# 语料库中的插入语标注研究

安娜[1]，侯敏[2]

（1. 中国传媒大学应用语言学系，北京 100024；2。中国传媒大学应用语言学系，北京 100024）

**摘 要**：本文通过对“传媒语言语料库”和部分人民日报语料中包含的插入语的分析，比较全面地考察了汉语插入语的使用情况，并试图按语义类型和语用功能将插入语进行分类，归纳出汉语插入语的主要特点，在此基础上从话语分析的角度确立了话语标记集来标注插入语。

**关键词**：插入语；标注；话语标记集

# Research on tagging the Parenthesis in corpus

An Na[1], Hou Min[2]

(1.Communication University of China, Beijing 100024; 2.Communication University of China, Beijing 100024)

**Abstract**: Through analyzing the parenthesis in Corpus,we knew the usage of Chinese parenthesis completely. Moreover, parenthesis are classified according to their pragmatic functions. We concluded the pragmatic features of these parenthesis. Based on the theory of discourse analysis, we set up a pragmatic annotation manual for parenthesis so that their pragmatic functions can be reflected.

**Keywords**: parenthesis, tagging, annotation manual

## 1. 引言

在加工生语料时，通常的做法是使用分词标注工具对原始语料进行标注。我们采用的是由中国传媒大学应用语言学系改进后的北大的分词标注系统。在对加工后的熟语料进行校对时，我们发现目前的插入语标记存在一定的问题。插入语的组成比较复杂，有些是词，有些是词的组合。对由词充当的插入语，北大的分词标注系统把它当作一个词汇单位来处理，只在句法层面上给插入语一个词性标记，如“据说/v”。对那些由词的组合充当的插入语，北大的分词系统对每一个组成成分进行了词性标注，如“按/v 道理/n”。我们认为插入语是在语用平面上需要研究的内容，应该根据我们的研究目的来决定是否要在词性标注阶段给插入语一个词性标记。

我们认为，插入语可以存在于句法分析之外，它们是不充当句法成分的，提供的只是语用层面的程序信息，而并不是句法层面的概念信息，对话语命题的真假不产生影响，在话语中的作用主要是语用的而非句法的。因此，在我们进行分词标注的过程中，主张不考虑插入语的词性标记，只给出特定的语用标记就可以。但究竟给插入语词性标记还是语用标记最终还是由分词目的决定的。

自然语言十分复杂，大规模真实文本在处理时存在的噪声很多，想得到一棵完整的句法树很困难。其中，插入语也属于画树时的噪声之一。插入语的构成非常复杂，在句法分析的过程中我们很难为它在句法树中找出一个合理的句法位置。因此，我们认为在句法分析的层面可以把插入语排除在句法分析的内容之外，这也可以作为剔

作者简介：安娜（1979-），女，山东青岛，在读博士 E-mail.yunxiner07@cuc.edu.cn

除文本噪声的一种手段。

## 2. 插入语的研究

插入语是独立语中的一种。在独立语研究领域中，由于研究的出发点和侧重点不同，出现了各种名称，如：独立成分、独立语、插说等等（本文研究的“插入语”只是独立语的一个组成部分）。以往的研究成果主要集中在独立语的性质、独立语的语义分类、独立语的功能这三方面。由于插入语是独立语的下位分类，因此插入语继承了独立语的大部分性质。

### 2．1 插入语的性质

关于插入语的性质，语言学家们基本认为插入语是句子中的特殊成分。插入语在句子中具有独立性，不和句子中别的成分发生结构上的关系，这一认识是人们共同的。但插入语是否充当句子成分呢？有人认为插入语应该作为特殊的句子成分，或者叫做独立成分，并且还指出了插入语是一种语用成分。有人认为插入语是比较特殊的成分，但不是句子的成分。我们认为，插入语具有独立性，是比较特殊的成分，它不和句子中的成分发生结构上的关系，因此不是句子的成分。

### 2．2 插入语的语义分类

目前，语言学家们对插入语做语义方面的研究，多为表意分类和定性的举例，具有一定程度的局限性。如：王力的插语法大致可分为 8 种，黄伯荣等的插入语涉及到 6 类，胡裕树认为独立成分在表意上有 7 种类型等等。

