中国学生英语口笔语语料库

Spoken and Written English Corpus of Chinese Learners

（1.0版）

文秋芳 王立非 梁茂成 编著

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

中国学生英语口笔语语料库

Spoken and Written English Corpus of Chinese Learners

（1.0版）

文秋芳 王立非 梁茂成 编著

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS
北京 BEIJING

图书在版编目(CIP)数据

中国学生英语口笔语语料库/文秋芳等编著. —北京:外语教学与研究出版社, 2005.6
ISBN 7-5600-4973-7

Ⅰ.中… Ⅱ.文… Ⅲ.英语—高等学校—教学参考资料 Ⅳ.H31

中国版本图书馆 CIP 数据核字(2005)第 069883 号

出 版 人: 李朋义
责任编辑: 胡伟春
封面设计: 邹 蕊
版式设计: 付玉梅
出版发行: 外语教学与研究出版社
社 址: 北京市西三环北路 19 号 (100089)
网 址: http://www.fltrp.com
印 刷: 北京外国语大学印刷厂
开 本: 787×1092 1/32
印 张: 4.375
版 次: 2005 年 7 月第 1 版 2005 年 7 月第 1 次印刷
书 号: ISBN 7-5600-4973-7

* * *

前　言

“中国学生英语口笔语语料库”（Spoken and Written English Corpus of Chinese Learners，下文简称 SWECCL）系南京大学“211 工程”二期子项目，由南京大学外国语学院和外语教学与研究出版社共同建设。本语料库收录了我国大学英语专业学生的口语和笔语语料 200 多万词，可供大、中、小学英语教师、大学生、研究生、社会各界英语爱好者开展英语教学研究和学习使用，也可作为英语教材编写、英语教学测试、英语师资培训、英语网络课程建设的参考依据。

本语料库包含三张光盘和一本手册，光盘内刻录了中国大学英语专业学生的口语和笔语语料。口语部分有经过词性赋码的文本，还配有语音样本及由语音转写而来的生文本；笔语部分也提供了生文本和经过词性赋码的文本数据。手册共分两部分。第一部分为第 1—3 章，介绍学习者语料库研究发展概况和 SWECCL 建库情况；第二部分为第 4—5 章，详细说明 SWECCL 工具的安装和使用方法。

本项目由文秋芳总体设计与审订，王立非负责语料库的建设，梁茂成负责语料库的技术设计。本手册由王立非撰写，梁茂成参与修改部分章节，文秋芳定稿。本语料库面广量大，参与其中的机构和人员很多。南京大学英语口语研究所为本语料库提供了全国英语专业四级口试考生的录音资料和 1996—2002 年的口试考题。全国有 9 所高等学校的英语专业学生为本语料库提供了书面作文语料。建库过程中许多师生参与了不同阶段的不同工作。他们有的参加了作文文本的收集，有的参加了文本转写、输入和校对，有的参加了磁带录音的转录，有的参加了文本的切分和文本头的标注。值得一提的是南京大学的朱叶秋和陈萱博士对部分口语语料中冠词和动词过去时的使用进行了专门的人工标注，这是一项很细致的工作。在语料库的建设初期，英国伯

明翰大学的 S. Hunston 教授、华南师范大学的何安平教授、上海交通大学的卫乃兴教授、河南师范大学的李文中教授、洛阳外国语学院的濮建忠博士，都对本语料库的设计提出了宝贵的建议。在此，我们要对给予语料库建设支持与帮助的所有单位和个人表示衷心的感谢。

本项目还得到了北京外国语大学中国外语教育研究中心的资助，光盘及手册的出版还得到了外语教学与研究出版社的大力支持，责任编辑对语料和手册提出了宝贵的修改意见，在此一并表示感谢。

