A Database of Multilingual Child Speech with Recordings from a Longitudinal Project for Multilingual Education

Paola Escudero[1], *Gloria Pino Escobar*[1], *Milena Hernández Gallego*[1], *Chloé Diskin-Holdaway*[2] and *John Hajek*[2]

[1] MARCS Institute for Brain, Behaviour and Development, Western Sydney University, New South Wales, Australia, [2]School of Languages and Linguistics, The University of Melbourne, Victoria, Australia

paola.escudero@westernsydney.edu.au, g.pinoescobar@westernsydney.edu.au, m.hernandezgallego@westernsydney.edu.au, chloe.diskinholdaway@unimelb.edu.au, j.hajek@unimelb.edu.au

Abstract

We introduce a project focused on developing a multilingual database of Australian children's English and Heritage Language (HL) speech, starting with Spanish and adding other HLs progressively. Leveraging the technology and expertise from the AusKidTalk corpus [1], our database currently comprises approximately 610 hours of speech from children aged 3-7 years who speak English only or English and Spanish. Data were collected through online testing sessions featuring eight psycholinguistic tasks designed to elicit both single-word, sentence, and short story productions. This paper outlines the key features, design, data collection and analysis method, as well as the repository storage for data management. The aim is to facilitate linguistic research on language development in monolingual and multilingual Australian children. In this paper, we showcase the database, discuss the analyses conducted so far, and outline projects that have already used the database as well as future related projects. Additionally, we will detail our current data management plan and share our vision for collaborating with the broader research community.

Index Terms: corpus linguistics, speech corpus, children's speech, Spanish-English bilinguals, child bilinguals.

1. Introduction

The understanding, modeling, and conceptualizing of language acquisition and development relies heavily on language datasets and corpora [2, 3]. Despite interest in the analysis of natural language use increasing in the last decade (see studies listed in [4]), there is still a shortage of adult and child speech corpora worldwide, including in Australia.

Child speech corpora are especially difficult to find, with currently fewer than 20 worldwide (exceptions include CHILDES and AusKidTalk [1, 5]). This may be attributed to the fact that child speech, particularly that of young children, is relatively difficult to collect and analyse [1]. Their scarcity notwithstanding, child corpora offer numerous advantages for the study of child bilingualism (see [2] for a more detailed description). For instance, the CHILDES corpus has facilitated more than 1,300 studies on language disorders, second language acquisition, literacy, among other disciplines [5].

Within the existing speech corpora, spoken ones remain a minority compared to written ones [6], with many more monolingual compared to multilingual corpora. Exceptions include the CHILDES Russian-German ZAS-MAIN Corpus [7] and the Leibniz-ZAS corpus of MAIN [8]. Importantly, the unpredictability of spontaneous child speech has led most child corpora to employ constrained protocols with limited tasks, and few are fully transcribed and annotated [9].

The current database aims at including fully annotated and transcribed child speech data resulting from recording sessions. It features eight different tasks, which could help the research community to answer a variety of linguistic and psycholinguistic empirical questions. The recordings of these psycholinguistic tasks are part of the data collected within a large longitudinal research program which aims at fostering multilingualism through play-based language immersion sessions in preschools [10, 11]. This longitudinal research program also includes recordings of language exposure sessions, electronic forms used to evaluate children's performance in those sessions, with results reported in in [10]. Here we report on the speech database from the psycholinguistic tasks in English and Spanish due to space limitations.

Our child speech database is intended to make a significant contribution to the study of child bilingualism as the first multilingual corpus of child speech in English and heritage languages (HLs) in Australia. It aims at filling in the gap in child speech corpora mentioned above with detailed documentation of how children's languages other than English (LOTEs) develop, which are learned alongside this societal language. This database is an invaluable tool to conduct research on children's linguistic, cognitive, and socio-emotional development, including collaboration among researchers with multidisciplinary expertise. The main characteristics of the data collection design and analysis which have been used in the first five articles written using these database [10-14] are as follows:

1. Online data collection for Australian preschoolers in English and Heritage Languages,
2. Data de-identification, transcription of recordings, curation and connection to linguistic background and demographic information,
3. Ongoing transcription of all collected data using available computational tools that enable speech-to-text for orthographic and phonemic segmentation of spoken utterances.
4. Availability of data to other researchers by writing to the first author and data custodian.

