教育部人文社会科学研究一般项目资助

基于SECOPETS语料库的中国学习者英语口语研究

主编 肖德法 向平

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

教育部人文社会科学研究一般项目资助(05JA740029)
《中国社会各阶层学习者的英语口头表达能力的现状调查及对策》

基于SECOPETS语料库的中国学习者英语口语研究

主　编　肖德法　向　平
副主编　邓耀臣　苏　勇

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

图书在版编目(CIP)数据

基于SECOPETS语料库的中国学习者英语口语研究 / 肖德法,向平主编. —上海:上海外语教育出版社, 2008
ISBN 978-7-5446-1110-7

Ⅰ. 基… Ⅱ. ①肖…②向… Ⅲ. 英语-口语-研究
Ⅳ. H319.9

中国版本图书馆CIP数据核字(2008)第176130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:徐 喆

印 刷:上海叶大印务发展有限公司
经 销:新华书店上海发行所
开 本:850×1168 1/32 印张13.25 字数332千字
版 次:2008年12月第1版 2008年12月第1次印刷
印 数:1 000册

书 号:ISBN 978-7-5446-1110-7 / H·0464
定 价:36.00元

序

基于语料库考察学习者的口头中介语是近年来第二语言习得研究中的一个热点，这种二语习得和语料库语言学交叉而产生的新方法，为我们深入观察第二语言发展的特点和规律提供了一种新视角。基于学习者的产出性数据，研究者可以从不同的角度和维度对口语与书面语的大样本开展研究，对比学习者与本族语者、母语与中介语、母语与目标语、不同母语背景的学习者、不同水平的学习者、学习者口语与书面语等。鉴于上述原因，国内外先后建成了多个学习者语料库，如比利时建成的国际英语学习者语料库（ICLE）、我国的中国英语学习者语料库（CLEC）和中国学生英语口笔语语料库（SWECCL）等。公共英语等级考试口语语料库（SECOPETS）就是最近完成的一个，该语料库容量为 100 万词，由 PETS 口试 1－4 级录音和转写语料组成，是十分珍贵的研究数据，与其他几种类型的语料互为补充。作为早期参与过语料库设计的人员之一，我相信，该语料库将极大地推动我国的外语教学研究走向深入，基于语料库的各种研究成果将会不断出现。本书就是这种成果的具体体现，收集了肖德法院长和他领导的研究团队的最新研究发现，让我们从语音、句法、语篇、交际策略、话语结构、言语

错误等不同层面和维度加深了对中国人学英语的特点和难点的认识。我期待有更多这样的科研成果问世,与读者分享。

对外经济贸易大学教授、博士生导师

王立非

2008.8.18

前 言

“说”是外语表达的核心技能之一，但也是学习的困难所在。由于口语表达的言语行为比书面语表达更加难以收集和分析，因此，国内外对二语学习者口语表达的研究也相对薄弱。自 20 世纪 90 年代以来，在引介国外二语口语研究成果和方法的同时，国内部分学者也开始致力于学习者的二语口语研究，特别是在口语流利性、口语交际策略的使用、影响口语能力发展的因素以及在英语口语标准化测试的可行性方面取得了丰硕的成果。这些研究成果对我国英语学习者口语表达能力的认识和口语教学有着重要的理论意义和指导价值，给外语教育的改革和决策提供了重要的依据。但是，以往研究大多局限于在校学生样本，样本的社会背景单一，对多背景的英语学习者的口头表达能力的现状调查研究不多，尚未能反映出中国社会不同阶层的英语口语表达能力的全貌。另外，由于以往研究所基于的语料基本上是静态的，动态性的很少，对于揭示中国学习者英语口语能力的发展规律和阶段显得软弱无力。正是本着应对这一问题的目的，在教育部人文社会科学研究资助项目的推动下，我们产生了建立 SECOPETS（Spoken English Corpus of Public English Test System）的想法并得以建成目前的规模。该语料库语料涵盖 PETS 口试 1－4 级不同级别的语料，具有动态性，更适合开展中国学习者英语口语发展特征和规律以及发展阶段的研究。在 SECOPETS 语料库的基础上，我们初步开展了一系列的研究。

