文章编号：1003-0077（2011）00-0000-00

汉语口语互动分级语料库的构建 *

王跃龙

（华侨大学文学院，福建 泉州 362021）

摘要：本文介绍了一个汉语口语互动分级语料库的构建工作。该语料库为国内首个汉语口语互动分级语料库，记录了测试环境下学生口语互动的实际情况。语料库由超过 1200 名学生的对话录制而成，时长超过 3000 分钟，样例分布范围从小学一年级到高中三年级。该语料库能为口语互动研究者提供经过转写和标注的真实语料，在语料调查的基础上可实现对口语互动的量化分析。另，该语料库回避了通常根据任务难易度进行分级的做法，而是根据会话特征进行互动分级，以供研究者参考。这对口语互动分级标准的确立和互动教材的编纂等也将有参考意义。

关键词：口语互动；分级语料库；分级标准

中图分类号：TP391　　**文献标识码：**A

The Construction of Spoken Interaction Corpus of Mandarin Chinese

WANG Yue-long

(College of Humanities of Huaqiao University, Quanzhou, Fujian　362021, China)

Abstract: We introduced the construction of a spoken interaction (SI) corpus of Mandarin Chinese, which is the first hierarchical corpus of SI in China, in this paper. This corpus is a valuable language resource which the spoken interactions among more than 1200 students are recorded in, and the duration time of all the date is more than 3000 minutes. The range of the example students is from Grade 1 in primary school to Grade 3 in high school. This corpus can provide the materials with transcriptions and annotations for researchers, by which the quantitative analysis for SI can be realized. In addition, exampler grading according to Conversation Analysis （CA）are also provided for the reference of researchers. Therefore, textbook compiling and the establishing of SI grading standard can both benefit from this corpus.

Key words: Spoken interaction; grading corpus; grading standard

1 引言

口语互动（Spoken interaction 简称 SI）不同于口语。口语互动是对口语各方面知识的综合运用，是使用口语进行社会交际的过程。如果把口语各方面的知识（语音、词汇、语法等）看作是建筑用的沙石泥土和钢筋，那么口语互动就是使用这些材料进行建设的过程。相同的材料最后可以形成不同的建筑形式，便显示出不同使用者建筑水平的差异。口语互动能力的差异也类似于此。两人或多人的对话与单人的说话之间具有不同的特点，最明显的特点即是对话时总是期望得到听话者的回应（Stenström 1994）。

在教学领域，学生即使学会了句子和词语也并不意味着能够很好地使用这些词语和句子来进行交际。因此，人们从关注口语研究逐渐延伸到对口语互动问题的研究上，如 Sinclair & Coulthard（1975）、Berry (1981)等。口语互动的研究已成为近年来一个新的关注热点（Mille and Mcllveenny 2000），学者们对口语互动的各方面进行了深入的研究（Ford & Wagner 1996, Och et al. 1996）。进而有学者进行了不同语言之间或同一语言不同群体之间互动特点的比较研究，如 Stenström (2002)，Aijmer & Simon-Vandenbergen (2003)等。

互动能力非常重要，应成为整个教育计划的一部分。目前，口语能力和口语互动能力的

* **收稿日期：**2015-8-10　**定稿日期：2015**

基金项目：教育部留学回国人员科研启动基金资助项目（Z1534014）；华侨大学高层次人才科研启动费项目（13SKBS219）

培养是混杂在一起的，在教学中并没有做明确的区分。学生的互动能力可以说是在无意识当中得到的提高。明确提出互动能力的培养，具体描写口语互动的等级差异，进而建立口语互动的等级标准，将有利于指导学生口语互动能力的提高，从而得到更加全面的发展。

国外的研究者已经开始了一些口语互动等级的制订工作。最典型的如《欧洲语言共同参考框架》（2008）开始设置了英语口语互动的分级标准，把口语互动分为三大类六个小类的等级。新加坡也正在着力于设立华语的口语互动标准。一些中文学者（Zhu 2011）也开始致力于华语口语互动标准的研究。

