

TOCFL 学习者语料库的偏误标记*

张莉萍

(台湾师范大学)

1. 研究背景

对学习语料库而言, 载有偏误信息是非常重要的特征, 目前学习者语料库偏误标记的方式, 大致上有两大系统 (Tono, 2003), 一是语言学分类系统 (linguistic category classification), 以下简称 LCC; 一是与目标语差异系统 (target modification taxonomy), 以下简称 TMT。

早期 TOCFL 学习者语料库为了便于研究者分析 (张莉萍, 2013), 也做了简要的偏误标记工作, 如 (1) 中的 [R] (redundancy) 指的是冗赘, 这里显示学习者多用了“了”; (2) 的 [M] (missing) 指的是缺漏, 这里学习者少用了“的”。

- (1) 老师一看到他就很生气了[R]。不仅是这样, 他的功课还没做完,
- (2) 很美 [M] 风景

这种标记方式采用的就是 TMT, 研究者从中可以知道学习者语言与目标语言差异的地方。但没有进一步的信息可以分析学习者多用了什么或少用了什么。一般而言, 越精细的分类, 越有助于二语习得研究或偏误分析, 然而, 分类太细, 也有不利之处, 例如, 标记人员对标记的掌握能力将影响标记工作的一致性。但是, 从上面例句的做法也可以看出, 过度简化的标记, 实际上要做大量观察或量化分析, 并无法达到研究者

* 本研究感谢“科技部”计划 (MOST 103-2410-H-003-043-MY2)、“科技部”“跨国顶尖研究中心计划” (NSC 103-2911-I-003-301)、“教育部”“迈向顶尖大学计划”以及台湾师范大学“华语文与科技研究中心”支持。关于本研究所使用的工具, 笔者特别感谢台师大曾元显研究员所带领的团队: 李龙豪、李奎庆、廖柏舜等人开发比对软体、偏误编修软体, 供研究团队使用。至于偏误标记工作, 笔者特别感谢郑婉铃、杨雅筑、杨志盛、刘美君同学的协助。

的需求,也导致研究者必须针对不同研究主题,重新检视文本,甚至重新标记新的信息,如此不断的重新检视、标记着实耗费了大量精力与时间。因此经过一年的实作,我们重新思考如何让语料库加注偏误标记的这项工作,能做到全面性与有效性。

所谓全面性,是希望这套偏误标记可以涵盖学习者偏误的各种表现。所谓有效性,有两个层面,一是指通过偏误标记,使用者可以很有效地分析语料;二是指标记结果的一致性和精确性。要达到这两个目标不容易,与偏误标记的设计、标记工作的步骤、流程息息相关。

因此接下来第2节将梳理偏误分类的相关文献,进行探讨;第3节则说明 TOCFL 语料库偏误标记系统;第4节说明进行偏误标记的步骤流程;最后做一总结。

2. 偏误分类文献探讨

关于偏误分类的相关文献,我们分三部分来谈,一是从理论层面来探讨偏误分类的问题,二是从实务层面来看语料库偏误标记的分类方法,三是检视目前华语学习者语料库的偏误标记工作。

2.1 偏误分类理论

讨论学习者偏误类型的文献相当多,较早有系统性讨论的当属 Dulay, Burt & Krashen (1982, Ch. 7), 他们从语言学分类、表面策略(surface strategy)、比较分析、交际效果四个方面进行了讨论。

语言学分类即依据语言学概念来为偏误做描述性的分类,例如,偏误可以分成构词层面或句法层面。句法中可以再分为定词偏误、名物化偏误等类型分项叙述。定词下可以再依遗漏、错用等项目来叙述。

表面策略分类则是从学习者句子的表面结构和目标语差异的状况将偏误分成四类,分别是缺漏(omission)、误加(addition)、形式错误(misformation)和错序问题(misordering)。

比较分类法则是基于 L2 偏误结构和某些其他类型结构的比较,例如,和目标语习得结构或学习者母语结构的比较。通常可以分为:(a)发展性偏误(developmental errors),这类偏误类似于目标语儿童习得语言阶段性的错误;(b)语际性偏误(interlingual errors),这类偏误起因于学习者母语结构迁移;(c)歧义性偏误(ambiguous errors),指的是偏误来源可能是前述的发展性偏误,也可能是语际性偏误;(d)其他类,也

就是不能归因于前述偏误来源的。

交际效果的分类法,不同于前三项分类法,所关注的是会产生沟通问题的偏误和不会产生沟通问题的偏误。Burt & Kiparsky (1972) 据此分为两大类,一是整体性偏误(global errors),偏误主要类型如句子主要成分语序的偏误和句连接词的错用、遗漏和错序,例如“*English language use many people. *He started to go to school since he studied very hard.”等;二是小范围偏误(local errors),是句中单一成分的偏误,例如“*Why we like each other?”此类偏误不影响沟通。

Dulay 等人指出,使用表面策略分类法来分析偏误可以使研究者系统性地关注学习者的偏误,因为从表面结构的差异,可以比较清楚地看到学习者在重建一个新语言时认知的过程。我们也可以看到后来研究者多引用这类分析法,例如,广为引用的 James (1998),只是将 Dulay 等人的 surface strategy 这类改称为 target modification taxonomy。因为他认为这类基本上就是表达中介语和目标语之间的差异,另外在四类之外,加了混合(blends)这个小类,定义为不止一个可能的目标语形式,如“*according to Erica's opinion”可以修正为“according to Erica”,也可以修正为“in Erica's opinion”(p. 111),因此形成五小类:缺漏、误加、错用(misselection)(原 Dulay 的 misformation)、错序和混合。文献显示,Dulay 等人在 20 世纪 80 年代已经为偏误分类做出了全面性的描述,经过 James 进一步的修正,这种分类法以及术语已经为多数学者或语料库建设者沿用至今(详见 2.2 节)。

不过,对于 James 加了“混合”这个小类,笔者认为并不合适。因为我们实际去看学习者的真实语言就可以发现,他们很多的产出在标的语中都可能不是只有一种表达方式(Ellis & Barkhuizen, 2005: 340; 肖奚强, 2011; 张莉萍, 2013),如果要标记者去推测这个偏误可以对应到标的语两个以上的表达方式,则是非常危险的;采用这种方式来标记,也无益于量化与分析,反而容易沦为标记时的垃圾桶。

2.2 语料库偏误标记的类型

最早的学习者语料库是 20 世纪 80 年代晚期由朗文出版集团所建立的朗文学习者语料库(Longman Learners' Corpus),约 1 000 万词规模,为商业用途,Turton & Heaton (1997) 也利用语料库编了学习者常见偏误辞典。另外有剑桥学习者语料库(Cambridge Learner Corpus, CLC),约 4 000 万词,语料来源是剑桥考试中心所举办的写作考试文本,考生是

