温志军、胡瑰玲，2001，开发利用世界上最大的儿童语料库--CHILDES，《外语教学与研究》，第5期：374-377。

Wen, Zhijun. & Hu, Guiling. 2001. Kaifa Liyong Shijieshang Zuidade Ertong Yuliaoku--CHILDES. (Exploiting the World's Largest Child Language Database--CHILDES".) *Waiyu Jiaoxue yu Yanjiu. (Foreign Language Teaching and Research.)* 33(5): 374-377

开发利用世界上最大的儿童语料库—CHILDES *

广东外语外贸大学 温志军 胡瑰玲

提要：本文简要介绍世界上最大的儿童语言语料库 CHILDES（儿童语言数据交流系统）。CHILDES 的重要性已为越来越多的语言学家、心理学家和认知科学家所确认。迄今为止，该数据库已囊括了 25 种语言的数据；已有 1,000 多篇研究论文使用了 CHILEDS 数据。本文呼吁中国的专家、学者也来开发利用这一资源，推动我国的儿童语言研究；同时也来贡献数据，使这一数据交流系统日趋完备。

关键词：儿童语言、语料库

1. 引言

当今的语言学、心理学和认知科学都密切关注儿童语言习得方面的研究。儿童语言不仅是儿童发展心理学(developmental psychology)的主要研究对象，也是哲学、语言学、心理学和认知科学经常争论的焦点。牵动着许多研究者心弦的语言先天禀赋论与后天培育论（Nature vs. Nurture）的纷争正是围绕儿童语言习得而展开的。争论的双方都试图从儿童语言获得过程的实例中寻找证据。以 Chomsky 为代表的一派认为语言是不可能被学会的。在他们看来，儿童之所以能够使用语言是因为在人的大脑里有一个固有属性（innate property），即语言官能（language faculty）；这一固有属性决定了语言的出现。他们声称可以用一个人体器官（即所谓 mental organ）的发育来比喻语言的发展；意思是说，语言就象一个人体器官一样由基因控制自然生长出来而非通过学习获得（Chomsky, 1975, 1988；Piatelli-Palmarini, 1989；Pinker, 1994）。认知科学家和心理学家通过对儿童语言的大量研究发现儿童的语言不是与生俱来的，而是通过学习才获得的；人类具有出色的学习能力，这种学习能力是先天的，但绝非语言本身（Bates, 1999; Bates & Elman, 1996）。

由此可见，儿童语言研究具有重要意义。但是，研究儿童语言并非易事。研究者首先必须获得真实自然的儿童语料。我们可以通过录音、录像把儿童在自然状态下使用语言的情况记录下来。然后，我们还需把录音或录像转换为文字材料。这样才能进行分析和研究。把录音或录像转换为文字是一件费时费力的工作。一个钟头的儿童语言现场录音或录像依据研究者的目的往往需要花费十到十四个钟头的时间来录写（transcribing）；因为作为研究

* 本文是作者在 Center for Research in Language, University of California at San Diego 作访问学者时写成的。其写作得益于 Dr. Elizabeth Bates 的帮助，在此向她表示感谢。此外，本文第一作者感谢中国国家留学基金委对此项研究的资助。

用途的录写应尽可能记录较全面的信息，光是记录话语是不够的，还要记录话语的语调、说话者的表情、说话的环境以及话语是否连贯等信息。过去从事儿童语言研究的科学家总是单独收集和使用数据。这些数据往往在用完之后便置之高阁，极少被公开或重复使用，造成极大的浪费，也不利于其他科学家对数据及结论的真伪进行鉴别（Bates & Carnevale, 1993；MacWhinney, 1995b）。基于上述原因，一些深谋远虑的科学家开始酝酿起要建立一个儿童语言语料库的想法。1983 年，心理学家 Elizabeth Bates、Brian MacWhinney 和 Catherine Snow 等人讨论了利用美国麦克阿瑟基金会(MacArthur Foundation)的资金建立这样一个语料库的可能性。同年，Brian MacWhinney 和 Catherine Snow 向该基金会提出了资金申请。1984 年，麦克阿瑟基金会批准了该申请。这样，这一世界上最大的儿童语言语料库便在 Brian MacWhinney 和 Catherine Snow 的负责下开始正式筹建。该语料库的名称为儿童语言数据交流系统（Child Language Data Exchange System），简称 CHILDES（MacWhinney, 1991, 1995a）。

