Four Decades of Open Language Science: The CHILDES Project

Vera Kempe[1]*, Patricia J. Brooks[2], Steven Gillis[3]

[1]Abertay University, Dundee, UK
[2]CUNY Graduate Center and College of Staten Island, USA
[3]University of Antwerp, Belgium

Abstract

The Child Language Data Exchange System (CHILDES), created by Brain MacWhinney and Catherine Snow in 1984, is one of the earliest Open Science and data sharing initiatives in child language development research, and probably in developmental psychology and the behavioral sciences more generally. It is the cornerstone of TalkBank—a repository of transcripts, audio, and video files of natural language samples. Here we highlight how the CHILDES Project served as a trailblazer for the language development research community by being the first initiative to introduce a Big Data approach, encouraging and facilitating crosslinguistic data collection and championing science collaboration through open access to data and analysis tools. We conclude with an outlook on the future of CHILDES and suggestions for where child language development researchers might turn their attention when collecting and donating observational data. Understanding the many paths to language will require expanding CHILDES to increase representation of culturally and neurally diverse populations, finding solutions to the challenge of promoting Open Science practices while safeguarding participant agency and privacy, and leveraging AI tools for automated transcription and data analysis.

Keywords: *Open Science, CHILDES, Big Data, Child Language, Data Sharing*

[1]Introduction

Child language development research has a rich history of collecting observational data. Known efforts date back to the beginning of the 20th century and consist of detailed records

---

* Corresponding author.

by parents of their children's early language production (e.g. Stern & Stern, 1907, for German, Pavlovitch, 1920, for Serbian, Isaacs, 1930, for English, and Gvozdev, 1948, 1949, for Russian; see Slobin, 1968, for a review). Given the potential bias and selectivity of diary entries, once portable recording equipment became accessible and affordable, researchers and other interested adults began recording children's vocal productions and transcribing them for further analysis. Many of the early efforts to document child language development focused on longitudinal observations of a small number of children (e.g., Bloom, 1970; Brown, 1973). The initial emphasis on case studies greatly constrained generalizability of findings, prompting the need for efforts to pool data across laboratories and languages to further understanding of variability in language development trajectories.

The visionary contributions of Brian MacWhinney and Catherine Snow were to lay the groundwork for a data exchange system that would allow researchers to grow a database of authentic language samples collected from many children and to make the data freely accessible to researchers, students, and the interested public. The idea for the CHILDES project emerged in 1981 at a conference organized by Dan Slobin at the Max Planck Institute for Psycholinguistics in Nijmegen on the topic of crosslinguistic influences on language development (IASCL Child Language, 2023), and developed further in conversations with Susan Ervin-Tripp and Willem Levelt (MacWhinney & Snow, 1985). At the time, copying technologies like the mimeograph had made it possible to share paper transcripts of child language, albeit at a limited scale, to interested researchers. As an early example, Roger Brown's transcripts of Adam, Eve, and Sarah's early language development were circulated via paper copies, with just 12 copies available in total (IASCL Child Language, 2023). With the advent of the personal computer, the group meeting in Nijmegen came up with the idea of entering the data from the paper transcripts onto computers and making the digital files available to language development researchers around the world. The early efforts to share CHILDES transcripts and data analysis programs relied on floppy disks, CD-ROMs, and the postal service (IASCL Child Language, 2023). Thus, CHILDES preceded by several years the launch of the Internet in 1983 and the development of the World Wide Web—an invention that would make it possible for researchers (and the lay public) to use their personal computers to access and transfer data. A couple of years later, in 1984, the Child Language Data Exchange System (CHILDES) was born with funding provided by the MacArthur Foundation (MacWhinney & Snow, 1985).

CHILDES corpora are sets of transcripts and supporting media (when available) collected by researchers with varied research aims and donated to the repository. From the start, the CHILDES Project included data from a variety of languages in addition to English, such as Danish (Plunkett, 1986), Dutch (Elbers, 1985; Gillis, 1984), French (Suppes et al., 1973), German (Wagner, 1985), Hebrew (Berman, 1990), Hungarian (MacWhinney, 1974), Italian (Volterra, 1972), Polish (Weist et al., 1984), Spanish (Linaza et al., 1981), Tamil (Narasimhan, 1981), and Turkish (Slobin, 1982). Efforts to build the database of media often involved digitizing original audio files delivered via nine-track tapes or cassettes (e.g., Hall et al., 1984; Nelson, 1989). At the time of writing, 42.0% of the corpora had accompanying audio files and 11.7% had accompanying video. The difference in the availability of audio vs. video recordings in CHILDES has not changed markedly over the years, and likely reflects considerations related to protecting the privacy of participating children and their family members. Privacy

considerations may also require certain recordings to be embargoed; the description of CHILDES we provide below therefore refers only to its publicly accessible part.

To facilitate analysis of the written transcripts, MacWhinney and Snow (1990) developed standardized conventions for transcribing utterances (CHAT: Codes for the Human Analysis of Transcripts) and dedicated software for data analysis (CLAN: Computerized Language Analysis). CLAN encompasses dozens of different commands. As examples, *freq* generates frequency counts of words, parts of speech, and other coded information, *kwal* (key word and line) searches for specific words or coding categories, and *combo* searches for specific combinations of elements (words, codes, etc.). Other commands calculate summary statistics, e.g., *MLU* for mean length of utterance and *VOCD* for vocabulary diversity. These and other CLAN commands have been developed and refined over the years, with up-to-date manuals and programs available on the CHILDES website (https://childes.talkbank.org/). Additionally, with Yvan Rose and others, MacWhinney developed Phon, a software tool for examining phonological development (Rose et al.., 2006) and PhonBank (Rose & MacWhinney, 2013), a repository of child phonology data that is now part of the TalkBank system (MacWhinney, 2019). Phon automates coding of segments, syllables, and other phonetic and phonological features of children's speech production and supports direct comparison of adult-produced (i.e., target) and child-produced forms. CHAT transcripts and associated media files interface with a number of other programs including Praat for phonetic analysis (Boersma & Weenink, 1995-2023) and ELAN for annotation of audio and video files (Auer et al., 2010), and with scripting languages like Python to allow users to pipe data from one tool to another. Together, these tools allow diverse corpora, organized in a large and coherent database, to be analyzed using similar procedures.

