基于语篇的中介语语料库回指偏误标注研究[①]

高玮
中国传媒大学 对外汉语教育学院 100024
marygao22@sina.com

摘要：本文从篇章角度对中介语语料库中回指偏误进行了分类，并在此基础上确立了基本标记和专用标记，提出了标记组合方式和标记的规则等具体实现方法。

关键词：回指；偏误；标记

A Corpus-based Analysis of Anaphoric Errors' Tagging in Discourse

Gao Wei
Communication University of China 100024

Abstract: This paper first establishes the analytic framework for anaphoric errors' annotation. It then divides the annotation into the basic and specialized types. Based on this, the author puts forward the combining forms of the anaphoric errors and some practical annotation rules.

Key words: anaphora; errors; annotation

0 引言

在对外汉语教学我们常常发现“学生在表达时，常常是一些简单句式的相加，而不是富有逻辑关系的语段。有些学生虽然具有组词造句的能力，但缺乏话语能力和篇章能力，极大地影响学生交际能力的提高。”（孙瑞珍，1995）而“汉语是语段取向的语言”（曹逢甫，1998），因此我们有必要从大于句子的篇章层面对偏误进行分析。本文在对三十万字中介语语料库中的偏误进行分析的过程中发现，回指偏误是其中出现频率比较高的一种偏误，虽然一些回指形式如人称代词、指代词等，在孤立的句中静态地考察时，其偏误显而易见，但是将其放在上下文语境中进行考察时，则呈现出一种比较复杂的动态变化，同一句话，从孤立的句子看可能是一种偏误，结合上下文语境，可能偏误就不同了，有时甚至会发现原本在单句中无法发现的偏误。这也正是语篇偏误标注的难处所在。如下例：

（1）走出教一楼，右边有一条宽的道路，这条路正好贯通校园的中间，还从东门到西门连接，这路的两边上有很多老的树，很美。（路）

从句子层面看，“这”和“路”之间缺失了量词“条”，当我们从语篇的角度来分析

时，发现如果没有“这”，这段话会更加连贯，因此对偏误的认定就由“量词的缺失”变成了“代词的多余”。

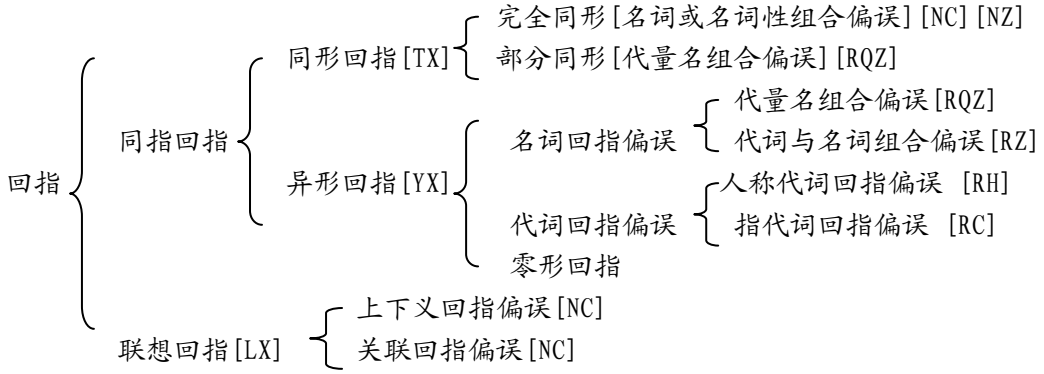
在实际的语料中我们还发现，指称作为篇章衔接的纽带，是相互影响的，有时单看一个句子没有问题，是完全可以接受的，但联系上下文就发现其可接受度大不相同了。请看下例：

（2）那时我真不知道怎么感谢她。因为在北京这本辞典很难找，而且……（这种）

如果只看这个句子，“这本词典”完全没问题，而问题在于从篇章整体来看，上文说的是作者自己的《越汉辞典》丢了，朋友费了很多周折才帮“我”买到，作者是要强调这种辞典的稀缺性，所以“很难找到”，这里用“这种”更恰当。本文将从篇章角度探讨中介语语料库中回指偏误的标注问题。

