

UD-English-CHILDES: 用于儿童语言互动的金银级别通用依存树收集资源

Xiulin Yang¹ Zhuoxuan Ju¹ Lanni Bu¹ Zoey Liu² Nathan Schneider¹

¹Georgetown University

²University of Florida

{ xy236, zj153, lb1437, nathan.schneider } @georgetown.edu
liu.ying@ufl.edu

Abstract

CHILDES 是一个广泛使用的儿童和针对儿童谈话的转录资源。本文介绍了 UD-English-CHILDES, 这是第一个正式发布的来自已注释的 CHILDES 数据的通用依存 (UD) 树库, 使用一致和统一的注释指南。我们的语料库统一了来自 11 位儿童及其看护人的注释, 总计超过 4.8 万句。我们在 UD v2 框架下验证了现有的黄金标准注释, 并提供了额外的 100 万银标准句子, 为计算和语言研究提供了一个一致的资源。

1 介绍

儿童语言数据交换系统 (CHILDES) (MacWhinney, 2000) 长期以来一直是语言习得研究、儿童语言计算模型和自然语言处理 (NLP) 工具评估的关键资源。然而, 许多分析依赖于不同的语法假设 (e.g., Pearl and Sprouse, 2013; Szubert et al., 2024; Liu and Prud'hommeaux, 2021; Gretz et al., 2015; ?), 因此采用不同的注释框架或标准。虽然大多数现有的注释使用句法依赖关系——部分原因是注释和解析相对简单, 以及对通用依赖关系 (UD) 框架 (Nivre et al., 2016, 2020) 的广泛采用——但注释实践在不同数据集之间依然不一致。这主要是由于缺乏统一的注释儿童语言的指南, 儿童语言具有独特的挑战, 这些挑战尚未在现有的 UD 文档中得到充分解决。

随着 UD 树库在 NLP (例如, Jumelet et al., 2025; ?) 和语言习得研究 (例如, Clark et al., 2023; ?) 中变得越来越宝贵, 使用工具如 stanza (Liu and MacWhinney, 2024) 解析 CHILDES 数据的努力也在增加。然而, 得出的注释质量通常不一致且无法得到保证。在本文中, 我们编译、协调并手动修正了 CHILDES 数据的主要 UD 风格注释, 使其成为一致、统一的 UD 格式, 生成了一个包含 48K 句子和 236K 标记 (包括例如 Figure 1 中的树) 的黄金标准树库。此外, 我们构建了一个更大的含有 1M 句子和 6M 标记的银标准树库, 并报告了解析器的准确性估计。我们公开发布了这两个

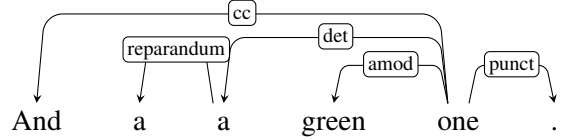


Figure 1: 莉莉 (普罗维登斯语料库, 句子 ID=16916280) 的儿童话语的 UD 树

数据集。¹

2 相关工作

2.1 CHILDES 语料库

CHILDES 在语言习得研究和 NLP 工具开发中扮演了重要角色。除了专门的语料库外——如临床数据集 (Gillam and Pearson, 2004)、自然家庭互动 (Gleason, 1980) 和控制的实验室研究 (Newman et al., 2016)——CHILDES 支持发展语言学的多种方法。它的许多语料库为语言习得的基础理论提供信息, 特别是刺激贫乏假说 (Chomsky, 1976)。研究人员经常使用 CHILDES 中的儿童导向语音来量化这些理论中核心的语言结构分布, 例如想要收缩 (Getz, 2019)、指代的 one (Foraker et al., 2009; Pearl and Mis, 2011)、辅助动词前置 (Perfors et al., 2011) 和句法岛屿 (Pearl and Mis, 2011)。它也被用于语言习得的计算模型 (例如, ?) 中。

CHILDES 也已成为 NLP 工具基准测试和语言模型预训练的重要资源。在 Huang (2016) 的工作之后, 诸如 Liu and Prud'hommeaux (2023) 的研究突出了 UD 解析器在应用于儿童导向语音时所面临的挑战, 显示出与成人数据相比存在显著的性能差距。CHILDES 还支持关于预训练动态 (Feng et al., 2024) 和高效语言模型发展的最新研究, 包括在类似 BabyLM Challenge (Choshen et al., 2024; Charpentier et al., 2025) 这样的倡议中。