Example (1) is an excerpt from a transcript within the MacWhinney (1991) corpus illustrating how CHAT is used to format the transcript. In CHAT, metadata containing information about participants, available media, date, situation, and other comments are listed on lines that start with the symbol @. The *main tiers* of the transcript start with the symbol * followed by a three-letter abbreviation of the participant's role (*FAT, *CHI). These lines contain a standardized orthographic transcription of the recorded speech, with specific conventions for marking overlapping speech, interruptions, pauses, repetitions, special words (e.g., onomatopoeia), unintelligible speech, and the like.

(1) Excerpt from MacWhinney (1991) transcript 020718c, available for download at: https://childes.talkbank.org/access/Eng-NA/MacWhinney.html

```
@Begin
@Languages: eng
@Participants: CHI Ross Target_Child , FAT Brian Father
@ID: eng|MacWhinney|CHI|2;07.18|male|TD||Target_Child|||
@ID: eng|MacWhinney|FAT||male|||Father|||
@Media: 020718c, audio
@Date: 19-AUG-1980
@Types: long, toyplay, TD
@New Episode
```

@Tape Location: tape22 , side b , 260
@Situation: Mark's making a pie of his breakfast
*__FAT:__ look Marky is making a mess !
*__CHI:__ yeah .
*__FAT:__ isn't that nice Mark .
*__CHI:__ that's nice .
*__FAT:__ that's real nice .
*__FAT:__ he's making a beautiful mess .
*__FAT:__ what's he doing .
*__CHI:__ he's making a mess .
*__CHI:__ not nice (.) Mark .
@End

Utterances may be coded in CHAT by inserting one or more *dependent tiers* under the main tier, with each dependent tier starting with the symbol %. Depending on the aims of the researcher, dependent tiers may contain phonetic or phonemic transcriptions, morphosyntactic coding, coding of grammatical relations in accordance with the universal dependencies (UD) framework (de Marneffe et al., 2021), or extralinguistic information (e.g., gestures, actions). Example (2) shows dependent tiers with morphosyntactic (%mor) and grammatical relations (%gra) coding of one of the child's utterances in transcript 020718c (MacWhinney, 1991); note that these coding tiers were auto generated using CLAN tools (MacWhinney & Fromm, 2022). Researchers may create their own dependent tiers to code features of child language that are of relevance to their interests.

(2) Example line from MacWhinney (1991) transcript 020718c showing dependent tiers
*__CHI:__ he's making a mess .
%mor: pro:sub|he~aux|be&3S part|make-PRESP det:art|a n|mess .
%gra: 1|3|SUBJ 2|3|AUX 3|0|ROOT 4|5|DET 5|3|OBJ 6|3|PUNCT

CHILDES as a Big Data Initiative

With the launch of Dataverse Network in 2006 (King, 2007), the Open Science Framework in 2012 (Spies, 2013), the Center for Open Science in 2013 (Foster & Deardorff, 2017), and Databrary in 2014 (Adolph, 2016; Gilmore et al., 2016), sharing of scientific data has become increasingly commonplace. In this context, it is easy to overlook the foresight of MacWhinney, Snow, and their collaborators who anticipated a Big Data approach to language development research over 40 years ago. Not only did the creators of CHILDES provide manuals with technical information (MacWhinney, 1991), but they also produced an edited volume demonstrating how to use CHAT coding and CLAN programs in specific, well-defined research projects on a variety of topics (Sokolov & Snow, 1994). Topics covered in this initial volume included parental use of diminutives (e.g., *froggy*, *doggy*), child and parental use of superordinates, children's acquisition of Spanish determiners, and the availability of direct and indirect forms of negative evidence pertaining to the grammaticality of children's utterances. In this section, we take stock of how CHILDES has evolved over this period and mention future directions that data-sharing efforts have taken within the broader TalkBank project.

Vera Kempe, Patricia J. Brooks, Steven Gillis

The summary statistics presented in this section give an indication of how the CHILDES project transformed over a period of 40 years from a small repository that fit onto a CD-ROM into a rich database with sibling projects now known as TalkBank. At the time of writing, CHILDES comprised 436 accessible corpora containing a total of 73,958,859 words embedded in 19,908,190 utterances produced by 16,382 children and their caregivers living in diverse societies around the world. Please note that corpora are being added continuously, some of which remain embargoed for a variety of reasons. Table 1 provides various measures of the current database. Columns are organized by corpus type, determined by developmental status (typical vs. non-typical), language context (monolingual vs. bilingual/multilingual), and elicitation method (observational vs. narrative). Within the latter categorization, observational data include recordings of conversations (dialogue) collected at home or in the lab, in free-play contexts or during specific activities like mealtime or book reading. The corpora include both longitudinal and cross-sectional designs, with the majority recording interactions involving young children (< 5 years of age) with their family members. Narrative corpora are elicited monologues, with adults providing support but generally taking on a secondary role. In more than 30 of the narrative corpora, children were asked to tell a story about a frog, using a wordless picture book as a prompt (Mayer, 1969). Analyses of children's frog stories representing various language contexts and age groups were published as an edited volume (Berman & Slobin, 1994).

Table 1

Parameters Indicating the Amount of Data Contained within the Various CHILDES Corpora Grouped by Developmental Status (Typical vs. Non-Typical), Language Context (Monolingual vs. Bi-/Multilingual), and Elicitation Method (Observational vs. Narrative)

	Corpus Type				
	non-typical, monolingual, observational	typical, monolingual, observational	typical, monolingual, narrative	typical, bi-/multilingual, observational	typical, bi-/multilingual, narrative
# of languages	9	47	14	16	7
# of children	2,346	9,385	3,499	374	776
# of corpora	51	304	32	7	41
transcript only	30 (58.8%)	129 (42.4%)	19 (59.4%)	5 (71.4%)	16 (39.0%)
audio available	16 (31.4%)	134 (44.1%)	13 (40.6%)	2 (28.6%)	18 (43.9%)
video available	5 (9.8%)	39 (12.8%)	0 (0%)	0 (0%)	7 (17.1%)
# of words	4,978,779	60,929,540	1,393,252	6,154,120	492,791
children	1,378,320	18,847,945	994,650	1,793,347	418,875
	(27.7%)	(30.9%)	(71.4%)	(41.1%)	(85.0%)
caregivers	3,600,459	42,081,595	398,602	4,360,773	73,916
	(72.3%)	(69.1%)	(28.6%)	(58.9%)	(15.0%)
# utterances	1,499,551	16,334,646	252,214	1,712,189	107,096
children	538,411	6,242,607	165,702	599,661	79,990
	(35.9%)	(38.2%)	(65.7%)	(35.0%)	(74.7%)
caregivers	961,140	10,092,039	86,512	1,112,528	27,106
	(64.1%)	(61.8%)	(34.3%)	(65.0%)	(25.3%)

