汉语儿童口语语料库的建立及语料初步统计分析

张碧川 王小捷 徐文智 刘冬雪
北京邮电大学智能科学与技术中心 北京 100086
E-mail: bugnec@gmail.com

摘 要：我们认为基于儿童语言习得的过程可以帮助建立一种语言的计算模型。研究儿童语料资源在语言习得及其计算模型的研究中是不可或缺的，本文将 CHILDES 语料中汉语语音进行转录及词性标注，得到一个儿童口语语料库。并比较了儿童语言，儿向语言和成人语言之间的字层句层特点，我们基于实验结果讨论了语言习得理论。

A Preliminary Statistics of a Chinese Children Corpus

Bichuan ZHANG, Xiaojie WANG, Wenzhi XU, Dongxue LIU
Center for Intelligence science and technology research, BUPT, Beijing 100086
E-mail: bugnec@gmail.com

Abstract: The research on children corpora is helpful for us to understand human language acquisition. For Chinese, there is currently no Chinese character based children corpus. This build a Chinese character based children corpus by transliterating a pinyin corpus in CHILEDS. After tagging POS information on the corpus, we then give a preliminary comparison among child speech, child directed speech (CDS) and a adult corpus. We find some interesting results which match well with some existed linguistic conclusions on language acquisition.

1 前言

当前，大部分的语言计算模型都是一种既得型的模型，它们或者是直接由人们构造的，或者是直接依据成人语料来进行训练而获得。而实际上，作为具有语言计算能力的人类，其语言能力存在着从简单到复杂的发展过程，语言的计算模型是逐步习得的。与直接从成人语料训练获得语言计算模型不同，我们更倾向于同意，基于儿童语言习得的过程可以帮助建立一种语言的计算模型，在这种思想下，针对儿童语言的计算研究就显得十分重要了。

已有不少基于语言发展计算模型的研究, (Briscoe and Carroll 1997)提出了一种从文本语料中习得次范畴词典的方法; (Korhonen 2003)使用信息瓶颈和最近邻方法，进行次范畴框架聚类，主要工作围绕多义动词的聚类; (Buttery 2006)从 CHILDES 中获取儿童与儿向语言的次范畴框架，与成人语料相比较，并提出一个范畴语法学习器，这些工作是针对英语为母语的儿童语言，在其他语言，(Schulte im Walde 2002)应用 K-means 算法，应用动词次范畴信息，对德语中的动词进行语义聚类; (Sarkar and Zeman 2000)使用三种不同的统计模型从捷克语中提取次范畴框架，比较了三种方法的性能; (Gamallo et. al 2002)在习得方法中使用 co-composition，进行葡萄牙语的句法和语义次范畴习得。

汉语语言习得的研究主要来自于语言学家， (李宇明 1995)从语言角度初步构建了儿童语言学研究的理论框架，详实的记录和解释了汉族儿童语言的发展规律， (周国光 2001) 对儿童语言中的基本句法结构的结构类型、语义关系、句法功能进行了全面细致的描述, 讨论和分析了儿童语言习得的机制、手段、发展阶段等相关问题。

中国的语言学家对汉族儿童习得汉语基本句法结构的状况也进行了考察和研究。这些研究还主要停留在对一些案例的定性研究，其目标也不是为了建立计算模型，同时，还缺少公共的语言资源。

儿童语料资源在语言习得及其计算模型的研究中是不可或缺的，Buttery 等的研究采用了CHILDES (B MacWhinney 1995)，在 CHILDES 中有一些汉语语料，但是是以拼音的形式出现，对于以汉字为代表的汉语结构的习得研究，基于汉字进行研究可以避免由于大量同音字存在而造成的额外负担，为此，我们认为应该把直接基于汉字语料来进行。为此，本文进行了 CHILDES 种汉语语音语料到汉字的转换。由于儿童对话中使用的词汇结构与一般词典的差距较大，现有的音字转换都是基于成人语言的，因此，我们对儿童语音的转录难以完全自动进行，其中存在着很多问题。在转录完成之后，我们进行了初步的加工，并把这个语料和目前常用的标准成人汉语语料进行了对比分析。