2. CHILDES 简介

CHILDES 的出现促进了与儿童语言研究有关的各种研究活动的开展和信息交流，是一件具有远见卓识的举动。截止到 1995 年初，有关 CHILDES 的科研论文已达三百多篇（MacWhinney, 1995a）。迄今为止，该数据系统已囊括了包括普通话和广州话在内的二十五种语言的数据；而利用 CHILDES 数据所撰写的文章已有 1,000 多篇（MacWhinney, 2000）。这些文章有的是关于语法发展的（如 Eisenberg, 1989; Slobin,1994），有的是关于儿童如何接受语言输入的（如 Van Houten, 1988; Anderson & Shirai, 1994），有的是关于儿童如何学习词汇的（如 Clark & Carpenter, 1989; Au & Song, 1994），有的是关于儿童音位发展的（如 Wijnen, 1988; Bernstein-Ratner, 1993），等等。不仅如此，这样的大型数据库完全可以免费使用。本文就针对如何使用 CHILDES 系统做一简要介绍。

CHILDES 由三部分组成。第一部分是儿童语言数据库(The database)；第二部分是名为 CHAT 的录写系统；第三部分是做语言数据分析之用的程序，称为 CLAN 。这三部分均有 CD 版、FTP 版和万维网络（World Wide Web）版。其中 FTP 版和万维网版都可免费获取。需要 CD 版的，可同以下地址联系订购：

Lawrence Erlbaum Associates
365 Broadway, Hilldale, NJ 07642
U.S.A.
订购电话：(201) 666-4110。

需要 FTP 版的，可到下面的 FTP 服务器地址去获取：

<poppy.psy.cmu.edu>。

由于大多数读者更熟悉万维网，下面我们就介绍万维网络版的 CHILDES 数据库、CHAT 录写系统和 CLAN 语言分析程序。

2.1 万维网络版的 CHILDES 数据库

在网络浏览器的地址栏键入网址<http://childes.psy.cmu.edu>直接进入 Carnegie Mellon University 心理学系的 CHILDES 主页。也可以在网络浏览器的地址栏先键入<http://www.psy.cmu.edu>进入 Carnegie Mellon University 的心理学系主页。然后在心理学系主页上的 Affiliated Research Groups & Projects 下选 CHILDES System，由此进入 CHILDES 主页。

进入 CHILDES 主页后，我们可以看到不同的栏目，包括 System、Links、Theory、

Procedures、Program & Data、Manuals 和 Special Tools。在 System 栏目下有使用 CHILDES 的基本规约（Ground rules）。基本规约包括下面内容：

（1）在使用该数据库之前，应该和 Brian MacWhinney< macw@cmu.edu>联系成为 CHILDES 成员，向他提供自己的地址、电子邮件地址、电话号码及使用 CHILDES 数据的原因。

（2）建议参考 CHILDES 第三版的使用手册：MacWhinney, B. (2000) *The CHILDES Project: Tools for Analyzing Talk.* Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates。美国境内可打电话 1-800-926-6579 进行订购。美国境外没有说明，应该和 Brian MacWhinney 联系。

（3）任何使用 CHILDES 数据撰写的文章，必须在参考文献目录中列出数据的出处及 MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

（4）使用手册中和你所用数据有关的章节应仔细阅读，因为那里列出的一些文章是使用该部分数据时必须引用的。

（5）CHILDES 鼓励数据使用者把自己及其同事的研究数据提供给 CHILDES 以便利数据共享。

（6）遵守上面基本规则的成员可免费使用 CHILDES 提供的程序和数据。

CHILDES 主页上的 Program & Data 栏下有数据库(The database)和 CLAN 程序。在主页的 Manuals 栏下有数据库使用手册（The database manual）、CLAN 程序的使用手册和 CHAT 录写系统的使用手册。这些均可下载。数据库（The database）分别备有 Mac 和 Windows 两种版式可供下载。由于中国电脑用户多使用 Windows 操作系统，我们在此就只介绍如何获得 Windows 版的数据。选 Program & Data 栏下的 Windows 则可进入数据库页面。

进入数据库页面后，访问者最好先访问数据库文献管理（Database documentation）。该文件和数据库使用手册（The database manual）实际上是同一个文件。阅读数据库文献管理文件需使用 Adobe Acrobat Reader 这个软件。数据库文献管理主要介绍 CHILDES 数据库中所有数据的目录和数据收集的背景。