Note: *The single twin corpus combining data from typical and non-typical children is omitted from this table. Information about data type was unclear for two typical, monolingual, observational corpora, which are not included in the breakdown by media type.*

CHILDES also expanded to include language samples of children experiencing non-typical developmental trajectories due to autism (e.g., Bang & Nadig, 2015; Rollins, 1999), Down's

syndrome (e.g., Hooshyar, 1985; Rondal, 1978), Williams syndrome (Diez-Itza et al., 1998), epilepsy (e.g., Berl et al., 2005; Steinberg et al., 2013), delayed language development ("late talkers"; Moyle et al., 2007; Rescorla, 2011), developmental language disorder (formerly known as specific language impairment; e.g., Conti-Ramsden & Dykins, 1991; Paradis et al., 2013), prenatal exposure to cocaine, alcohol, and other substances (Malakoff et al., 1999), brain injury (Keefe et al., 1989), and hearing loss with/without cochlear implants (e.g., Ambrose, 2016; Szagun & Schramm, 2016). These clinical datasets represent various languages besides English, though relatively few include children growing up in bilingual or multilingual environments; see Tribushinina et al. (2017) for an exception (not included in Table 1). In the case of autism, datasets involving children acquiring Dutch, English, French, Greek, Mandarin, and Spanish are organized within a new ASDBank repository in TalkBank. In addition, TalkBank now includes FluencyBank—a repository of data from children and adults with fluency disorders (stuttering). Other repositories within TalkBank include datasets from adults with traumatic brain injury (TBIBank), right hemisphere damage to the brain (RHDBank), dementia (DementiaBank), and aphasia (AphasiaBank). Another repository is ClassBank, which contains transcripts of filmed interactions in a variety of classroom settings including, for example, basic geometry lessons with third graders (Lehrer & Curtis, 2000) and problem-based learning sessions with medical students (Loschmann & LeBaron, 2022). The entire TalkBank system uses CHAT format for transcription and CLAN programs for analysis, allowing child and adult corpora to be analyzed under a uniform framework and procedures.

CHILDES as a Crosslinguistic Repository

Despite growing appreciation of cross-cultural differences in childhood experiences, child-rearing practices, and child-directed speech, research efforts are still dominated by studies of children learning English (Kidd & Garcia, 2022). From its inception, CHILDES has been a major driving force behind the crosslinguistic expansion of language development research. To date, the monolingual and bilingual/multilingual CHILDES corpora encompass transcripts and recordings from 48 languages (for simplicity, we count Mandarin, Cantonese, and Taiwanese Hokkien as distinct languages). Figure 1 illustrates the rank-ordered frequency of languages by number of corpora, and Figure 2 shows the rank-ordered frequency of languages by number of children. While both indicators confirm an Anglocentric bias in existing research, they also show the success of collective efforts to diversify language development research, with progress most evident for Mandarin, Spanish, Dutch, French, German and Russian.

Vera Kempe, Patricia J. Brooks, Steven Gillis

Figure 1
Rank-Ordered Distribution of Languages by Number of Corpora in CHILDES
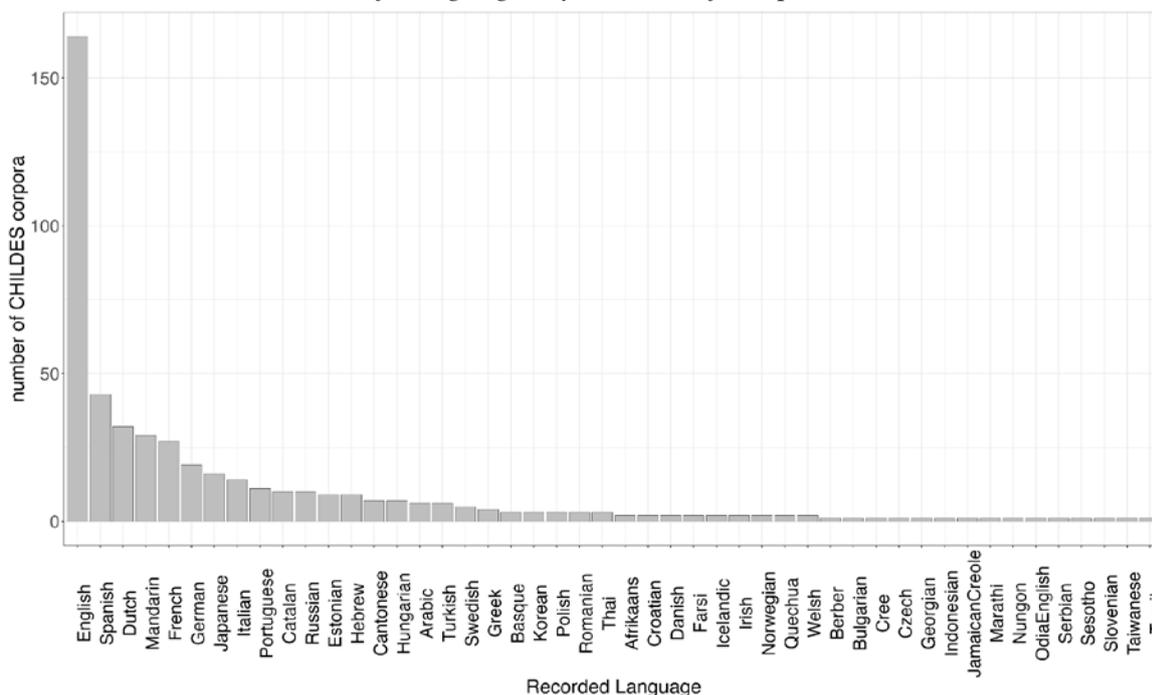

