中国学习者英语口语语料库建设与研究

Construction and Data Analysis of a Chinese Learner Spoken English Corpus

主编　杨惠中　卫乃兴

上海外语教育出版社
SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS
外教社

中国学习者英语口语语料库建设与研究

Construction and Data Analysis of a Chinese Learner Spoken English Corpus

主编　杨惠中　卫乃兴

外教社

上海外语教育出版社

SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS

图书在版编目(CIP)数据

中国学习者英语口语语料库建设与研究/杨惠中，卫乃兴编著. —上海：上海外语教育出版社，2005
ISBN 7-81095-784-8

Ⅰ. 中… Ⅱ. ①杨… ②卫… Ⅲ. 英语-口语-教学研究 Ⅳ. H319.9

中国版本图书馆 CIP 数据核字(2005)第 097714 号

出版发行：上海外语教育出版社
（上海外国语大学内） 邮编：200083
电　　话：021-65425300（总机）
电子邮箱：bookinfo@sflep.com.cn
网　　址：http://www.sflep.com.cn　http://www.sflep.com
责任编辑：李振荣

印　　刷：常熟市华顺印刷有限公司
经　　销：新华书店上海发行所
开　　本：787×1092　1/16　印张 13.5　字数 320千字
版　　次：2005年 10月第 1 版　2005年 10月第 1 次印刷
印　　数：3 100 册

书　　号：ISBN 7-81095-784-8 / H · 310
定　　价：27.00 元 (附CD-ROM)

《中国学习者英语口语语料库建设与研究》

主编：杨惠中　卫乃兴

编著者：

卫乃兴　李文中　濮建忠

项目组撰稿人：

第一章：卫乃兴
第二章：李文中
第三章：张　霞　甄凤超
第四章：甄凤超
第五章：卫乃兴　高　霞
第六章：卫乃兴　王晓婷
第七章：薛学研　卫乃兴
第八章：冯　友　濮建忠
第九章：卫乃兴
第十章：甄凤超

The ability to examine large text corpora in a systematic manner allows access to a quality of evidence that has not been available before.

J. M. Sinclair

In the final analysis if linguistics is not about language as it is actually spoken and written by human beings, then it is about nothing at all.

Michael Stubbs

序　言

中国学习者英语口语语料库(COLSEC)是国家哲学社会科学基金项目，是继“中国学习者英语语料库”(CLEC)之后建成的又一学习者英语语料库。建库的目的是为了对我国大学生英语口语能力的现状进行系统的描写，为提高英语教学质量服务。

随着改革开放的不断深化，我国大学生迫切要求提高英语口语能力，以便直接参与国际交流。改革开放二十多年来，我国大学英语教学取得了飞速的发展，大学生的英语阅读能力仅从阅读速度来看就已经从以前的每分钟不到20词提高到了每分钟70～150词，基本上能够满足“通过阅读获取本专业所需信息”的要求。但是，听说能力的发展则相对滞后。从1999年开始实施的我国大学英语四、六级口语考试，目前考生规模已达到每年近10万人次。考试结果表明，凡获得A等和B等证书的学生已经具备“就熟悉的题材用英语进行口头交际”的能力。这样大规模的英语口语考试在全世界并不多见，说明我国大学生的英语口头交际能力已有了很大程度的提高。但就全国每年入学的数百万大学生而言，大部分学生的英语口语能力仍然不强。这一现象已引起教育主管部门和社会的重视，一系列教学改革的举措已经出台。强化语言表达能力，特别是注重英语口语交际能力的培养，已成为当前和今后一个时期我国英语教学的重点和发展方向。