全书共七章。第一章重点介绍了该库建设背景、语料收集、抽样、语料转写以及语料标注原则和言语失误赋码的方案并概述了已有研究成果。第二章内容包括中国社会成员英语口语水平及发展趋势、二语发展阶段、考生年龄和级别分布及成绩分析、性别与口试成绩等研究结果。第三章是有关经济发展与社会成员英语口语水平关系的研究成果。第四章报告了基于 SECOPETS 语料库所进行的中国学习者英语口语中言语失误的分析结果，重点介绍了中国学习者英语口语中常见的言语失误类型及其所揭示的发展特征和模式。第五章涉及交际任务对交际策略的影响、交际策略与口试成绩的关系和交际策略的年龄差异等研究内容。第六章介绍了中国学习者英语口语中加音和语音错误的研究结果。第七章探讨了 PETS 口试的话轮与会话结构。

本书后有三个附表。附表 1 是 SECOPETS 语料库及其子库词形、词长、频数及累积覆盖率对照表，附表 2 是 SECOPETS 词目、频数及累积覆盖率对照表，附表 3 是用于对比的 853 个 SECOPETS 词目表。

本书可供外语教师和第二语言习得与外语教学研究者参考查阅。希望本书的出版会对二语口语特别是二语口语的发展阶段、规律与特征的研究起到一定的推动作用。

本书所基于的 SECOPETS 语料库建设和开发的研究只是初步的，已开展的研究是探索性的，研究的结果有待未来大量相关研究结果验证。语料库的容量需进一步扩大，语音、词汇、句法、语义、语用、文化等层面的诸多研究有待进行。

受本书作者的水平以及研究的时间和人力物力所限，书中疏漏和错误在所难免，敬请广大读者和同行批评指正。

肖德法

2008 年 8 月

目　录

第一章

SECOPETS 语料库的设计原则及研究综述

肖德法　邓耀臣

一、研究背景

语料库语言学和第二语言习得研究的有机结合促进了学习者语料库研究的迅猛发展。近年来，为了满足不同研究的需要，各种学习者语料库在世界各地相继建成问世。较为著名的包括 ICLE 语料库（International Corpus of Learner English）、中国的 CLEC 语料库（Chinese Learners' English Corpus）、COLSEC 语料库（College Learners' Spoken English Corpus）、SWECCL 语料库（Spoken and Written English Corpus of Chinese Learners）、匈牙利的 JPU（Janus Panniu University）语料库、瑞典的 USE（Uppsala Student English）语料库、香港科技大学的 HKUST（Hong Kong University of Science and Technology）语料库等。这些语料库的建成和应用无疑为研究者多纬度、多层面地揭示中介语的特征提供了客观、真实和丰富的语料，“在一定程度上拓展和丰富了第二语言习得研究的内容，并对外语教学理论和实践提供了颇具价值的反馈与指导”（卫乃兴，2007：235）。

然而，现有学习者语料库除 COLSEC 语料库和 SWECCL 语

料库中的 SECCL 子库外，大都以学习者书面英语为基础，主要可供针对学习者书面英语的使用特征的研究使用，而可用于探讨学习者英语口语特征的语料库较少。即使 COLSEC 语料库和 SECCL 语料库也仅包含处于某一特定学习阶段的特定群体的英语口语语料，语料在本质上是静态的，缺乏动态性。这两个语料库可以为研究者提供丰富的言语失误信息，通过和本族语口语语料库的比较，可以有效地揭示课堂教学环境下大学生英语口语在语音、词汇、语法、语篇等各个层面的使用特征，但它们不能准确地反映同样学习环境下处于不同语言发展阶段的学习者以及不同学习环境下处于不同语言发展阶段的学习者的英语口语特征，很难较系统地折射出学习者英语口语发展的过程和规律。正是在这种背景下，我们产生了建设 SECOPETS 语料库（Spoken English Corpus of Public English Test System）的想法并经过几年的努力建成了目前的规模。SECOPETS 语料库语料涵盖了 PETS 口试 1－4 级不同级别的语料，因此具有明显的动态性，更适合用来研究处于语言发展不同阶段的学习者英语口语特征及其发展过程和规律。本章从总体设计方案、语料的收集、语料的转写和标注以及语料处理工具的开发四个方面简要介绍该语料库的设计原则，并报告了该语料库的总体统计信息和目前的研究结果。