来自不同国家的超过 16 万的英语学习者, 建库者对其中半数语料进行了偏误标记 (Nicholl, 2003)。而目前国际上最有计划的广泛收集语料的, 也最为人所知的是国际英语学习者语料库 (International Corpus of Learner English, ICLE) (Ellis & Barkhuizen, 2005: 338), 主要负责人是比利时的 Sylviane Granger 教授, 语料是由每篇大约 700 字的进阶级作文所组成。这些作文基本上都是从不同国家的课堂上取得的。目前有 17 个合作伙伴国加入, 语料约有 200 多万词, 也对语料做了偏误标记。

关于对语料库偏误标记的工作, Diaz-Negrillo 和 Jesus Fernandez-Dominguez (2006: 87) 两人回顾了 1990—2005 年间学习者偏误标记系统的发展, 归纳结果为 LCC 类标记是每个语料库系统都标示的, 如比利时鲁汶大学所开发的 FreeText 系统有 11 个主要的词类、54 个次分类、12 个标点符号类, 共计 100 多个标记符号; 日籍学习者英语语料库 (NICT JLE, Izumi et al., 2005) 有 11 个词类, 共 46 个标记; 剑桥学习者语料库 (CLC) 有 9 个词类, 总共有 88 个不同标记 (Nicholls, 2003)。以下是 CLC 偏误标记的例子:

(3) *He died<#MP>|.We</#MP> buried him the next day.* (引自 Nicholls, 2003: 574)

<#MP> 是 omission of punctuation 的缩写, 表示标点遗漏, 并将标点补上。这个标记体现了两个概念, 一是遗漏, 属于 TMT 概念; 一是标点, 属于 LCC 概念。但不是每一个偏误标记的系统都完整呈现这两个概念, 例如, NICT JLE 几乎都是 LCC 分类系统, 没有 TMT 概念。^① 例子如下:

(4) *I belong to two baseball <n_numcrr="teams"> team</n_num>.* (引自 Izumi et al., 2005: 75)

其中 n 表示名词类错误, num 表示数量类偏误, 在引号内的是更正后的信息。

从现今国际大型学习者语料库的偏误标记工作可以看出, 偏误标记基本上采用的是自 Dulay et al. (1982), James (1998) 以来的分类架构, 大都涵盖了 TMT 和 LCC 两大概念。TMT 的分类没有太大弹性, 就是遗漏、冗赘、错序、词语选用错误四大类, LCC 则因各系统对语言学分类

① 在无法归入词类的偏误时, 才设置了其他类, 例如, 错序。这样的标记仅有 5 个。详见 Izumi et al. (2005) 的表 1。

的粗细,数量差异相当大,从几个到100多个标记都有,当然标记的粗细,视各语料库的研究性质而有所不同。

从各语料库的标记系统,我们也可以看到多数的标记集(tagset)都有阶层性的概念。例如例(4)中,n是主要类别,属于数量的num则是它下面的次类。ICLE的标记集也是如此设计,主类有七大类,其中G代表语法类,GV则是语法类中的次类“动词”,GVT代表更低阶层的时式偏误类(Dagneaux, Dennes & Granger, 1998: 166)。在第3节,我们可以看到TOCFL的标记集也有阶层性,目的是为了让标记人员在判断时能更清楚脉络,增加标记时的效率与一致性;另一方面,也让标记集更有弹性,也就是主类不会变动,但次类可以因为研究的需求而具有扩充性。

2.3 华语学习者语料库偏误标记情况

上述所描述的学习者语料库都是英语为标的语,关于汉语带有偏误标记的学习者语料库则以北京语言大学的“HSK动态作文语料库”为代表(张宝林、崔希亮、任杰,2004)。这个语料库的内容是母语非汉语的外国人参加高等汉语水平考试(HSK高等)所写的作文,于2003年开始建设,2006年底试行开放网络版本1.0,至2009年正式公开1.1版本。这个语料库搜集了1992—2006年部分外国考生的作文答卷,共计11569篇,424万字。从字、词、句、篇、标点符号等角度分类,约50种错误标记。

该语料库依据偏误所在位置的 language 成分分为字、词、句、篇章等不同层次,似乎层次分明,但实际执行时却受到考验。如例句(5)的标记,^①显示考生遗失了“在”字,语料库标示为缺词(CQ),也就是判断这个偏误是在“词”的层次上。而例句(6)则标记为病句(CJ),也就是在“句”的层次上,而非“词”。事实上考生在(6)中将“认为”这个词误用为可以拆的词语结构。如果我们把“在……下”当成一个结构来处理,例(5)的偏误反而成了句式层次上的偏误。

(5)……有的农民{CQ在}不使用化肥和农药的情[B精]况下
[BD,]养农作物……

(6)……现代社会应当认“保护自己,尊重其他人”为口号{CJgd}……

另外,像句成分残缺或多余(CJ-)的标记之下又以主语、谓语、状

^① 例句(5)~(7)都是摘录自 HSK 动态作文语料库中的“语料标注与代码说明”部分的例子(<http://202.112.195.192:8060/hsk/help2.asp>)。

语等功能分为八小类,如例(7)标记为句层次(CJ),状语成分残缺(zy),而非词层次的缺词问题,也值得商榷。只因为“都”可以做状语,所以不做缺词(CQ)处理。但这样的处理没有太大意义,汉语可以充当状语的词语太多了,标记状语缺失没有太大意义,偏误的关键还是在于学习者不能掌握“都”的使用,该用而没用。

(7)……我认为这种现象……在很多人身上 {CJ-zy 都} 会有的。

从上面这些例子,可以看到 HSK 语料库偏误标记基本上还是基于语言学分类的考量较多,例如, {CJ-zy} 中 CJ 是句层次, zy 是状语,都是 LCC 概念。只是 LCC 的概念多呈现在句的层次,如上所述,有主语、状语、宾语、定语等等,另外也有句式的标记,如把字句、被字句、存现句等等,并没有词类的细致标记。部分标记则涵盖了 TMT 概念,例如, {CQ} 指的是缺词(即遗漏)。虽然标记工作到达篇章范围,但实际上关于篇章层次的标记只有一个,即 {CP…P}, 意义并不大。因此我们认为,在标记的上层划分词、句、篇章等类别是否有必要,值得商榷。

不过,这个语料库对于大部分的偏误语料都做了更正,也就是同时可以找到正确的表达形式,有利于精确检索。

3. TOCFL 语料库的偏误标记理念

在前言中,我们提过 TOCFL 语料库在早期做过简化的偏误标记,经过一年的实作,我们体会到这项工作的复杂与辛苦。虽然我们一开始也训练了人员对偏误标记的认识,以及标记的原则。但当时囿于人力与时效,没有采用两人同时标注一份档案的方式,没有使用互相校对的程序。之后我们进行检视,发现最大的问题就是标记不一致,以及太过简化的标记使得量化结果过于粗糙。这让我们重新思考如何才能让语料库加注偏误标记的这项工作做到全面且有效。以下我们从这个带有偏误标记的语料库所针对的使用对象、偏误标记的内涵、标记的范围等方面来说明这套偏误标记集。

3.1 标记的内涵与范围

对于学习者语料库带有偏误标记的必要性,见仁见智,主要看使用者的目的。对一般使用者而言,可能原始的文本与简单的检索功能就可以符合需求,可以大致观察学习者产出的语言是什么样子;对于进阶的