1 留学生回指偏误状况综述

首先，我们对指称形式的界定是整个名词性成分，而不只是中心语部分。我们前期对指称偏误的分析和标注过程中，发现留学生在不同阶段出现的指称偏误情况不同，既有初中级阶段同形的名词回指偏误，也有在中高级阶段出现较多的对上文描述的情形的回指，或者上文所述内容的总括性回指，如“这样”在篇章回指中的偏误。因此，我们对回指偏误的分类是根据实际的偏误情况而确立的，同时也参考了廖秋忠（1992）对指同的表达分类和徐赳赳（2003）的“汉语名词回指框架”，我们确立的“回指偏误标记框架”具体如下，方括号中是对该项偏误的标记。



其中，代量名组合偏误在留学生作文中出现的频率比较高，由于与数量有关，许多学者（孙朝奋 1994；许余龙 2004）研究表明数量词语的使用往往与语篇中话题的延续有直接的关系，因此我们把这些偏误单列出来，还区分了同形和异形；联想回指的偏误在留学生作文中的数量不是很多，而且多为名词偏误，因此我们不作更细的区分。

此外，陈平(1987)把汉语的回指形式分成零形回指(zero anaphora) 、代词回指(pronominal anaphora) 和名词回指(nominal anaphora) 三种。汉语中零形回指是经常使用的一种回指形式，也是留学生常常出现偏误的地方。我们在标记中通过回指的“多余”和“缺失”标记，把该用零形回指而没用的和不该用却用了的这两类偏误都能标记出来，因此，对零形回指就不使用特别的符号进行标记了。

2 标记的确立

在标记确立的过程中,我们参考了目前已有的、比较权威的语料库标记规范，以便提

供一个相对一致的平台，便于今后进行对比研究。

考虑到篇章偏误涉及的范围比较广泛，指称只是其中的一个方面，所以，我们借鉴以往语料库的标注经验，将标记分为基本标记和专用标记。基本标记主要提供词类标记、偏误类型、句法位置这些篇章标注中最基础，且最常用的标记，这些标记一般比较稳定，作为篇章标记的基本组成元素而存在，在各种不同的篇章偏误标记中都将或多或少地被使用；专用标记则是针对篇章偏误研究的具体情况而确立的，相对比较灵活多变，如回指偏误的标记，针对性强。

2.1 基本标记

这一层的标记内容在具体的标注中是广泛使用的，无论从什么角度对篇章进行分析，都可能涉及到这些基本的信息，而且基本标记作为一个相对稳定的标记元素，与其他标记之间有很强的结合能力，可以根据具体的研究需要对这些标记进行有效地组合。由于篇幅所限，我们在此不列出全部的基本标记，只是列出一些与回指相关的标记。

2.1.1 词类标记

我们首先借鉴现有的标记规则，以《现代汉语语法信息词典》作为基本参照，对词类进行标记，由于我们主要关注的是名词性成分，一般来说指称主要是名词性成分，以及相关的代词、数量词。因此，我们主要选取了以下的标记。

名词[N]取自英语 NOUN 的第一个字母；数词[M]取自英语 NUMERALS 的第三个字母；量词[Q]取自英语 QUANTITY 的第一个字母；代词[R]取自英语 PRONOUN 的第二个字母。②

2.1.2 偏误类型

我们的偏误标记是基于客观语料而进行的，因此把偏误分为多余、缺失、错用和语序错误这四类③，之所以采用这种分类术语，是因为我们的标注是以客观语料为基础的，这四类偏误都是针对语料中出现的偏误情况而确立的，是从客观的角度对偏误进行的区分。这种分类更具客观性，具体标记如下：

多余[&]；缺失[+]；错用[#]；语序错误[%]

以上的偏误标记与语料库原始文本中字词偏误的标记一致，有利于记忆和提取。

2.1.3 句法位置

一般名词性指称多出现在主语、宾语和定语中，我们根据偏误的具体情况，把一些偏误出现的句法位置进行了细分化，使所标记的内容更加精确。如在最初标注时把诸如“希望、知道、明白”等后面小句的主语统一都归入“主语”之列，标为“ZY”，但随着研究的深入，我们觉得两者有必要区分一下，因此，就改用“zY”来标记小句宾语中的主语；又如介词宾语，因为与一般宾语的位置不同，常常出现在动词之前，而形式上的不同必伴有语义上的差异，因此我们增加了“PY”，标记介词宾语。具体如下：

主语[ZY]；[zY]；宾语[BY]；定语[DY]；分别取自“主语、宾语、定语”的拼音首字母；介词宾语 [PY]中的“P”取自英语 PREPOSITION 的第一个字母。