就人类语言的发展进程来说，书面语的出现远比口语要晚，没有书面语的语言为数不少，没有口语的语言则不可想像。从儿童语言的个体发展进程来看，也是先学会口语，后学习书面语。因此，就语言的本质来说，口语是第一性的，书面语是从属于口语的。但是，口语和书面语作为两种不同的交际方式服务于不同的交际目的，不可能互相取代。书面语和口语，两者各有自己的特点。首先，书面语是静态的，不受时空的限制，作者有充分的时间谋篇布局，仔细推敲。事实上，在书写过程中，语言使用的过程也常常是思想成形的过程，因此，书面语可以做到用词比较准确，思路比较严密；另一方面，书面语虽然也是作者与读者之间的一种交流，但交流的双方不在现场，而且读者对象往往是不确定的，又没有实时反馈，没有语境信息可以利用，作者不得不力求做到所指明确，才能避免歧义。因此，书面语常常是比较严谨的正式语体，口语则是动态的、即兴的、受时间制约的。一般来说，在口语交流中说话人没有时间进行充分的构思，没有时间反复推敲。因而，口语的话语结构往往比较松散，重复多，停顿多，填充语多，遣词造句远没有书面语严谨。即使是语言能力很强的人，如果把所说的话一字一句转写成书面语，不合语法的地方肯定很多。因此，口语常常是非正式语体。

口语交际的本质特征之一是交互性，参与者就一定的话题你一句我一句地相互交流。这一过程涉及what，why，when，where，who和how六个方面。成功的口头交际不但要求参与者掌握必要的语言知识，包括词汇、语法和语音语调方面的知识，掌握必要的背景知识，而且还要懂得如何使讲话得体，要懂得一定的会话策略。实时的会话具有不可预测

性，否则就变成照着剧本念台词，不是真实的言语交际。因此，参与者要懂得怎样利用反馈信息使话语正常展开，懂得怎样提问，怎样反驳，必要时可以要求澄清，自己也可以补充说明等等。会话策略还包括怎样开始交谈，怎样切入话题，怎样恰当地终止交谈等等。这些能力是需要培养的，我们的学生已经掌握了这些能力吗？口语教学中包括了这些能力吗？

这里所讨论的大学生英语口语能力不是指“你好”、“我好”之类简单的口头寒暄，而是指就熟悉的材料进行实时的、即兴的口头交际的能力。我们知道，英语在我国是作为外语进行教学的，大学英语教学又是针对成年人进行的英语教学。这一特点，加上历史的原因，使我国大学生的英语能力在发展上显得不平衡。就目前情况而言，我国大学生的英语领会能力总体上强于表达能力，阅读能力强于口语能力。在英语阅读能力达到一定的实用程度后如何迅速提高英语口语能力，怎样才能培养起具有实际应用价值的英语口头交际能力？除了必要的词汇、语法知识外，还要掌握什么才能用英语进行口头交际，做到既准确，又流利，又得体？这是摆在我国外语教学工作者面前的迫切课题，教师需要认真研究，学生也十分关心。为此，必须首先了解我国大学生的英语口语能力现状，这正是建立中国学习者英语口语语料库的目的和初衷。

语料库是语言学研究的强有力工具，学习者语料库则是研究学习者中间语的强有力工具。我们把学习者英语叫做中间语。中间语是一个连续统，从初学者的英语到掌握接近本族语者的地道英语，这是一个漫长的发展过程。学习者语料库一旦建成，研究人员就可以开展基于语料库的实证研究，多层面、多纬度地深入研究学习者语言的特征和发展模式，丰富语言学理论和二语习得理论，提出有针对性的、有利于提高我国外语教学质量的建议。但是，语料库的建库过程则是一项艰苦的、工程浩大的基础建设，一般人往往望而却步，尤其是学习者口语语料库的建设，不但涉及语料的采样、收集、校对，而且涉及口语语料的转写、标注、赋码等等繁重的工作。语料库建设除了语言学的层面，还有语料转写的技术标准、在线查询软件开发等技术方面的难题需要一一攻克。因此，中国学习者英语口语语料库的开发和建设是一项文理工相结合的跨越多个领域的科研项目，其困难和艰巨的程度可想而知。幸好有一批学者，不甘寂寞，不计报酬，孜孜以求，埋头苦干，坚持了好多年。除此以外，上海交通大学、解放军外国语学院、河南师范大学的部分博士生、硕士生也承担了该语料库的转写和标注工作，付出了艰辛的劳动，终于使这个容量达到 70 万英语词的中国学习者英语口语语料库得以建成问世，为研究我国大学生的英语口语特征提供了坚实的、可以共享的第一手资源。