Below we explain the specific features of the corpus data collection, including participants and testing protocol, which we hope to continue as more data is collected. The goal is to streamline both collection and transcription as data entries are curated using the same effective and efficient procedure.

2. Method

The part of the database reported here includes child speech that allows the evaluation of multiple linguistic features in a single, child-friendly recording session. Different utterance types, e.g., single words, sentences, and narratives, were elicited to facilitate research into monolingual and bilingual language production and development. Crucially, demographic metadata that were collected for each participant have been linked to the recordings for essential understanding of each child's unique language experience.

The eight tasks were conducted online in a single Zoom session which lasted for approximately 45 minutes (session duration range: 45-60 mins), reducing the likelihood of participant dropout. The recording procedure enabled easier participation than in most previous infant and child experiments that take place in laboratories, as families are unlikely to readily travel long distances to university-based laboratories. Importantly, the suitability and reliability of using audio recordings made with Zoom for linguistic studies has been confirmed by [15-17]. Tasks were presented to parents and children as games and activities with the aim of eliciting different types of speech samples while maintaining children's engagement during the session (11-15).

Parents were asked to complete a Qualtrics online survey which included important demographic and sociolinguistic information, which is indispensable for linguistic analysis because failure to connect data to metadata would lead to "nothing but disconnected words of unknowable provenance or authenticity" [18]. We thus ensured that all data entries were correctly indexed with their corresponding child speech data (point 5), and extra data checks were conducted for verification.

2.1 Participants

As shown in Table 1, the current data set includes sessions from 64 children who resided in Australia, with no diagnosis of language or developmental disorder (M age = 4.60 years, range 3-7). They were recruited from a database of parents who had volunteered to participate in research at an Australian university laboratory (n = 27) and from a bilingual preschool located in Sydney, Australia, where parents had volunteered to participate in the larger longitudinal program mentioned above [10] for HL maintenance and additional language (AL) learning (n = 37). Of the participants, 18 were English monolinguals, 39 were Spanish-English bilinguals, and 7 were bilinguals of English and another language. The Spanish-English bilingual group was further divided into HL simultaneous bilinguals (n = 19), who had acquired Spanish at home, and children who had acquired Spanish as an AL in their childcare setting (n = 20).

Following the recruitment procedure from previous studies [11,15], parents received an initial email with a link to the study's information sheet, consent form, and demographic survey, all hosted on Qualtrics (https://www.qualtrics.com). Participation was voluntary, with parents providing written consent and children giving oral assent prior to their participation. Parents completed the demographic questionnaire after consenting online. The survey included questions about the child's language background, such as weekly language exposure and use of English, Spanish, and any additional languages, along with their daily routines. Caregivers also reported their own language background, proficiency, and language use with their child. Parents then received a follow-up email with instructions to schedule their child's online session via Calendly (https://calendly.com/). The study was approved by the Western Sydney University Human Research Ethics Committee (approval number: H11022), with the first author serving as the data custodian. All data was anonymized to protect children's identities.

122 files with a duration of 45-60 minutes each have been organized and prepared for transcription to date. There are currently 67 recordings in English and 55 recordings in Spanish. The demographic survey shows that all the multilingual children were acquiring two languages, and some three, as shown in a recent study reporting on one of the psycholinguistic tasks with the same participants [11].