但是，目前这些会话分级的标准都是以任务的复杂程度为判定标准的。任务复杂程度可以作为口语互动分级的重要参考指标，但并不应是最重要或唯一的指标。首先，任务复杂程度的评判具有很大的主观性，一些话题对某些人可能很复杂，对另一部分人却可能很简单。缺乏操作上的统一性，因此并不是绝对的标准。第二，任务复杂程度属于语言外部的观察角度，缺乏对话轮内部结构的分析，因而并不是一个完善的分类标准。需要引入会话分析的内部指标来观察分析口语互动，才能完善口语互动分级的研究。内部和外部两者相结合，才是完善的口语互动分类标准的制订原则。

新世纪以来，口语互动的研究越来越多地借助于语料库的资源。2003 年在加拿大多伦多召开的国际语用学协会会议（International Pragmatics Association congress）上还专门讨论了在语料库基础上的口语互动研究方法。但是，现有口语语料库并不适用于分级的研究。在目前可见的口语语料库中，口语互动的内容所占的比例虽多，但是互动的场景却是复杂多样的，缺乏统一性。这就使话题的分布比较分散，容易使得到的数据过于稀疏，不利于进行统计分析，也不利于样例之间的横向比较。众所周知，印欧语口语语料库的构建工作开始的比较早也比较多，而且规模较大，如最早的英语口语语料库 London-Lund 口语语料库（Svartvik Quirk 1980, Svartvik ed. 1990），London-Lund 口语语料库中 50 万词的语料中涉及对话涉及的场景包括面对面交谈、电话交谈、讨论、采访、辩论等多种情景，不适宜进行直接的横向比较。因此，并不是理想的对口语互动做深入研究的语料库。另外如 British Academic Spoken English Corpus 、The Michigan Corpus of Academic Spoken English、Corpus of Spoken Professional American-English、香港英语口语语料库（The Hong Kong Corpus of Spoken English）等都存在类似的问题。而且，这些语料库中的语料样例在年龄分布上也不全面，虽可以比较相同或差异之处，但不能反映互动能力动态发展的过程。

目前人们已认识到这些问题，开始尝试进行一些专门的互动语料库和互动资源的建设工作，如专门的医患门诊口语语料库（Belvin and May 2004）的建立，Holmes（2003）进行的工作场合言语（The Language in the Workplace）项目等。但能全面反映互动能力发展过程的分级语料库还没有见到。

汉语方面，汉语口语语料库的构建工作起步比较晚，针对特定场景的汉语口语互动的语料库也比较少。目前可见的有台湾中央研究院的现代汉语口语对话语料库（Sinica MCDC）（曾淑娟 刘怡芬 2002）等一些零星的语料库构建，有些还属于个人所有并不开源，如医患会话语料库（杨石乔 2011）等，远不能满足口语互动研究者的需要。更为重要的是，可用来研究学生口语互动差异情况的口语互动分级语料库还没有见到。同时，网络技术的飞速发展使得音视频文件的在线查询与传输成为可能。对音视频文件进行转写只能无限逼近但始终无法取代现实场景的互动，音视频与转写资料的同步呈现可以弥补这种不足，使得研究的基础更加真实可信。因此，建立一个同时有音视频语料和转写语料的，记录不同年龄的学生口语互动情况的分级语料库将具有重要的意义。

2 语料库构建过程

我们构建了一个经过语料转写和会话标注的汉语口语互动分级语料库（Spoken Interaction Corpus of Mandarin Chinese, SICMC），以提供详实记录学校环境下学生口语互动

实际的语言资源供口语互动研究者使用。

2．1 语料库的规模

语料在形式上包括音视频文件、语料转写文件和会话标注文件三个部分。音视频文件大小约 750 个 GB，转写文本约 57 万字。语料在内容上有两人互动和多人互动两大类，由分层抽样的 1200 名中小学学生的对话录制而成，话题方面比较集中，总时长超过 3000 分钟，因此是一个较为大型的语料库。

