AusKidTalk: Developing Transcription Guidelines for Continuous Australian English Child Speech

Tünde Szalay[*†], **Zheng Nan**[†], **Renata Huang**[*‡], **Mostafa Shahin**[†], **Sirojan Tharmakulasingam**[†], **Kirrie Ballard**[*], **Beena Ahmed**[†]
[*]The University of Sydney, [†]University of New South Wales, [‡]Macquarie University
Sydney, Australia
{tuende.szalay, kirrie.ballard}@sydney.edu.au, renata.huang@mq.edu.au
{zheng.nan, m.shahin, s.tharmakulasingam, beena.ahmed}@unsw.edu.au

Abstract

Guidelines are required for accurate and consistent transcription of speech corpora, especially when they contain more challenging, e.g. spontaneous or under-resourced speech. This paper presents a workflow and guidelines for transcribing spontaneous and under-resourced child speech in AusKidTalk, the first Australian English child corpus. Speech samples were elicited using a story-telling task and are 3.5 minutes long per child on average. Orthographic transcriptions were generated using automatic speech recognition (ASR) tools and corrected manually. A novel hand-correction protocol consisting of guidelines, hand-correction interface, and ground truth transcriptions together with consistency metrics were developed. Nine annotators submitted hand-corrections for 261 children's story-telling task, and 25 ground truth tasks. Manual correction was 11-fold of speech time with a 3.5-minute-long story-telling task corrected in approximately 40 minutes. Efficiency is attributed to the quality of automatic transcription with 23% word error rate. Manual correction was accurate with annotators achieving consistent results on 15/25 ground truth submissions. Most inconsistent ground truth submissions were caused by a single, challenging ground truth task. These results show that our workflow yields efficient and accurate transcriptions, although transcriptions of potentially more challenging narrative tasks (e.g., elicited from younger children) might require further corrections.

Keywords: child speech corpus, corpus building, automatic speech recognition (ASR), ASR-assisted transcription

1. Introduction

Speech corpora, the large collections of transcribed and aligned audio recordings, are crucial resources in speech technology and science, required for training automatic speech recognition (ASR) tools and for revealing speech variation in large datasets (Liberman, 2019; Sobti et al., 2024). Developing accurate transcription, a key part of every speech corpora, is resource intensive (Glenn et al., 2010), therefore corpora are often developed using read speech, removing the need for orthographic transcription (e.g., Garofolo et al., 1993; Burnham et al., 2011; Panayotov et al., 2015; Kressner et al., 2024). Spontaneous speech, which provides more life-like samples, requires labour-intensive orthographic transcription prior to its use to train ASR models or conduct more detailed analysis (e.g., phonetic transcription), increasing annotation time and cost (e.g., Pitt et al., 2005; Grønnum, 2009; Schuppler et al., 2017; Mereu and Vietti, 2021).

ASR-assisted transcription augmented by manual correction is a known method to reduce transcription time (Bazillon et al., 2008; Cieri, 2009). For example, the approximate time-need for manual transcription with manual verification is 50× the length of speech time, while automatic transcription with manual verification is 15× the length of speech time depending on the level of detail in the transcription (Cieri, 2009). Yet, details of ASR tools and guidelines of hand-correction are not always reported in detail, introducing challenges for reliability and replicability of orthographic transcription (transcription guidelines not reported: Pitt et al. 2005; Grønnum 2009; Gao et al. 2012; Mereu and Vietti 2021, transcription guidelines reported: Schuppler et al. 2017; Pérez-Espinosa et al. 2020; Rumberg et al. 2022).

This paper addresses this gap by providing guidelines for systematic integration of manual correction with automatically generated orthographic transcriptions, when transcribing semi-spontaneous, continuous child speech in AusKidTalk, the first Australian English (AusE) child corpus (Ahmed et al., 2021). This will enable future researchers to adapt the guidelines for other corpora containing spontaneous speech as well as the techniques to evaluate the quality of both human and automated annotations, offering practical methods to enhance annotation workflows.

1.1. Automatic corpus transcription

When building novel corpora for high resource languages, ASR tools reduce transcription time, and thus costs, by generating orthographic transcription automatically, allowing for manual correction instead of transcription (Bazillon et al., 2008;

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5794–5804
11–16 May 2026.

Glenn et al., 2010; Szalay et al., 2025). For French scripted speech, manual transcription time is reported as 9 times the length of audio duration and ASR-assisted as 4 times the length of audio duration (Bazillon et al., 2008). For French spontaneous speech, however, manual transcription time was 10 times and ASR-assisted was 9 times the speech time (Bazillon et al., 2008). Larger gains were reported for longer manual transcription times, with manual transcription times being as high as 25 to 50 times the speech time and ASR-assisted correction being 5 to 15 times (Cieri, 2009).

ASR transcription accuracy, however, is reduced when the training corpora differ from the new speech data on which ASR tools are applied (Shivakumar and Georgiou, 2020; Szalay et al., 2022b; Wassink et al., 2022; Sobti et al., 2024). ASR models trained on General American English perform less well on the marginalised ethnolects of African American and Native American Englishes, as well as on Australian English, the de-facto official language of Australia (Wassink et al., 2022; Szalay et al., 2022b). ASR systems trained on adult speech perform 2–5 times worse on children's speech relative to adults' due to age-dependent anatomical, articulatory and language variations (Shivakumar and Georgiou, 2020).