2.2 CHILDES 依存树库

早期关于英语 CHILDES 数据的依赖解析研究使用了一个自定义的语法关系目录 (GR; Sagae

¹pseudo-link-for-childes-UD

Corpus	Children	Gold Annotation	Speakers	UPOS	Feats	Utterances		Tokens	
						Gold	Silver	Gold	Silver
S+24	Adam	Trees, UPOS; features from original CHILDES	Adults	Gold	Convrtd	17,233	0	91,114	0
LP21	Eve	Trees; others unspecified	All	Silver	Silver	2,207	108,044	8,497	532,319
LP23	10 Children	Trees; others unspecified	All	Silver	Silver	34,530	1,101,061	168,284	6,461,084
UD-English-CHILDES	11 Children	Trees, UPOS	All	Gold	N/A	48,183	1,197,471	236,941	6,892,314

Table 1: 本文中编译的基于 CHILDES 的 UD 树库概览。源语料库标签 (S+24、LP21、LP23) 在 §3 中定义。注意在 Adam 数据中有重叠: S+24 的数据是原始数据集中的统计数据; 在我们的版本中, 这些数据经过过滤以避免重复, 并与相应的 LP23 语句合并。

Child	Corpus	Child age range	Gold sents	Gold toks	Silver sents	Silver toks
Laura	Braunwald (Braunwald, 1971)	1;3-7;0 (1;3-7;0)	4,622	21,079	41,862	205,427
Adam	Brown (Brown, 1973)	1;6-5;2 (1;6-5;2)	16,736	84,643	93,315	452,348
Eve	Brown	1;6-5;1 (1;6-5;2)	2,207	8,497	108,044	532,319
Abe	Kuczaj (Kuczaj, 1977)	2;4-5;0 (2;4-5;0)	4,167	22,437	38,630	230,489
Sarah	Brown	1;6-5;2 (1;6-5;2)	5,347	23,233	104,926	517,654
Lily	Providence (Demuth et al., 2006)	0;11-4;0 (0;11-4;0)	1,499	6,337	79,573	422,245
Naima	Providence	1;3-3;11 (0;11-4;0)	2,534	14,360	236,350	1,422,543
Violet	Providence	0;11-4;0 (0;11-4;0)	721	1,857	32,801	164,975
Thomas	Thomas (Lieven et al., 2009)	2;0-4;11 (2;0-4;11)	4,240	20,333	313,550	2,039,132
Emma	Weist (Weist and Zevenbergen, 2008)	2;2-4;10 (2;1-5;0)	2,423	13,730	74,825	474,460
Roman	Weist	2;2-4;9 (2;1-5;0)	3,653	20,557	73,595	467,633

Table 2: 每个孩子的详细统计数据, 包括金标注和银标注的数量及其对应的年龄范围 (月)。银标注语料库中的年龄显示在括号内。有关语料库来源的网址, 请参见 ??。

et al., 2004; ?)。这些目录逐渐演变以应对 CHILDES 特定的挑战 (Sagae et al., 2007), 并使用一个监督解析器应用于整个英语 CHILDES 语料库 (Sagae et al., 2010)。

最近, CHILDES 中引入了 UD 风格的注释。Liu and Prud’hommeaux (2021) 使用半自动方法将以前基于 GR 的注释转换为 UD 树, 重点是来自 Brown 语料库中 Eve 数据 (18-27 个月儿童) 的儿童语言。随后, Szubert et al. (2024) 通过自动转换 GR 注释并手动校正它们, 开发了黄金标准的 UD 注释。他们的数据集包括来自 Brown 语料库中 Adam 数据和希伯来语 Hagar 语料库的儿童导向语音, 解决了诸如重复和非标准词汇等口语特有的现象, 并进行了语义映射。

在这些研究的基础上, Liu and Prud’hommeaux (2023) 显著扩大了 UD 标注的范围, 包括来自 10 名年龄在 18 至 66 个月的儿童 (布朗语料库中的亚当以及其他语料库中的 9 名儿童) 的话语, 涵盖了儿童和看护人的语音。他们的工作处理复杂的口语特征, 包括语音修复和重启。