Table 1: *Parental report in the demographic survey for the participants included in the corpus (n=64)*

	English mono-lingual	Other bilingual	Spanish-English HL	Spanish as an AL
N	18	7	19	20
Mean age (range)	4.4 (3-5)	4.4 (3-5)	4.7 (3-7)	4.8 (4-6)
Mean English exposure % (range)	97.40 (70.7-100)	45.13 (5.2 -65.0)	42.9 (14-80)	75.8 (45-100)
Mean 1st LOTE exposure % (range)	2.4 (0-29.0)	47.43 (14.0-95)	52.9 (20-86)	21.8 (0-52)
Median Principal Carer Relation	Mother (n=14)	Mother (n=6)	Mother (n=18)	Mother (n=19)
Principal Carer Education	Universi-ty degree 85%	Universi-ty degree 85%	Universi-ty degree 85%	University degree 85%

The research project was designed as a pre- and post-intervention study, and the children's proficiency was measured both in Spanish and English, over three timepoints: two prior (2021) and one posterior (2022), to the delivery of Spanish language immersion sessions reported in [10], as the children transitioned from preschool to primary school. Given that some participants did not complete all pre and post sessions, some researchers may consider the data pseudo-longitudinal (cf. Corpus of Learner German (CLEG13) [2]) and cross-sectional.

The current dataset includes 92 recordings resulting from the first timepoint of data collection (45 in Spanish and 47 in English). 22 participants were tested at both the first and the second timepoint (15 in Spanish and 7 in English). Many participants have continued their participation in the longitudinal project after graduating from preschool, with data

from these subsequent sessions awaiting organisation and connection to their correspondent demographic information. The dataset mainly includes Spanish and English sessions from the first session, as most participants were only tested in Spanish because their English proficiency was comparable to that of monolingual peers [11]. Additionally, the aim of the longitudinal project was to investigate the development of HL proficiency, due to the demonstrated difficulty in maintaining HL proficiency in Australia [19].

2.2 Online data collection protocol

Recordings were conducted online following the online testing procedures described in [11, 15]. The sessions took place via Zoom and a link was sent to parents prior to the session for use with a home device (preferably a laptop or desktop), with an experimenter guiding the children through eight tasks assisted by one of the child's parents for session setup [15]. The tasks included an initial exposure to an audiovisual eBook, followed by a retelling and comprehension task [11-14], a digit span task [20], a nonword repetition task [21], two verbal fluency tasks [22], a receptive vocabulary task [23], and an episodic memory task [24], in that order. Each session was recorded for response data reliability [15, 16] and for speech database building.

The audiovisual eBook was a 12-page electronic storybook featuring colourful 2D line drawings, which was narrated by a female native speaker of Australian English. The story depicted two children sharing fruit and toys at school. After exposure to the eBook, children completed a retelling task where they recounted the story in their own words [23, 24], followed by a comprehension task that involved answering questions with visual aids [11, 14, 25, 26]. The retelling task was always presented first to avoid bias from comprehension questions, as studies have shown that comprehension does not influence retelling accuracy [27, 28]. The digit span task measured short-term verbal memory with children asked to recall sequences of numbers that increased in length [20]. The nonword repetition task measured phonological memory as the ability to store and reproduce unfamiliar sound sequences in the target language [21]. The verbal fluency tasks assessed lexical access and cognition by asking participants to retrieve words from either a specific category or beginning with a particular sound, within a 1-minute time limit [22]. In the receptive vocabulary task, children listened to a word and identified its corresponding image, demonstrating their understanding of the word's meaning [23]. Finally, in the episodic memory task, children recalled a sequence of actions previously presented, such as the steps involved in celebrating a birthday [24]. These eight tasks provided speech data in the target language.

The present corpus did not involve additional specialised recording apparatus as the data was obtained via online Zoom recordings following previous studies [11, 15]. Using the Zoom subscription that most universities have, our online procedure represents a time and cost-efficient option for large-scale longitudinal projects such as the one in [11].

3. Speech data transcription, analysis, organisation, and storage

So far, we have partially transcribed and annotated the current dataset, using the following procedure: the audio recordings sampled at 16kHz or above were first extracted as WAV files from Zoom sessions run on desktop/laptop computers. These files underwent orthographical transcription at the utterance level using OpenAI's Whisper [29] an automatic speech recognition (ASR) system that is trained on approximately 680,000 hours of diverse, multilingual data sourced from the internet. Transcription can be done much more quickly and with fewer human resources when ASR tools such as Whisper are used.

