儿童语言发展与教育研究丛书

[美]布赖恩·麦克维尼（Brian MacWhinney） 著
许文胜　高晓妹　译

GUOJI ERTONG YUYAN YANJIU FANGFA

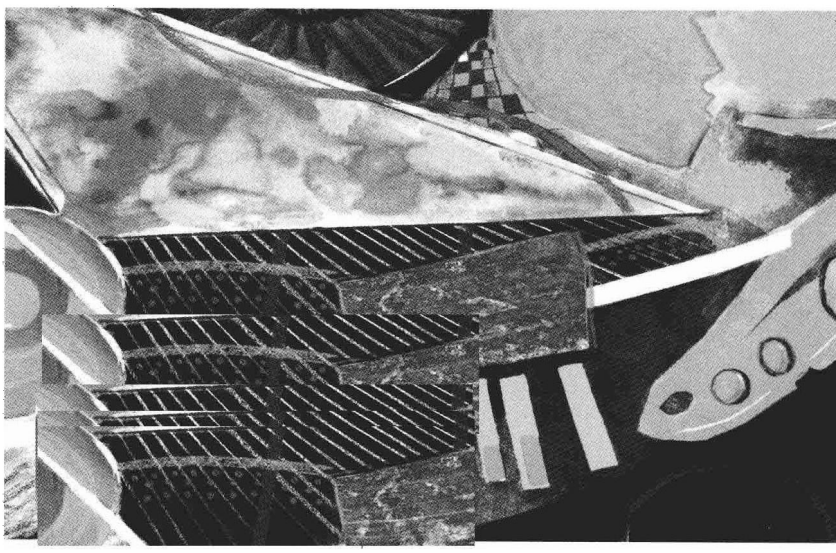
国际儿童语言研究方法

——CHILDES国际儿童语料库数据储存和分析系统

教育科学出版社
ESPH Educational Science Publishing House

[美]布赖恩·麦克维尼（Brian MacWhinney） 著
许文胜　高晓妹　译

GUOJI ERTONG YUYAN YANJIU FANGFA



国际儿童语言研究方法

——CHILDES国际儿童语料库数据储存和分析系统

教育科学出版社
·北　京·

责任编辑　王春华
版式设计　杨玲玲
责任校对　张　珍
责任印制　曲凤玲

图书在版编目（CIP）数据

国际儿童语言研究方法：CHILDES 国际儿童语料库数据储存和分析系统/（美）麦克维尼（MacWhinney，B. M.）著；许文胜，高晓妹译. —北京：教育科学出版社，2010.6
（儿童语言发展与教育研究丛书）
书名原文：Children's International Language Data Exchange System-CHILDES
ISBN 978-7-5041-4596-3

Ⅰ. 国…　Ⅱ. ①麦… ②许… ③高…　Ⅲ. 儿童语言—研究　Ⅳ. H003

中国版本图书馆 CIP 数据核字（2009）第 135198 号

北京市版权局著作权合同登记 图字：01-2009-7729 号

出版发行	教育科学出版社		
社　　址	北京·朝阳区安慧北里安园甲 9 号	**市场部电话**	010-64989009
邮　　编	100101	**编辑部电话**	010-64989584
传　　真	010-64891796	**网　　址**	http://www.esph.com.cn
经　　销	各地新华书店		
制　　作	北京金奥都图文制作中心		
印　　刷	北京人卫印刷厂	**版　　次**	2010 年 6 月第 1 版
开　　本	169 毫米×239 毫米　16 开	**印　　次**	2010 年 6 月第 1 次印刷
印　　张	17.75	**印　　数**	1-3 000 册
字　　数	309 千	**定　　价**	38.00 元

如有印装质量问题，请到所购图书销售部门联系调换。

中文版序

非常高兴看到 CHILDES 指南被翻译成汉语。感谢周兢博士和许文胜博士出色的工作。汉语 CHILDES 指南的出版，使中外致力于语言学习的众多学生和研究者，得以更加方便地进行儿童语言发展研究。通过直接阅读中文版 CHAT 和 CLAN 的规则和步骤，研究者对于语言学习的理解将更加深入、直接。从已有的日语版、西班牙语版、法语版和意大利语版指南来看，译本对于更为广泛的人群获得 CHILDES 系统的语料和方法产生了无可估量的作用。我期待中文译本的出版对于汉语语言学习研究也同样产生积极的影响。

值得庆幸的是，现在我们的 CHILDES 语料库中已有几项重要的汉语语言学习研究。来自于这些研究的转录文本和音频资料，可以为那些希望了解第一语言学习复杂过程的学生和研究者提供具体的资料以及案例，如 Thomas Lee、Colleen Wong 和 Samuel Leung 的广东话纵向研究语料库，香港中文大学 Virginia Yip 和 Stephen Matthews 的广东话—英语纵向双语语料库，后者还提供语音及音频链接。而在汉语普通话的语料库中，有周兢教授主持的几个语言研究项目所提供的转录文本，由 Twila Tardif 于北京采集和张鑑如于台北采集的叙事复述语料，还有 Twila Tardif 提供的家庭对话语料。