建设本语料库耗时费力，特别是口语录音的转写全部靠人工完成，因此，疏漏与错误在所难免，我们恳请广大读者和同行提出批评和意见。

编者

2005 年 6 月于南京

缩略语

1. ARG: Argumentation（议论文）
2. DVF: Digital Voice File (数码语音文件)
3. DVP: Digital Voice Player（数码语音播放器）
4. EFL: English as a Foreign Language（英语作为外语）
5. EXP: Exposition（说明文）
6. GR: Grammar（语法）
7. ICLE: International Corpus of Learner English（国际英语学习者语料库）
8. IL: Interlanguage（中介语）
9. LC: Learner Corpus（学习者语料库）
10. L1: First Language（母语）
11. L2: Second Language（第二语言）
12. NAR: Narration（记叙文）
13. SECCL: Spoken English Corpus of Chinese Learners（中国学生英语口语语料库）
14. SWECCL: Spoken and Written English Corpus of Chinese Learners（中国学生英语口笔语语料库）
15. TEM4: Test for English Majors Band 4（英语专业四级考试）
16. TL: Target Language（目标语）
17. SP: Spelling（拼写）
18. STTR: Standardized Type/Token Ratio（标准化类形符比）
19. TTR: Type/Token Ratio（类形符比）
20. WCOMP: Written Composition（书面作文）
21. WECCL: Written English Corpus of Chinese Learners（中国学生英语笔语语料库）

目 录

图表 …… Ⅵ
缩略语 …… Ⅷ

第一部分 学习者语料库介绍 …… 1
第一章 绪论 …… 1
1.1 SWECCL 的建设背景 …… 1
1.2 SWECCL 语料库概述 …… 4
1.3 SWECCL 项目组概况 …… 6
1.4 SWECCL 的安装 …… 7
1.5 学习者语料库的研究方法 …… 9

第二章 SECCL 的设计与建立 …… 11
2.1 SECCL 的建库程序 …… 11
2.2 语料的来源 …… 12
2.3 样本的选择 …… 15
2.4 转录与转写 …… 17
2.5 规模与结构 …… 19
2.6 统计描述 …… 20
2.7 文本头标记 …… 24
2.8 口语特征标注 …… 27
2.9 部分语法错误标注 …… 28
2.10 词性赋码 …… 29

2.11 小型专题研究语料 …… 31

第三章 WECCL 的设计与建立 …… 33

3.1 WECCL 的建库程序 …… 33

3.2 样本的选择 …… 35

3.3 样本的采集和录入 …… 36

3.4 语料库的规模 …… 36

3.5 语料库的结构 …… 37

3.6 统计描述 …… 39

3.7 样本的标注 …… 41

3.8 词性赋码 …… 44

第二部分 SWECCL 的工具简介 …… 46

第四章 SWECCL 的检索工具 …… 46

4.1 WordSmith Tools 概况 …… 46

4.2 WordSmith Tools 的安装与运行 …… 48

4.3 WordSmith Tools 程序的处理单位 …… 48

4.4 WordSmith Tools 文件菜单的操作 …… 49

4.5 WordSmith Tools 的设置 …… 51

4.6 WordSmith Tools 的主要功能及操作 …… 52

4.7 WordSmith Tools 快捷键的使用 …… 71

4.8 WordSmith Tools 常用符号 …… 72

第五章 SECCL 的播放工具 …… 73

5.1 数码语音播放软件的安装 …… 73

5.2 DVP 的运行与操作 …… 73
5.3 Windows 媒体播放器插件的安装 …… 75

参考文献 …… 77
附录一 1996—2002 年 TEM 4 口语试题 …… 78
附录二 SWECCL 语料样本 …… 90
附录三 WECCL 命题作文题目 …… 107
附录四 用于专题研究的小型口语语料说明 …… 112
附录五 SWECCL 词性赋码集 …… 117
附录六 SWECCL 词性赋码工具 CLAWS4 的操作说明 …… 126