研究者而言,大量的语料如果没经过整理、分类,在使用上仍是非常不便,也很难看到全貌。例如,学习者缺漏的语言成分无法经词语检索得到。因此当我们决定进行这项偏误标记工作时,我们设定的对象是要提供给进阶研究者。然而,这又牵涉另一项议题,毕竟每个人研究的兴趣不同,要做到理想的全面性标记谈何容易。例如,对于语体、语用或篇章研究者而言,现阶段的标记内容仍不足以提供这方面的信息。

那么怎么让这标记系统有扩充的可能性,能顾及研究者的个别需求呢?我们采取阶层式的分类方式,不同于HSK语料库的标记方式,我们第一个阶层是TMT分类方法。也就是说,先根据学习者的语言和目标语的差异分成四类:遗漏、冗赘、选用错误和语序错误,分别以M、R、S、W这四个符号为代表。这个阶层的标记是固定不变的。第二个阶层则是LCC分类法,也就是每一个偏误在给定了主类后,根据语言学概念给予次类。这个阶层的标记则是具有扩充性的,现阶段我们依据一年来的实作经验,就已经观察到的现象,规范了36个次类。这36个次类分布在词、语、句型和其他这四个范围内。如下所示:

1. “词”的范围(16类)^①,如动作动词(v)、能愿动词(aux)、状态动词(vs)、名词(n)、代名词(pron)、连接词(conj)、介词(p)、数词(num)、指示词/定词(det)、量词(cl)、句末助词(sp)、时貌助词(asp)、副词(adv)、助词(de)、疑问词(que)、复数词缀(plural);

2. “语”的范围(11类),如主语(sub)、宾语(obj)、名词短语(np)、动词短语(vp)、介词短语(pp)、修饰语(mod)、时间词语(time)、地方词语(loc)、及物性(tran)、[数/定+量]结构(dm)、离合结构(vo);

3. 句型范围(7类),如复杂名词子句(rel)、特殊句型(把ba、被bei、让rang、是shi、有you)、其他句式(pattern)。

4. 混合范围(2类),形式错误(form)、表达方式错误(sentence)。

之所以细分为36类,是为了未来能提供细致分析与统计。前面我们也提到,有的学习者语料库的标记符号多达100多个,主要看研究者的兴趣或范围。基本上,我们的标记说起来也是100多个,因为假设每个

① 我们的标记中有一小类是要标示学习者使用词缀“们”的问题,由于只有此类属于词缀,为简化呈现,我们将它归入“词”的范围。

主类之后都跟着 36 个次类,那么就有 144 个不同的标记。但是对标记者或使用者而言,只要在第一阶段区分四大类,再根据词性(16类)或语类、句型等等给予次类,把 36 个类别掌握好即可,不需要记忆 144 个类别。这正是阶层式分类的好处,除了便于记忆,也具有系统性、扩充性。(8)~

(11)的例子即是展示不同主类下的同个次类。

(8) 公共汽车没有空的椅子,所以两个人都得站 [Masp]。

→公共汽车没有空的椅子,所以两个人都得站着。

(9) 我朋友也找工作找了很久才找到了 [Rasp]。

→我朋友也找工作找了很久才找到。

(10) 他爬山爬的 [Sasp] 半天。

→他爬山爬了半天。

(11) 我放假的时候去欧洲了 [Wasp]。

→我放假的时候去了欧洲。

这四个句子都是关于时貌助词的偏误表现,分别是少用、多用、选用错误和语序错误的问题。

3.2 次类的设计理念与标记原则

针对 LCC 分类(即次类)的精致性,视研究的需求而定。在这个系统中,我们不标记学习者用字的问题,因为这个语料库的来源是电子文本,也就是说学习者都是经由键盘输入中文,汉字问题与研究手写汉字偏误的类型不同,多是“选字”错误引起,关于汉字的问题,为了不影响自动断词的准确率,我们在建设语料库前就做了特殊的前置处理(张莉萍, 2013: 146)。在偏误标记时,我们略过字的偏误,标记的最小的语言单位是词缀,如“们”[plural]。大部分的偏误都是在词的范围,这时除了判定四个主类外,需要进一步标记出该词的词类,如上面(8)~(11)的例子。词类究竟要分到多细?我们在邓守信(2009)的八大类(名词、动词、副词、助词、连词、介词、限定词、量词)的概念之外,另外特别标记出时貌助词、结构助词、句末助词、代名词、疑问词等等,主要因为这些虚词或带有特别功能的词语,是学习者的难点,也是研究者热衷观察分析的语言点。

实际上,学习者所产出的语言不是只有选用词语这样单纯的问题。例如,学习者写出“我庆祝会不能参加”这样的句子,正确的表达应该是“我不能参加庆祝会”,因为在前文并没出现过“庆祝会”的信息,

此时在语料库中的标记方式就如(12)。这时如果因为“庆祝会”是名词而标记 [Wn] 则失去偏误分析的信息。因为我们知道这是学习者无法掌握宾语是否前移的问题,而非个别名词的问题。这也是为何在词类之外,标记也涵盖了“语或词组”的范围。

(12) 我庆祝会 [Wobj] 不能参加。

→我不能参加庆祝会。

(13) 我们到现在都是以写信为联络 [Mnp]。

→我们到现在都是以写信为联络的方式。

例(13)也是词组范围的一个标记例子,学习者缺失了“的方式”,这时缺失的单位大于名词,则选择次类名词组[np]的标记。在词组范围内,除了主语、宾语、介词组、名词组等标记外,还有一个比较不常见的标记,即修饰语 [mod], 主要是 [X+ 的] 结构,这也是二语习得研究者所关心的主题,在我们的系统中特别标示出学习者在这一类的偏误。如例(14)~(16)所示。如果仅标示 NP, VP 也会失去偏误信息,因此,我们设计了 [mod] 这个次类。研究者可以针对修饰语的多用、少用、错序等进行观察分析。

(14) 现在我知道什么样的 [Wmod] 买东西是值得的。

→现在我知道买什么样的东西是值得的。

(15) 因为你要到我的城市来看看 [Mmod] 生活怎么样。

→因为你要到我的城市来看看这里的生活怎么样。

(16) 我希望你改掉这两个你的 [Rmod] 生活习惯。

→我希望你改掉这两个生活习惯。

在华语教学中,特殊句型也是教学的重点项目。在这个系统中,我们针对把字句、被字句、让字句、是字句、有字句这五种句型做了标记。在标记原则中,则需要说明偏误涉及这些句式时,不论是多用、少用、错序或错用都要优先标示 ba, bei, rang, shi, you, 而非词类标记。另外,在句的范围内,还有一个问题也是语言学者所热衷探讨的,那就是复杂名词子句或称关系子句,我们标记为 rel, 如下面例子所示。这类偏误如果没有经过人工标记出来,很难进一步进行量化、质化分析。