指南的译本和实证研究语料，将促进在单语和双语情境下汉语第一语言习得研究方面的进步。当然，这仅仅是个开始。对于汉语普通话而言，我们依然缺少类似现有的广东话以及多种欧洲语言的细致的纵向研究。在各种可用的汉语语料库中，仅 Tardif 从北京采集的叙事语料以及香港中文大学的双语语料库有音频链接。这一点非常重要。因为使用音频、视频链接的文本，能够使研究者对语言学习中词汇、语音形式、语法结构以及互动模式方面的背景获得更加丰富的理解。有兴趣探讨这些可能性的研究者，可以浏览 http://childes.psy.cmu.edu/browser，查看链接的语料库。

汉语研究方面的另一重要进步，是该系统已经创造性地为广东话和普通话提供了准确性极高的自动词性标注。借助这一功能，我们已经为超过一半的广东话和普通话的语料文本全面标注词性。这方面下一个阶段的工作，是语法关系标注的自动系统建构。我们已经为英语、希伯来语和日语建构了这些系统。为汉语语言建构语法分析系统的工作不久将开始进行。

这本指南上描述的工具不仅对于理解语言学习非常重要，对于理解更为广泛的语言互动亦是如此。这些项目和转录格式，现在正用于研究成人在课堂、职场、法庭、演讲以及聚会中的对话。CHILDES 和 TalkBank 数据库内容丰富，包括教师关于物理学原理的讨论、失语症患者访谈等录音。我非常希望这些工具能够不断拓展实证语料库的发展，用来研究汉语在多种语境下的使用。

在我写这些文字时，耳畔是办公室窗外直升飞机和汽车的轰鸣。世界经济发展中占据重要地位的主要国家元首，参加今天在匹兹堡举行的20国集团峰会。这次会议使我想起在过去的20多年，我们在经济全球化方面取得的成就。与此同时，它也使我想起在语言学习的研究上加快全球化的必要性。我们需要全球化，需要了解不同语言和文化背景下的学习如何不同或表现出什么样的共性。我们还需要全球化，使语言学习研究向所有学术团体公平开放。为此，我们清楚地看到，中国的研究者应该在其中扮演日益重要的角色。

感谢许文胜、高晓妹、徐翠芹、陆晓兵、曲喆和王超……感谢你们在翻译过程中付出的辛勤劳动！

■ 布赖恩·麦克维尼 2009 年 9 月 24 日于匹兹堡

目　录

上卷：国际儿童语料库数据储存系统
CHAT OF CHILDES

下卷：国际儿童语料库数据分析系统
CLAN OF CHILDES

2 学习指南

3 编辑器

4 功能（features）

5 分析命令

上卷：国际儿童语料库数据储存系统

CHAT OF CHILDES

1 导 言

语言习得研究的发展基于从自然情境下自发的互动中收集而来的数据。你可以听磁带或看录像，你会在不经意间积累起数十甚至数百小时的自然交互数据。但是收集数据仅仅是更大任务的开始，因为转写和分析自然样本是异常耗时而且并不那么可靠的。在此我们将介绍一套旨在增强转写可靠性、使数据分析程序自动化以及可以促进数据共享的计算工具。这些新的计算工具为儿童语言的研究方法带来了革命性变化。除此以外，它们还对第二语言习得、成人会话、社会学内容分析和失语症患者语言恢复的研究有着同样革命性的影响。尽管这些工具有着广泛的适用性，但是此处的重点是其在儿童语言研究领域的使用。我们希望其他领域的研究者可以在其各自领域进行类似的研究与应用。

在详细说明这一系统之前，我们最好先回顾一下早期语言习得研究的数据收集方法及其发展中的主要事件。我们将分5个主要历史时期来叙述。

1.1 印象主义观察时代

首次试图理解语言发展进程的尝试出现在《圣·奥古斯汀的忏悔》（*The Confessions of St. Augustine*）中一段著名的话中。在这段话中，奥古斯汀描绘了他是怎样学习语言的。

> 记得从那时起，我就已经注意到我是怎样学习说话的了。这并不是大人们用什么固定方法来教我单词（以及稍后的其他知识），而是我通过哭喊发出断断续续的音节以及四肢的动作表达我的想法，以期得到满足。不过我无法完全表达出我想要表达的意思，也无法对我想要的人表达出这些。但是，通过上帝赋予我的理解力，我在记忆中练习那些声音。当大人们说到一件事物，并转过身面对着这样东西时，我看到了这些并记住了他们所说并所指的那样东西的名称。他们的身体动作清楚地表明，他们确实指着这样东西而不是其他。这种包括面部表情、目光、肢体动作和语音语调等在内的体态语，是各国共有的自然语言。它是人们所思所想、拒绝或回避等情绪的结果。因此，通过不停地听他们在各种句子中说出的单词，我逐渐明白了它们的意思；当我说出这些单词时，我也