也有语言学家是在大规模语料库的基础上对插入语做定量的描述和意义的分类，如邢红兵通过对“现代汉语研究语料库系统”中包含插入语的句子的分析，按照意义将插入语分为 17 类。

## 3. 语料库中插入语的语用功能分类

### 3．1 语料库中插入语的自动识别

插入语的一般形式特点是：口语中，前后都有语音停顿；书写时常用标点与其他成分隔开。目前，我们还无法借助韵律标注的手段来自动识别插入语，因此以只能以标点符号作为自动识别插入语的浅层依据。

#### 3.1.1 插入语单位的浅层识别方法和技术路线

1)浅层识别方法

本研究首先运用语言学规则和语言材料对大规模真实文本进行字符串过滤，然后根据插入语的性质对利用浅层规则识别出的字符串进行评估，将符合插入语特点的字符串收入插入语单位表。语言材料主要是语料库存储的带有标点符号的生语料。语言学规则，包括语法特征和语义规则。

插入语既然不能充当句子成分，因而也就不具备一般词或短语所具有的语法特征。如：“你说”作为主谓短语，它的语法特征包括可以带宾语，主语和谓语之间可以插入状语，谓语动词“说”可以带动态助词。插入语一般不具备这些语法特征。

2)技术路线

本研究主要采用语言学规则对字符串进行删选的技术路线，处理过程包括如下内容：

(1)选定语料。

本研究选用的是传媒语言语料和人民日报语料，约 8000 万字。

(2)自动识别。

我们充分利用插入语的形式特点来确定识别插入语的界限。根据以往插入语的研究，我们把插入语字符串的长度确定为 2～5 个汉字。利用插入语前的标点符号冒号、逗号或句号，插入语之后的标点符号逗号自动识别插入语。由于插入语之后出现的中心句一般为陈述句，字符串之后出现的第二个标点符号如果为问号或叹号，视为不合格的插入语，这类插入语将不被识别。不可否认，这种方法只能识别出部分插入语，并且识别结果中有一部分不属于插入语，这时我们采用人工干预的方法来排除那些非插入语字符串。

(3)数据格式转换

经过第二步的自动识别，我们得到一些形式上完整的 2 字串、3 字串、4 字串、5 字串，我们将其转化为数据库格式，分别存入相应的列表。目的是方便计算、比较、过滤。

(4)选取插入语出现的语段

将单位表中的插入语按频率顺序输入到 CCRL（北京语言大学研制开发的语料检索工具）中，检索出前后带有 40 个汉字的含有插入语的语段。将这些语段按插入语的频率顺序保存为 txt 文本格式。再使用改进后的北大分词标注系统对这些语段进行切分标注。这样做的主要目的是得到插入语的切分标注结果，分析出现在插入语前后句子的句法和语义特征，便于我们完善插入语自动识别工具、兼用插入语的消歧研究以及进一步做插入语的功能分类。

## 3．2 插入语的语义和语用功能类型

我们知道一个完整的句子，它的意思应该是句子结构意义和非句子结构（是指在句子中不充当句法成分的结构）意义之和。这种表示非句子结构意义的手段有多种，其中插入语就是表示非句子结构意义的一种常用的手段。因此无论是自然语言处理还是语言教学，插入语所表示的这种非句子结构意义都是不容忽视的。下面我们根据插入语表达的意义和功能进行分类，并将表达意义相同或相近的插入语归类如下：

1）表示引起或结束话题的插入语

这类插入语表示说话人在后面或前面的话语中提出或结束一个新的话题，引起听话人的注意。比如：

**我要讲的是，话又说回来，顺便问一下，就这样，到此为止**

这些插入语经常被用在引起或结束一个话题，偏离一个话题或重新回到以前的话题。引起一个新话题包括在对话开始时提出第一个话题或开始一个新的话题。结束一个话题就是在引入新话题或结束整个对话前结束旧的话题。

2）表示话语来源的插入语

这类插入语表示消息的来源或是引用常理来说明一个道理。比如：

**据我所知，有人说，听说，俗话说，按你所说，据说**

虽然这类插入语都表示消息来源或是引用常理，但是在具体的上下文中它们也有不同的功能。“据我所知”是以说话者为出发点的，表示下文所说的仅是说话人知道的，如果说话人所说的话不全面的话，听话者可以通过这个插入语帮助自己理解说话者所说的话。“听说”经常被当作话语策略使用，表示说话者所说的并不是自己的观点，在人际层面上可以减轻说话者的面子损失。