These shortcomings have demonstrated negative impact on the use of ASR in corpus transcription (Gorisch et al., 2020; Mereu and Vietti, 2021; Liu et al., 2023; Kempe et al., 2024; Szalay et al., 2025). For example, Dialogic ItAlian, a corpus of Italian conversational speech, was transcribed using free ASR tools provided by YouTube then hand-corrected (Mereu and Vietti, 2021). However, the accuracy of automatic transcription is reduced on non-native speakers compared to native speakers of Italian (Mereu and Vietti, 2021). In AusKidTalk, a single word production task by AusE-speaking children was transcribed using the UNSW ASR tool, then hand-corrected (Szalay et al., 2025; Shahin et al., 2020). Comparison of the automatic and hand-corrected transcriptions indicated that automatic transcription was less accurate for younger children (Szalay et al., 2025). Similarly, Batchalign, the custom-built ASR tool for annotating data in the Child Language Data Exchange System (CHILDES) corpus shows reduced performance on children relative to adults, in particular for younger children and children with mild to moderate speech disorders (Liu et al., 2023; Kempe et al., 2024). In these corpora, automatic transcription was applied to reduce transcription time; however, ASR tools performed less accurately on foreign-accented and younger speakers (Mereu and Vietti, 2021; Szalay et al., 2025; Liu et al., 2023; Kempe et al., 2024). Thus, limitations of existing ASR systems motivate the need for manual correction of automatically generated transcriptions while the uneven performance of ASR tools across varieties necessitates investing more resources into manually correcting speech from under-resourced varieties (Gorisch et al., 2020).

Commercially available ASR systems, such as IBM-Watson, Google, and Amazon, might yield higher accuracy than custom ASR tools (e.g., UNSW ASR, Batchalign) due to larger training data. However, they require uploading speech data to the developers' server, potentially preventing researchers from using them due to constraints of ethical research data storage and sharing. The orthographic transcription system developed for speech researchers, OCTRA, also relies on commercially available ASR tools or ASR tools run on servers located on academic service platforms in various European countries, making data upload challenging due to different data protection principles between countries and regions (Draxler and Pömp, 2022).

1.2. Manual corpus transcription

When ASR tools are not available or not suitable, orthographic transcription is developed manually (e.g., Pitt et al., 2005; Grønnum, 2009; Gao et al., 2012; Schuppler et al., 2017; Rumberg et al., 2022); although manual transcription is more time-consuming and prone to disagreements in transcription, especially for spontaneous speech (Bazillon et al., 2008; Glenn et al., 2010). To improve the accuracy of manual transcription, manual transcription may be conducted by linguistically trained annotators with a background in phonetics or speech language pathology, or a professional transcription company (Rumberg et al., 2022). However, annotators may require multiple hours of corpus-specific training to achieve consistency (Schuppler et al., 2017). The need for expert annotators, combined with the time needed for training them, further increases transcription costs.

To reduce costs, crowd-sourced transcriptions by non-experts were tested against expert transcriptions and reached comparable accuracy for English, German, and Italian; however, crowd-sourcing was not possible for Flemish due to lack of transcribers (Sprugnoli et al., 2017). For crowd-sourced transcriptions, quality control using ground truth or automatic screening methods becomes crucial. (Sprugnoli et al., 2017; Lee and Glass, 2011).

Corpora differ in the details of orthographic transcription and annotation, with more detailed transcriptions needing longer time (Bazillon et al., 2008; Cieri, 2009; Esteve et al., 2010). Some corpora report transcribing back-channels or non-linguistic affirmations, fillers or filled pauses (Schuppler et al., 2017; Pérez-Espinosa et al., 2020), repetitions and broken words (Schuppler et al., 2017), and unde-

fined vocalisations (Pérez-Espinosa et al., 2020). Some child corpora reported annotating developmental speech errors (Rumberg et al., 2022; Szalay et al., 2022a), while another did not (Pérez-Espinosa et al., 2020). Other corpora did not report the level of detail in manual orthographic transcription and annotation (Pitt et al., 2005; Gao et al., 2012; Mereu and Vietti, 2021).

The more detailed the manual transcription, the lower the agreement between transcribers (Bazillon et al., 2008; Esteve et al., 2010). To improve transcription consistency and agreement on spontaneous speech, some corpora used an iterative consensus procedure with multiple annotators working on the same data (e.g., Gao et al., 2012; Rumberg et al., 2022). In the KidsTALC corpus, German continuous child speech was transcribed consecutively by two speech language pathologists and by a professional transcribing company, discussing transcriptions until a consensus was reached (Rumberg et al., 2022). A Mandarin corpus of child speech was first transcribed by a team of junior transcribers and then reviewed by a senior transcriber; however, details on the transcribers' research background and training were not reported (Gao et al., 2012).

1.3. Transcribing AusKidTalk

Transcribing AusKidTalk, the first AusE-speaking child corpus, shares many of the challenges described above (Ahmed et al., 2021; Szalay et al., 2022a, 2025). Efficient transcription of AusKidTalk required ASR tools as manual transcription is prohibitively cost-intensive due to the size of the corpus (Szalay et al., 2022a, 2025). However, due to the limitations of current ASR tools on AusE-speaking children, manual correction was required (Szalay et al., 2022a, 2025). Therefore, a semi-automatic workflow was developed for transcribing the single word production task in AusKidTalk consisting of an ASR toolkit generating automatic transcription and a protocol for hand-correction (Szalay et al., 2022a, 2025).

Our goal was to (1) apply our existing AusKidTalk ASR toolkit to generate automatic transcriptions for a semi-spontaneous narrative speech task in AusKidTalk that reduced transcription costs and (2) develop novel guidelines and methods for hand-correction to increase transcription accuracy. To achieve consistency between annotators, a novel hand-correction protocol was developed consisting of annotation guidelines, corpus-specific training materials, a custom-built hand-correction interface, and ground truth consistency checks.