就可以表达我的意愿了。然后，在父母及大人话语权威性思想的基础上，我与周围的人交流着现在使用着的这类词，更深入地了解纷繁复杂的人类生活的交流。

奥古斯汀对于早期词汇学习的概述，使人们注意到注视、指向、语调和相互理解是语言学习的基本线索。现代词汇研究（P. Bloom，2000）为奥古斯汀分析的每一点都提供了证据，包括其强调的儿童意志的作用。从这个意义上说，奥古斯汀关于自己语言习得富有想象力的回忆，在整个中世纪甚至文艺复新时期的儿童语言研究领域中都是高水准的。这种研究方法取决于人对于童年早期的真实回忆的能力——很不幸，这种能力只赋予我们中间为数甚少的几个人。

1.2 婴儿传记时代

研究语言产生的第二种主要技巧，是由达尔文率先提出的。利用卡片和实地记事本，跟踪加拉帕哥斯群岛和印度尼西亚等地数百物种和亚物种分布的情况，达尔文得以收集到大量的自然数据，以支持他的自然选择和进化论。在研究儿子手势发展的过程中，达尔文（1877）证明了同样的自然观察工具也可以用于研究人类的发展。通过每日的详细记录，他指出研究者可以建立日志形成传记，从而真实地记录人类发展的各个方面。从达尔文起，Ament（1899）、Preyer（1882）、Gvozdev（1949）、Szuman（1955）、Stern & Stern（1907）、Kenyeres（1926，1938）和 Leopold（1939，1947，1949a，1949b）等学者建立了里程碑式的传记，详细记录了他们自己孩子的语言发展情况。

达尔文的传记法，对研究成人失语症也产生了影响。Loh（1931）、Pick（1913）和 Wernicke（1873）等人据此对某些病人或症状的语言情况进行了研究。

1.3 转写时代

日志法的局限性是显而易见的。即使是受过严格训练的观察家，也无法记下大量的言语。拿着笔和本子跟在儿童后面记录时，你发现漏掉了许多细节，而且作笔记会干扰当时的相互交流。

20 世纪 50 年代后期，录音机的引入打破了这些限制，语言研究由此进入了第三阶段。录音机对语言研究领域的影响，就像其对民族音乐的研究所产生的影响一样。这使像 Alan Lomax（Parrish，1996）等研究者顿时可以制作高音质的现场录音磁带。这一阶段的特点，是研究者在 2—3 年中收集了大量的不同主体的录音数据。录音和录音转写，使大量原始数据的获得成为可能，

这直接推动了20世纪60年代儿童语言研究的发展。

原始数据的增加还导致了一种结果，不过这一点很少被提及。在婴儿传记时期，最终出版的文字极为接近卡片上的原始数据。就此而言，观察的数据和公布的数据之间没有太大的差别。在录音转写时期，这种差别加大了。20世纪60年代和70年代产生的大量录音，不可能全部出版。因此，研究者们不得不对别人没法看到的数据进行高水平分析，并公布由此得出的结果。这使得原始经验数据仅保存在个人手里，无法得到公开检验。人们在母盘空白处写下评论，贴上标签，然后又生产了新的甚至是更难懂的备份。每位研究者都有一套适合本研究的转写系统和项目编码。当我们开始对照手写稿和打印稿时，转写使用方法、编码方案和不同研究者的信度等问题就会更明显。

Roger Brown 认识到了这个问题，他率先做出了努力，与其他学者共享他对 Adam、Eve 和 Sarah 的转写（Brown，1973）。这些转写被刻在了模板上并复印了好几份，然后散发到其他研究者手中，请他们进行分析。这些研究者把副本带回家，制定并应用自己的编码方案（通常是直接用铅笔在转写上编码），写一篇论文。然后，如果非常礼貌的话，再送一份拷贝给 Brown。其中有一些报告（Moerk，1983），甚至试图反驳 Brown 从数据中得出的结论。

在早期，不同编码方案之间的关系始终笼罩上一层神秘色彩。编码系统不稳定特性的一个幸运的后果是，研究者非常小心保存自己的原始数据，即使已经编码，也不会丢弃。Brown 本人在以下段落中，对即将到来的向计算机的过渡作出了评论（Brown，1973，p. 53）。

> 有一个很明智也是我们常被问及的问题："为何不能将句子按重要语法特征加以编码并输入计算机，这样，任何人都可以非常便捷地从事研究？"我总是这样回答：我不断发现会话记录中包含的各种新信息，从未确信完整编码应该如何。事实的确如此。实际上可以这样说，从1962年开始的整整10年，调查者继续从自由对话中不断发现推论语法和语义知识或能力的新方法。但是坦白地说，就本人而言，我必须补充一点，那就是研究风格。我耐心不足，总是想不断开始。一位更加出色的科学家可能更有计划性并使用计算机。他今天也可以那样做，并对于哪些内容进行编码信心十足。

在具备30多年计算机分析的经验基础上，我们现在知道，将儿童语言数据缩减到一套编码，然后抛弃原始数据的想法根本就是错误的。与之相反，我们的目标必须是以这样一种方式将数据计算机化，即方便我们继续使用新的编码和注解来不断提高它。幸运的是，Brown 保存的转写数据，可以让人们