图表

图

1. 图 1.1 SWECCL 语料库的总体设计结构 …… 6
2. 图 2.1 SECCL 口语库的结构图 …… 21
3. 图 3.1 WECCL 作文语料收集流程 …… 34
4. 图 3.2 WECCL 语料库的结构 …… 38
5. 图 4.1 WordSmith Tools 3.0 的主界面 …… 47
6. 图 4.2 File 菜单选项 …… 50
7. 图 4.3 选择文本 …… 50
8. 图 4.4 WordSmith Tools 的设置 …… 52
9. 图 4.5 选择工具 …… 53
10. 图 4.6 准备检索 …… 53
11. 图 4.7 检索词的输入 …… 54
12. 图 4.8 检索结果 …… 55
13. 图 4.9 对检索行进行排序 …… 56
14. 图 4.10 WordList 工具 …… 57
15. 图 4.11 准备开始 …… 58
16. 图 4.12 单词列表 …… 59
17. 图 4.13 词表（F） …… 60
18. 图 4.14 词表（S） …… 60
19. 图 4.15 单词列表参数 …… 61
20. 图 4.16 停止词表（stoplist） …… 62
21. 图 4.17 启用停止词表 …… 63

22. 图 4.18 启用削尾列表 …… 64
23. 图 4.19 削尾处理 …… 64
24. 图 4.20 词表的保存 …… 65
25. 图 4.21 主题词工具窗口 …… 66
26. 图 4.22 选择词表 …… 67
27. 图 4.23 主题词工具运行结果 …… 67
28. 图 4.24 设置 Clusters 的长短 …… 69
29. 图 4.25 赋码文本 …… 70
30. 图 4.26 设置忽略赋码 …… 70
31. 图 5.1 数码语音播放器的快捷方式 …… 73
32. 图 5.2 数码语音播放器运行时的主界面 …… 74

表

1. 表 1.1 国内外主要的英语学习者语料库 …… 3
2. 表 1.2 学习者语料研究的多种比较 …… 10
3. 表 2.1 全国英语专业四级口试评分标准 …… 14
4. 表 2.2 对 SECCL 样本选择的描述 …… 17
5. 表 2.3 SECCL 数码语音文件描述 …… 18
6. 表 2.4 SECCL 语料库的文本语料总容量 …… 19
7. 表 2.5 SECCL 总样本与 1996—2002 年各子样本的单词列表统计结果 …… 23
8. 表 2.6 SECCL 语料库标注符号一览表 …… 25
9. 表 3.1 WECCL 作文语料库的容量 …… 37
10. 表 3.2 WECCL 单词列表统计结果 …… 40
11. 表 3.3 WECCL 语料库标记一览表 …… 41
12. 表 4.1 WordSmith 常用符号一览表 …… 72

第一部分
学习者语料库介绍

第一章 绪论

1.1 SWECCL 的建设背景

世纪之交，正值计算机技术与第二语言习得研究迅速发展时期，一种基于计算机技术的第二语言习得研究的新方法——学习者语料库研究正在崛起，这种以概率和频数为基础的二语习得研究的全新方法，拓宽了应用语言学研究的视野，为外语教学研究提供了一种崭新的哲学思维方式。

学习者语料库指经过计算机处理的外语学习者的口笔语数据库（Leech，1998）。借助于计算机强大的存储和处理语言信息的功能，语料库可以进行词性赋码、错误赋码、语义赋码、话语赋码和句法标注，这些语料不仅为深入研究学习者的二语语音、词汇、语法、语篇、语用、交际能力的发展提供了充分的数据，而且还为研究学习过程、自主性学习以及教材编写、测试提供了重要的反馈和支持。