2.2 语料库的构建

此语料库的构建过程可以分为四个阶段：

第一， 准备阶段

此阶段的主要任务为进行取样方面的论证工作，具体包括：文献梳理，确定会话分级的参照点和基本分类原则；确定语料库规模，采样的数量和单个时长；确定测试题目和会话的方式；准备相关文件（如学生信息表、家长同意书等）。

1）会话场景的选择

互动分级的研究，首先需要考虑的是语料之间可比较性的问题。只有在形式和内容上具有类似性的语料才存在比较和分级的可能，这就需要限定对话的形式和主要内容。另外，还需考虑外部因素对对话的影响，如声音、光线、时间、位置等。理论上来说，有全面消除外部影响因素或者保持相对一致的外部影响因素两种选择可能。但由于外部影响不可能完全消除，因此使外部影响因素做到统一是一种可行的方案。

经过讨论，我们确定会话发生的场景为统一的测试场景，这样可以使外部影响的因素做到相对一致，以实现相对的公平。各样本录制的时长也大致相同。这样得到的语料可以进行不同的会话之间的横向和纵向比较，研究其相同和不同之处。

2）话题的选择

话题的选择需要做到既不影响学生口语互动能力的充分发挥，又要能保证进行横向比较的可操作性。因此，不宜选用多样化的话题，而是需要在谈论的话题上相对比较集中，避免数据稀疏，并且要使各个年龄的学生都有话可说。经过讨论和实验取样，我们确定了两个相对集中的测试话题，如下表所示：

表 1 口语互动测试话题

话题	内容
话题一	日常寒暄
话题二	对课外辅导班的看法

两个话题有类别上的差异：话题一侧重于日常生活当中的口语互动；话题二则侧重于辩论性质的口语互动。而且，两个话题还照顾到了各个年龄段的参与者，基本能做到都有话可说。

话题集中可以有效地解决以往口语语料库存在的数据稀疏问题，同时也可以减少话题难易度差别对口语互动水平发挥的影响。具体来说，统一的话题既便于进行样本之间横向的比较，发现不同的互动之间的共同之处，也便于进行样本之间的纵向比较，发现不同的互动之间的差异之处，从而可以在此基础上确立口语互动分级的标准。因此，我们的分级标准是按照自下而上的步骤自然聚类的结果，是基于归纳方法的研究，与以往的以任务困难度为标准的分级做法是两种不同的处理思路。

3） 测试形式的选择

测试形式也有两种选择形式：一种是师生对话，一种是生生对话。目前，得以普遍使用的考察形式多为第一种，即师生对话的形式。显然，这样的考察形式是受口语测试传统的影响。师生对话的形式在对学生的口语能力考察方面，比如词汇量的考察等方面是非常有效的，但是在对口语互动的考察中却是弊大于利的。原因主要有以下几个方面：

第一，由于考察者的互动能力相对于被考察者要高出很多，很容易划定口语互动的框架，

进而控制话题的推进。而作为被考察者的学生往往落入被动的应答者局面，不能充分发挥和展示自己的互动能力。

第二，在考察的具体话题上，由于考察者与被考察者年龄存在较大的差异，他们对同一个话题感兴趣的点也会不同，不利于话题的拓展深入。

第三，由于考察者的强势地位，很容易让被考察者产生紧张和畏惧心理，从而压抑其互动能力的发挥。

所以，我们采用的是第二种测试的形式，即使用生生对话的形式。生生对话的形式便于学生的自由发挥，表现出真实的互动状态。

互动测试的时长平均控制在 8-10 分钟之间，个别低年级（小学 1、2 年级）控制在 5 分钟左右。这样既能使学生有充足的时间互动，又避免时间过长导致的节奏散漫。而且，统一的时间控制便于语料之间的比较。