(17) 我觉得餐厅在你家旁边 [Srel] 好像好吃。

→我觉得在你家旁边的(那家)餐厅好像很好吃。

(18) 希望我们的计划要见面 [Srel] 没有什么问题

→希望我们要见面的计划没有什么问题。

(19) 我觉得野餐是最好的办法认识新朋友 [Srel]。

→我觉得野餐是认识新朋友的最好办法。

上面我们讨论的标记大致涵盖了词、词组、句子的范围，还有两个次类——[form]，[sentence] 很难纳入这些范围内，姑且称这两类为其他类或混合类。属于形式 (form) 的偏误，我们归纳了三种情况：

a. 选用的词没错，只是形式错误 (例如重叠式、尝试貌)，如例 (20) (21)；

b. 可以理解语义，只是表达形式错误，如例 (22)~(24)；

c. 遗漏了固定格式中的某部分，如例 (25)。

(20) 他在路上走着想想 [Sform] 他的功课。→他在路上走着想他的功课。

(21) 谁都不知明天也 [Radv] 自己的工作或公司还会 [Sform] 存在。

→谁都不知明天自己的工作或公司还会不会存在。

(22) 幸运 [Mform] 我住的地方有越南餐厅。→幸运的是我住的地方有越南餐厅。

(23) 李文平敬于 [Sform] →李文平敬上

(24) 我觉得不舒服得不得了 [Sform]。→我觉得非常不舒服。

(25) 可是对他们 [Mform] 是开心的一天吧！

→可是对他们来说是开心的一天吧！

a 类的标记方法有别于学习者选用词的标记 (例如, Sv 或 Sn)，因为这类偏误重点并非选用了错误的词，而是词的构形问题。b 类的用意则是标记出学习者使用的错误形式，而这个错误形式根本不是一个词，如例 (23) 的“敬于”。c 类则有别于单纯标记缺失了名词组或动词组，因为这里所缺失的是固定格式中的某个成分。毋庸置疑，[form] 这一类搜寻出来的偏误语料会较纷杂，但因为初始标记时，无法预测各小类的数量，又不想让标记符号扩张得太多，造成标记者的负担。于是，我们采用单一标记来记录这几小类。

[sentence] 这个标记是指学习者的语言需要做整体性的修改。前提是标记者 (母语者) 能理解意思，如果通过上下文，标记者仍不得其门而入，只能放弃不标。通过这个标记，研究者至少可以进一步观察分析，

学习者采用何种策略来产出想要表达的句义。如例(26)(27)。

(26) 合格了交换学生 [Ssentence] 以后 → 符合了交换学生的资格
以后

(27) 她是最好的家教一样的感觉 [Ssentence] 。 → 感觉她是最好的
的家教一样。

经过上面的叙述, 我们知道标记人员除了需要受训练外, 还需要一份非常清楚的操作手册, 里面记载了标记的类别、意义、例句以及标记的大小原则、标记选用的优先顺序等等, 有兴趣者请详见附录一。

4. 偏误标记的步骤与结果统计

在进行偏误标记工作之前, 研究者必须招募人员、训练人员、制作偏误标记说明手册。人员的素质、训练的过程与手册的说明决定了未来工作的成败。在人员部分, 理想的人选需要具有语言学知识背景以及华语教学经验, 因为标记者既要有汉语语言学概念, 也要对学习者的语言有初步的体认。

至于人员的训练, 如何确定相关人员经过研究人员的讲解, 在实际标记时能达到一定的准确性是我们关心的问题, 这关系着标记结果的有效性。然而, 偏误标记这项工作, 既然是人工进行, 就难免带有主观性, 虽然我们制作操作手册就是为了把主观性降至最低, 但不同能力的学习者的偏误表现仍远超过手册所能规范。理想的步骤应该是两人共同标记一个文档。但是偏误标记是一项费时费力的工程, 在有限的时间与人力下, 这个理想几乎是遥不可及的。于是我们采取了折中的做法, 让人员在研读手册、接受训练之后, 每人标记同样的 50 篇作文, 而这 50 篇作文研究者事先做好了标准版本, 程式设计人员为我们设计了一个工具, 可以标示出每人和标准版本的不一致处, 这些不一致包括对偏误的认定以及对偏误标记的指定, 也就是可能有人认为是偏误, 给了标记符号, 但另一个人认为不是偏误, 没给标记; 也可能是两人都认定是偏误, 但给的标记类别不一样。这样的不一致结果输出给训练者检视, 训练者再针对新进标记者对标记的认知模糊处, 进行再教育。这样的训练期大概需要两周的时间, 这两周需要训练者、标记者与研究者之间密切地反复讨论, 而这个过程也让研究者看到操作手册的不足之处, 从而再次地检视修订其内容。

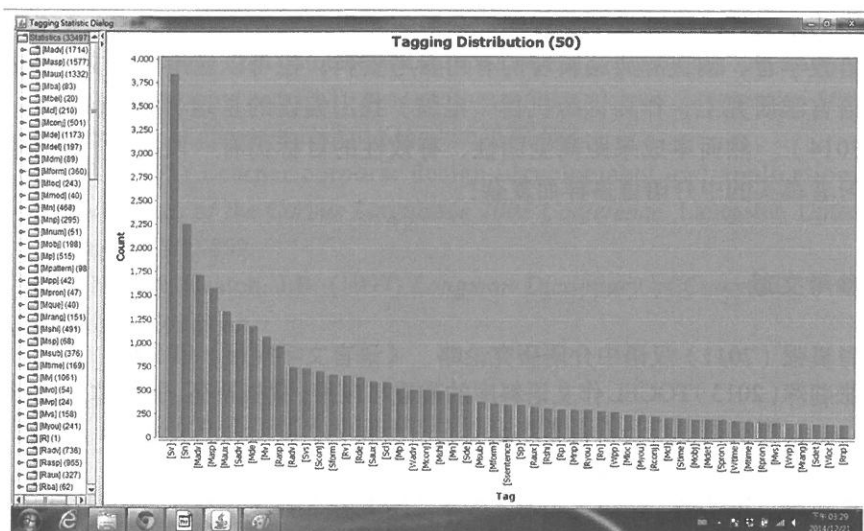


图2 偏误标记工具统计画面

表1 出现频率最高的前10名偏误次类

次类	v	adv	n	asp	aux	de	conj	p	vs	form
次数	5 546	4 145	3 124	2 740	2 338	2 328	1 437	1 162	1 072	1 017

从表1可以看到三大词类(名词、动词、副词)居偏误的前三名,这个不难理解,因为这三个词类中词的总数量多,犯错的次数自然也高。不过,时貌助词和结构助词分别居第4名和第6名,则可以说明这两类的确是学习者的难点,因为这两类词数量很有限。另外,搭配图2的统计图来观察,也可以知道动作动词和名词的偏误,呈现在四大主类中最多的是选用错误这一类;副词偏误则是最多分布在缺失类。这些都仍是粗略的数据,研究者未来可以从这些标记过的语料中再缩小范围进行深入的观察分析。

5. 结语

现阶段 TOCFL 语料库的偏误标记语料近 100 万字,未来即将提供给研究者检索使用。根据使用者的回馈反映,再决定是否持续进行更多语料的标记工作。因为人工标记偏误非常耗时费力,如果对于汉语作为第二语言的教与学的助益有限,研发者必须重新思考进行偏误标记这项工

作的必要性。目前这项工作的成果至少可以应用在语言教学上,作为二语教学者、偏误辞典编辑者的有用参考资料;也可以应用在中文计算机语言学研究上,作为训练机器(电脑)找出偏误的基础材料(Lee et al., 2014)。然而本成果距离全面性、有效性的目标仍有一段距离,毕竟学习者真实的中介语言多样而复杂!