This paper reviews details of the narrative speech task and the existing ASR toolkit designed for AusKidTalk, then discusses guidelines, new tools, and results of hand-correction. Our workflow can be adapted for other corpora containing spontaneous speech and provides techniques to evaluate the quality of both human and automated annotations, offering practical methods to enhance annotation workflows.

2. The AusKidTalk toolkit

2.1. AusKidTalk data description

At the time of writing, AusKidTalk comprises data collected from 620 AusE-speaking children aged 3–12. Speech samples were collected using three scripted (single word production, sentence repetition, non-word repetition) and two semi-spontaneous tasks (story-telling, emotional speech elicitation) (Ahmed et al., 2021). All tasks were prompted using visual and/or audio stimuli to allow for the inclusion of pre-literate children as young as 3 years old. The story-telling task, Task 3, was designed to elicit semi-spontaneous continuous speech by presenting a cartoon video and then prompting the child to tell the story in sentences using picture prompts (Ahmed et al., 2021). The cartoon depicted a green-skinned boy on a skateboard finding a large egg, crashing into the egg, and becoming friends with the green dinosaur that hatched (Doggy Dog Cartoons, 14:37–15:27). The video contained pictures and background music without any speech. Having watched the video, children were invited to retell the story using their own words while viewing a series of 13 key-event picture prompts one-by-one.

The five AusKidTalk tasks and prompts were presented on an Android tablet while speech was recorded onto a PC using a lapel microphone and directional microphones (Ahmed et al., 2021). As there was no direct synchronisation between tasks and prompts on the tablet, and audio recorded onto the PC, the Android app played a 1s high-frequency tone to mark the start of each task and recorded timestamps at the start and end of each task and prompt (Szalay et al., 2022a, 2025). The entire session was recorded with each audio file containing the five tasks, background noise, and multiple speakers producing task-related and conversational speech: the child and the interviewer (e.g., "What's in this picture?") (Szalay et al., 2022a, 2025).

Children's responses combined picture descriptions and non-responses ("I don't know"). The content of picture descriptions was expected to be related to the story and contain some key words (e.g., "boy, dinosaur"); however, spontaneous speech was characterised by false starts ("di- dinosaur"), partial and/or unintelligible words, and filler words for hesitation (e.g., "erm") inherent to spontaneous child speech. Morphological and syntactic errors typical to child speech, such as non-adult like use

of pronouns or auxiliaries were present, as were incomplete sentences (Lee and Canter, 1971). The combination of spontaneous speech and two speakers – the child and the interviewer – increased the difficulty of orthographic transcription.

2.2. Automatic transcription using the existing toolkit

To reduce annotator burden, time-aligned orthographic transcriptions were generated using existing but suboptimal tools (Szalay et al., 2025).[1] Task 3 was identified in the audio using high-frequency tone detection and time-aligned with the 13 picture prompts based on the timestamps recorded during data collection (Szalay et al., 2025). Diarisation was conducted using NeMo and orthographic transcription with the UNSW ASR (Szalay et al., 2025; Shahin et al., 2020). The output, a Praat textgrid with time-aligned word-level orthographic transcription (Fig. 1), was screened for quality, with high-quality audio-textgrid pairs passed on for hand-correction (Fig. 2).[2]

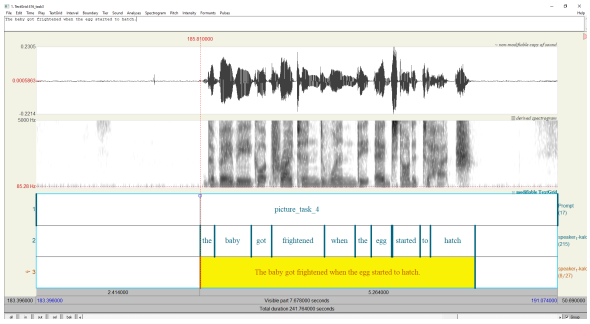

Figure 1: Output of the automatic annotation workflow with picture prompt (tier 1), automatic word-level transcription (tier 2), and child turn (tier 3).

3. Methods of hand-correction

3.1. Developing hand-correction guidelines

Hand-correction guidelines for the automatically generated transcriptions were developed to assist annotators with accuracy and consistency. Annotators were instructed to correct automatic transcription of the child's speech in five steps. In Step 1), annotators identified child-turns produced during each picture prompt: the child speaking without adult interruption was a turn (Table 1). The turn ended when the adult started talking. For every picture prompt, annotators generated one turn for every time the child spoke without adult interruption. Turns were identified instead of sentences as the goal was transcribing all continuous child speech in each picture-prompt interval instead of providing linguistic analysis of the child's use of sentences.

In Step 2), annotators corrected the orthographic transcription of the child's speech. Annotators transcribed the child speech only, capturing what the child said as opposed to what they thought the child meant. That is, transcriptions were to reflect syntactic and morphological errors. For example, "he smile" was transcribed as-is, instead of being corrected to "he smiles/he smiled", avoiding assuming either present or past tense. Apostrophes were used to avoid assuming underlying grammar (e.g., the transcription "the boy's" was preferred over assuming "boy is" or "boy has"). Colloquialisms were transcribed as pronounced (e.g., "gonna"). Every complete word was transcribed irrespective of sentence quality (e.g., in "A dino... the dino", both instances of "dino" were transcribed). Phoneme-level errors (e.g., [botu] for "bottle") were not transcribed and the intended word was transcribed using standard AusE orthography. Nonwords, such as unintelligible and partial words (e.g., "din") were transcribed as XXX. Hesitations (e.g., "uhm") were not transcribed. Annotators were instructed not to use any punctuation marks other than capitalising the first turn of the picture-interval and ending the last turn with a period. When a child named the protagonist of the story – the green stop-motion boy – as "baby Hulk", 'Hulk" was capitalised as a proper noun.