本文是这样安排的，下一节介绍 CHILDES 汉语语料和其他一些汉语儿童语料库，在第三节介绍将 CHILDES 语料中汉语语音进行转录及词性标注的工作，第四节将转录后的基于汉字的儿童语料库与成人语料库进行一些列对比，最后给出我们的结论，并提出今后的研究工作。

2 CHILDES

2.1 CHILDES 语料库

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。牵动着许多研究者心弦的语言先天禀赋论与后天培育论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。

1984 年，Brian MacWhinney 和 Catherine Snow 合作筹建儿童语言语料库，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。经过近 20 年的建设，已成为目前世界上最大的儿童口语语料库。

2.2 CHILDES 汉语语料库

在 CHILDES 中的汉族儿童口语语料库有：

1)香港粤语儿童语料库(Cancorp) (Fletcher et al. 2000)

语料库收集了儿童与成人之间的口语片段，是 8 名(4 男 4 女)1 岁 5 个月到 2 岁 8 个月讲粤语儿童在一年半时间内的语言发展的纵向纪录。每个文本文件均有汉字、拼音和相应的词类标注。

2)HKU

该语料库包括两部分，一部分直接来源于 Cancorp，共 128 个文件，另一部分纪录了调查人员与 70 个 2 岁半到 5 岁半儿童的对话。

3)Chang(Chang 1998)

该语料库由国立台湾师范大学 Chienju Chang 建立。共 24 份文件，记录了 12 个 4 岁儿童和 12 个 6 岁儿童在 Chang 提供的场景下所说的话。语料包括录音和录像两种格式，所得的文本语料均有繁体汉字、拼音和相应的词类标注显示。

4)Beijing 和 Context(Tardif 1993, 1996)

这两个语料库均由香港中文大学心理系 Tardif 建立，前者共 50 个文件，记录了 10 个 20-22 月的孩子在玩玩具、看电视、与邻居玩耍和在公园玩等自然状态下的语言。Context 语料库记录了 25 个说普通话的儿童和 25 个说英语的儿童在自然状态下的语言。两个语料库均只用拼音转写。

5)Zhou

该语料库由华东师范大学周兢提供，共 50 个文件，记录了 50 对母子在预设情境下的对话。

除了 CHILDES 的基于拼音描述的儿童语音语料之外，还有一些汉语方言的语料库。

6)闽南语儿童语料库(Taicorp) (Jane, 2008)

该语料库是台湾中正大学语言学研究所蔡素娟教授建立。语料库共收集了 330 小时的儿童口

语语料，共431盒磁带，最终形成的文本语料有230万字。

为充分利用CHILDES的汉语语料库，避免拼音中过多同音字造成的额外负担，我们需要将其基于拼音的记录转写为成汉字，以对语言结构习得进行研究，而忽略音到字的习得过程。拼音与汉字不是一一映射，汉语的主要特征体现在汉字而不是拼音，常用汉字有7000个左右，共有汉字更有90000之多，而汉语中只有二百五十个标准的带声调的汉语拼音音节。同音异形在汉语中是大量的存在，比如“声誉”和“生育”，“ 同化”和“童话”，还有“城市”“乘势”“成事”“程式”等等，它们具有完全一样的发音即拼音完全一样，但是不 论汉字形态还是语义，这些词都是完全不同的，因此，将CHILDES中基于拼音的汉语语料转写为基于汉字的语料，将需要处理大量的同音现象。下一节描述该语料的转录和加工。

3 CHILDES汉语语料库的转写和POS标注

我们对CHILDES汉语儿童语料中的Beijing、Chang、Context、Zhou语料库进行了转写，由于拼音描述是有词间空格的，我们转写后的词也基本以此为准，只在少数与汉语词差别较大的地方进行了修改。在此基础上，对词进行了POS的标注。以下分别描述。