4）评分体系

关于生生对话的评分体系，我们使用的是首先整体评价，然后分别给予两个参与者一个层级评定的做法。这是因为口语互动并不仅仅是说话者个体的行为，而是一个复杂的社会实践过程（Fairclough 1992）。在不同的情况下，同一个互动参与者表现出的水平是不同的，其中很大的原因在于参与对话的对话方不同。因此，在评分中需要充分考虑参与对话的不同对话方因素，先对整体的互动做出评价，然后再分别计算各自的互动表现。目前普遍采用的对对话参与者各自的互动行为进行评分的做法是不够全面的，忽视了互动参与者的影响因素。这也可以看出受口语评测形式影响的痕迹明显。

5） 口语互动分级标准

在本语料库中的语料存在有两个分级的标准：一个是自然分级，即按照学生所处的年级高低所做的分级，由于各年级的学生年龄大致相同，因而也可以说是按照年龄的自然分级；另一个则是根据口语互动水平的差异所做的分级。

两种分级形式可以在语料库中实现自由切换，便于不同的研究者做不同的研究使用。自然分级的数据便于研究者分别考察各年级（年龄）段学生的口语互动情况，而差异分级则便于研究者了解学生口语互动能力差异的情况。

确定口语互动分级的标准；互动分级目前没有像口语分级那样明确的操作标准，需要在数据观察的基础上分析总结。我们把一半的数据作为观察数据，然后用另一半数据做对比数据，验证总结的特征，这样总结得出全部汉语口语互动的分级标准。共分为六级，下表为一、二级口语互动的特征描述示例:

表 2 口语互动特征描述示例

	话轮掌控	互动合作	互动步骤	话题内容	伴随策略
一级	1.话轮开始方式单一，多为疑问句； 2.维持话轮手段贫乏； 3.话轮让渡方式单一，多为疑问句； 4.无反馈手段；	1.遵守合作原则； 2.遵守礼貌原则，但回答直接单一；	1.直接提问的方式开始； 2.回应简单； 3.没有自我纠正；	1.限周围熟悉的学习生活的社会语境； 2.话题内容前后完整一致，但使用的手段单一；	1.话轮构成简单，多为单句，一般不超过 3 句； 2.词汇使用简单，偶有使用错误； 3.无缓和话语的手段；

二级	1.话轮开始方式稍多，能够运用连词和感叹词； 2.维持话轮手段稍多，能够运用连词或重复对方的话； 3.话轮让渡方式稍多，能够运用疑问或反问句； 4.有反馈手段；	1.遵守合作原则； 2.遵守礼貌原则；	1.间接提问的方式开始，单一； 2.回应并能够提供相关细节内容； 3.能够自我纠正；	1.限周围熟悉的学习生活的社会语境； 2.话题内容前后完整一致，使用的手段多样；	1.话轮构成简单，多为单句，但数量增多； 2.词汇丰富，偶尔使用不当； 3.有缓和话语的手段；

第二，数据收集阶段

主要任务包括：进入采样学校，根据制订的测试话题，按年级收集互动数据，用摄像机记录互动的场景。

我们调查了泉州市 15 所中小学的学生口语互动情况，按照地理位置分布，照顾到各个方位的学校。数据分布如下图所示：

表 3 样例分布情况表

	调查数量（间）	学生人数	男女比例	样例数目
小学	5	600	1:1	300
中学	5	300	1:1	150
高中	5	300	1:1	150

第三，数据加工阶段

1) 视频选取与切割；

选取合适长度（小学 1、2 年级约 5 分钟，其他年级约 8-10 分钟）和内容的视频文件，设定切分时间点，使用 Adobe Premiere 软件对录制的视频进行分割。个别时间较长的对话，截取其中 8-10 分钟内容较完整的部分备用。然后对分割后的视频文件统一命名以备使用。

2) 音视频的分离与对齐；

使用 Adobe Premiere 软件进行音视频的分离工作。全部视频我们录有同步的清晰度更高的音频文件，可进行音频的替换。最后需要进行时间点对齐的工作，使最后呈现的结果能实现音视频的同步调用。