参考文献

- 肖奚强(2011)汉语中介语研究论略,《语言文字应用》第2期。
- 张莉萍(2013)TOCFL作文语料库的建置与应用,载崔希亮、张宝林主编《第二届汉语中介语语料库建设与应用国际学术讨论会论文集》,北京:北京语言大学出版社。
- 张宝林、崔希亮、任杰(2004)关于“HSK动态作文语料库”的建设构想,《第三届全国语言文字应用学术研讨会论文集》。
- 邓守信(2009)《对外汉语教学语法》(修订2版),台北:文鹤出版社。
- Burt, M. & Kiparsky, C. (1972) *The Gooficon: A repair manual for English*. Newbury House. Rowley, MA.
- Ellis, Rod & Barkhuizen, Gary (2005) *Analysing Learner Language*. Oxford: Oxford University Press.
- Dagneaux, E., Denness, S. & Granger, S. (1998) Computer-aided error analysis. *System: An International Journal of Educational Technology and Applied Linguistics* 26(2): 163-174.
- Díaz-Negrillo, Ana & Fernández-Domínguez, Jesús (2006) Error tagging systems for learner corpora. *Resla*, 19: 83-102.
- Dulay, Heidi C., Burt, Marina K., & Krashen, Stephen D. (1982) *Language Two*. Oxford: Oxford University Press.
- Granger, S. (2003) Error-tagged learner corpora and CALL: a promising synergy. *CALICO Journal* 20/3: 465-480.
- Izumi, E., Uchimoto, K. & Ishahara, H. (2005) Error Annotation for Corpus of Japanese Learner English. *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora (LINC 2005)*, 71-80.
- James, Carl. (1998) *Errors in Language Learning and Use: Exploring Error Analysis*. London: Addison Wesley Longman.
- Lee, Lung-Hao, Yu, Liang-Chih, Lee, Kuei-Ching, Tseng, Yuen-Hsien, Chang, Li-Ping & Chen, Hsin-Hsi (2014) A sentence judgment system for grammatical error detection. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING'14)*. Dublin, Ireland,

67-70.

- Nicholls, D. (2003) The Cambridge Learner Corpus: error coding and analysis for lexicography and ELT. *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster, United Kingdom, 572-581.
- Tono, Y. (2003) Learner corpora: design, development and applications. *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster, United Kingdom, 800-809.
- Turton, N.D. & Heaton, J.B. (1997) *Longman Dictionary of Common Errors*. Harlow: Longman.

附录一 TOCFL 学习者语料库偏误标记说明手册

一、四大主类

先从句子的表面结构和目标语的差异分成四类：

类型	与目标语的差异
遗漏错误 (M)	少用了语言成分
冗赘错误 (R)	多用了语言成分
选用错误 (S)	错用了语言成分
语序错误 (W)	语言成分错序

二、三十六次类

给定了主类后,根据语言学概念给予次类。这个次类又分为几个层次,包括:

(1) “词”的层次(16类),如动作动词(v)、能愿动词(aux)、状态动词(vs)、名词(n)、代名词(pron)、连接词(conj)、介词(p)、数词(num)、指示词/定词(det)、量词(cl)、句末助词(sp)、时貌助词(asp)、副词(adv)、助词(de)、疑问词(que)、复数词缀(plural);

(2) “语”的层次(11类),如主语(sub)、宾语(obj)、名词短语(np)、动词短语(vp)、介词短语(pp)、修饰语(mod)、时间词语(time)、地方词语(loc)、及物性(tran)、[数/定+量]结构(dm)、离合结构(vo);

(3) 句型层次(7类),如复杂名词子句(rel)、特殊句型(把ba、被bei、让rang、是shi、有you)、其他句式(pattern)。

(4) 混合类(2类),形式错误(form)、表达方式错误(sentence)。

三、标记大原则

1. 学习者的偏误表现，应该是具有系统性、具有解释性的。不过在标记过程中，有时很难判定是学习者的偏误还是失误，因此有时我们所标记的涵盖“失误”及“偏误”。如例（1），无法确定是输入时的失误将“静”打成“间”，还是偏误，将“夜深人静”误用为“夜深人间”。

（1）我并不反对你夜深人间时，玩线上游戏。

如果是明显的打字或编辑问题，不做错误标记，而是直接订正原文。例如“他给我跟我介姊姊绍了他的朋友”显然是考生文字处理时（注：TOCFL 是电脑输入作文）的错误。我们将原文订正为“他给我跟我姊姊介绍了他的朋友”即可。

2. 做病句修正时，有两大原则。（a）尽量不改动作者想表达的意思；（b）尽量保留该阶段学习者能力所具备的词语（可能对目标语人士不是那么自然，但合乎语法）。

3. 标记时，先决定主类，再选择次类。如例（2），“真”多余，因此是 R 类；“真”是副词，因此选择 adv。

S（选用偏误）的次类标记选择则与其他主类不同，如例（3）“去”错用，正确用法应该是“要”，此时次类标记应该选择 aux。也就是说，S 的次类是以应使用的正确成分为标记选择的。

（2）他是真 [Radv] 很好的人。→他是很好的人。

（3）李台生建议他不去 [Saux] 一个人爬山。→李台生建议他不要一个人爬山。

4. 原则上，标记以词为最小原则（除了“们”以外），因此像例（4），要将“不良性”当成一个词来看，而非标记“性”为多余成分（注：如果不确定“词”的切分，可上平衡语料库查询 <http://140.109.19.114/>）。例（5）也是将用错的成分视为一个词，词间的成分不对，尤其是四字格问题，标记为选用错误，次类为形式问题 form。