3.1 转写

自动转写

我们首先使用基于隐马尔科夫的音字转换模型进行自动转写。

定义拼音输入A，对应的汉字串W，最佳的输出W^*由下式决定：

$$W^* = \arg\max_{w} P(W/A) = \arg\max_{w} P(A/W)P(W)$$

其中$P(A/W)=1$，因此：$W^* = \arg\max_{w} P(W)$令输出字串$W = w_1, w_2, ...w_i, ...w_n$，

依据N-gram语言模型，有：$P(W) = P(w_1, w_2, ...w_i, ...w_n) = \prod_i p(w_i / w_1, ..., w_{i-1})$

本文使用三元模型，有：$P(W) = P(w_1, w_2, ...w_i, ...w_n) = \prod_i p(w_i / w_{i-2}, w_{i-1})$

没有儿童语料为音字转换提供训练语料，而使用成人语料的效果很差，为次，我们使用bootstrapping方法实现上述模型，实验步骤如下：

初始：较少拼音汉字对应的手工标注语料作为起始训练语料D_0，建立一个初始语言模型L0。

迭代：采用Li-1对部分语料D_i，进行标注后，进行人工校对，将校对后的语料加入初始训练语料，训练新的模型Li。

结束：模型达到预设精度，或者已加入全部语料。

在模型的训练过程以及最后的转写结果基础上，我们进行人工校正，由于语料是口语，又是儿童语和儿向语，训练语料的规模较小，自动转写产生很多错误，需要人工校对。我们进行了仔细的人工校正工作，在校正工作中，主要发现以下错误，主要有以下5种情况：

1)同音异形

因为汉语同音字同音词很多，平均每个音节对应5 个汉字，有些音节对应汉字达100多个。

2)原语料分词标准不一致

原拼音语料是经过切分的，但是其中的切分与目前一些基于汉字的切分标准并不统一。语料中的切分单元，与目前常用的北大词典(俞士汶等，2002)，有部分不同。

3)未登录词

由于训练语料规模较小，即使语料中的词汇量不大，也存在不少在训练语料中没有出现的词。

4)方言或口语

语料记录的是日常的口语，常见一些方言和口语，引起自动转写发生错误，比如“zen4me”，会自动转写“怎么”，正确的结果是“这么”，口语中的“leng1”，实际上是“扔”等等。

5)儿童依呀学语时的一些特殊词

这点是儿童语言的特点所在，有些儿童的语音，无法确定对应的汉字，我们转写为拟声词。

6) 其他

人名是比较难以处理的一种错误，还有汉语拼音中的 ü 在语料库中用 u 替代造成错误，等等。

标注者是母语汉语，并且是发音标准，熟练掌握拼音和汉字的 NLP 专业的研究生，标注者每天只工作 6 小时，并在休息 15 小时以上再工作。

3.2 词性标注

为了对语料进行更深入的分析，我们对转写之后的语料进行了 POS 标注。我们采用北大 POS 标注集(俞士汶等，2003)。

在 POS 标注时，我们仍然是先进行自动标注，之后进行人工校对。自动词性标记把每个词的所有可能 POS 都标记上，标注者需要在多个词性中，结合上下文，选择唯一正确的词性标记。

经过转写和词性标注，我们得到了一个基于汉字得汉语儿童语料库，我们把其中的儿童语言和儿向语言分别开来为儿童语言和儿向语言两部分。

4 比较实验

4.1 概述

我们在儿童语料和成人语料间，进行了一系列比较，比较其中的差异性，并试图将比较数据和儿童语言习得的相关结论进行比对。

我们选择三部分语料库进行比对试验，它们包括：

1)儿童语言，由儿童语料库抽取得到的 1 万句儿童语料，儿童有男有女，年龄在 1 岁至 5 岁，包含从短到较长的句子，从简单到较为复杂的语言现象，以及大量的拟声词。