Figure 2
Rank-Ordered Distribution of Languages by Number of Children in CHILDES
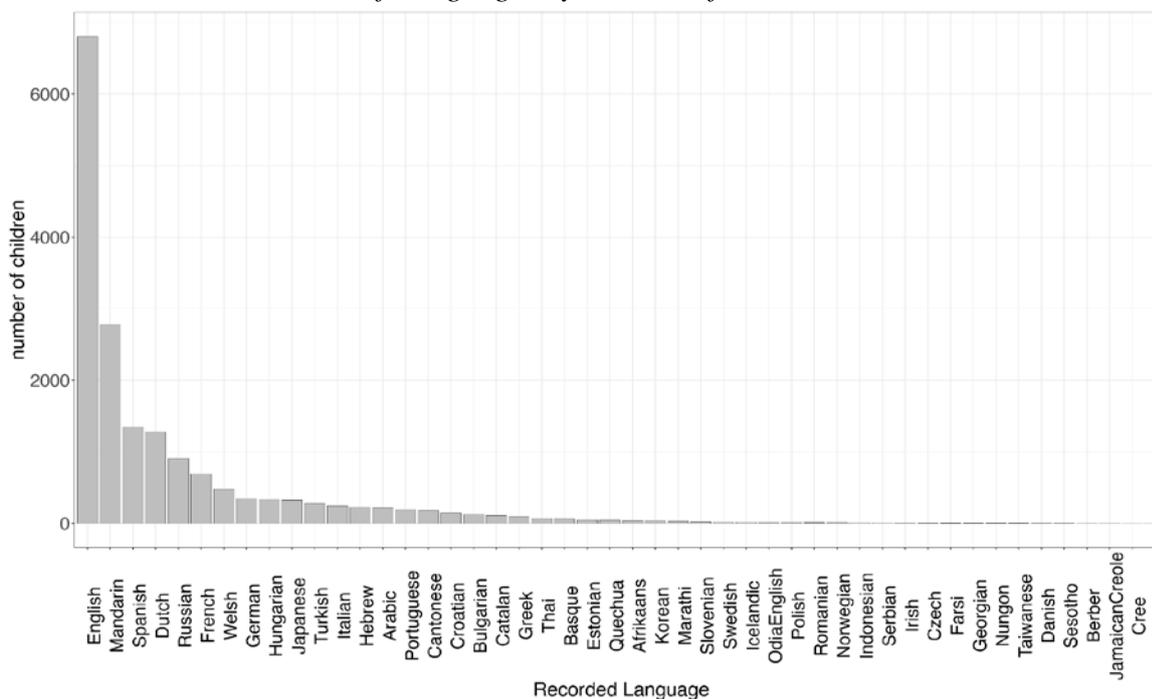

Figure 3 presents a network graph showing the frequency of the different language combinations (37 in total) represented in the bilingual and multilingual corpora. This visualization includes datasets involving sequential bilinguals, as in Guthrie's (1983) study of 14 Cantonese-speaking children learning English in school, as well as simultaneous bilinguals, as in Bailleul's (2017) case-study of a child learning Russian and French in accordance with

the one parent, one language approach (Grammont, 1902). Note that the language combinations refer to the child's language competence, as described in the documentation provided with each corpus, rather than the language of observation. As an example, the edge that connects "Various" to "English" refers to the Paradis (2005) corpus, which recorded productions of bilingual children speaking English as a second language. The dataset did not include recordings of children using their first languages: Arabic, Cantonese, Dari, Farsi, Japanese, Korean, Mandarin, Cantonese, Romanian, Spanish and Ukrainian. The network graph also disregards information about language dominance and includes datasets involving heritage speakers whose language dominance often shifts at school entry (e.g., Mai & Yip, 2017). As a summary of extant CHILDES corpora, the network graph indicates that most bilingual and multilingual language combinations include English, confirming its dominance in language development research. At the same time, it illustrates the growth of efforts to expand crosslinguistic studies of bilingual language development, as indicated by numerous language pairings that do not include English. Supplementing the CHILDES bilingual/multilingual corpora are BilingBank and SLABank within the TalkBank project, which provide mostly adult corpora for research on multilingualism and second language acquisition, respectively.

Figure 3
Language Combinations in the Bilingual and Multilingual Corpora
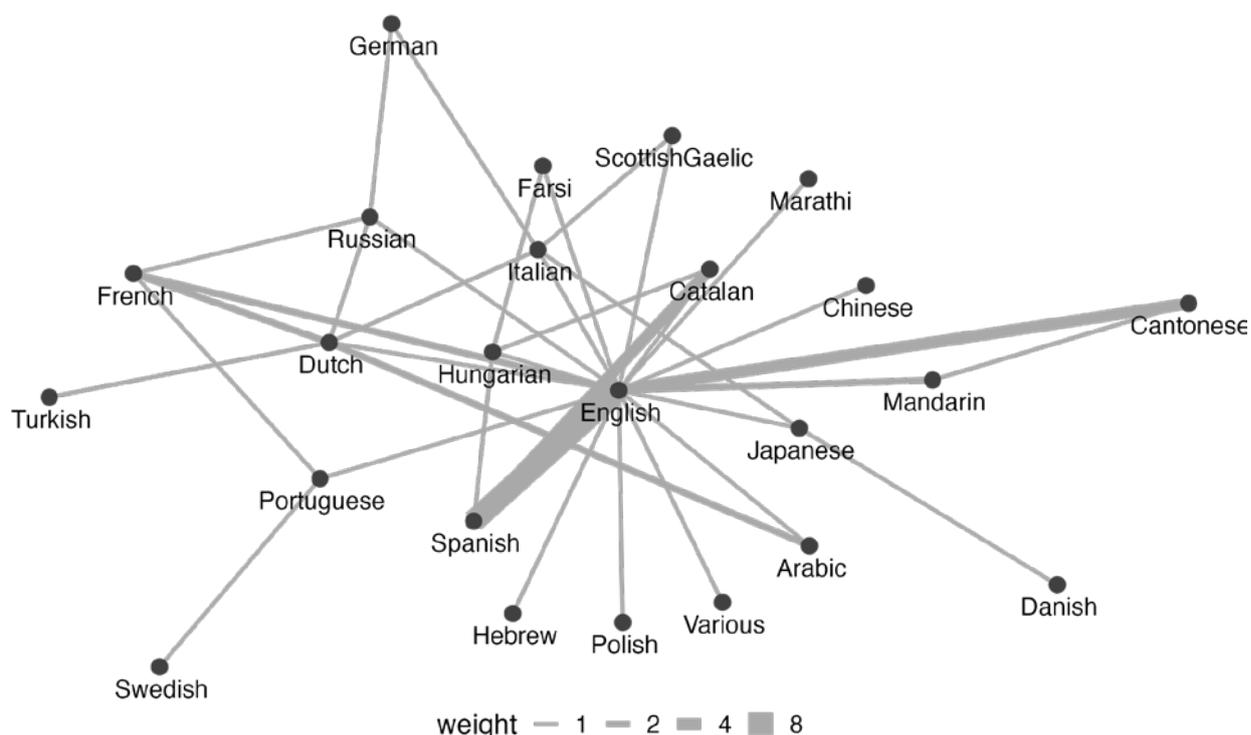
Note. Number of corpora is indicated by edge thickness.