The text transcriptions, linguistic variable labelling, and time information were then converted into Praat TextGrid files using MATLAB scripts. Research assistants manually verified the correctness of these transcriptions. Subsequently, the audio and transcription data underwent forced alignment using WebMAUS [30] with the Australian English model, to implement annotation at both the word and phoneme levels. The research team manually verified these boundaries and annotations, adhering primarily to the automated suggestions unless errors were evident. Aside from time efficiency. compared with manual transcription, Whisper ASR followed by manual verification substantially decreases the human resources required for transcription. OpenAI technology to transcribe the recorded data thus facilitated automated pre-processing and segmentation, and reduced the time required for manual phoneme-level annotation. Units of speech such as words, utterances, and errors at the word level were further annotated for phonetic, phonological, semantic, and morphological analysis [31].

For phonetic analysis, research assistants also labelled segments and syllables depending on the specific topic, e.g., stressed and unstressed syllables [13] or types of voiceless plosives [14]. All scripts and tables generated from this segmentation procedure are part of the documentation and metadata of the database with fully de-identified data that can be made available to researchers via the project's principal investigator and data custodian (first author of this paper).

All files were then named including date of collection, length of the file, participant type, ID number, and age in the file name, following the standard file naming system for the Talkbank and CHILDES (Child Language Data Exchange System) projects, called the Codes for the Human Analysis of Transcripts (CHAT) [32] and indexed in a Microsoft Excel spreadsheet. All data have been stored along with their corresponding metadata in a WSU online data repository, with full access for the current research team, and with future access available through the principal investigator and data custodian. Additionally, this data repository ensures that sensitive data are protected by security safeguards against loss, unauthorised access, misuse, or disclosure. The five articles and any new articles using parts of the dataset will be connected to the data repository.

Our transcription procedure has been followed for the already reported studies using this dataset [11-14]. We will finish transcription and annotation of the recordings for the 64 participants and 122 files following the above-mentioned procedure and included in the current corpus. Further data collected within the longitudinal study will follow the same procedure. As was mentioned in a previous presentation of the database, it includes data collected in language exposure sessions for HL and AL children that have been delivered in French, Mandarin, Spanish, and Vietnamese [33], which are part of the longitudinal research program [10]. Additionally, ongoing research will report on psycholinguistic data including datasheets with participant scores from the Spanish retelling and comprehension tasks, the memory, verbal fluency, and sequential memory tasks, along with the already included

dataset with the results of the English retelling and comprehension task reported in [11].

4. Discussion and Conclusions

This project makes a significant contribution to the study of child bilingualism and language acquisition by developing a multilingual and multimodal database of Australian children's English and Spanish speech. This database has already allowed for a broad range of research topics across multiple linguistic domains, such as phonetics, phonology, lexical acquisition, and narrative discourse, including cross-linguistic comparisons. The variety of tasks included, such as story retelling, auditory and phonological memory, and verbal fluency, among others, ensures that researchers can explore different aspects of language development, from sentence structure to vocabulary acquisition. By incorporating demographic metadata, the database offers a robust contextual framework for examining individual differences in bilingual language acquisition. This comprehensive approach enables a deeper understanding of how factors such as age, language exposure, and social context shape linguistic development, with opportunities for other researchers to harness this important resource.

The use of online experimental sessions and modern transcription technologies, such as OpenAI, highlights the efficiency and innovation of the project. By providing de-identified, fully transcribed and annotated data, the database saves researchers time and reduces transcription biases, allowing for more accurate analyses from diverse theoretical perspectives. The project's commitment to transparency and accessibility through the data custodian sets a strong precedent for future linguistic research, promoting collaboration and open science practices, while at the same time protecting participants' anonymity.

In addition to its focus on child bilingualism, this corpus contributes to the broader research agenda on LOTEs in Australia and beyond. It provides valuable insights for educational strategies and language learning programs, and it holds potential for informing clinical interventions for children with language disorders. The detailed speech data can also help improve automatic speech recognition (ASR) systems, particularly in recognizing speech patterns in spontaneous conversational settings in young multilingual populations, thus addressing a key technological challenge [1, 34].

