COLLECTING AND ANNOTATING NATURAL CHILD SPEECH DATA – CHALLENGES AND INTERDISCIPLINARY PERSPECTIVES

Hanna Ehlert[1], Edith Beaulac[1], Maren Wallbaum[1], Lars Rumberg[2], Christopher Gebauer[2], Jörn Ostermann[2], Ulrike Lüdtke[1]
Leibniz Universität Hannover,
[1]Institut für Sonderpädagogik, Abteilung für Sprach-Pädagogik und –Therapie
[2]Institut für Informationsverarbeitung
hanna.ehlert@ifs.uni-hannover.de, maren.wallbaum@ifs.uni-hannover.de

BACKGROUND: Data builds the basis for the training of machine learning algorithms. The amount of data required to train a model depends on the intended task and the property of the data. In the case of automatic child speech recognition the latter is extremely variable, increasing the distance to adult speech and its heterogeneity with decreasing age [1, 2]. At the same time collecting and manually processing representative child speech data for software development is a challenging task [3]. Poorly transcribed speech data can have far-reaching consequences [4]. In this paper we share experiences from our speech language therapy background and the TALC-project (Tools for Analyzing Language and Communication) where we explore the application of machine learning models for linguistic and speech therapy purposes in an interdisciplinary team.

COLLECTING CHILD SPEECH DATA: From a speech and language therapy perspective collecting and analyzing child data has a long tradition as a method for researching and assessing language development [5]. Although language sampling is generally a non-standardized procedure, a number of aspects contribute to obtaining representative and comparable samples across different children and age groups. These deliberations can guide the collection of speech data for the development of ASR software as well. Due to their typical insecurity in unfamiliar contexts usually resulting in a lack of compliance or restricted communicative interaction, collecting natural speech samples from children may be more challenging than collecting them from adults. At the same time, child speech samples collected in constrained contexts, such as sentence repetition or picture naming may be much less representative of unconstrained, natural child speech than is the case with adults. Examples from the kidsTALC corpus [6] will illustrate this.

Components such as location, materials used, elicitation method, and the conversational style of the person interacting with the child have been shown to have a direct influence on the success of collecting a speech sample as well as on its properties and reliability [7,8,9]. For example, while play-based activities may be appropriate to elicite more and more complex speech from younger children, in older children story telling may be the favourable context [8]. Table 1 provides an overview of recommended sampling contexts to collect continuous speech at different ages.

Table 1 – Recommended language sample contexts by age [5]

	Preschool	Schoolage	Adolescents
Freeplay	**X**		
Picture description	**X**	**X**	
Story telling / retelling	**X**	**X**	
Expository discourse		**X**	**X**
Persuasive discourse			**X**
Free conversation / dialogue	**X**	**X**	**X**

Additionally, considerations from an information science perspective should complement guidelines for collecting child speech data. These may also address the location (e.g. in terms of background noise), the materials used (e.g. in terms of its noisiness), the elicitation method (e.g. in terms of its ability to elicit longer and complete utterances), and the conversational style of the person interacting with the child (e.g. in terms of speech overlap). The latter refers to the next steps of processing the collected data. Facilitating the transcription of the data can already be considered by a skilled researcher during data collection.

ANNOTATING CHILD SPEECH DATA: In terms of annotating child data, we will share our continuous process of balancing the need for standardization, technological possibilities and initial but also evolving requirements in the TALC project. Before data can be processed several decisions have to be made, such as the mode of transcription (orthographic or phonetic; standard or verbatim). If phonetic transcription is desired by the project goals, agreement should be achieved in terms of the detail of this transcription (e.g., using only selected IPA symbols instead of the whole IPA). Audio metadata for child speech should always include age, language status (e.g., monolingual/multilingual, typical developing/language impaired) and sampling context (e.g., elicitation method). Ensuring communication between researchers collecting the data, those annotating (e.g., transcribing) and those training the ASR model is central for the further processing of child data. For example, in addition to the metadata, each audio should be furnished with notes on child specifics during data collection, such as health status (Does the child have changes in pronunciation and voice quality due to having a cold?) or developmental speech errors (which maybe sometimes missed without notification). To reach standardization and an acceptable inter transcriber agreement, which is generally lower for child data and specifically in phonetic transcriptions, training of and communication between transcribers is of utmost importance [10]. In our TALC-project we have established a training consisting of several tasks and rounds of feedback to complete if new transcribers are to be integrated into the project. Typical characteristics of developmental child speech should be addressed in the training. Emerging disagreement and uncertainty of transcribing specific audio parts should be resolved via discussion among transcribers. Consensus should be integrated into a continuously updated annotation manual.

CONCLUSIONS: An interdisciplinary approach to collecting and annotating natural child speech data for training automatic speech recognition models is beneficial. A background in child language development or speech and language therapy should guide data collection in order to obtain robust and representative data. Communication between researchers of each discipline and working on different aspects of a project is central to addressing the challenges of automatic child speech recognition.

LIST OF REFERENCES

[1] FENSON et al.: Variability in Early Communicative Development. Monographs of the Society for Research in Child Development, In *JSTOR*, 59–185, 1994.

[2] A. POTAMIANOS et al.: Robust recognition of children‘s speech. *IEEE Transactions on Speech and Audio Processing.* 2003.

[3] HAZEN: Automatic Alignment and Error Correction of Human Generated Transcripts for Long Speech Recordings. INTERSPEECH 2006 - ICSLP, 1606–1609, 2006.

[4] KNIGHT et al.: Clinicians' views of the training, use and maintenance of phonetic transcription in speech and language therapy. In *International journal of language & communication disorders* 53 (4), 776–787, 2018.

[5] NIPPOLD: Language Sampling with Children and Adolescents, 3rd ed., Plural Publishing, 2021.

[6] RUMBERG et al.: kidsTALC: A Corpus of 3- to 11-year-old German Children’s Connected Natural Speech. INTERSPEECH 2022, 5160-5164, 2022.

[7] MILLER: Assessing language production in children: Experimental procedures. Baltimore, MD: University Park Press, 1981.

[8] KLEIN et al.: Influence of Context on the Production of Complex Sentences by Typically Developing Children. In *Language, Speech, and Hearing Services in Schools*, 41, 289–302, 2010.

[9] SOUTHWOOD et al.: Comparison of Conversation, Freeplay, and Story Generation as Methods of Language Sample Elicitation. In *Journal of Speech, Language, and Hearing Research*, 47, 366–376, 2004.

[10] RAMSDELL et al.: Predicting phonetic transcription agreement. Insights from research in infant vocalizations. In *Clinical Linguistics & Phonetics,* 21 (10), 793–831, 2007.