3）表示推理的插入语。交流从本质上讲是基于推理的。在交流中使用这类插入语是表示推理或总结关系的一种手段。比如：

**概括起来说，总的看来，总体上看，总而言之，如此说来，这么说，可以说，原来如此**

从认知的角度来看，标记此类插入语的其中一个目就是指导话语理解，减轻听话者在理解话语过程中的负担。

4）表示解释或补充说明的插入语。这类插入语之后的话语都是说话者或听话者的重新陈述、详述或是补充。比如：

**换句话说，就是说，你的意思是，这样说吧**

这类插入语之后的内容与之前的在语言形式上是不同的，但是在表达的意义上是相同的。

5）表示说话内容真实性的插入语。这类插入语表示了说话的方式和说话者的态度。比如：

**恕我直言，简单而言，说穿了，说真的，确切地说，客观地讲，不瞒你说**

通过这些插入语，听话者可以意识到说话者的说话方式和态度。这样的插入语可以修饰说话者的话语行为或说话者的行为方式。在话语开始之前，这类插入语的使用表明了说话者的方式或态度，并且说话的方式和态度是伴随说话者的整个话语过程的。话语交际取得的成果也是与说话者的方式和态度紧密相连的。

6）表示对比关系的插入语。这类插入语之后出现的内容经常与之前的内容形成对比。插入语之后出现的话语内容或是对前面内容的否定或是与之前出现的话语形成对比。比如：

**不然，相反，尽管如此，事实上，不过**

7）表示自我评价的插入语。在交际中，我们经常会使用一些表示个人意见或看法的插入语，比如：

**按道理，应该说，一般而言，依我之见，照我看，依我看**

这些插入语表示说话者对事物的个人意见或评价，一般是以说话者为出发点的。这类插入语使用的目的是强调说话者自己的评价或意见，而这些意见在插入语之前的话语内容中已经出现过，插入语之后的内容是说话者对之前所说的话做一个总结。

8)表示言语行为的插入语。比如：

**你说，你告诉我，你知道，我要说，我问你，大家想想，我劝你**

我们所确定的插入语是以其前后出现的标点符号为划分依据的，因此这类插入语后面如果没有标点符号，这样的情况我们暂不考虑。这类插入语有两种情况，一种是以说话者为中心的说话者的言语行为表达方式，如：我要说，跟你说吧等等；另一种是以听话者为中心的，把说话的机会和话语理解的空间留给听话者的言语行为表达方式，如：你说，你想，你说说看，大家想想等等。这种表示言语行为的插入语使用得非常普遍，它们的语用功能包括突出重点，拉近说话者之间的心里距离，引起听话者的注意等等。