2)儿向语言，从儿童语料库中抽取的超过 4 万句成人对儿童的话语，从简单的单句, 到一般复句, 再到比较复杂的复句, 包含的丰富的汉语语言现象和句型描述，又不同于成人间的对话。

3)《人民日报》语料库(俞士汶等，2002)，是北京大学计算语言学研究所与富士通公司(Fujitsu)合作的产品，加工 2700 万字的《人民日报》语料库。

我们对以上儿童、儿向和成人语料，在字层、词层、POS、句子简单特点上，做了一系列基于语料的比较。

4.2 各种比较数据及分析

词层特点

以下对儿童语、儿向语并对比人民日报语料进行词层信息的统计分析如表 1。

口语，尤其是儿童语言和儿向语言，与书面新闻用语，有很大的不同，由表 1 可以看出，儿童语言由儿童模仿父母的语言而得，很相近，并且有这样几个特点：

1)早期的语言中最多的是关于事物属性方面的词语，如形状，大小，质地等，如“颜色，好，大，小”。

2)表示积极意义的“是，好，大”比消极意义的“不，小”更多使用。

3)在儿向语言中出现最多的词，也基本上是儿童语言中出现最多的词，如“个，是，了”等。

为分析儿童词性习得，我们取 Zhou 中三组文本，一组儿童年龄为 20 个月，一组为 26 个月，一组为 32 个月。统计三组文本中儿童掌握词类的变化如表 2。

汉语的习得过程中，相比较于英语等曲折语言，缺少形态变化，形态语素所表达的语法意义由相应的虚词来表达。在词语法阶段，儿童语言中出现的都是实词，动词名词占绝大多数，而随着儿童语言越来越丰富，实词的组合已经不能满足意义表达的需要，虚词也就出现，并丰富起来。

儿童词语习得特点与儿童词汇习得的研究结论是吻合的(周国光 2001)。

儿向语		儿童语		人民日报	
词语	次数	词语	次数	词语	次数
个	1451	个	351	我	189508
是	1189	是	208	你	175142
什么	1026	的	165	了	113578
的	939	了	142	的	92886
好	776	这	90	不	69828
不	711	一	80	是	63298
呀	708	颜色	76	在	53700
了	687	要	75	一	48548
这	560	不	66	好	47360
一	469	有	63	有	41158
啊	428	好	62	没	40352
你	396	画	54	就	34706
来	363	大	52	天	34008
有	362	我	50	去	33874
看	361	呀	48	么	33724

表 1 三个语料库的最高频词表

年龄	20 个月	26 个月	32 个月
	词数	词数	词数
名词	366(41.2)	287(28.8)	208(25.7)
动词	299(33.6)	354(35.8)	237(29.3)
形容词	62(6.9)	55(5.5)	62(7.7)
副词	88(9.9)	102(10.3)	96(11.9)
代词	41(4.6)	145(14.6)	151(18.7)
连词	6(0.7)	7(0.7)	12(1.5)
数词	11(1.2)	14(1.4)	5(0.6)
象声词	9(1)	4(0.4)	4(0.5)
语气词	6(0.7)	27(2.7)	33(4)
总计	888	995	808

表 2 不同阶段词性统计表

语料库	儿童	儿向	人民日报
句子总数	3014	8972	95202
平均句长	3	5	11
总词频	10006	45486	超过 110 万
总词数	788	1592	54664
平均词频	13	29	20
平均词长	2	4	2
最长词	5	5	12
单字词	8552	36657	2956
双字词	1329	8489	30735
多字词	125	340	2097

表 3 句层和词语层对比结果

句层和词语层的对比结果如表 3，通过对表 3 分析，我们不难得到：

• 无论是句子总数还是最长句子儿向语都多于儿童语，说明儿童还处于语言习得阶段；

• 相比常用的书面语，儿童语料库的平均句长较小，说明儿童语和儿向语比较简洁，没有复杂繁冗的句子；

• 儿向语词频与儿童语词频差距很大，成人经常重复某些词以便儿童模仿；

•单字词、双字词、多字词的纵向比较，可看出单字词和双字词基本占据了对话语言的 99%，而成人语言中，单字词的比例并不高，在儿童语言习得期间，语言并无复杂词语，以简单词语和习语居多；