In Step 3) annotators marked the location of any hesitations between the corrected words using the position of the words in the turn. Hesitations were marked when the child paused noticeably between two words (unfilled pause) or when the child inserted a filler, such as "erm" between two words (filled pause). In Step 4), they identified words containing overlap between speakers or overlap between the child and background noise (Table 1). In Step 5), they reviewed their corrections.

3.2. Developing a hand-correction interface

To streamline the hand-correction procedure, a custom Praat interface was developed that prompted annotators to carry out Steps 1)–5) for every picture prompt using a series of pop-up windows.[3] The custom interface loaded the audio and the textgrid for every picture prompt one-by-one, automatically generating speaker-turns by merging the words identified as belonging to the child during diarisation. Having loaded the first picture-prompt, the in-

[1] The technical description of these tools is reported in detail in Szalay et al. (2025).

[2] A similar hand-correction workflow for AusKidTalk's single word production task is reported in (Szalay et al., 2022a).

[3] Code available via https://github.com/rbtbecontinued/auskidtalk

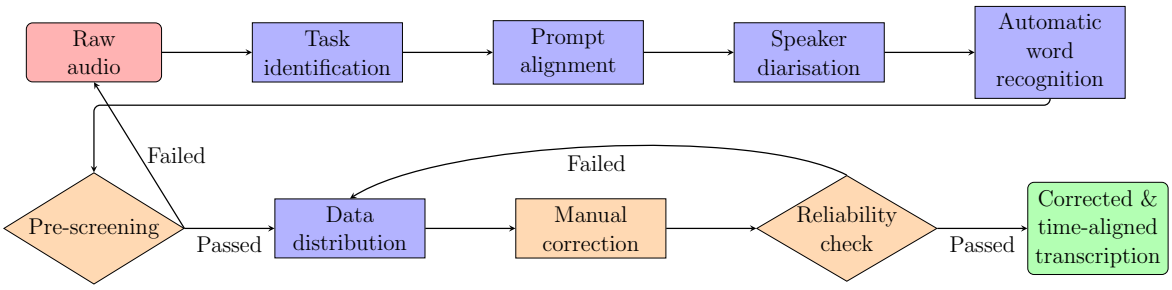

Figure 2: AusKidTalk annotation pipeline. Top row: Pre-processing. Bottom row: Hand-correction. Red: raw audio containing five speech tasks. Blue: automated procedure. Yellow: manual procedure. Green: orthographic transcription.

Term	Definition	Marker
Picture prompt	Audio recorded during the presentation of a picture	Boundaries
Child turn	Child speech produced without adult interruption	Boundaries
Nonword	Unintelligible or partial word	XXX
Pause, filled	A hesitation with a non-word	{X}
Pause, long	A hesitation over 1s	{..}
Pause, short	A noticeable gap in speech shorter than 1s	{.}
Overlap	Child speech concurrent with adult speech or noise	{O}

Table 1: Terms, definitions, and their transcription markers used during hand-correction.

terface first prompted annotators to carry out Steps 1) and 2) simultaneously by editing turn-interval boundaries and transcription.

After completing orthographic transcription, the interface prompted the annotator to mark hesitations (Step 3). The possible positions before and after each word where hesitations may occur were automatically numbered in consecutive order in a pop-up window. For instance, as in "{1} He {2} is {3} riding {4} his {5} skateboard {6}.", every number corresponds to a possible hesitation. Annotators were prompted to identify the location and evaluate the type of hesitation: a) filled pause (e.g., "erm"), b) long unfilled pause (i.e., silence of 1s or more), or c) short unfilled pause (i.e. a noticeable pause shorter than 1s). Then, the script automatically inserted hesitation markers into the transcription as {X} for filled pause, {..} for long pause, and {.} for short pause (Table 1). Consequently, after Step 3), the transcription might read, "He {X} is riding {..} his skateboard."

In Step 4), the interface listed all words in the turn, and annotators were prompted to select any words that were produced concurrently with the interviewer or background noise (e.g., tapping, coughing). Once an overlap was flagged, an overlap marker {O} was automatically inserted after the selected word in the transcription. For instance, the resulting transcription could be "He {X} is {O} riding {..} his skateboard {O}." indicating that the child's speech overlaps with background noise or adult speech during "is" and "skateboard".

In Step 5), the interface prompted annotators to review their transcription and listen to the speech before continuing to the next picture prompt to be corrected. Once the annotator completed corrections to the first picture prompt, the interface automatically saved the annotator's work and loaded the next one.

3.3. Reliability checks

To evaluate annotator reliability, a ground truth (GT) method was chosen (Szalay et al., 2022a). To prepare GT files, six files were transcribed and annotated by expert annotators with a background in speech pathology and phonetics in an iterative consensus procedure starting from ASR-generated transcriptions (Szalay et al., 2022a). Audio-textgrid pairs, including both novel and GT files were assigned to annotators automatically using a web interface allowing for downloading uncorrected files and uploading corrected files (Szalay et al., 2022a). GT files were assigned automatically to every annotator as the first two files for training and as every 10th file after training. Annotators were blinded as to which files were GT files except for the first two files used as training.

When annotators uploaded corrections to a GT file, their work was compared to GT transcription and an evaluation report was generated automatically to assist the supervisor. The report contained

information on 1) number of child speech turns identified; 2) alignment accuracy; 3) transcription accuracy; and 4) hesitation and overlap marking accuracy, comparing these metrics to pre-set passing thresholds, as follows.