2.2 专用标记

专用标记是针对我们具体的研究需要而设立的，这些标记可以随时添加，但不是随意的，一方面尽可能地沿用基本标记中已提供的符号, 另一方面要避免与已有的标记重合

或冲突。

2.2.1 指称形式

首先，有必要将“词”和“词语组合”进行区分：

词语[C]取自“词”的第一个拼音字母；词语组合[Z]取自“组”的第一个拼音字母。

然后，将这些标记与所属的词类相结合，形成各种指称表达式的标记：

名词[NC]；人称代词[RH]；指示代词[RC]；其他代词组合[RZ]；代量名组合[RQZ]；数量名组合[MQZ]；领属性组合[LZ]。

2.2.2 回指方式标记

回指方式分为：（1）同指回指分为：同形回指[TX]分别取自“同形”的第一个拼音字母；异形回指[YX]分别取自“异形”的第一个拼音字母。

（2）联想回指[LX]分别取自“联想”的第一个拼音字母。

2.2.3 回指范围的区分和语义内容的标记

指称形式虽然是以词或词语组合的方式呈现的，但是指称的范围有时并非只限于简单的名词性成分的回指，因此我们有必要对回指范围进行不同层次的区分，并对其中使用率最高的名词性回指进行语义上的区分。

我们按照回指的范围，主要分为三个层次：

（1）词语或词语组合回指，按照语义内容可分为以下四种：

人称[H]取自“HUMAN”的第一个字母；时间[T]取自“TIME”的第一个字母；地点[S]取自“SPACE ”的第一个字母；事物[M]取自“MATTER”的第一个字母，除以上各类之外的名词。

（2）情形回指[Q]取自“情”的第一个拼音字母，多为前面描述的情形或状况，回指的可能是动宾结构、小句或句子组合。

（3）总括性回指[Z]取自“总”的第一个拼音字母，总括上文所叙述的内容，可能是多句组合，也可能是成段的内容，或是前面所有的叙述。

3 标记的组合与标记规则的确立

3.1 指称偏误标记组合方式

标记组合原则：遵循从词或词语组合到句法位置再到篇章关系这样一个从小到大、从具体到抽象的顺序进行标记。

组合方式分为两大类：

（1）名词回指标记：这是回指偏误中数量最多，情况最复杂的一类，因此，需要比较细致地标出相关内容，并进行有效区分，主要分为以下三个组成部分：

指称形式+偏误标记组合→句法位置→回指方式+语义内容

（2）情形回指和总括性回指的标记：比名词回指标记简单，不标句法位置。

指称和偏误标记组合→指称方式[异形回指]+情形回指/总括性回指

3.2 标记规则

我们在反复调整之后，采用了“整体描写，凸现偏误”的标记方法。如果我们只是关注具体的偏误，这样可以比较准确地标出偏误，也能减少很多的标注工作量，但由于标记的范围比较小，就是名词偏误、数词偏误、量词偏误等，这些语法上的偏误，无法

反映与这些偏误直接相关的各种篇章因素，难以提供统计学意义上有价值的数据，如有些复杂组合可能只是中心语的偏误，我们如果只标出名词，就无法提供整个指称的信息。而另一方面，如果把偏误所在的整体指称形式都标出来，又会使偏误不明显，因此，为了凸现对偏误的标记，我们在对偏误所在的整个指称形式进行标记的基础上，还通过句法位置、语义内容等对具体的偏误进行限定，这样就可以避免顾此而失彼的情况发生。具体限定如下：

3.2.1　区分整体偏误和部分偏误

以下都是领属性组合的偏误，而且都标的是多余偏误，但具体情况不同，有的是整个指称的多余，如“我妈”，有的是定语的多余，如“我的”。我们通过句法位置标记就可以把偏误情况凸现出来了。这样就能有效地区分是整体偏误还是部分偏误。

（3）我来中国以后，我妈妈常常给我打电话，[LZ&-ZY-TXH][&我妈]说：“……”

（4）我非常生气，对[LZ&-DY-YXH][&我的]妹妹说：……

3.2.2　以偏误为导向确定最佳标记

有时同一个偏误，可能有多种不同的标法，我们以真实反映偏误情况为目标，来确定我们的标记。如下面的偏误，可以标为“介词宾语”，也可标为“定语”，标定语就能比较明显地知道这是定语多余的偏误，因此我们就标为“定语”。