5. Acknowledgments

This study was supported by three grants from the Australian Research Council awarded to the first author (CE140100041, FT160100514) and to the first and two last authors (LP210300631). Special thanks go to Dr. Weicong Li for his assistance with automatic transcription and initial and ongoing database organization, and to our dedicated team of research assistants for their invaluable help with data collection and the manual refinement of automatic transcription. We extend our sincere thanks to the children who participated in this study and to their parents for providing consent.

6. References

[1] B. Ahmed et al., "AusKidTalk: an auditory-visual corpus of 3-to 12-year-old Australian children's speech", Proceedings of the 22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, 3680-3684, 2021,

[2] S. Granger, G. Gilquin, and F. Meunier, The Cambridge Handbook of Learner Corpus Research. Cambridge University Press, 2015.

[3] B. MacWhinney, "Understanding spoken language through TalkBank," Behavior Research Methods, vol. 51, no. 4, pp. 1919–1927, Dec. 2018.

[4] M. Planelles Almeida, J.A. Duñabeitia and A. Doquin de Saint Preux (2022). "The VIDAS Data Set: A Spoken Corpus of Migrant and Refugee Spanish Learners". Frontiers in Psychology, vol. 12, Jan. 2022.

[5] B. MacWhinney, The Childes Project. Psychology Press, 2014

[6] J. Fernández and T.R. Davis, "Overview of Available Learner Corpora" in The Routledge Handbook of Second Language Acquisition and Corpora, N. Tracy-Ventura and M. Paquot,1st Ed, New York, U.S. Routledge, 2021, Country: Abbrev. of Publisher, year, pp. 147-159.

[7] N. Gagarina, "Narratives of Russian–German preschool and primary school bilinguals: Rasskaz and Erzaehlung," Applied Psycholinguistics, vol. 37, no. 1, pp. 91–122, Dec. 2015

[8] N. Topaj, Z. Rizaeva, A. Sternharzy N. Gagarina, "Leibniz-ZAS corpus of MAIN". Zenodo, abr. 28, 2021.

[9] N. F. Chen, R. Tong, D. Wee, P. Lee, B. Ma, and H. Li, "SingaKids-Mandarin: Speech Corpus of Singaporean Children Speaking Mandarin Chinese," In Proceedings of Interspeech 8 Sep. 2016.

[10] P. Escudero, G. Pino Escobar, C. Diskin-Holdaway, and J. Hajek. "Enhancing Heritage and Additional Language Learning in the Preschool Years: Longitudinal Implementation of the Little Multilingual Minds Program." OSF Preprints. September 16, 2024. https://doi.org/10.31219/osf.io/rvjc

[11] G. Pino Escobar and P. Escudero. "Vocabulary, Comprehension and Retelling in Multilingual Children: Age and Input Tell the Whole Story." OSF Preprints. September 16, 2024. https://doi.org/10.31219/osf.io/8e4nf

[12] M. Hernández Gallego, G. Pino Escobar, P. Escudero. "English lexical productivity and diversity in Spanish-English bilingual children in Australia." Proceedings of the 19th Australasian International Conference on Speech Science and Technology, SST 2024.

[13] Escudero, P., Li, W & Diskin-Holdaway, C. Have four-year-olds mastered vowel reduction in English? An acoustic analysis of bilingual and monolingual child storytelling. Proceedings of the 19th Australasian International Conference on Speech Science and Technology, SST 2024.

[14] Diskin-Holdaway, C., Li, W & Escudero, P. Bilingual preschoolers' phonetic variation keeps up with monolingual peers: The case of voiceless plosives in Australian English. Proceedings of the 19th Australasian International Conference on Speech Science and Technology, SST 2024.

[15] P. Escudero, G. Pino Escobar, C. G. Casey & K. Sommer, "Four-year-old's online versus face-to-face word learning via eBooks", Frontiers in Psychology, 12, 450, 2021.

[16] C. Zhang, K. Jepson, G. Lohfink, & A. Arvaniti, A, "Speech data collection at a distance: Comparing the reliability of acoustic cues across homemade recordings", The Journal of the Acoustical Society of America, 148(4_Supplement), 2717-2717, 2020.