First, the number of turns identified by the annotator was counted. Agreement of 100% with the GT file was required to pass. Alignment between annotators' turn and GT turn was calculated (Equation 1, Paulo and Oliveira 2004; Gonzalez et al. 2020). Turn alignment of <0.95 was considered an error, with a passing threshold of >90% agreement for the turns in a file.

$$Overlap = \frac{DurShared}{DurGT + DurCurrent - DurShared} \quad (1)$$

To evaluate transcription accuracy, overall word error rate (WER) in the entire file, as well as WER of each turn was calculated. Nonwords marked by XXX were included when calculating WER, to deter annotators from overusing this marker. For example, a disagreement between an annotator transcribing "din dinosaur" and the GT transcription containing "XXX dinosaur" was coded as an error. Differences in the use of contractions ("boy's" vs. "boy is") were considered errors as the disagreement reflects both a phonetic difference (/z/ vs. /əz/) as well as potential grammar difference. Overall WER, as well as minimum, mean, median, and maximum WER of turns were reported. The passing threshold was <5% overall WER and <5% real word to XXX ratio.

Agreement values for hesitation markers and speaker-overlap markers between annotators' and GT transcription were evaluated. Ratios were calculated for the short, long, and the filled pause markers, as well as for the speaker-overlap marker (Table 1). While there was no required minimum for identifying markers, a ratio closer to 1 was expected for long and filled pause markers and speaker-overlap markers than for short pause markers. Additionally, an annotator who did not identify any hesitation or overlap markers received a warning from the supervisor. The location of hesitation and speaker-overlap markers was not compared and not evaluated.

When annotators reached passing rates on all metrics, they were allowed to download a new audio-textgrid pair for hand-correction. When the passing rate was not reached, annotators received written feedback and were assigned additional training until they passed.

4. Results and Discussion

4.1. Applying the existing toolkit to the narrative task

Out of the 620 children's Task 3 files, data from 429 were successfully pre-processed (Fig. 2). The remaining 191 failed pre-processing for multiple reasons. For example, 74 children's data were recorded using a paper protocol, thus time-stamps were not recorded, making pre-processing impossible, while for others, automatic high frequency tone detection failed.

The 429 audio-textgrid pairs were pre-screened for quality with 297 (297/429, 69%) passing pre-screening. Audio-textgrid pairs were excluded when the audio was too noisy (31 files, 7%), included speech from tasks other than Task 3 (87 files, 20%), contained diarisation errors (13 files, 3%) or the picture prompts were misaligned with the audio (1 file, 0.2%).

Pre-screening revealed that on average, 1.9 speakers (standard deviation = 0.7) per sample were identified in 297 audio files, slightly lower than the expected two (interviewer and child). Consistent with this, NeMo diarisation failed to separate the child from the adult in 60 files. As 60 files were deemed to be a too large subset to exclude from hand-correction but too small to spend time on adjusting NeMo's parameters, they were distributed for hand-correction (Fig. 2). Relatively low diarisation quality is attributed to the short duration of Task 3 (3.5 minutes on average) and to the limited amount of speech produced by the interviewer. These results on data and diarisation quality from 431 audio-textgrid pairs are consistent with preliminary results examining data and diarisation quality using a subset of the data (201 files) (Szalay et al., 2025).

4.2. Quality of automatic transcription

At the time of writing, of the 298 accepted audio-textgrid file pairs, 261 (261/298, 88%) were hand-corrected, including the six GT files. The quality of automatic transcription and alignment was evaluated by comparing automatically generated textgrids to annotator's hand-correction in data from these 261 children.

4.2.1. Turn-identification

In the sample of 261 hand-corrected files, 3261 turns were identified automatically, with an average of 12.5 turns per child (standard deviation = 1.3), whereas annotators identified a total of 3965 turns with an average of 15.2 turns per child (standard deviation = 4.2). Out of the 3261 automatically identified turns, 2724 (83.7%) matched

to a single manually-identified turn, indicating a 83.7% agreement rate between automatically generated and manually-identified turns. 451 automatically generated turns (13.8%) corresponded to multiple manually-identified turns, consistent with turns being missed during automatic diarisation. 86 of the automatically generated turns (2.6%) were removed. An additional 134 turn-intervals were missed by the automatic alignment and added manually by the annotators.

The relatively high accuracy of turn identification (83.7%) compared to the poor speaker diarisation is attributed to interviewer-behaviour: interviewers were trained to only prompt the child at the start and the end of picture prompts. Therefore, when automatic diarisation included adult speech in the child's turn, the adult speech typically fell at the beginning and/or the end of the turn. Thus, to remove adult speech, the child's turn had to be shortened rather than split into multiple turns.

4.2.2. Turn overlap

There were 3175 automatically generated turns corresponding to one or more manually-corrected turns, yielding 3831 turn-pairs. Mean overlap rate for all turn-pairs was 0.7 (s.d. = 0.37); the median however, was considerably higher at 0.9. When only turn-pairs in which the automatic turn matched onto a single manually-identified turn were considered, mean overlap rate was considerably higher at 0.9 (s.d. = 0.2).