现已初步建成的中国学习者英语口语语料库，语料来源主要是全国大学英语四、六级口语考试的实景录像语料。由于该项口语考试在方案设计上已经考虑到口语交际的交互性，使考生在真实交际的语境中表现自己的口语能力，因此能够真实反映我国大学生的实际英语口语能力。语料包括三种类型的口语交际：教师—学生型晤谈、学生—学生型自由讨论和教师—学生型讨论。每场考试 20 分钟，围绕着某个给定的话题展开。语料的转写力求做到真实、完整和准确。交流过程中的话轮转换、停顿、犹豫、语音语调以及非言语交际等信息逐一标注记录，因此本语料库可以成为研究我国大学生英语口语能力现状的有价值的数据资源。

中国学习者英语口语语料库的建成提供了一个坚实的平台，为开展一系列研究提供了条件。例如，学习者笔语和口语的共享特征和差异分析、学习者口语的话语结构分析、学习者口语的词语特点分析、学习者英语的语用特点分析、语音语调分析等等，这些研究都以学习者实际语言运用数据为基础，其研究结果一定会对我国英语教学，尤其是口语教学提供有价值的信息，揭示改进教学的方法和方向，帮助我国大学生掌握地道的英语口语，获得口语交际能力。

《中国学习者英语口语语料库建设与研究》一书不但包含了对中国学习者英语口语语料库的描写性信息，包括建库原则、采样方案、技术标准、各种相关统计信息等，而且也有一些初步的研究成果，这些研究成果至少在三个方面对改进我国大学英语口语教学有启示意义：

1. 关于会话结构。自然的口语交际构成一段会话，会话有一定的结构，东拉西扯，甲说甲的，乙说乙的，不可能构成有意义的会话。Sinclair 等首先采用起始—回应—后续（I-R-F）话语模式来分析会话结构，认为这三个话步（Move）构成一个回合（Exchange），每个话步包含若干个话目（Act）。话目是最小的话语单位，话步是最小的可以自由运用的话语单位。这种分析模式可以十分有效地分析会话结构，是一种强有力的分析工具。另外，话步用来实施一定的语用功能，以 F 话步来讲，或表示对对方的回应，或实施话轮交替等等，这就使交谈的双方有可能对会话进程实施有效的动态管理，进行有效的口语交际。这些能力是属于较高层次的能力。在英语口语教学中，不但要向学生传授必要的相关知识，更重要的是要教会学生实施这些话语功能的语言形式，基于 COLSEC 语料库的分析，可以发现哪些地方在教学中存在着不足，以便采取有针对性的措施来加强教学；

2. 关于词块。愈来愈多的语料库语言学研究成果发现词块在语言运用中起着很重要的作用。所谓词块，有的语言学家称为预构成词组，有的叫做词语预制件，不论叫什么名称，它们大多是大于单个词的固定话语片断或可扩展的半固定的话语片断，这些话语片断以一个个单位存储在记忆中，以便在说话需要时可以随时调用。

例如：

开始交谈时为了引起对方注意，可以说：

Excuse me, Sir ...

I'm sorry, but ...

在正式场合为引起听众注意，可以说：

May I have your attention, please?

Listen to me, please.

没有听清楚，要求对方重复，可以说：

I beg your pardon?

Would you please say it again?

Excuse me, but I didn't quite follow you.

不同意对方的意见，进行礼貌的反驳，可以说：

I can't agree more with you, but ...

What you said is quite all right, however ...