二、语料库的设计原则

1. 总体设计方案

SECOPETS 语料库是为了完成教育部人文社会科学研究资助项目《中国社会各阶层学习者的英语口头表达能力的现状调查及对策》而设计开发。主要研究目的是：力求从我国社会不同背景、不同地区、不同学历、不同职业、不同年龄、不同性别的英

语使用者的口语表达数据中发现他们的口语水平现状和特点，从而较全面地勾画出中国不同阶层学习者的英语交际能力。基于这一目的，SECOPETS 语料库的设计总规模为 100 万词，包括 SECOPETS1、SECOPETS2、SECOPETS3、SECOPETS4 和 SECOPETS5 五个子库，设想分别代表学习者英语发展的五个不同阶段，每个子库设计容量为 20 万词。在语料的收集和加工过程中，我们遵循了以下四条原则：

1）力求所收集的语料具有广泛的代表性；

2）入库的话语尽可能具有自然性和互动性；

3）对入库语料进行真实、准确地转写；

4）对入库语料进行客观、实用地标注。

2. 语料的收集

语料库建设过程中，研究者首先面对的问题就是如何确定和选取入库语料。实践证明，语料的选取不能依据语料的内部标准，即语言学标准。这样做只能导致语言研究的自循环，必将降低语料库研究结果的可信度。因此，和其他口语语料库的建设一样，SECOPETS 语料的选取也依据了语料的外部标准，即根据“文本或话语的交际功能”（Sinclair，2005）而确定的一些社会语言学标准，包括话语的任务场景、话语类型、话题、学习者的背景等因素。

SECOPETS 语料库的语料收集过程包括采集和选取两项任务。该库的所有语料皆取自 2004 年至 2007 年全国公共英语等级考试口试录音。公共英语等级考试是一种以考查考生的语言交际能力为核心的多级别英语考试体系，根据难度由低到高分为五个级别。目前语料主要来自山东、辽宁、江苏、浙江、湖南、河南等 10 个省市的 30 个考点抽取的 2000 个样本。为了保障所选取的语料具有较广泛的代表性，在选取阶段，研究者重点考虑了语

料提供者的性别、年龄、受教育程度、职业、所属地市的经济状况、学习英语的年限以及学习者英语水平等社会因素。基于这些外部标准，研究者最终确定 1200 个样本作为入库语料（由于五级考生较少，目前整理的语料仅包含 1－4 级考试样本）。

SECOPETS 语料库建设的主要目的之一是研究学习者运用英语进行口头交际的能力，特别是接受能力、产出能力和互动能力。全国公共英语等级考试口试部分的三种任务类型都涉及考官－考生、考生－考生或考官－考生－考生之间面对面的直接交流，形式包括简单的问答、就某一特定话题进行无准备的讨论或辩论。因此这种任务场景在一定程度上保障了入库语料的话语交互性。另外，SECOPETS 语料库语料的采集都是在考生不知道自己的谈话在被录制的情况下完成，这就尽可能地保障了交谈的自然性。

3. 语料的转写和标注

和书面语语料库建设相比，口语语料库建设面对的一大难题是如何有效地将声音信息转化为文字，即如何进行语料转写。与结构分明、条理清晰的书面语不同的是，口语产出过程中，讲话者会出现重复、口误、省略部分甚至整个单词的现象。有时他们还会暂时中止话语并重新组织要表达的信息，或者一句话未讲完而停止，或者打断别人讲话。要将这些特征在纸面或屏幕上显示出来确实是一项有挑战性的任务。为了保障计算机检索结果的一致性和可靠性，进行 SECOPETS 语料转写时，研究者就一些话语特征如停顿的时间长短、重复、插话的方式、讲话的语调等确定了明确的转写原则。另外，对于单词拼写体系和非言语声音表示方法的选用，研究者也作了明确规定。有关单词拼写的规范包括：（1）所有单词采用英式拼法；（2）对于考生发音有误的单词，只要转写者能确认该词，在转写时都采用正确拼写；（3）缩略词全部用大写；（4）所有数字都要转写成英语单词，禁止

使用罗马或阿拉伯数字等。表 1 列出了各种非言语声音的转写标记以及所代表的意义。

表 1 非言语声音转写标记及表达的意义

标记	表达的意义
er, erm, mm, en	to indicate hesitation
uh-uh	to indicate a negative answer
uh-huh	to agree with or show understanding of what has just been said
uh-oh	to indicate that a person has made a mistake or done something wrong