4.2.3. Word Error Rate

Evaluating WER on all turn-pairs yielded a high WER of 166.9% (s.d. = 690%), with WER exceeding 100% due to inclusion errors. Inclusion errors were caused by incorrect diarisation, resulting in automatic transcription of adult speech, which was then removed by the annotators. Low overlap rate between automatically generated and hand-corrected turns caused by poor diarisation correlated with high WER (Spearman's rho for monotonic relationship between overlap rate and WER = -0.78). When only turn-pairs in which the automatic turn matched onto a single manually-identified turn were considered, mean WER was considerably lower at 35.9% (s.d. = 53.1%, Spearman's rho for WER-Overlap correlation = -0.45). When only turn-pairs with 1:1 overlap were considered, mean WER was 23.3% (s.d. = 21.8%). These WER results are consistent with errors in transcription being caused by imperfect diarisation. Thus, having to remove transcription of the adult speech is likely to have increased annotator workload, showing the importance of accurate diarisation in transcription.

WER by children's age and sex were analysed using turn-intervals with 1:1 overlap rate from 252 children; one child was excluded for missing age information and eight for not having any turn-pairs with a 1:1 overlap rate. The 252 children were divided into three age groups: 83 children aged 3-5 years (F = 42, M = 41), 78 children aged 6-8 years (F = 36, M = 42), and 91 children aged 9-12 years (F = 49, M = 42). WER decreased as age increased, but there were no considerable differences between female and male children, potentially increasing hand-correction time for younger children due to lower quality ASR (Table 2). These results are consistent with preliminary results on automatic transcription of the story-telling task in AusKidTalk (Szalay et al., 2025). However, due to the 23% WER, hand-correction is still required, necessitating investing time and resources in hand-correction as well as in the development of hand-correction software (Gorisch et al., 2020).

Age Group	Female	Male	Both sexes
3–5	32 (28)	32 (24)	32 (26)
6–8	19 (19)	24 (23)	22 (21)
9–12	17 (17)	21 (18)	19 (18)
Total	21 (22)	25 (22)	**23 (22)**

Table 2: Mean WER as percentage (s.d.) by Age group and Sex. Numbers in bold show overall WER.

4.3. Reliability of hand-correction

A team of nine annotators with a background in linguistics and phonetics submitted 261 transcriptions for the story-telling task (Figure 3). They hand-corrected one file in approximately 40 minutes on average. Hand-correction time decreased considerably from their first to second file, with annotators spending up to three hours on their first file, and completing their second in 45 minutes. Average file duration was 3.5 minutes, indicating that hand-correction, together with marking disfluencies and nonwords, was 11x longer than audio duration. As the time-need of transcription greatly varies with transcription method (manual vs. ASR-assisted), speech type (scripted vs. spontaneous), and annotation detail (content words only vs. partial words, hesitations, disfluencies), the time cost of a comparable task was selected as a benchmark, reported as 15 times the length of audio for manual verification of automatic transcription with incomplete marking of partial words, filled pauses, disfluencies, and background noise (Cieri, 2009). Our result – correction time 11 times longer than audio – is more efficient than the cost of transcription time being x15 longer than audio time reported for comparable annotation (Cieri, 2009).

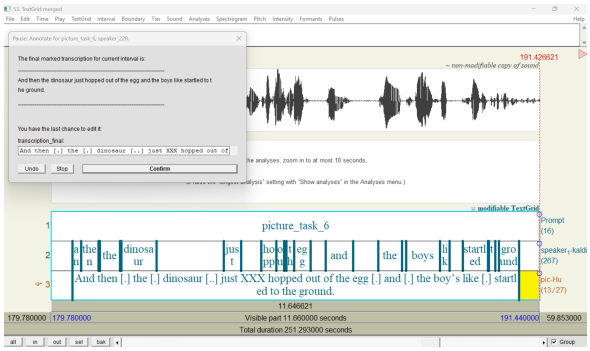

Figure 3: Hand-corrected transcription with picture prompt (tier 1), automatic word-level transcription (tier 2), and hand-corrected child turn (tier 3) with a pop-up window prompting the annotator to review their transcription.

Seven annotators submitted a total of 25 ground truth files, of which 15 (60%) reached a passing score. Out of the 10 failed GT submissions, six failed on the same GT file from Child 651, a 5-year old typically developing male. No annotator achieved passing score on Child 651. This suggests that the 60% pass rate is partially caused by one GT file being more difficult than the others, as with Child 651 not included, the pass rate is 78% (15/19). Reviewing annotators' work between the last passed and a failed Child 651 submission showed that annotators submitted accurate transcriptions between the two, demonstrating accurate and reliable transcription. However, the 100% fail rate on Child 651 questions whether transcriptions of more difficult speakers (e.g., younger or with speech disorders) can be accepted at face-value without further revision.

One of the seven annotators was removed from the project due to consistently failing consistency checks on easier GT files. The remaining two annotators contributed less than ten novel files to the project after passing their training sessions; therefore they did not reach the first GT checkpoint.

Although using automatic transcription as a starting point sped up hand-correction time, it might have biased annotators towards accepting the ASR's transcription. To counter this potential bias, annotators were explicitly instructed to practice awareness of the bias, and to make independent judgements for transcription. Instruction to use XXX to transcribe unintelligible utterances rather than relying on the ASR output combined with regular monitoring of annotator's work and feedback sessions proved to be an effective strategy, as reflected in the pass rates observed in the GT files. For instance, when in one GT file, the ASR transcribed a word ambiguous between "when" and "went" as "when", multiple annotators transcribed the ambiguous word as XXX, replacing the automatic transcription with the nonword marker.

The overall AusKidTalk reliability (60%-78%) appears to be high; however, the failed consistency checks are due to genuine mistakes by trained annotators. In contrast, when crowd-sourced transcriptions of adult conversational speech were evaluated using ground truth methods, the passing rate was lower (26%-73%), but failed annotations were often clear cases of spammers submitting repeated and nonsense words, while genuine annotations by untrained transcribers were of good quality (Sprugnoli et al., 2017). As in AusKidTalk, even trained annotators made genuine mistakes due to the difficulties associated with child speech and the level of annotation required, crowd-sourcing for untrained annotators is not recommended for transcribing continuous, semi-spontaneous child speech.