3 注释

本工作利用了三个现有的 UD 树库: Szubert et al. (2024) (以下简称 S+24)、Liu and

Prud’hommeaux (2021) (LP21) 和 Liu and Prud’hommeaux (2023) (LP23), 在 Table 1 中总结。这些树库已经进行了标注, 我们的人力标注工作主要集中在纠正错误和协调语料库之间的标注。我们在 Tables 1 and 2 中展示了编译后的统计数据。Table 1 总结了完整的语料库及其来源贡献, Table 2 提供了每个子项的统计数据。

3.1 标注流程

根据 Liu and Prud’hommeaux (2023), 我们使用 `childesr`² R 包收集 CHILDES 语料库, 并确定每个句子是否已用 UD 树进行标注。对于已经存在标注的句子, 自动使用 `stanza`³ (Qi et al., 2020) 进行通用词类标注 (UPOS), 而未标注的句子则被分配 UPOS 和依存树。我们当前的工作集中在修正先前人工标注的数据。为了确保符合 UD 指南, 我们通过 UD 验证工具⁴ 处理所有标注的句子, 并手动修正那些未通过验证的句子。这项修正工作由三名接受过 UD 标注训练的语言学研究生完成。许多错误源于 UPOS 标签和依存标签之间的不匹配 (如 LP21

²<https://langcog.github.io/childes-db-website/>

³`stanza 1.9.2`

⁴<https://github.com/UniversalDependencies/tools/blob/master/validate.py>

```
# sent_id = 17235906
# childes_toks = who's that
# corpus_name = Providence
# gold_annotation = True
# speaker_age = 37.49358303045237
# speaker_gender = male
# speaker_role = Mother
# type = question
# text = Who's that?
1-2   Who's _ _ _ _ _
1     Who  who  PRON  WP  _  0  root  0:root  _
2     's   be   AUX  VBZ  _  1  cop   1:cop  _
3     that that PRON  DT  _  1  nsubj 1:nsubj  SpaceAfter=No
4     ?    ?    PUNCT ?   _  1  punct 1:punct  _
```

Figure 2: 来自 CHILDES-Providence 语料库的金标准标注 CoNLL-U 语句示例。

和 LP23 使用自动 UPOS 标注)。此外，我们还处理格式问题，例如多词标记、空格不匹配（例如，SpaceAfter），以及当前 UD 指南不支持的已弃用的依存关系（例如，compound:svc，obl:about_like，nmod:over_under）。最常见的五个语言问题如下：

advmod 被标记为 ADP 这个错误通常出现在动词短语中，如 get up 和 take over。原始标注将 advmod 分配为带有词性标记 ADP 的动词短语的依赖关系。我们将其修订为 compound:prt，以符合 UD 对短语粒子的处理。

像 be 和 have 这样的助动词经常被错误地分类为主要动词或小品词。在某些情况下，词元也被错误标记——最显著的是，缩写形式如's 的词元被错误地指定为's 而不是适当的助动词 be。在这些情况下，我们修正了词性的标记和词元的标注。

标记为 PUNCT 的词汇项 stanza 解析器经常将自发言语中不流利的单词片段误标记为标点符号（例如，OK/INTJ Adam/PROPN ride/VERB dat/ PUNCT ./PUNCT。我们基于上下文和说话者意图重新分配这些标记的合适 UPOS 标签，通常作为感叹词。

限定词的含糊或减少形式——例如“de”——经常被误识别为专有名词（PROPN）。我们手动审查这些情况，并在适当时重新标注为 DET。

功能词作为中心词并带有从属成分 在先前的树库中，出现在功能关系中的词语如 case、mark 和 aux 被指派了子节点，这违反了 UD 对这些词语应作为叶节点的限制。我们将这些错误依赖重新分配到适当的内容中心，以确保结构符合 UD 的投射性和功能词限制。

3.2 协调

由于每个树库遵循其自己的标注指南，我们进行了系列标准化步骤，将它们统一成一致的格式。我们的统一格式主要基于 LP23，并在下文中描述了几个改动。

在我们归一化的 CoNLL-U 文件中，我们包括了以下元数据字段，例子在 Figure 2 中提供：sent_id（通过 childes R 包获取的语句 ID）；childes_toks（词元化的语句）；corpus_name（原始语料库名称）；gold_annotation（指示句子是否为手动注释）；speaker_gender、speaker_role 和 speaker_age（发言者元数据）；text（文本与树对齐），以及 type（句子类型）。Table 3 总结了主要句子类型的分布，并将它们与多体裁英文语料库 GUM 的 UD 2.15 版本中的句子类型进行比较。值得注意的是，问题在 CHILDES 对话中出现的频率要高得多——几乎是陈述句的一半（45%），而在 GUM 中只有 9%。

