中介语语料库建设中的语言错误标注方法

李　斌
(南京师范大学文学院，江苏，南京　210097)

[关键词] 中介语语料库建设；语言错误标注；可扩展置标语言

[摘　要] 中介语语料库的建设是对外汉语教学研究中的重要内容。留学生的各种语言错误，尤其是偏误信息，可以为研究者提供可靠的统计数据。然而，针对留学生文本中标注各种错误的方法尚没有较好地研究。本文从语料库加工流程的角度，探讨了这一问题，并借助 XML (可扩展置标语言) 提出了错误标注的具体实现方法。

[中图分类号] H08　[文献标识码] A　[文章编号] 1671 - 5306 (2007) 03 - 0055 - 05

Error Tagging Method in Inter - language Corpus Construction

LI Bin
(*School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu* 210097, *China*)

Key words: inter-language corpus construction; language error tagging; XML

Abstract: The inter-language corpus construction is an important issue in the study of teaching Chinese as a second language. The language errors of foreign students, especially the interlingual transfer errors are the reliable statistical data for investigation. However, there existed little research done on the method of tagging language errors in text. This paper is to discuss the problem and put forward a practical method of tagging language errors in corpus tagging procedure by XML (Extensible Markup Language).

0. 引言

1995年，北京语言大学发布了中国大陆第一个留学生中介语语料库 ——汉语中介语语料库系统 (储诚志、陈小荷，1993)。[1]从此，基于中介语语料库的对外汉语教学研究拉开了帷幕 (崔永华，2003[2]；高立群，2001[3]；李大忠，1996[4]；马跃，2003[5]；孙德金，2002[6]；肖奚强，2002[7])。此后 10年间，国内加工建设了一批中介语语料库，如南京师范大学留学生语料库[7]、哈萨克族学生汉语中介语语料库系统 (武金峰，2002)[8]、HSK动态作文语料库 (张宝林等，2004)[9]等等。综观这些语料库，在语料的抽样方法、篇章属性、录入整理方面都较为严整，为教学研究提供了坚实的基础。

作为基础资源，留学生的各种语言错误，尤其是偏误信息，可以为研究者提供大量的统计数据。如汉字的书写错误、词汇错误、句法格式错误等等。然而，在对语料中留学生语言

[收稿日期] 2007 - 03 - 06

[作者简介] 李斌，(1981 -)，男，南京师范大学文学院博士生，主要研究方向为计算语言学。

错误的标注上，各系统基本还停留在对错误的简单索引上，缺乏深层次的标注规范。本文就语料库中的语言错误的标注问题展开讨论。

为了方便论述，这里把语料错误（录入过程中的错误）和语言错误（留学生所犯的各种语言错误，包括偏误）作一个区分。在语料库的建设过程中，首要的任务是以原始材料为依据，对生语料，即仅仅是录入而没有做特殊处理的语料，进行校对和质量检查，把语料错误降到最低。然后再在原始的基础上进行各种加工处理，如分词、词性标注、句法结构标注等等。留学生在文本中各层面所犯的各种错误，如字词层面（错别字、缺字、拼音字、用词不当等）、句法层面（把字句、被字句、动补结构等），是中介语语料库特殊的、最为重要的标注内容，影响到语料库的使用价值。语言错误中既有常见的笔误，也有母语影响造成的偏误，两者都是教学研究中的重要对象。

1．现有的语言错误标注方法

中介语语料库因其特殊的言语作者，存在着大量的语言错误。错误标注，是进行自动标注（分词、词性标注、句法分析）的先决条件。现有的中介语语料库，都有语言错误的人工标注。这里，以北京语言大学的中介语语料库为例。

该语料库收录了 740位留学生的 1731篇语料，共 44218句，1041274字，进行了分词、词性标注及一些特殊的语言学标注。全部语料均登录有作者姓名、性别、年龄、国别、是否华裔、第一语言、文化程度、所学主要教材、语料类别、写作时间、提供者等 23项属性。“汉语中介语语料库系统”对抽样语料按照字、词、句、篇等不同层次进行了加工和标注，对语料样本的非规范形式（例如：错字、别字、繁体字、拼音字、非规范词等）作出索引标记，并登录其相应的规范形式。

可见，在该语料库中，针对字词级别的语言错误研究，可以直接检索使用。而对于较为复杂的句法错误调查，如对“把”字句的调查，则往往需要提取出语料中所有的“把 /p”（/p表示词性为介词），再进行人工鉴别。如果要研究更为复杂的句法格式，如动补结构，则更是如此。如果可以在语料库中直接对语言错误进行错误标注，则会大大方便检索和统计。