自从 20 世纪 90 年代初期比利时学者 Sylviane Granger 牵头筹建国际英语学习者语料库（International Corpus of Learner English）以来，国内外已建成了一批颇有影响的学习者语料库（见表 1.1）。我国英语学习者语料库的建设起步不算晚。1996 年广东外语外贸大学桂诗春教授和上海交通大学杨惠中教授主持筹建“中国英语学习者语料库”（简称 CLEC），并于 2003 年 1 月由上海外语教育出版社正式出版（见桂诗春、杨惠中，2003）。该语料库拥有英语专业、非英语专业以及中学英语三类学习者的 100 万词书面语料，并对语料中的错误进行了标注。同期，香港科技大学建成了规模更大的学习者语料库，但其涵盖

的英语学习者类型较为单一。

	名 称	语料来源	容量	特征
国外	International Corpus of Learner English (ICLE) (Granger et al. 2002)	欧洲 11 个国家大学英语专业 3—4 年级学生的课内限时作文和课外非限时作文，主要是议论文体。	200 万词	带简单文本头；无标注。
	The Louvian International Database of Spoken English Interlanguage (LINDSEI)	欧洲 5 个国家、亚洲两个国家的大学英语专业 3—4 年级学生的口语语料，计划与 ICLE 书面语相匹配。	200 万词	正在建设中。
	Cambridge Learner Corpus (CLC)	8 种剑桥书面测试，参加的考生有 50,000 名，说 100 种不同的母语，来自 150 个国家。	2,000 万词	800 万词的语料进行了错误标注；开放式语料库。
	The Longman Learners' Corpus (LLC)	由世界各地的学生和老师主动提交的英语书面作文。	1,000 万词	没有标注的生语料；可在网上购买。
	The Standard Speaking Test Corpus (SSTC)	日本英语学习者参加英语口语标准测试的录音以及转写的文本。	100 万词	全部进行词性赋码和错误标注。

（续表）

名 称		语料来源	容量	特征
国内	Chinese Learner English Corpus (CLEC) (桂诗春、杨惠中，2003)	中国大陆的中学、大学非英语专业学生和英语专业学生在测试环境下写出的作文，也包括部分课外作文。	100 万词	全部进行了错误标注。
	Middle School Student Writing (MSSW) Middle School Student Speaking (MSSS) (何安平，2003)	1997—1998 年全国高考广东省 21 市区 3,200 名考生的英语作文。 包含中国初中生和高中生的口头英语。	40 万词 47.6 万词	对（话轮等）部分口语特征进行了标注。
	The HKUST (Hong Kong University of Science and Technology) Corpus of Learner English	香港大学预科生和 1 年级新生的作文。	2,500 万词	全部进行了词性赋码；部分进行了错误标注。

表 1.1 国内外主要的英语学习者语料库

对于拥有世界上英语学习者人数最多的中国来说，仅有少数几个上百万词的语料库显然不能全面反映中国英语学习者的特点。何况，国内至今还没有大型的口语语料库。仅限于书面语语料的研究有很大的局限性，难以揭示学习者英语口语的特点与存在的问题。有鉴于此，南京大学于 2003 年开始筹建“中国学习者英语口笔语语料库”。这是一个既有笔语又有口语的大型学习者语料库。虽然我们起步晚，但在充分借鉴前人经验、广泛听取专家意见的基础上，我们采取了团队攻关战略，只用了两年时间就建成了整个语料库。我们在完成语料库科研项目的同时，也形成了一支学习者语料库的研究队伍。

1.2 SWECCL 语料库概述

SWECCL 语料库的设计总规模为 200 万词，分为两个子库：口语语料库（Spoken English Corpus of Chinese Learners，简称为 SECCL）和书面语语料库（Written English Corpus of Chinese Learners，简称为 WECCL）。

1.2.1 SECCL 简介

SECCL 的设计容量为 100 万词左右，主要的语料来源为我国大学英语专业学生参加全国英语专业四级口试的磁带录音语料。入选的录音磁带时间跨度为 7 年（1996—2002 年），共有 1,150 盒，转成的数码语音样本有 1,148 个，转写的电子文本有 1,148 个①，合计有 1,460,042 词。我们对每个转写文本语料都进行了文本头标注（header markup），并运用 CLAWS4 自动赋码器进行了词性赋码。此外，我们还按照复述、即席讲话和会话三种考试题型将口语语料切分为 3 个子库，并将 7 年的语料按照年份逐年存放。归纳起来，SECCL 语料库具有以下特点：