5. Conclusion: Generalisability and lessons learnt

We applied an ASR-assisted transcription method to a semi-spontaneous continuous speech production task (Fig. 2, Szalay et al. 2025). Automatic speech processing tools performed speech diarisation using NeMo and automatic word recognition using the UNSW ASR (Szalay et al., 2025; Shahin et al., 2020). Automatic transcription was greatly assisted by designing the data collection protocol with automatic annotation in mind: the collection of high tones separating tasks, time stamps for tasks and prompts, and interviewer training contributed to the implementation of automatic transcription. Automatically generated transcriptions were corrected manually by a team of annotators using a set of guidelines and an interface developed for the narrative task of the AusKidTalk project.

Comparing automatically generated and manually corrected transcription showed that the quality of automatically generated transcription was sensitive to diarisation errors. When NeMo diarisation failed, the adult's speech was transcribed as child speech, which then manually had to be removed by annotators, increasing annotation time and cost. Future work may compare other diarisation methods to NeMo, aim to adjust diarisation parameters of NeMo, or apply diarisation prior to task identification such that the entire recording is diarised.

ASR errors remained high after adjusting for errors caused by diarisation (23%); however, the UNSW ASR was trained on American English child speech rather than other parts of the AusKidTalk corpus (Shahin et al., 2020), thus accent differences between American and Australian English (e.g., the presence/absence of post-vocalic /ɹ/, differing vowel pronunciations) are likely to have negatively impacted ASR quality and increased hand-correction time (Tatman and Kasten, 2017; Sza-

lay et al., 2022b; Wassink et al., 2022). In future work, the currently transcribed speech samples can be used for fine-tuning open, state-of-the-art systems that can be run locally on the corpus for better performance. For example, Whisper ASR has achieved 7%-14% WER on the My Science Tutor corpus containing spontaneous child speech recorded from virtual tutoring sessions when it was trained on the same corpus (Southwell et al., 2024; Attia et al., 2024). Similarly, using Task 3 transcription from AusKidTalk to fine-tune ASR models might allow for greater accuracy in transcribing the emotion elicitation task, the other spontaneous task in AusKidTalk.

Working with annotators showed that despite annotators having a background in linguistics and phonetics, annotator agreement can be challenging to achieve for spontaneous speech, consistent with previous work (Glenn et al., 2010). However, semi-automatic consistency and reliability checks assisted with separating high-quality annotations from inconsistent and potentially inaccurate annotations, similar to recommendations for crowd-sourced annotations (Sprugnoli et al., 2017; Lee and Glass, 2011). The results of annotator agreement suggest that child speech transcription might not be suitable for crowd-sourcing. Despite the limitations noted here, the results show that accurate annotations can be achieved with regular consistency checks. Our approach promises more rapid data handling with a correction speed 11 times the speech time, together with techniques for quality control transferable to other corpora.

6. Acknowledgement

We would like to thank our participants without whom this project would not have been possible. This project was supported by the Australian Research Council LE190100187 and FT180100462 grants, as well as the University of New South Wales, The University of Sydney, Western Sydney University, Macquarie University and The University of Melbourne. This project was approved by The University of New South Wales Human Research Ethics Approval HC190320.

7. Bibliographical References

Beena Ahmed, Kirrie Ballard, Denis Burnham, Tharmakulasingam Sirojan, Hadi Mehmood, Dominique Estival, Elise Baker, Felicity Cox, Joanne Arciuli, Titia Benders, et al. 2021. AusKidTalk: an auditory-visual corpus of 3-to 12-year-old australian children's speech. In *Annual Conference of the International Speech Communication Association (22nd: 2021)*, pages 3680–3684. International Speech Communication Association.

Ahmed Adel Attia, Jing Liu, Wei Ai, Dorottya Demszky, and Carol Espy-Wilson. 2024. Kid-whisper: Towards bridging the performance gap in automatic speech recognition for children vs. adults. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 74–80.

Thierry Bazillon, Yannick Esteve, and Daniel Luzzati. 2008. Manual vs assisted transcription of prepared and spontaneous speech. In *LREC*.

Denis Burnham, Dominique Estival, Steven Fazio, Jette Viethen, Felicity Cox, Robert Dale, Steve Cassidy, Julien Epps, Roberto Togneri, Michael Wagner, et al. 2011. Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable black box. ISCA.

Strassel S. M. Cieri, C. 2009. Closer still to a robust, all digital, empirical, reproducible sociolinguistic methodology. In *NWAV 38: New Ways of Analyzing Variation*. University of Ottawa.

Doggy Dog Cartoons. Doggy dog.

Christoph Draxler and Julian Pömp. 2022. OCTRA - an innovative approach to orthographic transcription. In *INTERSPEECH*, pages 5217–5218.

Yannick Esteve, Thierry Bazillon, Jean-Yves Antoine, Frédéric Béchet, and Jérôme Farinas. 2010. The EPAC corpus: Manual and automatic annotations of conversational speech in french broadcast news. In *LREC*.

Jun Gao, Aijun Li, and Ziyu Xiong. 2012. Mandarin multimedia child speech corpus: Cass_child. In *2012 International Conference on Speech Database and Assessments*, pages 7–12. IEEE.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. 1993. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403.

Meghan Lammie Glenn, Stephanie M Strassel, Haejoong Lee, Kazuaki Maeda, Ramez Zakhary, and Xuansong Li. 2010. Transcription methods for consistency, volume and efficiency. In *LREC*.