肖奚强[7](P306 - 307)在偏误用例的分类中指出：同样作为“也”，“罗马大学也上汉语课的时间不多”，“今年也庄稼长得不错”两句中第一个“也”，把原来修饰名词短语的“也”误用作修饰动词短语了。第二个“也”，是将有关联作用的副词简单地等同于连词。如果在属性标注的时候予以区分，便可以在语料库统计的时候得到“也”字的偏误类型和某种偏误类型中出现“也”的情况。这样就避免了到底应该把同样的一个错误的句子归在哪一类的难题。

此外，在分词和词性标注的过程中，由于现有的分词和词性标注系统并没有自动校对或者查错的功能，因此，语言错误会带来自动标注的困难。如“学习”写成了“xue习”或“字习”。而研究者对中介语语料库的一大期望就是可以提供偏误数据。如果一个带标语料库不能提供这样的功能，这样的语料库的实用价值必将大打折扣。因此，现有的中介语语料库都对语言错误进行了人工标注，但具体的做法各异。有的是在自动分析前进行标注，有的是在自动分析后进行标注和索引。

从语料库的加工流程来看，如果先对原始语料进行人工校对，更正留学生的词汇和句法错误，再使用软件进行自动标注的话，势必要求对标注软件进行调整和改进，以避免对更正信息进行标注。这样做也不便于自动标注软件的更新与替换。而且，标注软件的正确率是有限的，自动标注以后还得进行人工校对。一个简单而实用的步骤是，先用标注软件进行自动标注，然后进行人工校对。校对的内容有两大项：一是校对出自动标注的错误。如，汉语分词中常见的交集型歧义错误，如“她从小学钢琴。”；兼类词词性标注错误，如“为 \ [v] 他 \ [r] 高兴 \ [v] ”。二是对语料当中留学生犯的各种错误进行标记。

2．XML 与语言错误标注

2.1　错误标注的基本要求

错误标注的基本要求为：标明错误的语句；标明错误类型 (大类、小类)；标明正确形式；便于标注、检索、统计。XML 可以满足这个要求，应用于语言错误的标注。

2.2　XML 简介

可扩展置标语言 XML (Extensible Markup Language) 是一种置标语言，它依赖于描述一定规则的标签和能够读懂这些标签的应用处理工具来发挥它的功能。XML 提供了一个标准，利用这个标准，用户可以根据实际需要定义自己的新的置标语言，并为这个置标语言规定其特有的一套标签。准确地说，XML 是一种源置标语言，它允许用户根据它所提供的规则，制定各种各样的置标语言。比如，可以给文本中的姓名加上标签。看下面这个例子：

起始标签 对象 结束标签

<姓名 >张三 < 姓名 >。

标签可以根据用户的需要，进行自定义。需要注意的是，标签必须是成对的，而且必须有相应的检索软件来支持该标签系统。我们可以制定一套标签，应用于中介语文本的标注。比如，对于一个生文本，内容为一篇留学生的作文，可以做如下标注：

< title >我的爸爸 < / title >
< author >
< name >崔永海 < / name >
< nationality >韩国 < / nationality >
. . . . . .
< / author >
< content >我的爸爸是一个 . . . . . < / content >

这里仅仅是一个例子，说明 XML 的标记功能。对于现有的语料库，不必如此进行标注，只需对文本中的语言错误进行专门的标记。

3. 利用 XML 进行错误标注

利用 XML 提供的标记功能，就可以对中介语语料库中的错误进行标注了。首先要设计一套标签，标签的内容可以是语言错误的类别，如，针对词语错误，可以设计为 <word >标签；标点错误，可以置为 < punc >标签；句法错误可以置为 < syn >标签等等。

3.1 **错别字**

一般地，在原始语料录入的过程中，都已基本解决了错别字的表示方式。如，不成字的字、笔画部件不完整或多余的字可专门造字或用形近字进行代替。下面，对已经处理好的原始语料中的错别字进行标注。如，“我字习汉语一年了。”现有的标注软件只能完成下面的标注（中科院计算所的 ICTCLAS[①]标注结果）：“我/r字/n习/vg汉语/nz一/m年/q了/y。/w”。对于这个结果，我们不能说错。但是这样的语料对于检索来说，既无法找到“学习”，又会在检索“字”的时候出现问题，因此有必要引进偏误标注。我们可以以多种方式进行标注。如常用的 XML 标注方法：“我/r <word>字/n习/vg </word=学习/v> 汉语/nz一/m年/q了/y。/w。”<>里面是错误所属的大类，并且要成对标记，内容为“错误修正/错误小类”。