第一，口语语料来源于随机样本，具有广泛性和代表性；

第二，口语语料按照 7 年的时间跨度分年存放，为研究者考察我国学生口语能力的发展提供了可能；

第三，口语语料按照不同类型的任务加以分类，为考察任务类型变量对口语产出的影响提供了可能；

第四，运用自动标注器 CLAWS4 对所有的文本进行了词性赋码，便于研究中国学生口语中的词法和句法的变化规律；

第五，所有的文本语料都有相对应的语音文件，计算机可以直接对其进行读取和播放。这样研究者既可以做基于文本语料的口语研究，也可以对语音文件进行标注，开展基于语音语料的相关研究。

---

① 其中有 2 个语音样本因录音质量欠佳而无法听清。

第六，所有文本语料的文本头中均标记出考生在小组（一般每组32—35人）内的成绩排名，便于研究口语水平变量对口语发展的影响。

1.2.2 WECCL 简介

WECCL 的设计容量为 100 万词，与 SECCL 大体相等。书面语料主要从国内 9 所不同层次的高校英语专业 1—4 年级的学生中采集，以保证所选语料具有广泛的代表性。语料内容为若干不同题目的英语作文，文体为议论文，也有少量的记叙文和说明文，长度为 200—800 词不等，写作条件为课堂限时和课外非限时两种。作文篇目总数为 4,077 篇，其中议论文 3,059 篇，记叙文 529 篇，不同文体的 1—4 年级作文 489 篇（278 篇议论文、90 篇说明文、121 篇记叙文）。我们对所有作文都进行了文本头标注，并运用 CLAWS 自动赋码器进行了词性赋码。简言之，WECCL 语料库具有以下特点：

第一，作文按照限时和非限时加以分类，为考察时间变量对二语写作的影响提供了方便；

第二，作文按照不同文体和年级加以分类，便于考察学生写作能力的发展情况；

第三，对所有的书面语料都进行了词性赋码，便于研究中国学生的中介语词法和句法发展的特点。

1.2.3 SWECCL 的总体结构

SWECCL 语料库光盘的总体设计结构包含三个部分：1）SECCL，2）WECCL，3）Tools，如图 1.1 所示：

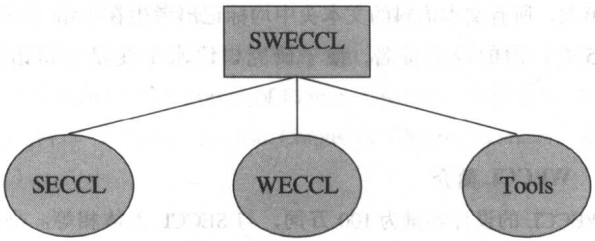

图 1.1 SWECCL 语料库的总体设计结构

1.3 SWECCL 项目组概况

SWECCL 是集体合作的成果，全国多所高校的部分师生参与了语料库的建设，包括提供语料以及直接参与语料库建设各个阶段的工作。

1.3.1 项目组的构成

SWECCL 项目组涉及的单位共有 8 个，具体如下：

项目承担单位：南京大学外国语学院

项目合作单位：外语教学与研究出版社

项目资助单位：南京大学、外语教学与研究出版社、北京外国语大学中国外语教育研究中心

项目国际顾问：Prof. Susan Hunston (University of Birmingham, UK)

项目组组长：文秋芳（北京中国外语教育研究中心、南京大学外国语学院）

项目组副组长：王立非（南京国际关系学院、南京大学中国语言文学博士后流动站）

项目组技术指导：梁茂成（徐州师范大学、南京大学外国语学院）