CHILDES as a Data and Tool Sharing Platform

Judging from the more than 9000 published papers in which CHILDES corpora and/or CLAN programs were used (Liu et al., 2023), the CHILDES project has proven to be an invaluable database and tool-sharing resource. CHILDES corpora have been used to trace developmental trajectories of children's speech production (e.g., Marcus et al. 1992, Xu et al., 2023), to

describe features of the language input provided by caregivers (e.g., Cameron-Faulkner et al., 2003; Chouinard & Clark, 2003), to link these features to various language learning outcomes (e.g., Che et al., 2018; Ninio, 2014; Saxton, 2000), and as input to computational models testing theoretical assumptions about learning mechanisms (e.g., Macaulay & Christiansen, 2019; Monaghan & Christiansen, 2010). While this list of study aims is undoubtedly incomplete, it serves to illustrate how this rich data source has been used to advance understanding of child language development.

In this section, we will describe ways that researchers, instructors, students, and the interested public can interact with the CHILDES database and CLAN programs. CHILDES has a flexible interface allowing users to engage with corpora online via a standard internet browser or, alternatively, download corpora and CLAN programs to work offline. The browsable database is ideal for previewing datasets and for introducing students to CLAN because it does not require software installation; see Brooks et al. (2020) for sample lessons for using CHILDES in the classroom. Figure 4 is a screenshot of transcript 020718c from MacWhinney (1991) in the browsable database: https://sla.talkbank.org/TBB/childes/Eng-NA/MacWhinney/020718c.cha.

Figure 4

Screenshot of a Portion of MacWhinney (1991) Transcript 020718c as Viewed via the Browsable Database

The left column lists the files contained within the MacWhinney (1991) directory with the selected file (020718c) in bold. Below the file list on the left is a control panel for playing the audio file linked to the transcript, and a control panel for running CLAN commands. The right column provides documentation including the participants (Target_Child Ross at age 2 years, 7 months, 18 days; Father Brian). At the top of the transcript is an option to view the dependent tiers. The screenshot cuts off the transcript so that only the metadata are visible (i.e., the lines

at the top with the symbol @). Users can scroll through the entire transcript using the mouse or trackpad. In addition to browsing the transcript, users can annotate the transcript using the Collab button (top right corner). This button opens the Collaborative Commentary feature of TalkBank, allowing researchers and instructors to create groups of users who work together to code transcripts for recurring features, such as children's two-word combinations or specific features of child-directed speech like diminutives or explicit correction.

Figure 5
Screenshot of a Portion of MacWhinney (1991) Transcript 020718c with CLAN Output of the "Freq" Command

Taking a closer look at the CLAN control panel, as shown in Figure 5 (bottom left), one sees the *freq* command for running an analysis on the Target_Child's utterances. In creating this example, we selected *freq* from a menu of available commands. Note that +t*CHI instructs CLAN to select only the child's utterances and 020718c.cha designates the file to use. Clicking on the *Run* button executes the *freq* command in CLAN, generating a list of all the words Ross produced, the number of times he produced them, and summary statistics for word types, word tokens, and type-token ratio. Figure 5 (right column) provides a screenshot of the summary table at the bottom of the *freq* output, indicating that Ross produced 90 different words (types) and 177 words in total (tokens).

While the browsable database may be perfect for teaching purposes and initial scanning of datasets, it is less suitable for research because the interface does not save any CLAN output. Consequently, researchers interested in using the CHILDES corpora and CLAN programs should install CLAN on their computer. Once the software is installed, the CLAN editor may be used to prepare new transcripts in CHAT format with direct links to the original audio or video files. This functionality allows researchers to run CLAN analyses on datasets that are not yet part of CHILDES. As an example, Harvey and Brooks (2022) used CLAN to analyze digital text messages produced by American children enrolled in a Chinese language immersion program as an indicator of their second language (Mandarin) proficiency. Researchers can add dependent tiers to CHAT formatted datafiles, then run CLAN on the dependent tier (rather than on the main tier) to analyze the codes. For instance, Aldrich et al. (2011) coded children's use

of psychological state terms (e.g., *thinking, heard, looked, scared*) and explanations of psychological states in their narratives of another frog story (Mayer & Mayer, 1975), then used the *freq* command to tally the codes for analysis. To streamline data processing, the researcher can use wildcards to instruct CLAN to analyze all transcripts in a given directory as a batch and save the output to files. The output can be assembled into spreadsheets or via Python script. To further improve this process, Sanchez and colleagues (2019) developed childes-db—a mirror of the original CHILDES database that restructures datafiles for statistical analysis in the R programming environment (R Core Team, 2021), a dedicated Python library for accessing the database, and a *childesr* package to replicate some of the functionality of CLAN.

It has no doubt been a massive undertaking to maintain the compatibility of the CHILDES data interface and CLAN programs across a wide range of computer operating systems (e.g., Windows, MacOS, Unix) and an evolving range of personal computing devices (e.g., tablets, smartphones) over the past four decades. While dealing with a myriad of technological challenges, MacWhinney and his team made continuous improvements to expand software functionality and integration with other applications (e.g., Praat, ELAN). MacWhinney has also been the driving force in community-building efforts. He created (and moderates) the *infochildes* listserv, which unites the community of language development researchers around information exchange, and the *chibolts* listserv, which provides tips and advice for data transcription, coding, and analysis. Using *chibolts* also gives users rapid access to technical support from CLAN software developer Leonid Spektor and Brian MacWhinney himself. Users can subscribe to these and other Google Groups through TalkBank (https://www.talkbank.org/share/email.html).