这些预构成词块减轻了说话人的编码负担，熟练掌握这些词块使说话人有可能更多地关注言语内容的组织，调整语调节奏，计划更大的话语单位等等，这是保证口语交际自然、流畅、准确的必要条件。从这个意义上来说，孤立地死记词汇表，脱离上下文来学习语法是费时低效的方法，是学不好语言的；

3. 关于主题词。基于语料库的研究还发现词语的使用取决于题材，采用统计方法可以找出一篇文章的主题词，其根据就是所谓的关键性(Keyness)。如果对同一题材的一组文章做关键性计算，则可以找出表达某一题材所必需的关键主题词。关键主题词通过一定的语义联系构成语义网络，存在于人的心理词库中；表达某一话题的相关主题词，通过一定的语义场构成主题词网络图。这些语义场包括施事者、行为词、受事者、位置词、描述词、联想词等等。各个语义场的具体词语越多，心理词库越丰富，说起话来就越自然流畅，恰当得体。研究表明，通过主题网络来组织口语教学，当可提高口语教学的效率和效果。

语料库的开发只是一种基础建设，开发的目的是为了进行研究。中国学习者英语口语语料库的建成为我国广大外语教学研究工作者提供了强有力的工具和手段。上面提到的研究只是初步的成果，欢迎大家充分利用这些可以共享的数据，对我国学习者英语口语从不同角度、不同层次进行深入的研究和分析，提出改进教学和提高教学质量的建议，帮助我国大学生掌握地道、自然的英语，如果这样，COLSEC 语料库开发者的辛勤劳动也就得到了最好的回报。

中国学习者英语口语语料库也有一定的局限性，主要表现在语料的局限性、标注的局限性和解释的局限性方面。COLSEC 语料库所采集的语料，由于时间、人力、财力的限制，无论在题材、体裁、库容量方面都有一定的局限性，未能尽如人意，标注也必然带有一定的主观性。此外，目前所收集的语料主要是大学英语口语考试学生的言语产出，虽然口语考试的设计已经考虑到了真实性原则，但是在考试场合下的真实性只是相对的，考场里的语言使用是一种非常态的语言使用，这一点研究者必须要注意。

另一方面，当我们强调口语重要性的时候，不要从一个极端走向另一个极端，不要忽视书面语的重要性。在强调培养学生听说能力的时候，不要忽视继续提高学生的读写能力的必要性。事实上，要提高语言输出能力，首先要加大语言输入量。语料库统计表明，英语本族语者的书面语语料库 BNC 共涉及354 300多个不同的英语词目，而 BNC 口语语料库中常用的英语词目只有3 000 多个。可见，阅读是扩大词汇量，提高英语总体能力的重要途径，不能忽视。

建设学习者英语口语语料库在我国尚属首次，基于语料库的口语研究也刚刚起步，书中一定有不少欠妥甚至不正确之处，欢迎读者批评指正。

杨惠中

2004 年 12 月 30 日

目　录

图表一览表

第一章

中间语研究与外语教学研究的新资源与新途径

——学习者英语口语语料库研究的理念、方法与实践

1. 基于学习者语料库的中间语对比研究

学习者语料库是按照一定的设计标准和原则通过科学抽样方法采集、编制而成的学习者语言(口语或书面语)电子文本库。基于语料库的学习者中间语对比研究(Contrastive Interlanguage Analysis,Granger 1998:12,以下简称 CIA),始兴于 20 世纪 90 年代,由比利时鲁汶大学的 Sylviane Granger 教授及其同事首倡。其后迅速发展,在多个国家与地区方兴未艾,成为语料库语言学、第二语言习得(SLA)研究和外语教学研究等领域一支重要力量。CIA 以其独特的理念和思想、严谨的方法和强大的数据,正在拓展和丰富第二语言习得研究的诸多内容,并对外语教学理论和实践产生颇具价值的反馈与指导。