（4）身体立刻有不良性 [Svs] 的反应。→身体立刻有不良的反应。

（5）我并不反对你夜深人间 [Sform] 时，玩线上游戏。→我并不反对你夜深人静时，玩线上游戏。

5. 关于次类标记的选用是有优先顺序的，以下是几个原则：

A. 偏误牵涉特殊句式，也就是“把、让、有、被、是”，不论是多用、少用、错用、错序，应优先标示该次类标记（ba, rang, you, bei, shi），而非 p, pp, v, vp 等。

(6) 我现在没有时间写下来所有要告诉你的东西 [Mba]。→我
在没有时间把所有要告诉你的东西写下来。

(7) 录取那个课程的话 [Mbei] →被那个课程录取的话

(8) 这个 [Sdet] 话的印象非常深刻 [Mrang]。→这些话让我印
象非常深刻。

但是，下面情况中地方词语或时间词语的错序标记还是优先于特殊句式。

(9) 他在柜子里 [Wloc] 把衣服放。→他把衣服放在柜子里。

B. 在错序、遗漏、冗赘的情况下，名词或名词短语针对它们在句中的功能优先标示为主语或宾语，如例（10）；如果是名词错用，则标示 n，如（11）；主语或宾语中遗漏了部分成分，则标记为 np，如（12）。

(10) 我庆祝会 [Wobj] 不能参加。→我不能参加庆祝会。

(11) 然后从公司到等公车的地 [Sn] →然后从公司到等公车的
地方

(12) 我们到现在都是以写信为连络 [Mnp]。→我们到现在都是
以写信为连络的方式。

C. 表示地方或时间的介词短语、名词或名词短语，优先标示为 time 或 loc，如例（13）（14），而非标示为 n, np 或 pp。

(13) 进名校是在韩国 [Wloc] 非常重要的关键。→进名校在韩国
是非常重要的关键。

(14) 明天在我家下午 2 点 30 分 [Wtime] 集合。→明天下午 2 点
30 分在我家集合。

四、插入标记位置的原则

1. M 类：标记在该句缺少的偏误成分的空缺处，如例（15）。如果

属于句式漏用,则放句末,如(16)。

(15) 我真的 [Madv] 紧张要跟林美美见面。→我真的很紧张要跟林美美见面。

(16) 我现在没有时间写下来所有要告诉你的东西 [Mba]。→我现在没有时间把所有要告诉你的东西写下来。

2. R类: 标记在该句多余的偏误成分之后,如(17)。如果属于句式多用,则放句末,如(18)(19)。

(17) 我知道你又很 [Radv] 聪明又用功。→我知道你又聪明又用功。

(18) 而且我觉得她男朋友对她这样做是太过分的。 [Rpattern] →而且我觉得她男朋友对她这样做太过分。(标记在整个句式之后)

(19) 以免更多人被其食品中毒。 [Rbei] →以免更多人食物中毒。(标记在句末)

3. S类: 标记在该句选用错误的偏误成分之后。

(20) 我还不 [Sadv] 决定要在哪一家餐厅吃。→我还没决定要在哪一家餐厅吃。

(21) 我不会记得 [Sform] 到火车站怎么走。→我记不得到火车站怎么走。

(22) 下个星期什么空我都 [Ssentence]。→下个星期我都有空。(通常标记在句末)

(23) 如何可能你那么绝佳的同学 [Ssentence]。→你是一个如此聪明的同学。

(24) 然后他爬上山来了 [Sform]。→然后他开始爬山。

4. W类: 标记在该句出现语序偏误的成分之后,如(25)。优先标记属于状语功能的成分,如(26),标记要放在“一直”后面,不是放在“心里”后面。其次是名词(n)、主语(sub)、宾语(obj),如(27),不标记在“敬上”后面,而是标记在“李希平”后面。会标记在动词组(vp)后面情况的应该很少,大概是连动结构才会发生。

(25) 今天早上我起床六点半 [Wtime]。→今天早上我六点半起床。

(26) 一直 [Wadv] 心里觉得很痛苦。→心里一直觉得很痛苦。

(27) 敬上李希平 [Wsub] → 李希平敬上

5. 两个都是状语功能的成分错序时, 标记在第二个成分之后。

(28) 今天他先要 [Waux] 做功课再到学校去上课。→ 今天他要先做功课再到学校去上课。

(29) 吃饱以后再要 [Waux] 运动一下吧! → 吃饱以后要再运动一下吧!

(30) 明天在我家下午2点30分 [Wtime] 集合。→ 明天下午2点30分在我家集合。

6. VO 结构: W 类的标记, 放在整个结构后面, 如 (31)。M 类的标记, 放在 VO 动词之后, 如 (32) (33)。

(31) 他就停车了 [Wvo] 帮我。→ 他就停了车帮我。(“了”要插在 VO 中间)

(32) 他走路 [Mvo] 了三个小时。→ 他走路走了三个小时。(VO 动词要重复)

(33) 你很久以前找工作 [Mvo] 很辛苦。→ 你很久以前找工作找得很辛苦。

五、偏误标记举例 (以次类字母排序)

(一) 词的层次

1. 副词错误 (adv)

(34) 他们看了新电影就 [Madv] 高兴。→ 他们看了新电影就很高兴。

(35) 你是真的 [Wadv] 很好的人。→ 你真的是很好的人。

(36) 安卡拉不 [Sadv] 那么热。→ 安卡拉没那么热。

(37) 然后他一直走一直走就天快要 [Radv] 黑了。→ 然后他一直走一直走就天黑了。

(38) 所以你快一点 [Sadv] 找到了工作。→ 所以你很快地找到了工作。

(39) 我差不多 [Sadv] 忘了告诉你什么方式可以认得我。→ 我差点忘了告诉你什么方式可以认得我。

2. 时貌错误 (asp)

- (40) 公共汽车没有空的椅子, 所以两个人都得站 [Masp]。→公共汽车没有空的椅子, 所以两个人都得站着。
- (41) 我朋友也找工作找了很久才找到了 [Rasp]。→我朋友也找工作找了很久才找到。
- (42) 他爬山爬的 [Sasp] 半天。→他爬山爬了半天。
- (43) 我放假的时候去欧洲了 [Wasp]。(这时使用 W, 而不使用 R+M 方式标记) →我放假的时候去了欧洲。

3. 能愿动词错误 (aux): 能愿动词包括“要、想、愿意、肯、敢、应该、应当、应、该、得 (dei3)、能、能够、可以、可、得 (de2)、可能、会、打算”。

- (44) 你们得好好地念书才 [Maux] 上大学。→你们得好好念书才能上大学。
- (45) 你愿意 [Raux] 帮我好不好? →你帮我好不好?
- (46) 有一部电影我很要 [Saux] 看。→有一部电影我很想看。
- (47) 吃饱以后再要 [Waux] 运动一下吧! →吃饱以后再要再运动一下吧!

4. 量词错误 (cl)

- (48) 我真的高兴你找到一 [Mcl] 新的工作了。→我真的高兴你找到一个新的工作了。
- (49) 还有休息几个 [Rcl] 天。→还有休息几天。
- (50) 今天我上了一个 [Scl] 数学课。→今天我上了一堂数学课。

5. 连接词错误 (conj)

- (51) 我想说你要想在那个时间做那件事是正常 [Mconj] 不正常。
→我想说你要想在那个时间做那件事是正常还是不正常。
- (52) 要不然 [Rconj] 如果成绩不那么好大概会伤她的心。→如果成绩不那么好大概会伤她的心
- (53) 不论 [Sconj] 你自己觉得是小事情。→虽然你自己觉得是小事情。
- (54) 我们一般来说 [Wconj] 能从那个捷运站走路或者坐公车。→

一般来说我们能从那个捷运站走路或者坐公车。

6. “的、得、地”错误 (de)

- (55) 今天我决定我也要当一个很好 [Mde] 人。→今天我决定我也要当一个很好的人。
- (56) 今天他只有数学的 [Rde] 课。→今天我上了一堂数学课。
- (57) 时间过了 [Sde] 真快。→时间过得真快。
- (58) 突然我的 [Rde] 另外 [Mdm] 朋友打电话给我说。→突然我另外一个朋友打电话给我说。
- (59) 用我今年的 [Rde] 夏天放假 [Mde] 时间 (这时不用 W, 而用 R+M 方式标记) →用我今年夏天放假的时间