3) 视频内容的文本转写；

采用通用的口语转写符号对视频内容进行转写。转写的内容包括讲出的词语，发出的声音，听不到或者不能理解的声音和词语，沉默，重叠的话语和声音，讲话速度、拖音、重音、音量等。

转写标准我们主要参照的是美国会话分析学派使用的转写系统（Atkinson, J & Heritage, J. 1984），结合实际情况进行了简化，形成一个口语互动符号集。如下表所示：

表 4 口语互动转写常用符号集

符号	意义	符号	意义
201 202	说话者的表示	[	重叠起始点
// //	两个以上人同时说话	]	重叠终止点
=	紧密衔接话轮	<hhh>	表示吸气
（1）（2）（3）	分别表示 1 秒 2 秒 3 秒的停顿	hhh	表示呼气
？	表示升调	（hhh）	表示笑声
！	表示强调，是降调	（ ）	对动作行为注释

是.	“.”表示降调	…	表示不清楚
：	表示声音延长	[a]	方括号表示语音转写
是-	表示中断	（（））	研究者对现象的描述
是	表示重音	> <	表示语速较快
° 是 °	表示音量降低	< >	表示语速较慢
⁺是⁺	表示音量增大	→	提醒注意的地方
*	星号强调需讨论的现象	**是**	黑体字表示随后谈话会重复使用的话语

4) 口语互动的标注；

标注包括两个部分：第一部分是元信息（metadata）标注，即有关语料的非语言信息标注，包括语料的来源（学校信息）、录制时间、说话者代码、性别、年龄、年级、话题、等级等。元信息标注可以为语料库检索和分析提供不同的查询条件和依据。第二部分为互动标注，采用互动标记对口语转写的内容进行互动标注，标注的内容包括话轮、行为、话题、反馈、发起、保持、结束、强调等具体的会话特征。