1.1 理念、思想与方法

CIA 是语料库语言学与第二语言习得研究结合后新生的研究领域,是语料库语言学的发展与延伸。无论语料库语言学自 20 世纪 70 年代以来的复兴,还是基于语料库的 CIA 研究的兴起,都反映了语言学领域一种重要的学术思潮,代表了一种语言学理念,同时也揭示了一种研究方法。在哲学层面,语料库语言学隶属于经验主义的思想流派,它注重人的客观世界经验,强调语言使用者的实践,因而与理性主义相对。在语言学理论层面,语料库语言学认同于功能主义,与形式主义相对;它视语言为一种社会行为方式或社会现象,与心理活动结果或现象相对。视语言为一种社会现象,就意味着它的研究对象为外化语言,而非内化语言。外化语言是可观察的、可记录的、可描述的和可分析的文本数据。数据只能从自然交际活动忠实记录,而不能生造;本族语者虽有能力判断句子的合格与否,但个人的语言直觉有限;当一个语言学家为自己假设的理论模型自造证据时,会将其直觉证据扭曲,与真实语言交际活动的实例相去甚远。当自造数据不再合格和有效,由其支撑的理论模型之解释力便大打折扣。因此,真实数据,语言交际活动中的自然数据,而非语言学家头脑中的数据,是语言描述和语言学模型构建的生命所在。语料库语言学的这些理念和思想与乔姆斯基的理性主义思想泾渭分明。Sinclair 教授曾经一针见血地指出:

> Starved of adequate data, linguistics languished-indeed it became almost totally introverted. (Sinclair 1991:1)

在 Granger 教授主编的《基于计算机的学习者英语研究》一书中,Michael Stubbs 也题写道:

> In the final analysis, if linguistics is not about language as it is actually spoken and written by human beings, then it is about nothing at all.

这些都反映了语言学领域对理性主义某些做法的批判。然而,语料库语言学并非是对早期 John Firth, Leonard Bloomfield, Zellig Harris 等人经验主义思想和实证研究方法的简单承继和回归。借助于强大的计算机技术,语料库语言学在理念、思想和方法上都实现了许多超越,影响或开辟了许多新领域、新学科,促使相关研究活动兴起。

1.2 对比纬度与方法

"中间语对比研究"采用强大的真实语言使用证据,基于语言项目在语料库中的概率信息等量化数据,多纬度、多层面地对比分析至少两个语言使用者群体的言语行为,从而概括和描述学习者中间语的特征、模式及其发展规律,剖析和诊断中间语的错误及其原因所在。与经典 SLA 研究不同,CIA 不太预设研究假设,而采用开放的数据驱动方法;CIA 使用的语料库数据较之 SLA 惯用的内省数据(Introspective data)和诱导数据(Elicited data),更为客观、真实和海量。参照 Granger 的观点,根据我们的研究实践,CIA 以对比分析为其主要的特征方法,通常进行三类对比:

(1) NL *vs* IL
(2) IL *vs* TL
(3) IL *vs* IL

其一是学习者母语数据与学习者中间语数据的对比(Learner's Native Language *vs* Interlanguage),旨在发现中间语中存在的母语影响因素,包括母语迁移、学习者语言系统转换等。如基于中国学习者英语语料库数据和现代汉语语料库数据,研究汉语对中间语的影响等;其二是中间语数据和目标本族语数据的对比分析(Interlanguage *vs* Target Language),旨在发现中间语的非本族语特征(Foreignness)。目标本族语数据来自建构良好的本族语英语语料库,如 BROWN、LOB、COBUILD 和 JDEST 等;中间语数据一般来自具有良好代表性的学习者语料库,如"国际学习者英语语料库"ICLE (Granger, 1998: 9)、"中国学习者英语语料库"CLEC(桂诗春,杨惠中 2003: 3—9)等。研究的焦点可以是学生英语中的类联接、搭配、语篇、语义、语体等任何问题。其三是不同中间语数据间的对比分析(Interlanguage *vs* Interlanguage),旨在发现不同母语背景的学习者的异同,探究中间语的性质,如基于 CLEC 和 USE(Uppsala Student English)对比研究中国学习者与瑞典学习者的英语输出异同。由于这些特点,CIA 见长于数据对比、中间语特征描述和学习者行为概括(参见 Granger 1998:12)。