7. 指示词/定词错误 (det): 包括“这、那、当(天)、别(人)、另外、该(地)、这些、那些、有些、一点、一些、整(个)、每(人)、其他、隔(一天)、第、某、上、下、全、全部”。

- (60) [Mdet] 一天很有意思。→这一天很有意思。
- (61) 我会在第 [Rdet] 一楼等你。→我会在一楼等你。
- (62) 我可以给你那么 [Sdet] 钱。→我可以给你那些钱。
- (63) 也要花有些 [Sdet] 钱。→也要花一些钱。
- (64) 我们认为一边可以多知道 [Mdet] 中华文化, ……→我们认为一边可以多知道一点中华文化, ……

8. 名词错误 (n)

- (65) 因为泰国菜是世界上最美味佳肴 [Rn] 的食物。→因为泰国菜是世界上最美味的食物。
- (66) 回家的时候, 天气 [Sn] 已经黑了。→回家的时候, 天已经黑了。
- (67) 这里什么式 [Sn] 的食物都有。→这里什么种类的食物都有。

9. 数词错误 (num)

- (68) 是爱吃有机蔬菜的 [Mnum] 个消费者。→是爱吃有机蔬菜的一个消费者。
- (69) 小王是一个两 [Snum] 年级的高中学生。→小王是一个二年级的高中学生。

(70) 你不必全一 [Rnum] 天戴着。→你不必全天戴着。

10. 介词错误 (p)

(71) 你这样子做不但 [Mp] 你的健康好。→你这样子做不但对你的健康好。

(72) 我们跟 [Rp] 一起坐捷运。→我们一起坐捷运。

(73) 我想给 [Sp] 你直接说“恭喜!”→我想跟你直接说“恭喜!”

(74) 感动了 [Sp] 没话说。→感动到没话说。

11. 代名词错误 (pron): 为了和 sub, obj 区分, 如果是 M, W 类的优先给 sub, obj; 如果是 S, R 类 (通常是指称错误或回指问题) 则给 pron。

(75) 所以要见面之前就知道互相 [Spron] 的个性。→所以要见面之前就知道彼此的个性。

(76) 我们应该 [Mdet] 一个人练习说他们 [Rpron] 自己的部分。
→我们应该每一个人练习说自己的部分。

12. 疑问词错误 (que): 指“什么、怎么、几、怎么样、哪些、为什么、哪里”。

(77) 开始慢慢地教她这么 [Sque] 弹钢琴。→开始慢慢地教她怎么弹钢琴。

(78) 可是什么 [Wque] 买? →可是买什么?

(79) 在生活上你碰到 [Mque] 困难吗? →在生活上你碰到什么困难吗?

13. 句尾助词 (sp)

(80) 你可以告诉我最近的火车站在哪里 [Msp]? →你可以告诉我最近的火车站在哪里吗?

(81) 你可不可以送我到火车站吗 [Rsp]? →你可不可以送我到火车站?

(82) 如果你有空, 我想再开一个好吧 [Ssp]? →如果你有空, 我想再开一个好吗?

- (83) 您可不可以跟我们一起来参加好吗 [Rsp]? ① →您可不可以跟我们一起来参加?

14. 动词错误 (v): 包含动词补语错误, 此处视动补结构为一个词。这里采用“中研院”断词规则, 要注意动词+时态词“了、着、过”是切分的, 与大陆标准不一样。

- (84) 谢谢您请我 [Mv] 你的庆祝会。→谢谢您请我参加你的庆祝会。
 (85) 他们打算一块儿在 101 大楼去 [Rv] 看电影。→他们打算一块儿在 101 大楼看电影。
 (86) 下 [Sv] 了电影我就回家了。→看完了电影我就回家了。
 (87) 他一看 [Sv] 小安的汽车就放心了。→他一看到小安的汽车就放心了。

15. 状态动词错误 (vs)

- (88) 他的方法不太温柔, 可是我在很短的时间内进步得很 [Mvs]。
 →他的方法不太温柔, 可是我在很短的时间内进步得很快。
 (89) 以前已经给你我的大 [Rvs] 电话号码。→以前已经给你我的电话号码。
 (90) 日本的天气非常清楚 [Svs]。→日本的天气非常清爽。

16. “们”错误 (plural)

- (91) 不过我希望我们会变成好朋友们 [Rplural]。→不过我希望我们会变成好朋友。
 (92) 所有的同学们 [Rplural] 都是我的战友。→所有的同学都是我的战友。

(二) 语的层次

1. 地方词、方位词错误 (loc): 这个标记包括词与语的概念。如 (93)~(96) 是“上、里”等词语问题, (97) 则是“在日本”这个组的问题。

① 这里多余的是“好吗”, 应该是“语”的层次, 但考量到多数 Sp 都是词层次, 目前将其归类在“词”中。

- (93) 考试上 [Rloc] 的时候→考试的时候
- (94) 李台生就去火车站里 [Rloc] 坐公车回家。→李台生就去火车站坐公车回家。
- (95) 要是你在功课里 [Sloc] 有问题→要是你在功课上有问题
- (96) 我要带你去在同安街 [Mloc] 我常常去的餐厅。→我要带你去在同安街上我常常去的餐厅。
- (97) 如果我买那件衣服在日本 [Wloc] →如果我在日本买那件衣服

2. 修饰语错误 (mod): 名词前的修饰语属于 [X+ 的] 结构, 标记为 mod。

- (98) 现在我知道什么样的 [Wmod] 买东西是值得的。→现在我知道买什么样的东西是值得的。
- (99) 因为你要到我的城市来看看 [Mmod] 生活怎么样。→因为你要到我的城市来看看这里的生活怎么样。
- (100) 我希望你改掉这两个你的 [Rmod] 生活习惯。→我希望你改掉这两个生活习惯。

3. 名词短语错误 (np) (包括“的+N”的偏误)

- (101) 我跟同学一起住在房租 [Snp]。→我跟同学一起住在租的房子。

4. 宾语错误 (obj)

- (102) 他们说那个电影 [Wobj] 想看。→他们说想看那个电影。
- (103) 他请我教 [Mobj] 日文。→他请我教他日文。
- (104) 我不知道你可不可以帮我拿回来一瓶红酒 [Wobj]。→我不知道你可不可以帮我拿一瓶红酒回来。

5. 介词短语错误 (pp)

- (105) 他坐在教室前跟林美美 [Wpp]。→他跟林美美坐在教室前。
- (106) 希望你在太晚的时候 [Rpp] 早点去睡。→希望你早点去睡。
- (107) 不可以买给别人 [Wpp] 饭。→不可以买饭给别人。

6. 主语错误 (sub)