5) 视频分类；

根据年龄、性别、年级、时间长度、主题、等级等元信息标签对视频及相应的音频、文本文件进行分类并交叉索引。

第四，数据整合阶段

1）网页设计；设计最后呈现结果的调用平台，要求能够实现数据在线查询的功能。能够根据使用者的不同实现三级不同的处理权限。有视频、音频、文本三级同步的可视化窗口。

2）数据库整合；包括将标注文本进行整理和清洁工作，统一格式（字体、间距、字号、模板等），输入数据库。

整个语料库的构建流程可以用图示的方式表示如下：

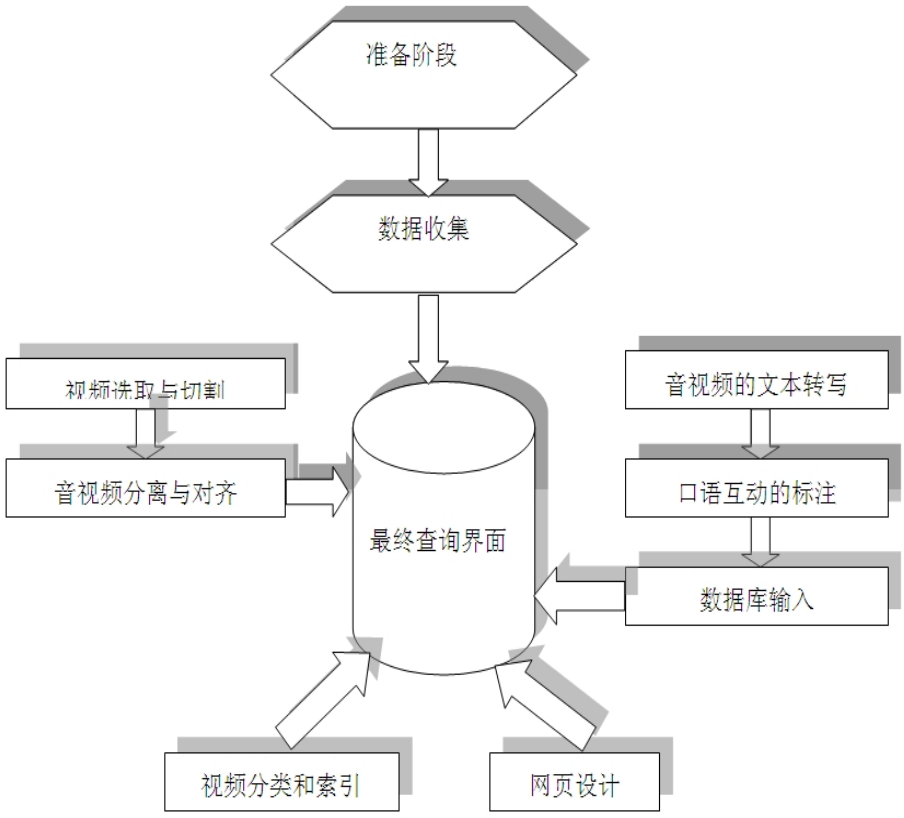

图 1 汉语口语互动分级语料库构建流程

2.3 语料库的呈现形式

目前，语料已经全部转写和标注完毕，正在做最后的核对和整合工作。最后的语料视频将使用面部遮挡的形式呈现，以保护访谈者的肖像权益。这些转写和标注的结果以文本形式保存在服务器后台，以供用户在线查询。我们提供了用户在线查询的网页界面，可以实现音视频与文本的同步调用。这包括三个层级的使用权限。

第一个层级的使用权限仅为学校内部及合作单位的研究人员。可以提供视频、音频和转写标注文本的同步对应检索，无条目数量的限定。可提供高级分类检索功能（根据词、短语、谈话类型、话题、学校种类、学生年龄、性别、互动时间、地点等），提供截图和文本下载功能。

第二个层级的使用权限为一般注册使用者，可提供音频和文本的同步对应检索，无条目数量的限定。可提供高级分类检索功能（根据词、短语、谈话类型、话题、学校种类、学生年龄、性别、互动时间、地点等），提供截图和文本下载功能。

第三级的使用权限为非注册用户，可以提供部分音频和文本文件样例的同步对应检索。有条目数量的限定（50 条）。可提供初级分类检索功能（话题、学校种类、学生年龄、性别），不提供任何下载功能。

3 语料库的构建意义

首先，可以为口语互动研究者提供经过了转写并标注了互动标记的实际语料。口语互动的研究是建立在对口语互动的录音录像及音视频的转写和标注基础之上的，这是一个非常费时费力的工作。由于个体研究的局限性，使得调查语料的范围很小，这会直接影响到所得结论的可信度。建立一个较大规模的口语互动语料库供大家使用，可以节约研究者单独研究时转写和标注的时间花费，从而节约大量的人力和财力。

第二，口语互动分级语料库的建立也可以实现标注标准的完全一致；语料的转写会受到诸多因素的影响，比如研究者的能力，研究者对原始材料的理解，研究者利用的转写体系，以及转写结果的阅读对象等。建立统一的转写语料，可以使得不同研究者之间具有相同的研究基础和对话的平台，从而使研究结论更加可信。而且，这些结论都是可以观测和验证的。

第三，口语互动分级语料库的建立可以实现对互动的量化分析。以往的互动研究多是基于个案的、比较笼统的定性分析，较少进行定量的分析。大型口语互动语料库的建立可以在更大规模上，实现各种互动特征的量化分析，使得分析的结果更加精细准确。有利于我们把定量研究和定性研究结合起来。

第四，有助于学校口语互动能力的培养。建立一个口语互动分级语料库，可以了解现阶段学生口语互动能力的实际状况，从而制订合理的教学目标。口语互动能力的教学与培养是目前教学一个突出的亮点。在新加坡已经有一些口语互动教材的编纂，在中小学也开始尝试进行口语互动的教学。这将是我们中小学教育一个新的努力方向。