(5)妈妈每天都和[LZ&-DY-YXH][&我的]妹妹锻炼身体，每天都吃药。很难受。

3.2.3　偏误标记细分化

对于该用名词却用了代词，或者该用代词用了名词的错用偏误，为了与其他错用进行区分，我们把这类替代错误以[/]标记，这样能提供一些所需的数据。一般情况下，多为该用代词或零形式而用了名词的偏误，但是偶尔也有相反的情况如例（7）。

(6) 孔子是怎么样的人呢？[NC/-ZY-TXH]孔子不高也不矮。

(7) 我家三口人。妈妈、弟弟和我。我的爸爸呢？我小学三年级的时候，他去世了。所以，我特别爱[RH/-BY-YXH]她。

此外，句法位置的细分化，指称形式的细分化等也都是根据实际语料而确定的。

3.2.4　细分化与简化相结合

由于与指称相关的内容很多，可标注的信息也多，如果我们全都标出来，将是一个庞大且复杂的标记体系，而过分简单就不能全面、真实地描述语篇偏误的复杂状况。因此，我们综合考虑了各种因素，对标记的内容进行了多次整合，选取最有说服力的语篇偏误信息进行标注，在细分的同时进行相应地简化，如前面句法位置部分和标记组合中的细分与简化相结合的做法，使标记具有科学性和合理性，力求为研究提供真实有效的数据支持。

3.2.5　规范性与开放性结合

我们在建立标注集的过程中,参考了目前已有的语料标注规范,同时结合我们研究的具体需要，尽力使基本标记和专用标记能有效地整合，以避免标记的过于庞杂、难以掌握和辨识，使标记准确规范且容易明白和掌握。同时，我们的标记集还具有开放性特征，特别是专用标记，可以根据标记过程中出现的新情况，随时添加，两种标记的划分既是为了规范标注体系，也是为了方便标记的这种动态调整。

4　存在的问题及小结

因为我们的研究本身是处在探索阶段，标注过程中会有很多新的偏误情况出现，要随时添加，使我们的标记体系不断地得到充实和完善。

4.1 主观性的干扰

由于语篇的偏误要从不同的角度进行分析，所涵盖范围比较广，不可能进行自动标注，必须是人工标注，标注过程难免会受主观性因素的影响，如个人的语感差异，个人对篇章理解差异等，不同的人可能会有不同的解释，得出不同的结论。

虽然语篇标注中人为因素的影响是无法避免的，但是在标注过程中我们尽量坚持统一的标准，以保持标注的一致性，从而减少人为的偏差，力求客观地反映学生语篇偏误的真实状况，使我们的标注更具客观性。

4.2 小结

语料库本身只是一个工具，并不自然生成我们所需的信息资源，需要对这些原始语料进行深度加工，才能给我们的教学和研究提供有价值的统计数据。一个比较全面科学的回指偏误标记集的建立不是一蹴而就的,要在实际的标注过程中不断地进行修改和调整，因此我们的标记体系也经历了一个不断完善的动态调整过程。

附注：

① 本研究得到国家社会科学基金的支持，项目批准号为 07CYY012。论文写作过程中得到导师侯敏教授的指导和帮助，特致谢忱。

② 因为有介词[P]取自英语 PREPOSITION 的第一个字母。

③ 关于偏误的分类问题，周小兵在《第二语言教学论》中将偏误类型分为:语序错误、搭配不当、词语残缺、词语误加、词语混用、句式杂糅六类。鲁健骥(1994)将二语习得中的偏误分为遗漏(mission) 、误加(Addition) 、误代(Overrepresentation) 、错序(Misordering) 四类。

参考文献

[1] 曹逢甫著 谢天蔚译.主题在汉语中的功能研究——迈向语段分析的第一步.第二版.北京：语文出版社,1998：39

[2] 陈平.汉语零形回指的话语分析.现代语言学理论研究——理论·方法与实事.重庆：重庆出版社，1991：181-209

[3] 廖秋忠.现代汉语篇章中指同的表达. 廖秋忠文集.北京:北京语言学院出版社,1992：45

[4] 孙朝奋.汉语数量词在话语中的功能.功能主义与汉语语法. 北京:北京语言学院出版社,1994：139-158

[5] 孙瑞珍.《高级对外汉语教学语法等级大纲》的研制与思考.中高级对外汉语教学等级大纲.北京：北京大学出版社,1995：345

[6] 徐赳赳.现代汉语联想回指分析.中国语文,2005（3）：195-204