- (108) 我听说 [Msub] 打算到法国。→我听说你打算到法国。

- (109) 害得 [Msub] 全家食物中毒。→害得我们全家食物中毒。
 (110) 所以决定我 [Wsub] 一个人去。→所以我决定一个人去。

7. 时间词语错误 (time) : 这个标记包括词与语的概念。

- (111) 可是现在 [Rtime] 最近我工作很忙。→可是最近我工作很忙。
 (112) 他们正在 [Stime] 还在电影院一边看电影一边聊天。→他们现在还在电影院一边看电影一边聊天。
 (113) 他平常起床七点钟 [Wtime]。→他平常七点钟起床。

8. 及物性错误 (tran) : 不及物动词后面加了宾语, 标记为 tran。

- (114) 有三个日本学生入学 [Wtran] 了我们的学校。→有三个日本学生在我们学校入学。
 (115) 见面 [Wtran] 她的老公时, ……→跟她的老公见面时, ……
 (116) 我实在不好意思 [Wtran] 父母。→我实在对父母不好意思。

9. 动词短语错误 (vp)

- (117) 他们去学校前的公车站等一等 [Svp]。→他们去学校前的公车站等公车。
 (118) 有问题的话, 他们可以到我们家来打电话 [Wvp]。→有问题的话, 他们可以打电话到我们家来。

10. 数量词错误 (dm)

- (119) 他还买给自己 [Mdm] 筷子。→他还买给自己一双筷子。
 (120) 因为我们班有各种 [Rdm] 各国的同学们 [Rplural]。→因为我们班有各国的同学。
 (121) 然后去一个 [Wdm] 台北附近的山。→然后去台北附近的一座山。

11. 动宾结构离合所造成的问题 (vo) : 可能是中插成分的位置问题, 如 (122) (123); 或是动词重复的问题, 如 (124) (125)。

- (122) 可是一次也没见过面 [Wvo] →可是一次也没见过面。
 (123) 写信也很多 [Wvo]。→也写了很多信。
 (124) 因为我们在网络聊天 [Mvo] 了快一年了。→因为我们在网

络聊天聊了快一年了。

(125) 她赚钱 [Mvo] 得很辛苦。→她赚钱赚得很辛苦。

(三) 句型层次

1. “把”字句错误 (ba)：这个标记包括词与句型的层次。有时指的是“把”这个词，有时指的是“把”字句。如果是该用“把”字句而没用 (Mba)，或不该用“把”字句而用了 (Rba)，也就是不是单单一个“把”的问题，标记要放在句末，如 (129)。

(126) 先去饭店 [Mba] 行李放在床上。→先去饭店把行李放在床上。

(127) 我希望你每天都让 [Sba] 我们的房间弄干净。→我希望你每天都把我们的房间弄干净。

(128) 之后把 [Rba] 我的中文能力也很快地进步了。→之后我的中文能力也很快地进步了。

(129) 下课以后常常把学校的事一起讨论 [Rba]。→下课以后常常 (跟他) 一起讨论学校的事。

2. “被”字句错误 (bei)：这个标记包括词与句型的层次。有时指的是“被”，有时指的是“被”字句。标记原则同“把”字句的标记原则。

(130) 她的脸被 [Rbei] 受伤流血→她的脸受伤流血。

(131) 每一个人都需要有一些好朋友被 [Rbei] 帮他们过人生的路
→每一个人都需要有一些好朋友帮他们过人生的路。

3. “让”字句错误 (rang)：这个标记包括词与句型的层次。

(132) 不但给自己开心也 [Mrang] 大家看得很快乐的。→不但给自己开心也让大家看得很快乐。

(133) 但是我的老师马上就把 [Srang] 我心里安静下来。→但是我的老师马上就让我心里安静下来。

4. “是”错误 (shi)：这个标记包括词与句型的层次。

(134) 下个星期五 [Mshi] 我弟弟的生日。→下个星期五是我弟弟的生日。

(135) 小明看出来老师是 [Rshi] 50 岁。→小明看出来老师 50 岁。

- (136) 我觉得一件白色跟蓝色格子的裤子想 [Sshi] 很好的礼物。
→我觉得一件白色跟蓝色格子的裤子是很好的礼物。
- (137) 我们看的电影好像 [Mshi] 最有名的。(这里标记 shi 较单纯, 而非 pattern) →我们看的电影好像是最有名的。

5. “有” 错误 (you)

- (138) 平常也不太常有 [Ryou] 下雨。→平常也不太常下雨。
- (139) 因为北海道是 [Syou] 很多好吃的东西和漂亮的风景。→因为北海道有很多好吃的东西和漂亮的风景。
- (140) 她还说学习的路很长 [Myou] 很多困难。→她还说学习的路很长有很多困难。

6. 复杂名词子句错误 (rel): 这一类只出现在 S 主类。大部分是关系子句的表达问题, 以及名词中心语 (Head) 位置错误。

- (141) 我觉得餐厅在你家旁边 [Srel] 好像好吃。→我觉得在你家旁边的餐厅好像很好吃。
- (142) 希望我们的计划要见面 [Srel] 没有什么问题→希望我们要见面的计划没有什么问题
- (143) 我觉得野餐是最好的办法认识新朋友 [Srel]。→我觉得野餐是认识新朋友的最好办法。

7. 其他句式错误 (pattern): 包括“是……的、连……也、除了……以外、一边……一边……”等有框架形式的句型。标记插入在框架之后, 如果没能找到框架, 大部分放在句末。

- (144) 我这么觉得 [Mpattern]。→我是这么觉得的。
- (145) 我在书面上 [Spattern] 已经了解你的个性。→我从书面中已经了解你的个性。
- (146) 他告诉那个人是他电话号码的 [Rpattern]。→他告诉那个人他的电话号码。

(四) 混合类

1. 形式错误 (form): 主要有三种情况。

- a. 选用的词没错, 只是形式错误 (例如重叠式、尝试貌), 如 (147)~

(149)。

b. 可以理解语义, 只是表达形式错误, 如(150)~(152) (可能造出的不是中文词)。

c. 遗漏了固定格式中的某部分, 如(153)。

(147) 上课的时候我跟爱文常常谈一谈 [Sform] 不知道的问题。

→上课的时候我跟爱文常常谈不知道的问题。

(148) 他在路上走着想想 [Sform] 他的功课。→他在路上走着想他的功课。

(149) 谁都不知明天也 [Radv] 自己的工作或公司还会 [Sform] 存在。→谁都不知明天自己的工作或公司还会不会存在。

(150) 幸运 [Mform] 我住的地方有越南餐厅→幸运的是我住的地方有越南餐厅。

(151) 李文平敬于 [Sform] →李文平敬上

(152) 我觉得不舒服得不得了 [Sform]。→我觉得非常不舒服。

(153) 可是对他们 [Mform] 是开心的一天吧! →可是对他们来说是开心的一天吧!

2. 句子表达错误 (sentence)

当学习者的句子中不只出现一个错误, 母语者虽然能理解其意思, 但该句错误太多, 需要做整体性的修改时, 我们标记为 [Ssentence]。或是连续三个及三个以上的错误, 即使可以归类为词层次偏误, 仍以 sentence 来标记。

(154) 合格了交换学生 [Ssentence] 以后→符合了交换学生的资格以后

(155) 她是最好的家教一样的感觉 [Ssentence]。→感觉她是最好的家教一样。