[17] C. Ge, Y. Xiong, & P. Mok, "How Reliable Are Phonetic Data Collected Remotely? Comparison of Recording Devices and Environments on Acoustic Measurements", Proceedings of Interspeech, pp. 3984-3988, 2021.

[18] L. Burnard, "Metadata for corpus work", in M. Wynne, Developing Linguistic Corpora: A Guide to Good Practice. Oxford: Oxbow Books, pp. 30–46, 2005.

[19] P. Escudero, C. Jones Diaz, J. Hajek, G. Wigglesworth, and E. A. Smit, "Probability of heritage language use at a supportive early childhood setting in Australia," *Frontiers in Education,* vol. 5, p. 93, 2020, doi: 10.3389/feduc.2020.00093.

[20] S. E. Gathercole, "The assessment of phonological memory skills in preschool children," British Journal of Educational Psychology, vol. 65, no. 2, pp. 155-164, 1995.

[21] S. E. Gathercole, “Nonword repetition and word learning: The nature of the relationship,” Applied Psycholinguistics, vol. 27, no. 4, pp. 513-543, 2006.
[22] M. Regard, E. Strauss, and P. Knapp, “Children’s production on verbal and non-verbal fluency tasks,” Perceptual and Motor Skills, vol. 55, no. 3, pp. 839-844, 1982.
[23] S. Weintraub, S. S. Dikmen, R. K. Heaton, D. S. Tulsky, P. D. Zelazo, P. J. Bauer, et al., "Cognition assessment using the NIH Toolbox," Neurology, vol. 80, no. 11 Supplement 3, pp. S54-S64, 2013.
[24] S. S. Dikmen, P. J. Bauer, S. Weintraub, D. Mungas, J. Slotkin, J. L. Beaumont, R. Gershon, N. R. Temkin, and R. K. Heaton, "Measuring episodic memory across the lifespan: NIH Toolbox Picture Sequence Memory Test," Journal of the International Neuropsychological Society, vol. 20, no. 6, pp. 611-619, 2014.
[25] J. Heilmann, J. F. Miller, A. Nockerts, and C. Dunaway, “Properties of the Narrative Scoring Scheme Using Narrative Retells in Young School-Age Children,” American Journal of Speech-Language Pathology, vol. 19, no. 2, pp. 154–166, May 2010, doi: https://doi.org/10.1044/1058-0360(2009/08-0024).
[26] N. Gagarina et al., “Assessment of Narrative Abilities in Bilingual Children,” Multilingual Matters eBooks, pp. 243–276, Dec. 2015.
[27] K. Kawar, E. Saiegh-Haddad, and S. Armon-Lotem, “Text complexity and variety factors in narrative retelling and narrative comprehension among Arabic-speaking preschool children,” First Language, p. 014272372211498, Feb. 2023.
[28] M. Silva and K. Cain, “The use of questions to scaffold narrative coherence and cohesion,” Journal of Research in Reading, vol. 42, no. 1, pp. 1–17, Oct. 2017.
[29] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey and I. Sutskever. Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning (pp. 28492-28518). PMLR, July, 2013.
[30] T. Kisler, U. Reichel, and F. Schiel, “Multilingual processing of speech via web services”, Computer Speech & Language, vol. 45, pp. 326–347, 2017.
[31] B. MacWhinney, “The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs”. Psychology Press, 2014.
[32] K. E. Squires, M. J. Lugo-Neris, E. D. Peña, L. M. Bedore, T. M. Bohman, and R. B. Gillam, “Story retelling by bilingual children with language impairments and typically developing controls,” International Journal of Language & Communication Disorders, vol. 49, no. 1, pp. 60–74, Aug. 2013.
[33] P. Escudero, G. Pino Escobar, M. Hernandez Gallego, C. Diskin-Holdaway & J. Hajek, “The Little Multilingual Minds corpus: Educators’ and children’s speech in heritage languages”, Workshop on community language corpora in Australia, Australian National University, Canberra, 2023.
[34] S.-Y. Yoon, L. Chen, and K. Zechner, “Predicting word accuracy for the automatic speech recognition of non-native speech,” Interspeech 2010, Sep. 2010.