成人大量地对儿童讲话，刻意去要求儿童掌握模仿、替换、扩展等习得技巧，试图教会儿童运用语言。成人对儿童的语言没有复杂的语法成分，简单句比较多。儿向语与儿童语的总次数差距明显低于总词频差距，表示成人重复地对儿童讲话，要儿童去模仿，掌握习得技巧。而儿童语的最长词和儿向语最长词相同，儿童运用模仿的技巧，会去模仿成人的语言，而此时还在习得过

程，还没有能够使用更复杂的，自己的语言。人民日报中，双字词的比例最大，说明汉语言双字词是最普遍的，而无论是儿童语还是儿向语，单字词占最多的比例，儿童语言是与成人语言不同的，在语言习得期间，儿童通过学习比较简单的字词，习得语言。

儿向语和儿童语言的变化

儿童语言学家一般认为，儿童在 1 岁至 4 岁是“真正的语言”阶段，这个时期是儿童进行语法习得的阶段，就已完成初级阶段，我们通过对此阶段儿童语言和儿向语言的变化，研究儿童语言习得的阶段性结果，统计儿童语言，儿向语如表 4 和表 5。

儿童语言中，平均句长、最长句子随着儿童年龄的增长呈增加的趋势；平均词频总体的来说随着儿童年龄的增长呈下降的趋势，成人对儿童使用的词，和儿童能使用的词都增长了；平均词长随着儿童年龄的增长呈上升趋势，儿童习得了较为复杂的词；

儿向语中，单字词占总次数的比例随着年龄的增长呈下降趋势，双字词比例呈增加趋势，多字词基本稳定。也是表明，父母对儿童使用的词，从简单变的复杂。

语料分析表明， 14-20 个月的儿童，其语言形式多是单词句、双词句和电报句；20-32 个月的儿童，具有了运用词组构成语句的能力。儿童语言形式的特点是出现了词组；32-48 个月的儿童，构成语言的单位除了前期的单词、词组以外，又增加了小句，出现了小句做句子成分的句子、逻辑语义关系正确的复句。

月份	14	20	26	32	48
句子总数	34	58	86	69	54
平均句长	2	4	5	4	5
最长句子	11	13	18	20	20
总词频	80	248	388	303	284
总词数	28	100	139	129	136
平均词频	3	2	3	2	2
平均词长	1	2	2	2	2
最长词	5	6	8	8	10
单字词%	55	62.9	61.1	61.1	67.6
双字词%	2.5	12.9	13.7	14.5	10.2
多字词%	42.5	24.2	25.3	24.4	22.2

表 4 儿童语变化

月份	14	20	26	32	48
句子总数.	72	74	36	94	23
平均句长	5	6	6	6	6
最长句子	15	18	18	21	22
总词频	376	412	218	548	130
总词数	128	166	117	195	90
平均词频	3	2	2	3	2
平均词长	3	3	4	5	5
最长词	7	6	5	6	6
单字词%	66	65.8	65.6	64.8	64.6
双字词%	14.1	15	17	17.2	17.7
多字词%	19.9	19.2	17.4	18	17.7

表 5 儿向语变化

5 结论

本文简要地介绍了我们目前在语言习得研究方面所做的一些工作，包括汉语儿童语料转写和标注工作，并在此语料的基础上，在汉字、词以及句子三个层面上，对儿童语言、儿向语言以及成人语言进行了对比统计。