Sentence Type	CHILDES	GUM (s_type)
declarative	31,996	7,695 (decl)
question	14,241	716 (q, wh)
imperative_emphatic	797	1,326 (imp, intj)
others	1,149	2,409

Table 3: 在金色 CHILDES 和 GUM 语料库中的句子类型计数。“其他”包含不太常见的类别：无声结束、中断、自我中断、引号下一行、自我中断问题、无声结束问题和中断问题。

为了使记录与书面英语习惯一致，我们将每个话语的第一个词大写，并根据元数据中提供的句子类型推断每个句子的句末标点符号。

每个树库为 reparandum 和 parataxis 关系定义了自己的子类型。例如，S+24 包括 parataxis:repeat，而 LP23 使用 parataxis:discourse，这些标签在当前的 UD 指南中不存在。类似地，LP21 和 LP23 使用子类型如 restart 和 repetition 来标记儿童语言中的特殊话语特征。为了确保树库之间的一致性，我们将所有此类子关系信息移动到 MISC 列。

由于

Metrics	Children's speech	Parents' speech	Overall
LAS	81.2	86.3	84.2
UAS	87.2	91.0	89.5

Table 4: 儿童讲话、家长讲话和整体表现的 LAS 和 UAS 得分。

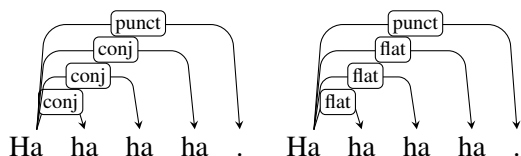


Figure 3: 标准化前（左）和标准化后（右）的树示例。

其他 S+24 是使用 UD 指南版本 1.0 注释的，我们使用一个脚本⁵和人工注释将其转换为 UD 版本 2.0。例如，在这些注释中，我们改变了 flat 的中心依存方向。

由于 S+24 和 LP23 在 Adam 语料库中有重叠，我们合并了这两个树库的标注。S+24 中有 3375 个句子是重复的。我们从语料库⁶中移除了这些句子。

为了确保更符合语言学的分析，我们在处理感叹词时也偏离了 Liu and Prud'hommeaux (2023)。我们没有将只包含感叹词的语句（例如，哈哈哈哈哈）标注为 conj，而是使用了 flat 关系，如图 3 所示。

3.3 银数据评估

为了创建银标准注释，我们对语料库的未注释部分应用 stanza。为了估计这些银标准注释的质量，我们在黄金标准的数据上评估解析器的性能。我们在 Table 4 中报告带标签和无标签的附件分数 (LAS/UAS)。解析器在总体上实现了 83.3 的 LAS。在成人语音上性能较高 (86.3 LAS)，而在儿童语音上则较低 (81.2 LAS)，这可能是由于成人的话语具有更大的句法规则性和更低频率的不流畅现象。

在本文中，我们介绍了第一个统一的 CHILDES UD 树库，涵盖了 11 个语料库和超过 48k 的来自儿童指向和儿童生成语音的句子。未来的工作将涉及对银标准数据的进一步修正和树库的扩展。我们邀请在这一持续努力中的合作。

References

Susan R Braunwald. 1971. Mother-child communication: the function of maternal-language input. *Word*,

⁵<https://github.com/UniversalDependencies/tools/tree/master/v2-conversion>

⁶从 S+24 提取的 883 个句子无法合并，因此也从我们的树库中移除。

27(1-3):28–50.

Roger Brown. 1973. *A first language: The early stages*. Harvard University Press.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Hu, Jaap Jumelet, Tal Linzen, Jing Liu, Aaron Mueller, Candace Ross, and 1 others. 2025. *BabyLM turns 3: Call for papers for the 2025 BabyLM workshop*. *arXiv preprint arXiv:2502.10645*.

Noam Chomsky. 1976. *Reflections on language*. Temple Smith London.

Leshem Choshen, Ryan Cotterell, Michael Y Hu, Tal Linzen, Aaron Mueller, Candace Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. 2024. *[Call for papers] the 2nd BabyLM challenge: Sample-efficient pretraining on a developmentally plausible corpus*. *arXiv preprint arXiv:2404.06214*.

Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger Levy. 2023. *A cross-linguistic pressure for Uniform Information Density in word order*. *Transactions of the Association for Computational Linguistics*, 11:1048–1065.