为了便于研究者进一步探讨中国学习者英语口语特点以及存在的问题，我们采用 XML 格式对 SECOPETS 语料库语料进行了两方面信息标注：基本信息标注和言语失误标注。基本信息标注包括 20 种赋码，按内容分为三类，分别代表：（1）话语结构；（2）考生信息；（3）话轮转换。表 2－4 列出了基本信息赋码集及说明。

表 2 话语结构赋码集

赋码	说明
< discourse > </discourse >	话语开始和结束
< filename > </filename >	文件名
< head > </head >	语料库头文件信息
< body > </body >	实际话语内容
< Task1 > </Task1 >	话语任务类型 1
< Task2 > </Task2 >	话语任务类型 2
< Task3 > </Task3 >	话语任务类型 3

表 3 考生信息赋码集

赋码	说明
< year > < /year >	话语时间
< spot > < /spot >	话语地点
< sexA > < /sexA >	考生 A 性别
< sexB > < /sexB >	考生 B 性别
< ageA > < /ageA >	考生 A 年龄
< ageB > < /ageB >	考生 B 年龄
< PETS_level > < /PETS_level >	报考级别
< scoreA > < /scoreA >	考生 A 成绩
< scoreB > < /scoreB >	考生 B 成绩

表 4 话轮转换赋码集

赋码	说明
< I > < /I >	考官话语
< A > < /A >	考生 A 话语
< B > < /B >	考生 B 话语

语料库头文件的样本：

< discourse >
< filename > L2-06-16-24. txt < /filename >
< head >
< year > 2006 < /year >
< spot > YANTAI < /spot >
< sexA > female < /sexA >
< sexB > male < /sexB >
< ageA > 19 < /ageA >
< ageB > 20 < /ageB >
< PETS_level > 2 < /PETS_level >

<scoreA >2 </scoreA >
<scoreB >3 </scoreB >
</head >

<body >
. . . . . .
</body >
</discourse >

4. 语料处理工具的开发

根据 SECOPETS 的标注格式，研究者运用 Visual Foxpro 语言开发了专用语料分析工具——SECOPETS 语料库分析软件包（Toolkit of SECOPETS Analysis，TOSA）。该软件包主要包括三大模块：语料标注、语料分析和数据统计。语料标注模块的主要功能是帮助研究者整理语料并根据既定标注方案对语料进行赋码；语料分析模块主要包括特定语料提取（如仅提取某一性别的学习者语料或成绩等）、词表生成、词串生成、句子切分、搭配抽取、搭配词检索和词形归并功能；数据统计模块可实现语料库形符和类符的统计、词长和句长的计算、原始频数的标准化处理、不同语料库之间频数差异的显著性检验以及语料库词汇增长率、覆盖率和重复率的计算。

三、研究结果

本节从语料库的容量、词汇和句法三个方面报告 SECOPETS 语料库的总体统计信息（所有统计结果不包括考官话语），并简单介绍迄今为止取得的相关研究成果。

1. SECOPETS 语料库的容量

表 5 统计结果显示，到目前为止业已转写完毕的 SECOPETS 语料库及其子库的样本数为 609 个，形符数为 316989 个，其中 SECOPETS1 分别是 163 个和 57112 个，SECOPETS2 是 122 个和 67969 个，SECOPETS3 166 个和 87583 个，SECOPETS4 158 个和 104325 个。

表 5 SECOPETS 语料库及其子库样本和形符总数

语料库	样本数	形 符
SECOPETS1	163	57112
SECOPETS2	122	67969
SECOPETS3	166	87583
SECOPETS4	158	104325
TOTAL	609	316989

2. SECOPETS 语料库的词汇统计

在词汇信息方面，除了对 SECOPETS 语料库的词汇总体进行了统计外，我们还从词长和词目两方面探索了中国学习者英语口语中词汇的使用特征以及随着学习者英语水平提高这些特征的变化规律。

1) SECOPETS 语料库词汇统计

表 6 显示的是 SECOPETS 语料库及其子库词汇统计结果。

表 6 SECOPETS 语料库总体词汇及其子库词汇统计结果

统计量	SECOPETS	SECOPETS1	SECOPETS2	SECOPETS3	SECOPETS4
形符	316989	57112	67969	87583	104325
类符	3897	782	1183	2287	2988