Conclusion

In organizing the CHILDES and TalkBank projects, MacWhinney created a high standard for open access to datasets and analytic tools that was well ahead of its time. It is important to underscore that the corpora were donated by researchers from around the world and were not collected as part of a coordinated endeavor. Yet, through MacWhinney's efforts, all the datasets are now accessible in a unified format just a few clicks away on an Internet browser. Further, given its XML compatibility, CHILDES datafiles can be read easily by many different programs, allowing its integration with new tools for corpus analysis as they become available. As an example, ALIGN is an open-source Python package for measuring linguistic alignment (i.e., semantic, syntactic, or lexical overlap) across conversational turns, which can run on CHILDES corpora (Duran et al., 2019).

Future efforts to expand the CHILDES database as a fully open-access resource face formidable challenge. This includes finding ways of reconciling the different international standards in regulating the privacy and self-determination rights of participants (especially non-consenting children) with the data-sharing ethos of the Open Science movement. Already, as the database grows, corpora are being password protected to safeguard privacy, as was evident with the creation of HomeBank, a repository for daylong (longform) recordings of home language environments (VanDam et al., 2016), and other TalkBank projects. Another challenge involves reducing the effort required for accurate transcription of audio files—a notoriously labor-intensive, expensive, and time-consuming process (Gillis, 2014). For purposes of automatic speech recognition (ASR), MacWhinney and his colleagues have developed an

automated pipeline called Batchalign that converts raw audio into CHAT files (Liu et al., 2023). Batchalign has shown promising results in recognizing and transcribing adult speech with a level of accuracy ($> 95\%$) sufficient for a first-pass transcription, but the tool needs refinement to process speech from young children. A key challenge in the continued development of these tools involves accurately transcribing interactions characterized by significant crosstalk and conversational overlap across speakers.

From the start, CHILDES was conceived in an effort to increase crosslinguistic research on language development. Such efforts need to be given priority in light of the endangered status of many of the world's languages (Moseley, 2010). The call to diversify research is urgent as the ongoing loss of human languages places irrevocable limits on our understanding of the many paths to language acquisition. Finally, given evidence that language use for most individuals involves multiple varieties, encompassing accents, dialects, and languages (Evans, 2017), increasing representation of multilingualism and multidialectism within the database, especially of children with typical and non-typical developmental trajectories, will help to improve understanding of the breadth of human language development with theoretical, but also clinical and educational, implications.

ORCID

https://orcid.org/0000-0001-5649-4351

https://orcid.org/0000-0001-8030-8811

https://orcid.org/0000-0002-2067-0576

Rights and Permissions
Open Access

References
Adolph, K. (2016). Video as data: From transient behavior to tangible recording. *APS Observer, 29*(3), 23.

Aldrich, N., Tenenbaum, H., Brooks, P. J., Harrison, K. & Sines, J. Perspective taking in children's narratives about jealousy. *British Journal of Developmental Psychology, 29*(1), 86-109. https://doi.org/10.1348/026151010X533238

Ambrose, S. E. (2016). Gesture use in 14-month-old toddlers with hearing loss and their mothers' responses. *American Journal of Speech-Language Pathology, 25*, 519–531. https://doi.org/10.1044/2016_AJSLP-15-0098

Auer, E., Russel, A., Sloetjes, H., Wittenburg, P., Schreer, O., Masnieri, S., et al. (2010). ELAN as flexible annotation framework for sound and image processing detectors. In N. Calzolari, B. Maegaard, J. Mariani, J. Odjik, K. Choukri, S. Piperidis, et al. (Eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation* (LREC'10) (pp. 890-893). European Language Resources Association.

Bailleul, O. (2017). *Aspects psycholinguistiques du développement du bilinguisme précoce: Une étude de cas d'un enfant bilingue français-russe de 2 à 4 ans.* Unpublished PhD Dissertation. University of Rouen Normandy.

Bang, J., & Nadig, A. (2015). Language learning in autism: Maternal linguistic input contributes to later vocabulary. *Autism Research, 8*(2), 214 – 233. https://doi.org/10.1002/aur.1440

Berl, M. M., Balsamo, L. M., Xu, B., Moore, E. N., Weinstein, S. L., Conry, J. A., ... & Ritter, F. J. (2005). Seizure focus affects regional language networks assessed by fMRI. *Neurology, 65*(10), 1604-1611. https://doi.org/10.1212/01.wnl.0000184502.06647.28

Berman, R. A. (1990). Acquiring an (S)VO language: Subjectless sentences in children's Hebrew. *Linguistics, 28*, 1135-1166. https://doi.org/10.1515/ling.1990.28.6.1135

Berman, R. A., & Slobin, D. I. (Eds.). (1994). *Relating events in narrative: A crosslinguistic developmental study.* Lawrence Erlbaum Associates.

Bloom, L. (1970). *Language development: Form and function in emerging grammars.* MIT Press.

Boersma, P. & Weenick, D. (1995-2023). Praat: Doing phonetics by computer. https://www.fon.hum.uva.nl/praat/

Brooks, P. J., Brodsky, J. E., & Che, E. S. (2020). Using open-source data from OSF, ICPSR, and CHILDES to scaffold quantitative reasoning in psychology coursework. In T. M. Ober, E. S. Che, J. E. Brodsky, C. Raffaele, C. & P. J. Brooks (Eds.) *How We Teach Now: The GSTA Guide to Transformative Teaching* (pp. 329-348). Society for the Teaching of Psychology. http://teachpsych.org/ebooks/howweteachnow-transformative

Brown, R. (1973). *A first language: The early stages.* Harvard University Press.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science, 27*(6), 843-873. https://doi.org/10.1207/s15516709cog2706_2

Che, E. S., Brooks, P. J., Alarcon, M. F., Yannaco, F. D., & Donnelly, S. (2018). Assessing the impact of conversational overlap in content on child language growth. *Journal of Child Language, 45*(1), 72-96. https://doi.org/10.1017/S0305000917000083

Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language, 30*(3), 637-669. https://doi.org/10.1017/S0305000903005701

Conti-Ramsden, G., & Dykins, J. (1991). Mother–child interactions with language-impaired children and their siblings. *British Journal of Disorders of Communication, 26*, 337–354. https://doi.org/10.3109/13682829109012019

De Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational Linguistics*, *47*(2), 255-308. https://doi.org/10.1162/COLI_a_00402

Diez-Itza, E., Anton, A., Fernandez Toral, J., & Garcia, M. L. (1998). Language development in Spanish children with Williams syndrome. In A. Koc, E. Taylan, A.S. Ozsoy, & A. Kuntay (Eds.), *Perspectives in language acquisition* (pp. 309–324). Bogazici University Press.