9）表示推测和估计的标记语

标明说话人对自己所说话语有所保留，是一种推测或估计。比如：

**看来、看样子、我想、说不定、充其量、少说、算起来**

这类插入语一般是以说话者为出发点的，插入语之后的内容是说话者的推测或是估计。作为一种会话策略，听话者听到这类插入语，可以自己判断说话者所说内容的真实性。

10）表示思维过程的标记语。

表示思维过程的标记语有两种情况。第一种是标明说话人言语过程中思维跟不上话语，需要思考，但又想保留话语权。比如：

**那么、然后、这个、嗯**

第二种是标明说话人演说过程中的某种习惯，俗称口头语。它与言语过程中的思维状况有关，在没有考虑成熟或比较紧张的状态下，这种标记比较多。比如：

**这个、那么、然后**

11）表示强调的标记语。

表示说话人对话语中某一部分的重视，以此提示听话人注意。比如：

**特别是、尤其是**

说话者用这类插入语提醒听话者这部分话语的重要性。在会话过程中，听话者的注意力不可能总是集中的。当听话者听到这类插入语时，会重新调整自己的注意力。

12）表示意料之外的标记语。

这类插入语表示没有预料到的结果。比如：

**别说、没想到、哪成想、孰料、谁知**

这些插入语之后的内容表示说话者没有意想到的结果，是以说话者为中心的。通过这些插入语听话者可以意识到这些意想不到的事情是说话者重点要陈述的。

## 4. 插入语的标注

上面我们分析了插入语的语义和语用功能特点，由于插入语是非常特殊的语言单位，所以对插入语的处理就有了一定的难度，难度最大的是如何在分词阶段，给出插入语合适的标记。具体地说，就是在分词阶段，不能简单地将插入语进行分词。因为插入语包括专用的插入语和兼用的插入语。专用的插入语，是指该结构除了在句中作插入语外，一般不能充当句子成分，而且，很少有与它同形的结构。这类插入语的特点是：意义比较固定，内部结构关系比较虚化。而兼用的插入语，是指有时它们可以在句子中充当句法成分，即这类插入语的意义还没有完全虚化。如果分词的话，专用插入语的意义就不完整，因为专用插入语不充当句法成分，它提供的只是语用层面的程序信息，而并不是句法层面的概念信息。而兼用插入语如果不做消歧研究的话，其意义就会被错误地理解。本文主要考虑的是专用插入语的标注问题，对于兼用插入语的标注问题我们还会进一步研究，找出好的方法来解决消歧的难题。

关于专用插入语的标注，我们采用的方法是：在语料库中自动识别插入语并且根据其语义和语用功能分类后，

我们确立了标注插入语的话语标记集。话语标记集共包括十二类话语标记语，每一类话语标记语都采用拼音首字母缩写作为其自动标注时的代码，见下表：

表 1 话语标记集

Tab.1 annotation set of discourse marker

| 话语标记 | 样例 | 代码 |
|---|---|---|
| 话题标记语 | 我想讲的是 | HTDM |
| 话语来源标记语 | 据说 | LYDM |
| 推理标记语 | 由此可以 | TLDM |
| 换言标记语 | 换句话说 | HYDM |
| 言语方式标记语 | 恕我直言 | FSDM |
| 评价性标记语 | 幸运的是 | PJDM |
| 言语行为标记语 | 大家想想 | XWDM |
| 思维标记语 | 这个 | SWDM |
| 强调标记语 | 特别是 | QDDM |
| 估测标记语 | 看起来 | GCDM |
| 意外标记语 | 哪成想 | YWDM |

自动识别的结果经过人工干预后，我们得到了 170 个插入语。这 170 个插入语包括专用的插入语和兼用的插入语。目前，我们已经把这 170 个插入语归到这 12 类中，建立了插入语标记词典。由于我们考虑的是专用插入语的标注问题，如果插入语是一个词，我们就在北大的基本词典中把这个插入语删除，只在插入语标记词典中保存它的话语标记。这样在自动标注词性时，插入语的标注结果是语用层面的话语标记而不是词性标记，这样会为下一步进行自动句法的标注扫清障碍。

## 5. 结语

我们主要从插入语的角度确立了这十二类话语标记语，但是这十二类话语标记语并不能涵盖所有的话语标记语，我们有必要在进一步研究的基础上继续扩大我们的话语标记集。插入语应该只是话语标记语中的一类，我们希望在对语料做进一步分析时，能够发现更多的话语标记语来补充已经确立的话语标记集。

本文解决的只是专用插入语的标注问题，而在插入语标注研究中，兼用插入语的处理是一个难题，尤其是在文本自动标注时如何识别一个插入语是专用插入语还是兼用插入语，当它是专用插入语时就给出属于语用层面的话语标记，当它是兼用插入语，在句子中的充当句法成分时就给出属于句法层面的词性标记。随着对插入语研究的加深，我们在兼用插入语处理方面还会作进一步的探讨。

**参考文献：**

[1] Daniel Marcu, A surface-based approach to identifying discourse markers and elementary textual units in unrestricted texts:2-3

[2]Sandrine Zufferey, Andrei Popescu-Belis, Towards Automatic Identification of Discourse Markers in Dialogs: The Case of Like:1-9

[3]冯光武，汉语语用标记语的语义、语用分析，《现代外语》（季刊）2004 年第 1 期：24－31

[4]冉永平，话语标记语的语用学研究综述，《外语研究》2000 年第 4 期：8－14

[5]冉永平， The Pragmatics of Discourse Markers in Conversation, 广东外语外贸大学博士论文，2000：43-56

[6]邢红兵，基于统计的汉语字词研究，北京：语文出版社，2005：103-114