1.3 学习者语料库建设现状与特征

自上个世纪90年代至今,学习者语料库建设及其相关研究方兴未艾。Granger教授及其同事建成的涵盖20余个国家英语学习者语料的ICLE,标志着学习者语料库建设和研究的一个重要里程碑。这一阶段的学习者语料库建设有如下几个特征:其一,人们首选建设的无一例外全是书面语语料库,口语语料库则几乎无人问津。这与口语语料库,尤其是学习者口语语料库建设难度大、耗时、费力等因素不无关系。其二,建库者的目的不仅相同,多数建库者的目的是学术性的,旨在第二语言习得研究和外语教学研究,如中国语言学家桂诗春、杨惠中主持建成的CLEC(Chinese Learners' English Corpus)语料库,匈牙利学者Jozsef Horvath主持建成的JPU(Janus Pannnius University)语料库,波兰Lodz大学与英国Lancaster大学合作共建的REPLICA语料库,瑞典研究者建成的USE(Uppsala Student English)语料库,香港科技大学建成的HKUST(Hong Kong University of Science & Technology)语料库等等。也有为商业目的而建库的,如剑桥大学出版社为开发新的ELT词典建成的CLC(Cambridge Learner Corpus)语料库,朗曼出版公司为开发表达词典"Longman Essential Activator"建成的LLC(Longman Learners' Corpus)语料库等。这从一个侧面反映出,无论是学术研究者还是商业经营者都趋于认同学习者语料库在学生语言研究和描述中的不可替代作用。其三,学习者母语背景由单一性向多样性扩展,涵盖多种母语背景学习者语言的语料库问世,包括ICLE和MELD(Montclair Electronic Language Database)。据不完全统计,截至2002年,部分重要的学习者语料库有10多个。这些语料库的类别(口语/书面语)、用途(商业/学术研究)、建设地和学习者母语背景等信息见表1.1(此表参照Pravec 2002:82-83提供的信息修订而成)。

表1.1 国际主要学习者英语语料库信息

Name of Corpus	Type of Corpus	Purpose	Location of Corpus	Language Background
CLC	Written	Commercial	England	Various
CLEC	Written	Academic	China	Mandarin Chinese
HKUST	Written	Academic	University of Science & Technology Hong Kong, China	Cantonese
ICLE	Written	Academic	University of Louvain, Belgium	Various
JEFLL	Written	Academic	Meikai University, Japan	Japanese
JPU	Written	Academic	University of Pecs, Hungary	Hungarian
LLC	Written	Commercial	England	Various
MELD	Written	Academic	Montclair State University, USA	Various
PELCRA	Written	Academic	University of Lodz, Poland	Polish

续表

Name of Corpus	Type of Corpus	Purpose	Location of Corpus	Language Background
TSLC	Written	Academic	Hong Kong University, Hong Kong, China	Cantonese
USE	Written	Academic	Uppsala University, Sweden	Swedish

2. 学习者口语语料库研究:语料库语言学发展之必然

2.1 口语与书面语:理论的反思

一般语言理论的研究和创立是基于口语语料还是基于书面语语料,反映出研究者的重要语言学立场。长期以来,尽管多数人都认同这样的观点,即口语是语言交际活动的首要方式和手段,书面语在人类语言系统发展进程中的出现相对要晚,与口语的重要性相比也居第二位。然而,相当多的研究者由于种种原因,一直将重点放在书面语而非口语上。以至于很多语言描述体系基于书面语建立,语言学模型依据书面语创设;反过来,人们又将这些体系和模型强加于口语语料之上,用于研究口语活动的特征,并继而指责后者"混乱"和"不合法"。近年,语料库研究的重要发现之一则是:自然语言的口语活动所使用的许多范畴同样可用于书面语描述,反之则不然。所以 John Sinclair 呼吁我们更多地注重口语研究,尤其是注重自然、即兴口语的研究,而不是相反(见 Sinclair 2004: 118)。语料库研究的发现促使人们对理论和实践作新的反思。如今,相当多的大型语料库建设项目都包括了口语语料部分,如 COBUILD、BNC 和 ICE 等等。