通过这些工作，我们在语料库建立和人工标注方面积累了经验。

基于数据的分析结果与已有的儿童语言习得结论基本吻合。

致谢

本文得到了高等学校学科创新引智计划（项目编号：B08004）、国家支撑计划项目(项目编号：2007BAHo5B02-04)的支持。

人工校正是一项繁琐而艰巨的工作，在这里，感谢实验室同学在转录与 POS 的人工标注付出的辛苦劳动。

参 考 文 献

[1] 赵艳芳,2001.认知语言学概论,上海外语教育出版社

[2] 周国光,王葆华,2001.儿童句式发展研究和语言习得理论,北京语言文化大学出版社

[3] 李宇明, 1995.儿童语言的发展,华中师范大学

[4] 俞士汶，段慧明，朱学锋，孙斌. 2002. 北京大学现代汉语语料库基本加工规范.中文信息学报，No.5，pp.49-64；No.6，pp.58-65

[5] 俞士汶,段慧明,朱学峰,孙斌,常宝宝,2003.北大语料库加工规范：切分·词性标注·注音,汉语语言与计算学报，16(6): 58–65.

[6] Briscoe, Ted and John Carroll, 1997. Automatic ex-traction of subcategorization from corpora. In Pro-ceedings of the 5th ACL Conference on Applied Natural Language Processing, Washington, DC.

[7] B MacWhinney. The CHILDES project: Tools for analysing talk. Lawerance ErlBaum Associates, Hillsdale, NJ, second edition, 1995.

[8] Chang, C. (1998). The development of autonomy in preschool Mandarin Chinese-speaking children's play narratives. Narrative Inquiry, 8 (1), 77-111.

[9] Fletcher, P., Leung, S. C-S., Stokes, S. F., & Weizman, Z. O. (2000). Cantonese pre-school language development: A guide. Hong Kong: Department of Speech and Hearing Sciences.

[10] Gamallo, P., Agustini, A. and Lopes Gabriel P., 2002. Using Co-Composition for Acquiring Syntactic and Semantic Subcategorisation, ACL-02.

[11] Guoguang ZHOU, Baohua WANG, 2001.The study of construction development in Chinese Children's Speech and the theory of Language Acquisition. Beijing Language and Culture Press

[12] Jane S. Tsay, 2008 .Acquiring causatives in Taiwan Southern Min, Journal of Child Language (2008), 35:467-487

[13] Korhonen, Anna, 2003. Clustering Polysemic Sub-categorization Frame Distributions Semantically. Proceedings of the 41st Annual Meeting of the As-sociation for Computational Linguistics, pp. 64-71.

[14] Milestones in the learning of spoken Cantonese by pre-school children. Language Fund, Hong Kong. Paul Fletcher, Thomas H.T. Lee, Samuel Leung, and Stephanie Stokes (1996-1999).

[15] M. Marcus, B. Santorini, and M. Marcinkiewiecz. 1993. Building a large annotated corpus of english. Computational Linguistics,19(2):313–330.

[16] P. Buttery and A. Korhonen. 2005. Large-scale analysis of verb subcategorization differences between child directed speech and adult speech. In Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes, Saarbrucken, Germany.

[17] P Buttery. 2006. Computational Models for First Language Acquisition. Ph.D. thesis, University of Cambridge.

[18] P. Schone and D. Jurafsky. 2001. Knowledge-Free Induction of Inflectional Morphologies. In Proceedings of NAACL-2001. Pittsburgh, PA, June 2001.

[19] Sabine Shulte im Walde, 2002. Inducing German Se-mantic Verb Classes from Purely Syntactic Sub-categorization Information. Proceedings of the 40st ACL, pp. 223-230.

[20] Sarkar, A. and Zeman, D. 2000. Automatic Ex-traction of Subcategorization Frames for Czech. In Proceedings of the 19th Interna-tional Conference on Computational Linguis-tics, aarbrucken, Germany.

[21] Tardif, T. (1996). Nouns are not always learned before verbs: Evidence from mandarin speakers' early vocabularies. Developmental Psychology, 32, 492–504.

[22] Weizman, Z. O. and Fletcher, P. A comparative study of language development: English and Cantonese pre-schoolers in Hong Kong. Committee on Research and Conference Grants, University of Hong Kong. (2000).