多中国文化。

从上面的例句可以看出，“旅游”的语法偏误主要表现为三种形式：一是其后直接加“场所名词”，如例句 1）和例句 2）；二是其前直接加“场所名词”，如例句 3）和例句 4）；三是词语误代，同时包含了些许语体偏误的微观信息，即“旅游”误代的是“之游”、“之行”两种结构形式，比如例句 5）的“香港之游”和例句 6）的“哈尔滨之游”。基于以上偏误信息，我们将单词“旅游”的语法认知难度系数标注为 3。

语义认知：确定“旅游”的语义认知难度为 2 的偏误信息如下：

1） *下周末，我想去旅游张家界。

2） *2008 年，我去北京中国游行。

3）*以后，我想去西安兵马俑旅行。

从以上三个例句可以看出，“旅游”在语义上常常跟“游览”、“游行”、“旅行”混淆，语义理解上的偏差极易诱发语法偏误。但是，我们也应当看到，这种偏误形式并不十分严重，因为“旅游”的语义透明度较高，语义的可理解度也较高，即使留学生用了其他混淆词，听话者也会迅速地进行内部语义调整，将混淆词的词义理解为当用词“旅游”的语义。正是根据这些信息，我们将“旅游”的语义认知难度确定为 2。

2. 3 迁移情况描写

词汇偏误的迁移情况描写是将偏误形成的正负迁移情况进行描写和解释。我们仍以“旅游”为例。因为韩国语没有声调，因此“旅游”的声调受语际因素的影响较大，是最难纠正的声调, 它常被韩国学生发成类似于“2+2”调值，而且很容易“化石化”。另外“ü”的发音，学生由于受到母语语音的影响，常常发得不标准，在初级阶段是比较难纠正的音；在汉字上，负迁移因素主要来自于汉语语内负迁移的影响；在语法和语义上，有的偏误形式主要是受韩国语母语语义和用法负迁移的影响，因为学生认为韩国语的“旅行”是及物动词，可以直接加宾语，于是形成了语言偏误最典型的产出模式——母语语义和用法诱发目的语偏误。如：

1）*三周前，我旅游釜山。

还有一种偏误形式是受到了韩国语语法负迁移的影响，比如：

2）*我北京旅游的时候，觉得很多地方真美丽。

名词“北京”位于动词“旅游”前面是韩国语语法的典型形式。而在语义上，“旅游”跟“游览”、“游行”“旅行”的混淆，反映的则是学生汉语习得的发展特点和过程特征。一般随着学习的进程，“旅游”跟其他词语混淆的可能性会大幅度降低。

2. 4 典型偏误展示

“典型偏误展示”呈现的是搜集到的典型偏误例句。以汉语程度副词“很”为例，其典型偏误例句主要有以下 13 种：

1）“比”字句中的“很”的偏误（*他比我很漂亮。）

2）祈使句中的“很”的偏误（*明天考试，你很早睡吧。）

3）“是”字句中的“很”的偏误（*我喜欢她的歌，因为她是有名的。）

4）“得”字句中的“很”的偏误（*他很哭了。 *我很多吃了。）

5）补语句中的“很”的偏误（*我很吃饱了。 *我很多吃了。）

6）主谓谓语句中的“很”的偏误（*他很努力学习，所以很好成绩。）

7）“太”、“真”与“很”的误代（*我喜欢那件太/真漂亮的衣服。）

8）一些词或结构不能跟“很”结合（很大量、又很…又很…）

9）词性因素引起的“很”的偏误（*他很变了。 *我的水平很提高了。）

10）母语负迁移因素引起的“很”的偏误（*今天天气好。 *我很多吃了。）

11）规则过度泛化引起的“很”的偏误（*我很腿疼。 *很一样）

12）违反“有界”原则引起的“很”的偏误（*先处理很难的事情，其他的以后再说。）

13）“很”位于某些介词结构前面（*妈妈很对我关心。）

课堂教学中，解决了以上 13 种偏误形式，留学生对汉语程度副词“很”的认知就会跃升到一个新层次，习得难度就会骤降。

3 汉语词汇偏误生态描写由“群体研究”向“个体研究”的转向

词语在语言中有生态性，一个词在目的语中有自己的语言生态，在中介语中同样如此。词语在中介语中的生态信息，既包含着跟目的语相重合的正确信息，也包含着跟目的语相悖逆的偏误信息。我们目前的当务之急是：深入细致地研究留学生生成的那些跟目的语相悖逆的词语偏误信息，对其精细分类，并条分缕析地描写表述出来，争取将每个词的偏误生态构拟出来。