学习者语料库建设对口语的忽视不能不说是个缺憾,因为从理论的层面看,学习者口语对一般语言理论的建立也极为重要。学习者语料库不仅仅是研究中间语错误的数据资源,它还是研究人类认知、习得和使用语言机制与规律的重要资源。即使是中间语错误,也不再被视为偏离常规的变异形式,而被看做学习者对第二语言系统积极假设并检验假设的尝试性结果。特征性错误往往揭示语言系统发展的阶段与规律。

2.2 第二语言习得研究与外语教学研究的需求

第二语言习得研究与外语教学研究同样呼唤学习者口语语料库的建立。仅凭书面语证据抽象和构建的习得理论往往失之偏颇,只有加上了口语语料,习得研究才能日臻完善。而且,实时、即兴的口语交际往往更能准确地反映出学习者的真实外语能力。由于当今人类活动日益全球化的趋势和国家外语教学方针的导向现实,口语教学变得日渐重要。有关的教学研究也从过去的仅注重接受技能研究向接受技能与表达技能并重转变,因而迫切需要有效、丰富和大量的学生口语语料证据。教学人员也急需来自学生口语语料的反馈与启示。据此,学习者口语语料库建设应是继书面语语料库之后的必然发展,语料库研究人员再也不能因其难度大和挑战性强而回避之。

3. 中国大学学习者英语口语语料库

3.1 动机与目的

基于理论与实践两方面的考虑，COLSEC(College Learners' Spoken English Corpus)项目组成员继 CLEC(桂诗春，杨惠中 2003)项目之后，启动了我国国内第一个学习者英语口语语料库的建设。项目组旨在建设可用于我国大学生英语口语能力研究的数据资源。CLEC 无疑为研究中国学生的英语书面表达提供了坚实而有价值的数据资源(见濮建忠 2000，李文中 2003，桂诗春 2004，等)。但是，要全面系统地研究和描述中国学生的英语表达，学习者英语口语语料库就显得不可或缺。再者，我们期望将 COLSEC 建成 CLEC 的姊妹库，以便进行口语和书面语表达的对比，并由此引发一系列的研究。不仅如此，项目设计者和主持人杨惠中教授尤其注重让我们摸索方法、积累经验，建立一套可资借鉴的建库原则、采样方案和技术标准，促进本学科的发展；以期能在某种程度上填补国内学习者口语语料库建设方面的空白，使国内的语料库研究逐步接近国际主流水平，便于国际交流。

3.2 语料特征与容量

COLSEC 的语料来源为全国大学英语考试口语考试部分的实景音像资料。语料涵盖口语考试的三部分内容：(1)教师—学生型晤谈；(2)学生—学生型自由讨论；(3)教师—学生型讨论。每场考试三个部分的会话围绕内容相关的几个具体话题展开。语料选择采用随机比例抽样方法，对考生的地区来源、专业、考试成绩(杨惠中，1999)、交谈话题等按比例选取。语料的转写按照"真实"、"完整"和"准确"原则，标注采用 XML 语言，用一系列符号对话轮转换、语音、语调、停顿、犹豫、打断、非言语交际等进行逐一标注。全库的主要统计信息包括：总形符数为 723 299 个；总类符数为9 192个；类符/形符比为 1.27；标准类符/形符比为 28.44；总话轮数 18 969 个；学生话轮数 11 994 个；教师话轮数 6 975 个；平均话轮长度 38.13 词次，涉及 39 个讨论话题等。

就容量而言，COLSEC 大体上属于中等规模的语料库。与国际上著名的两个口语语料库相比，它大于 ICE 的 719 578 词容，小于 BNC 的9 853 249词容。如果将国际上著名的学习者口语语料库(包括口语和书面语)放在一起考虑，COLSEC 基本上也是个中型语料库。下面的表 1.2 提供主要的用于学术研究目的的学习者语料库库容信息，供读者参照。

表 1.2　国际主要学习者英语语料库库容

Name of Corpus	Size of Corpus (words)
HKUST	25,000,000
TSLC	3,000,000
ICLE	2,000,000
CLEC	1,000,000