回到数据库页面，我们看到页面上列出了不同语言、不同类别的数据。在 CHILDES 数据库里，所有适合 Windows 使用的数据都用 WinZip 软件进行了压缩。访问者可以直接从数据库页面下载 WinZip 软件。有了 WinZip 软件便可以对下载的数据文件进行解压。访问者可以根据自己的研究需要选择不同语言的数据。我们在这里选择“双语者语料库（bilingual corpora）”的链接。这一链接把我们带到了双语者语料库页面。这里储存了不同研究者所录写的双语儿童语料。我们可以从中随意选择文件下载并存入我们的电脑。当我们把存在电脑中的文件解压之后，展现在我们面前的就是研究者为双语儿童语言所做的记录。我们会发现文件里有很多特殊小符号如@Begin、 %act 等。下面我们就介绍这些符号的意义以及如何使用这些符号来编写 CHILDES 数据。

2.2 CHAT 录写系统

收集儿童语言数据之前，应先设计好一个实验，找到合适的受试，把实验的过程录音或录像。然后把这些声音或图象信息转换成文字以供研究使用。上文曾谈到，详细、客观的录写不是一件简单的事情。有很多因素需要处理，如是否要记录手势和眼神等非语言行为（nonverbal behavior）、是否要记录停顿和音调、以及如何使用标点符号等。这些因素处理的好坏会直接影响记录的准确与否。关于录写原则可参考 Ochs（1979）。

正因为客观录写语言有如此大的难度，CHILDES 的设计者精心编写了一套录写符号，

以便能最大限度地将儿童语言使用情况客观、准确地反映出来。这一套录写符号就是 CHAT 录写系统的内容。CHAT 录写系统是 CHILDES 的一个重要贡献。它是一个看似复杂却十分灵活的多层次编码方案，专为电脑录写自然言语（free speech）设计。在 CHAT 未被开发出来之前，研究人员各自独立编码录写自然言语，致使语言数据无法共享。CHAT 的发展完善经历了五、六年的时间。在世界各地的研究人员共同努力之下，它如今已成为第一个世界上通用的自然言语录写编码系统。下面我们举一例说明如何使用该系统中的一些符号。

@Begin
@Participants: ROS Ross Child BRI Brian Father
*ROS: why isn't Mommy coming?
%com: Mother usually picks Ross up around 4pm.
*BRI: don't worry.
*BRI: she'll be here soon.
*ROS: good.
@End
(载自 MacWhinney, 1995a: 9)

@Begin 表示录写文件的开始。@Participants:指明实验参与者。在此例中，ROS 代表名叫 Ross 的孩子，BRI 代表名叫 Brian 的父亲。*ROS：表示 Ross 所说的话。*BRI：表示 Brian 所说的话。%com：后是录写者或研究者对背景等做的注释。@End 表示录写文件的结束。CHAT 共有一百三十多个符号。由于篇幅的考虑，我们这里不详细说明所有符号的意义，读者可以到 CHILDES 主页去下载 CHAT 的使用手册。

2.3 CLAN 数据分析程序

CHILDES 的另一重要贡献就是 CLAN 数据分析程序。CLAN 是 Computerized Language Analysis 的简称。它由 Carnegie Mellon University 的电脑编程人员设计，是用电脑来对 CHILDES 数据文件进行自动分析的工具。读者下载 CLAN 程序之前，应阅读 CLAN 的使用手册，了解其安装及使用方法。CLAN 有三十多个语句，其中最基本的语句及格式如下：

(1) CHECK

CHECK 语句用于检查整个文件的编写有没有格式上的错误。其命令行格式为：check 文件名。假如我们要检查文件 adam01.cha 中有没有编写错误，可键入：check adam01.cha。

(2) FREQ

FREQ 语句用于检查单词出现的频率。其命令行格式为：freq +s 需检查的单词+t*实验参与者代号 文件名。假如我们要知道 adam01.cha 文件中代号为 ADA 的孩子使用的所有代词的频率，可以键入：freq +spronouns+t*ADA adam01.cha。省缺“+s 需检查的单词”和“+t*实验参与者代号”两项时，输出的是实验参与者使用的所有单词的频率。

(3) MLU

MLU 语句用于计算平均话语长度。其命令行格式为：mlu +t*实验参与者代号 文件名。假如我们要知道 adam01.cha 文件中代号为 ADA 的孩子的平均话语长度，可以键入：mlu +t*ADA adam01.cha。省缺“+t*实验参与者代号”时，输出的是所有实验参与者的平均话语长度。