Katherine Demuth, Jennifer Culbertson, and Jennifer Alter. 2006. Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language and speech*, 49(2):137–173.

Steven Y. Feng, Noah Goodman, and Michael Frank. 2024. *Is child-directed speech effective training data for language models?* In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22055–22071, Miami, Florida, USA. Association for Computational Linguistics.

Stephani Foraker, Terry Regier, Naveen Khetarpal, Amy Perfors, and Joshua Tenenbaum. 2009. *Indirect evidence and the poverty of the stimulus: The case of anaphoric one*. *Cognitive Science*, 33(2):287–300.

Heidi R Getz. 2019. *Acquiring wanna: Beyond universal grammar*. *Language Acquisition*, 26(2):119–143.

Ronald Bradley Gillam and Nils A Pearson. 2004. *Test of narrative language*. Pro-ed Austin, TX.

Jean Berko Gleason. 1980. *The acquisition of social speech routines and politeness formulas*. In *Language*, pages 21–27. Elsevier.

Shai Gretz, Alon Itai, Brian MacWhinney, Bracha Nir, and Shuly Wintner. 2015. *Parsing hebrew CHILDES transcripts*. *Language Resources and Evaluation*, 49:107–145.

Rui Huang. 2016. *An evaluation of pos taggers for the CHILDES corpus*. *CUNY Academic Works*.

- Jaap Jumelet, Leonie Weissweiler, and Arianna Bisazza. 2025. [MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs](#). *arXiv preprint arXiv:2504.02768*.
- Stan Kuczaj. 1977. The acquisition of regular and irregular past tense forms. *Journal of verbal learning and verbal behavior*, 16(5):589–600.
- Elena Lieven, Dorothé Salomo, and Michael Tomasello. 2009. Two-year-old children’s production of multiword utterances: A usage-based analysis.
- Houjun Liu and Brian MacWhinney. 2024. [Morphosyntactic analysis for CHILDES](#). *arXiv preprint arXiv:2407.12389*.
- Zoey Liu and Emily Prud’hommeaux. 2021. [Dependency parsing evaluation for low-resource spontaneous speech](#). In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 156–165, Kyiv, Ukraine. Association for Computational Linguistics.
- Zoey Liu and Emily Prud’hommeaux. 2023. [Data-driven parsing evaluation for child-parent interactions](#). *Transactions of the Association for Computational Linguistics*, 11:1734–1753.
- Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press.
- Rochelle S Newman, Meredith L Rowe, and Nan Bernstein Ratner. 2016. [Input and uptake at 7 months predicts toddler vocabulary: the role of child-directed speech and infant processing skills in language development](#). *Journal of Child Language*, 43(5):1158–1173.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. [Universal Dependencies v1: A multilingual treebank collection](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).
- Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. [Universal Dependencies v2: An evergrowing multilingual treebank collection](#). *arXiv preprint arXiv:2004.10643*.
- Lisa Pearl and Benjamin Mis. 2011. [How far can indirect evidence take us? Anaphoric one revisited](#). In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.
- Lisa Pearl and Jon Sprouse. 2013. [Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem](#). *Language Acquisition*, 20(1):23–68.
- Amy Perfors, Joshua B. Tenenbaum, and Terry Regier. 2011. [The learnability of abstract syntactic principles](#). *Cognition*, 118(3):306–338.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A Python Natural Language Processing Toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.
- Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. 2007. [High-accuracy annotation and parsing of CHILDES transcripts](#). In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 25–32, Prague, Czech Republic. Association for Computational Linguistics.
- Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. 2010. [Morphosyntactic annotation of CHILDES transcripts](#). *Journal of Child Language*, 37(3):705–729.
- Kenji Sagae, Brian MacWhinney, and Alon Lavie. 2004. [Adding syntactic annotations to transcripts of parent-child dialogs](#). In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04)*, Lisbon, Portugal. European Language Resources Association (ELRA).
- Ida Szubert, Omri Abend, Nathan Schneider, Samuel Gibbon, Louis Mahon, Sharon Goldwater, and Mark Steedman. 2024. [Cross-linguistically consistent semantic and syntactic annotation of child-directed speech](#). *Language Resources and Evaluation*.
- Richard M Weist and Andrea A Zevenbergen. 2008. Autobiographical memory and past time reference. *Language Learning and Development*, 4(4):291–308.

在这项工作中，我们包含了以下语料库中的来源：