对中介语中词语的偏误生态进行构拟，可以分不同层面进行，比如语音层、文字层、语法层、语义层、篇章层、语用层和语体层等。对这些不同层面的词语偏误生态分别进行描写，汇集起来，就可以构拟出该词语偏误生态的总体样貌，从而为 CSL 课堂教学、教材编写和词典编纂提供全面、细致、有价值的信息。

正是基于以上认识，我们认为，对汉语词汇进行基于规模语料的偏误生态描写是一项非常具有前瞻性，也非常有意义的基础性研究工作，它对提高 CSL 课堂教学效果、提升教材编写质量、改善对外汉语词典内容的适用性都具有非常重要的价值。我们对汉语中介语中词语的偏误生态描写，应该逐步由过去的“多词一面的群体研究”向“一词一面的个体

研究”转向。

4 结语

语料库语言学“研究的目的是描述语言的使用，而不是语言的能力”。（ 黄昌宁、李涓子，2007：17）。语料库语言学的“学科交叉性”决定了语言理论研究方法的选择，它必然是代表实证主义语言观的描写取向，并且必须以应用价值为先导。（易绵竹等，2000）只有在对语言现象进行精细和穷尽描写的基础上，才有可能就某些语言现象做出形式化的解释。（ 郑定欧，1999：3-4）

学者崔健（2008：37-38）曾指出：面向韩国学习者的汉语偏误研究，起步较早，但是也存在一些问题，概而言之：始于举例、止于分类的研究多，缺乏系统的考察和梳理。为此，他呼吁要从语音、基本词汇、基本句式、常用功能词、篇章连接手段、范畴表达、语用规则、跨文化交际、双语对齐语料库等方面对汉语和韩语进行全面的对比、微观的考察和细致的描写，以便更好地服务于汉语作为第二语言的课堂教学。由此，我们认为，汉语中介语语料库的后续开发和应用，若想跟课堂教学实现更紧密的对接与融合，除了继续进行语料扩展、论文撰写、教材编写和词典编纂等工作以外，基于汉语中介语偏误反馈的国别化汉语词汇教学信息库的建设很可能是未来汉语语料库建设的方向之一。

参考文献

[1]崔　健. 关于加强国别化汉语教学的几点思考[A]. //郭鹏、赵菁主编. 汉语国际教育研究[C]. 北京：北京语言大学出版社, 2008.

[2]崔希亮、张宝林. 全球汉语学习者语料库建设方案[J]. 语言文字应用, 2011（2）：100—108.

[3]冯志伟. 从语料中挖掘知识[A]. //肖奚强，张旺熹. 首届汉语中介语语料库建设与应用国际学术讨论会论文选集[C]. 北京：世界图书出版公司, 2011:9—22.

[4]黄昌宁、李涓子. 语料库语言学（第二版）[M]. 北京：商务印书馆, 2007：3、20.

[5]易绵竹、薛奎恩、李　民. 网络背景下语言信息处理的理论研究[J]. 外语学刊, 2000（2）.

[6]郑定欧. 词汇语法理论与汉语句法研究[M]. 北京：北京语言大学出版社, 1999.

附录：信息库词汇基础附码一览表

1. 词性附码

n（名词）、v（动词）、vu（助动词）、a（形容词）、d（副词）、c（连词）、p（介词）、m（数词）、q（量词）、u（助词）、r（代词）、e（叹词）、o（拟声词）、lhc（离合词）

2. 词法附码

pz（偏正结构）、bl（并列结构）、bc（补充结构）、qz（前缀结构）、hz（后缀结构）、zw（主谓结构）、db（动宾结构）、jb（介宾结构）、cd（重叠结构）

3. 词调附码

1(一声)、2（二声）、3（三声）、4（四声）、0（轻声）、10（一声+轻声）、11（一声+一声）、12（一声+二声）、13（一声+三声）、14（一声+四声），以此类推