Simon Gonzalez, James Grama, and Catherine E Travis. 2020. Comparing the performance of forced aligners used in sociophonetic research. *Linguistics Vanguard*, 6(1):20190058.

Jan Gorisch, Michael Gref, and Thomas Schmidt. 2020. Using automatic speech recognition in spoken corpus curation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6423–6428.

Nina Grønnum. 2009. A Danish phonetically annotated spontaneous speech corpus (DanPASS). *Speech Communication*, 51(7):594–603.

Vera Kempe, Patricia J Brooks, and Steven Gillis. 2024. Four decades of open language science: The CHILDES project. *Language Teaching Research Quarterly*, 44:15–30.

Abigail Anne Kressner, Kirsten Maria Jensen-Rico, Johannes Kizach, Brian Kai Loong Man, Anja Kofoed Pedersen, Lars Bramsløw, Lise Bruun Hansen, Laura Winther Balling, Brent Kirkwood, and Tobias May. 2024. A corpus of audio-visual recordings of linguistically balanced, Danish sentences for speech-in-noise experiments. *Speech Communication*, 165:103141.

Chia-ying Lee and James Glass. 2011. A transcription task for crowdsourcing with automatic quality control. In *Twelfth Annual Conference of the International Speech Communication Association*. Citeseer.

Laura L Lee and Susan M Canter. 1971. Developmental sentence scoring: A clinical procedure for estimating syntactic development in children's spontaneous speech. *Journal of Speech and Hearing Disorders*, 36(3):315–340.

Mark Y Liberman. 2019. Corpus phonetics. *Annual Review of Linguistics*, 5(1):91–107.

Houjun Liu, Brian MacWhinney, Davida Fromm, and Alyssa Lanzi. 2023. Automation of language sample analysis. *Journal of Speech, Language, and Hearing Research*, 66(7):2421–2433.

Daniela Mereu and Alessandro Vietti. 2021. Dialogic ItAlian: the creation of a corpus of Italian spontaneous speech. *Speech Communication*, 130:1–14.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Sérgio Paulo and Luís C Oliveira. 2004. Automatic phonetic alignment and its confidence measures. In *Advances in Natural Language Processing: 4th International Conference, EsTAL 2004, Alicante, Spain, October 20-22, 2004. Proceedings 4*, pages 36–44. Springer.

Humberto Pérez-Espinosa, Juan Martínez-Miranda, Ismael Espinosa-Curiel, Josefina Rodríguez-Jacobo, Luis Villaseñor-Pineda, and Himer Avila-George. 2020. IESC-child: an interactive emotional children's speech corpus. *Computer Speech & Language*, 59:55–74.

Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. 2005. The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

Lars Rumberg, Christopher Gebauer, Hanna Ehlert, Maren Wallbaum, Lena Bornholt, Jörn Ostermann, and Ulrike Lüdtke. 2022. kidsTALC: A corpus of 3-to 11-year-old German children's connected natural speech. In *INTERSPEECH*, pages 5160–5164.

Barbara Schuppler, Martin Hagmüller, and Alexander Zahrer. 2017. A corpus of read and conversational Austrian German. *Speech Communication*, 94:62–74.

Mostafa Ali Shahin, Renée Lu, Julien Epps, and Beena Ahmed. 2020. UNSW system description for the shared task on automatic speech recognition for non-native children's speech. In *Proc. Interspeech*, pages 265–268.

Prashanth Gurunath Shivakumar and Panayiotis Georgiou. 2020. Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. *Computer speech & language*, 63:101077.

Rajni Sobti, Kalpna Guleria, and Virender Kadyan. 2024. Comprehensive literature review on children automatic speech recognition system, acoustic linguistic mismatch approaches and challenges. *Multimedia Tools and Applications*, pages 1–63.

Rosy Southwell, Wayne Ward, Viet Anh Trinh, Charis Clevenger, Clay Clevenger, Emily Watts, Jason Reitman, Sidney D'Mello, and Jacob Whitehill. 2024. Automatic speech recognition tuned for child speech in the classroom. In *Icassp 2024-2024 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 12291–12295. IEEE.

Rachele Sprugnoli, Giovanni Moretti, Luisa Bentivogli, and Diego Giuliani. 2017. Creating a ground truth multilingual dataset of news and talk show transcriptions through crowdsourcing. *Language Resources and Evaluation*, 51:283–317.

Tünde Szalay, Louise Ratko, Mostafa Shahin, Tharmakulasingam Sirojan, Kirrie Ballard, Felicity Cox, and Beena Ahmed. 2022a. A semi-automatic workflow for orthographic transcription of a novel speech corpus: A case study of AusKidTalk. In *Proceedings of the 18th Australasian International Conference on Speech Science and Technology*, Canberra. Australasian Speech Science and Technology Association.

Tünde Szalay, Mostafa Shahin, Beena Ahmed, and Kirrie Ballard. 2022b. Knowledge of accent differences can be used to predict speech recognition. *Proc. Interspeech 2022*, pages 1372–1376.

Tünde Szalay, Mostafa Shahin, Tharmakulasingam Sirojan, Zheng Nan, Renata Huang, Kirrie Ballard, and Beena Ahmed. 2025. AusKidTalk: Using strategic data collection and out-of-domain tools to semi-automate novel corpora annotation. In *Proc. Interspeech*, pages 4268–4272.

Rachael Tatman and Conner Kasten. 2017. Effects of talker dialect, gender & race on accuracy of Bing Speech and YouTube automatic captions. In *Interspeech*, pages 934–938.

Alicia Beckford Wassink, Cady Gansen, and Isabel Bartholomew. 2022. Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication*, 140:50–70.