3.2 **标点符号错误**

对于标点符号的错误，也可以按照以上方法来标注。如，“我吃了苹果，香蕉，和，梨。”，词性自动标注为：

我/r吃/v了/u苹果/n，/w香蕉/n，/w和/c梨/n。/w

人工标注：

我/r吃/v了/u苹果/n <punc>，/w </punc=“、/w”> 香蕉/n <punc>，/w </punc=“、/w”> 和/c梨/n。/w

3.3 **句法错误**

下面是对“把字句”错误的标注：

请你把这封信寄。[4](p134)

词性自动标注为：

请/v你/r把/p这/r封/q信/n寄/v。/w

人工标注：

请/v你/r <syn>把/p这/r封/q信/n寄/v </syn=“把字句”>。/w

标注内容可以是特定的句式类型，如“把字句”、“被字句”等。

3.4 **不易准确标注的内容**

还有一些不容易把握的错误类型。这首先需要从语料库本身的加工深度出发制定详细的标注规范，对难以把握或较为复杂的错误类型具体地规定其标注方法。如：

我们一起画蛇，先画蛇的人可以喝这壶酒。[4](p172)

人工标注为：

我们/r一起/d画/v蛇/n，/w先/d <syn>画/v </syn=“—结果补语”>蛇/n的/u人/n可以/v喝/vg这/r壶/q酒/n。/w

这个句子的错误是，“画”缺少结果补语。这样的错误标注难度较大，人工标注时可能会出现问题。或者一时看不出其错误类型，或者不同的标注者会把它列入不同的错误类型。解决的方案有两种：一种是简单地标记为句法错误，如只标记为“句法错误”而不做修改；

① 中科院计算所的 ICTCLAS软件 V1.0的下载地址为：http://www.nlp.org.cn/docs/download.php?proj_id=6&prog_id=1。

另一种是标记为详细的错误类型，并予以更正，如 < syn = “–结果补语 ”, “画 /v完 /v ”>。

虽然上面的例子还不完全，但已经可以看到使用 XML来标注语言错误是多么方便、实用。只要是语言错误或者不够地道的汉语句子都可以进行标注，前提是根据语料加工深度的需要来设计一套错误类型标签。一些论文已经对各种类型的语言错误进行了定性分析，划出了各子类，如字级、词级、句法级等等。另外，需要设计特殊的辅助校对软件，改进检索模块，对于用户选中的特定字符串进行 XML标注，并且可以随时调出加工规范的相应内容，以便参考。

4. 结语

本文主要讨论了如何实现对中介语语料库中的各种语言错误进行标注的问题，对语料的加工流程提出了改进，并采用 XML进行了一定的标注尝试，以期可以服务于语料库的深加工过程，为对外汉语教学研究有所帮助。

[参考文献]

[1] 储诚志，陈小荷 . 建立“汉语中介语语料库系统 ”的基本设想 [J]. 世界汉语教学，1993, (3): 199 - 205.
[2] 崔永华 . 汉语中介语中的“把 ··放 … ”短语分析 [J]. 汉语学习，2003, (1): 50 - 55.
[3] 高立群 . 外国留学生规则字偏误分析 ——基于中介语语料库的研究 [J]. 语言教学与研究，2001, (5): 55 - 62.
[4] 李大忠 . 外国人学汉语语法偏误分析 [M]. 北京: 北京语言文化大学出版社，1996.
[5] 马　跃 . 学生语料库与第二语言习得研究 [J]. 暨南学报，2003, (5): 87 - 92.
[6] 孙德金 . 外国留学生汉语“得 ”字补语句习得情况考察 [J]. 语言教学与研究，2002, (6): 42 - 50.
[7] 肖奚强 . 现代汉语语法与对外汉语教学 [M]. 上海: 学林出版社，2002.
[8] 武金峰 . 关于建立“哈萨克族学生汉语中介语语料库系统 ”的设想 [J]. 伊犁教育学院学报，2002, (4): 79 - 84.
[9] 张宝林，崔希亮，任　杰 . 关于“HSK动态作文语料库 ”的建设构想 . [A]. 中国应用语言学会编 . 第三届全国语言文字应用学术研讨会论文集 [C]. 香港: 科技联合出版社，2004. 544 - 554.

【责任编辑　蔡　丽】