第五，通过语料库可以了解各个年龄段学生口语互动能力的现状，也可以为将来制订口语互动等级的国家标准提供参考数据。

第六，口语互动教材的编纂将是口语互动语料库构建的受益者。互动教材编纂过程中可以借鉴语料库中收集的实际例子来设计场景、话题以及教学重点等。使得教材的设计更加贴近学生能力的实际情况，并且能够根据不同年级的特点更有针对性的设计教材以提高其口语互动的能力。

4 结语

本文简单介绍了一个汉语口语互动分级语料库的构建工作，包括话题选择、数据说明、数据加工和整合等具体的过程。该研究在一定程度上弥补了国内同类研究的不足，也可为以后类似语料库的构建提供参考。同时需要明确的是，本语料库对口语互动的分级只是一个初

步的尝试，具有实验的性质。受采样地域所限，样例在特征表现上会具有一定的趋同。而换个调查的地区，其表现则可能有所差异。因而，如果将来要从语料库研究制订具有国家水平的口语互动标准，需要扩大调查的范围。

参考文献

[1] Stenström, Anna-Brita. An introduction to spoken interaction[M]. Longman, London and New York , 1994.

[2] Sinclair, John. and M. Coulthard. Towards an analysis of discourse: The English used by teachers and pupils[M]. Oxford: Oxford University Press, 1975.

[3] Berry, Margaret. Systemic linguistics and discourse analysis: a multi-layered approach to exchange structure[A]. In: Coulthard, M., Montgomery, M. (Eds.), Studies in Discourse Analysis[C]. Kegan & Paul, London, 1981, pp. 120–145.

[4] Ford, Cecilia E., Wagner, Johannes,. Interaction-based studies of language: introduction[J]. Pragmatics, 1996, 6 (3), 277–279.

[5] Ochs, Ellinor, Schegloff, Emmanuel, Thompson, Sandra. Interaction and grammar[M]. Cambridge University Press, Cambridge, 1996.

[6] Stenström, Anna-Brita, Andersen, Gisle, Hasund, Ingrid Kristine. Trends in teenage talk: Corpus Compilation, Analysis and Findings[M]. Benjamins, Amsterdam, 2002.

[7] Aijmer, Karin, Simon-Vandenbergen, Anne-Marie,. The discourse particle well and its equivalents in Swedish and Dutch[J]. Linguistics, 2003, 41 (6), 1123–1161.

[8] Zhu Xinhua. Strategies and methods of authentic assessment in Chinese language[C]. In LEE Chi-kin, John, MA Hing Tong, KO Mo Lin (Eds). Building learning communities of Chinese language and mathematics: theories and practice. Nanjing: Nanjing Normal University Press.2011:67-69.

[9] Svartvik, Jan. and R. Quirk (eds) . A corpus of English conversation[M]. Lund: Lund University Press, 1980.

[10] Svartvik, Jan. (ed.) The London-Lund Corpus of spoken English: Description and research[M]. Lund: Lund University Press, 1990.

[11] Belvin, Robert. W. May, S. Narayanan, P. Georgiou, S. Ganjavi,. Creation of a doctor-patient dialogue corpus using standardized patients[A]. In Proceedings of the Language Resources and Evaluation Conference(LREC)[C], Lisbon, Portugal, 2004.

[12] Holmes, Janet. and M. Stubbe. Power and politeness in the workplace: a sociolinguistic analysis of talk at work[M]. London: Pearson, 2003.

[13] 曾淑娟、刘怡芬. 现代汉语口语对话语料库标注系统说明[R]，技术报告，2002。

[14] 杨石乔. 基于语料库的汉语医患会话修正研究[M]，广州：中山大学出版社，2011。

[15] Fairclough, Norman. Discourse and social change[M]. Polity Press, Cambridge. 1992:63.

[16] Atkinson, J. Maxwell. and J. Heritage, eds. Structures of social action: studies in conversation analysis[M]. Cambridge: Cambridge University Press. 1984.

作者简介：王跃龙（1979——），男，博士，主要研究领域为语料库语言学与中文信息处理。Email: wangyuelong_2001@126.com