3. 前景展望

上面，我们对 CHILDES 万维网络版的使用做了简单的介绍。我们希望从事有关研究的

专家、学者能充分使用这一工具，来推动我国的儿童语言研究。现在，CHILDES 已不再是一个单一的儿童语言数据库。它还开辟了儿童和成人失语症患者的语料库，其中也收有汉语失语症患者的语料。但是，该数据库中的汉语材料不论是正常人的还是失语症患者的都还极其有限，有待我们来丰富。这也正是加强中国儿童语言研究对世界范围儿童语言研究产生影响的一个良好机会。

CHILDES 的成功也导致了其他语料库的诞生。比如一些神经语言学领域的研究人员借鉴 CHILDES 开发出了一个失语症患者语言数据交流系统，即 ALDES (Aphasic Language Data Exchange System)。

CHILDES 的发起者不满足于目前的成就，他们下一步的方向是把 CHILDES 发展成为一个集文本、声音、图象为一体的多媒体数据库。使用者在阅读数据库中某一句话时，只要一点鼠标，就可以听到该受试讲这一句话的录音，看到受试说这一句话时的场景录像。到那时，CHILDES 对语言学、心理学和认知科学等研究领域所能作出的贡献就更大了。

参考文献

Anderson, R. & Shirai, Y. 1994. Discourse motivations for some cognitive acquisition principles. *Studies in Second Language Acquisition*. 16, 133-156.

Au, T. K. & Song, Y. K. 1994. Input vs. constraints: Early word acquisition in Korean and English. *Journal of Memory and Language.* 33, 567-582.

Bernstein-Ratner, N. 1993. Interactive influences on phonological behavior: A case study. *Journal of Child Language.* 20, 191-197.

Bates, E. 1999. On the nature and nurture of language. In R. Levi-Montalcini, D. Baltimore, R. Dulbecco & F. Jacob (Series Eds.), and E. Bizzi, P. Calissano & V. Volterra (Vol. Eds.), *Frontiere della biologia [Frontiers of biology]. The brain of homo sapiens* Rome: Giovanni Trecanni.

Bates, E. & Carnevale, G. F. 1993. New directions in research on language development. *Developmental Review*. 13, 436-470.

Bates, E. & Elman, J. 1996. Learning rediscovered. *Science.* 274, 1849-1850.

Chomsky, N. 1975. *Reflections on Language*. New York: Parthenon Press.

Chomsky, N. 1988. *Language and Problems of Knowledge*. Cambridge, MA: MIT Press.

Clark, E. & Carpenter, K. 1989. The notion of source in language acquisition. *Language.* 65, 1-30.

Eisenberg, S. 1989. *The Development of Infinitives by 3-, 4-, and 5-Year-Old Children.* Unpublished doctoral dissertation. CUNY Graduate Center.

MacWhinney, B. 1991. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. 1995a. *The CHILDES Project: Tools for Analyzing Talk.* Second Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. 1995b. Computational analysis of interactions. In P. Fletcher & B. MacWhinney (Eds.), *Handbook of Child Language*. Oxford: Blackwell. 253-244.

MacWhinney, B. 2000. Personal communication.

Ochs, E. 1979. Transcription as theory. In E. Ochs & B. Schieffelin (Eds.), *Developmental Pragmatics.* New York: Academic Press

Piatelli-Palmarini, M. 1989. Evolution, selection and cognition: From “learning” to parameter setting in biology and the study of language. *Cognition.* 31, 1-44.

Pinker, S. 1994. *The Language Instinct: How the Mind Creates Language*. New York: William Morrow.

Slobin, D. I. 1994. Passives and alternatives in children's narratives in English, Spanish, German and Turkish. In B. Fox & P. Hopper (Eds.). *Voice: Form and Function.* Amsterdam: Benjamins.

Van Houten, L. 1988. *Role of Maternal Input in the Acquisition Process: The Communicative Strategies of Adolescent and Older Mothers with Their Language Learning Children.* PhD Thesis. Harvard University.

Wijnen, F. 1988. Spontaneous word fragmentations in children: Evidence for the syllable as a unit in speech production. *Journal of Phonetics.* 16, 187-202.

通讯地址：510421 广东外语外贸大学培训部

Email: dwen@crl.ucsd.edu

Exploiting the World's Largest Child Language Database – CHILDES

By Wen, Zhijun & Hu, Guling

Abstract: This article gives an overview of the world's largest child language database—CHILDES(Child Language Data Exchange System). The importance of this language database has been increasingly recognized by researchers in linguistics, psychology and cognitive science. So far, the system has covered 25 languages. More than 1,000 articles have been written using this language database (MacWhinney, 2000). The present article calls on Chinese researchers to make use of CHILDES to promote child language research in China and also contribute to this language database to help it grow.