Duran, N. D., Paxton, A., & Fusaroli, R. (2019). ALIGN: Analyzing linguistic interactions with generalizable techNiques—A Python library. *Psychological Methods, 24*(4), 419–438. https://doi.org/10.1037/met0000206

Elbers, L. (1985). A tip-of-the-tongue experience at age two? *Journal of Child Language, 12*, 353–365. https://doi.org/10.1017/S0305000900006474

Evans, N. (2017). Did language evolve in multilingual settings? *Biology & Philosophy, 3*2(6), 905-933. https://doi.org/10.1007/s10539-018-9609-3

Foster, E. D., & Deardorff, A. (2017). Open science framework (OSF). *Journal of the Medical Library Association, 105*(2), 203-206. https://doi.org/10.5195/jmla.2017.88

Gillis, S. (1984). De verwerving van talige referentie. Doctoral dissertation, University of Antwerp.

Gillis, S. (2014). Child language data exchange system. In P. J. Brooks & V. Kempe (Eds.) *Encyclopedia of language development* (pp. 74-78). Sage.

Gilmore, R. O., Adolph, K. E., Millman, D. S., & Gordon, A. (2016). Transforming education research through open video data sharing. *Advances in Engineering Education, 5*(2). https://eric.ed.gov/?id=EJ1106045

Grammont, M. (1902). Observations sur le langage des enfants. In *Mélanges linguistiques. Offerts à M. Antoine Meillet par ses élèves* (pp. 61-82). Klincksieck.

Guthrie, L. F. (1983). *Learning to use a new language: Language functions and use by first grade Chinese-Americans*. ARC Associates.

Gvozdev, A. N. (1948). *Usvoenie rebjonkomzvukovoj storony russkogo jazyka, Voprosy izučenija detskoj reči.* Detstvo-Press.

Gvozdev, A. N. (1949). *Formirovanie u rebjonka grammatičeskogo stroja russkogo jazyka, Voprosy izučenija detskoj reči.* Detstvo-Press.

Hall, W. S., Nagy, W. E., & Linn, R. (1984). *Spoken words: Effects of situation and social group on oral word usage and frequency.* Lawrence Erlbaum Associates.

Harvey, R. & Brooks, P. J. Effects of text messaging using digital pinyin input on literacy skills of elementary school Chinese immersion learners. *Language Teaching Research.* https://doi.org/10.1177/13621688221099909

Hooshyar, N. (1985). Language interaction between mothers and their nonhandicapped children, mothers and their Down children, and mothers and their language-impaired children. *International Journal of Rehabilitation Research, 4,* 475–477.

IASCL Child Language (2023, May 4). *Interview with Prof. Brian MacWhinney* [Video]. Retrieved from: https://www.youtube.com/watch?v=jt1Ed8okk7o&t=170s

Isaacs, S. (1930). *Intellectual growth in young children.* Schocken Books.

Keefe, K., Feldman, H., & Holland, A. (1989). Lexical learning and language abilities in preschoolers with perinatal brain damage. *Journal of Speech and Hearing Disorders, 54,* 395–402. https://doi.org/10.1044/jshd.5403.395

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language, 42*(6), 703-735. https://doi.org/10.1177/01427237211066405

King, G. (2007). An introduction to the dataverse network as an infrastructure for data sharing. *Sociological Methods & Research, 36*(2), 173-199. https://doi.org/10.1177/0049124107306660

Koschmann, T., & LeBaron, C. (2002). Learning articulation as interactional achievement: Studying the conversation of gesture. *Cognition and Instruction*, *20*, 249-282. https://doi.org/10.1207/S1532690XCI2002_4

Lehrer, R., & Curtis, C. L. (2000). Why are some solids perfect? Conjectures and experiments by third graders. *Teaching Children Mathematics, 6*(5), 324-329. https://doi.org/10.5951/TCM.6.5.0324

Linaza, J., Sebastián, M. E., & del Barrio, C. (1981). Lenguaje, comunicación y comprensión. Conferencia A nual de la Sección de Psicología del Desarrollo de la British Psychological Society. *Infancia y Aprendizaje, 4*(Sup1), 195-197.

Liu, H., MacWhinney, B., Fromm, D., & Lanzi, A. (2023). Automation of language sample analysis. *Journal of Speech, Language, and Hearing Research, 66*(7), 2421-2433. https://doi.org/10.1044/2023_JSLHR-22-00642

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review, 126*(1), 1–51. https://doi.org/10.1037/rev0000126

MacWhinney, B. (1974). *How Hungarian children learn to speak.* Unpublished doctoral dissertation, University of California, Berkeley.

MacWhinney, B. (1991). *The CHILDES project: Tools for analyzing talk.* Erlbaum.

MacWhinney, B. (2019). Understanding spoken language through TalkBank. *Behavior Research Methods, 51*, 1919-1927. https://doi.org/10.3758/s13428-018-1174-9

MacWhinney, B., & Fromm, D. (2022). Language sample analysis with TalkBank: An update and review. *Frontiers in Communication, 7,* 865498. https://doi.org/10.3389/fcomm.2022.865498

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language, 12*(2), 271-295. https://doi.org/10.1017/S0305000900006449

MacWhinney, B., & Snow, C. (1990). The child language data exchange system: An update. *Journal of Child Language, 17*(2), 457-472. https://doi.org/10.1017/S0305000900013866

Mai, Z. & Yip, V. (2017, December 18). Acquiring Chinese as a heritage language in English-speaking countries and the Child Heritage Chinese Corpus. Paper presented at the *International Conference on Bilingualism: Language and Heritage*. Chinese University of Hong Kong.

Malakoff, M. E., Mayes, L. C., Schottenfeld, R. S., & Howell, S. (1999). Language production of 24-month-old inner city children of cocaine-and-other-drug-using mothers. *Journal of Applied Developmental Psychology, 20*(1), 159-180. https://doi.org/10.1016/S0193-3973(99)80009-4

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, *57*(4), i–178. https://doi.org/10.2307/1166115

Mayer, M. (1969). *Frog, where are you?* Dial Books.

Mayer, M., & Mayer, M. (1975). *One frog too many.* Dial Books.

Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language, 37*(3), 545-564. https://doi.org/10.1017/S0305000909990511

Moseley, C. (Ed.) (2010). *Atlas of the world's languages in danger.* UNESCO.

Moyle, M. J., Ellis Weismer, S., Lindstrom, M., & Evans, J. (2007). Longitudinal relationships between lexical and grammatical development in typical and late talking children. *Journal of Speech, Language, and Hearing Research, 50*, 508-528. https://doi.org/10.1044/1092-4388(2007/035)

Nelson, K. (Ed.) (1989). *Narratives from the crib.* Harvard University Press.

Ninio, A. (2014). Learning a generative syntax from transparent syntactic atoms in the linguistic input. *Journal of Child Language, 41*(6), 1249-1275. https://doi.org/10.1017/S0305000913000470

Paradis, J. (2005). Grammatical morphology in children learning English as a second language: Implications of similarities with specific language impairment. *Language, Speech and Hearing Services in the Schools, 36*, 172-187. https://doi.org/10.1044/0161-1461(2005/019)

Paradis, J., Schneider, P., & Duncan, T. S. (2013). Discriminating children with language impairment among English-language learners from diverse first-language backgrounds. *Journal of Speech, Language, and Hearing Research, 56*(3), 971-981. https://doi.org/10.1044/1092-4388(2012/12-0050)

Pavlovitch, M. (1920). *Le langage enfantin: Acquisition du Serbe et du Français par un enfant Serbe.* Doctoral thesis, Paris Librairie Ancienne Honoré Champion.

Plunkett, K. (1986). Learning strategies in two Danish children's language development. *Scandinavian Journal of Psychology, 27*, 64–73. https://doi.org/10.1111/j.1467-9450.1986.tb01188.x

Rescorla, L. (2011). Late talkers: Do good predictors of outcome exist? Developmental *Disabilities Research Reviews, 17*(2), 141-150. https://doi.org/10.1002/ddrr.1108

R Core Team (2021). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna*, Austria. https://www.R-project.org/

Rollins P. R. (1999). Pragmatic accomplishments and vocabulary development in pre-school children with autism. *American Journal of Speech-Language Pathology, 8*, 85–94. https://doi.org/10.1044/1058-0360.0802.181

Rondal, J. (1978). Maternal speech to normal and Down's Syndrome children matched for mean length of utterance. In C. E. Meyers (Ed.), *Quality of life in severely and profoundly mentally retarded people: Research foundations for improvement* (pp. 193-265). American Association on Mental Deficiency.

Rose, Y., MacWhinney, B., Byrne, R., Hedlund, G., Maddocks, K., O'Brien, P., & Wareham, T. (2006). Introducing Phon: A software solution for the study of phonological acquisition. In D. Bamman, T. Magnitskaia & C. Zaller (Eds.), *Proceedings of the 30th Annual Boston University Conference on Language Development* (pp. 489-500). Cascadilla Press.

Rose, Y., & MacWhinney, B. (2013). The PhonBank project: Data and software-assisted methods for the study of phonology and phonological development. In J. Durand, U. Gut & G. Kristoffersen (Eds.), *Handbook of corpus phonology* (pp. 380-401). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199571932.013.023

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods, 51*, 1928-1941. https://doi.org/10.3758/s13428-018-1176-7

Saxton, M. (2000). Negative evidence and negative feedback: Immediate effects on the grammaticality of child speech. *First Language, 20*(60), 221-252. https://doi.org/10.1177/01427237000200600

Slobin, D. I. (1968). *Early grammatical development in several languages, with special attention to Soviet research.* Technical Report No. 11. Language-Behavior Research Laboratory, University of California, Berkeley.

Slobin, D. (1982). Universal and particular in the acquisition of language. In E. Wanner & L. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 128-172). Cambridge University Press.

Sokolov, J. & Snow, C. (Eds.) (1994). *Handbook of research in language development using CHILDES.* Lawrence Erlbaum Associates.

Spies, J. R. (2013). *The open science framework: Improving science by making it open and accessible.* Unpublished doctoral dissertation, University of Virginia.

Steinberg, M., Bernstein Ratner, N., Berl, M. & Gaillard, W. (2013). Fluency patterns in narratives from children with localization-related epilepsy. *Journal of Fluency Disorders, 38*(2) 193-205. https://doi.org/10.1016/j.jfludis.2013.01.003

Suppes, P., Smith, R., & Leveillé, M. (1973). The French syntax of a child's noun phrases. *Archives de Psychologie, 42*, 207–269.

Stern, C. & Stern, W. (1907). *Die Kindersprache. Eine psychologische und sprachtheoretische Untersuchung.* Barth.

Szagun, G. & Schramm, S. A. (2016). Sources of variability in language development of children with cochlear implants: Age at implantation, parental language, and early features of children's language construction. *Journal of Child Language, 43*, 505-536. https://doi.org/10.1017/S0305000915000641

Narasimhan, R. (2013). *Modelling language behaviour. S*pringer-Verlag.

Tribushinina, E., Mak, W. M., Andreiushina, E., Dubinkina, E., & Sanders, T. (2017). Connective use in the narratives of bilingual children and monolingual children with SLI. *Bilingualism: Language and Cognition, 20*(1), 98-113. https://doi.org/10.1017/S1366728915000577

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., & MacWhinney, B. (2016). HomeBank: An online repository of daylong child-centered audio recordings. *Seminars in Speech and Language, 37*(2), 128-142. https://doi.org/10.1055/s-0036-1580745

Volterra, V. (1972). Prime fasi di sviluppo della negazione nel linguaggio infantile. *Archivio di Psicologia, Neurologia e Psichiatria, 33*, 16–53.

Wagner, K. R. (1985). How much do children say in a day? *Journal of Child Language, 12*, 475–487. https://doi.org/10.1017/S0305000900006565

Weist, R. M., Wysocka, H., Witkowska-Stadnik, K., Buczowska, E., & Konieczna, E. (1984). The defective tense hypothesis: On the emergence of tense and aspect in child Polish. *Journal of Child Language, 11*(2), 347-374. https://doi.org/10.1017/S030500090000581X

Xu, Q., Chodorow, M., & Valian, V. (2023). How infants' utterances grow: A probabilistic account of early language development. *Cognition, 230*, 105275. https://doi.org/10.1016/j.cognition.2022.105275