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

Hanna Ehlert[1]*, Edith Beaulac*[1]*, Maren Wallbaum*[1]*, Christopher Gebauer*[2]*, Lars Rumberg*[2]*, Jörn Ostermann*[2]*, Ulrike Lüdtke*[1]

Leibniz Universität Hannover,
[1]*Department of Speech and Language Therapy and Inclusive Education,*
[2]*Institute for Information Processing*
hanna.ehlert@ifs.uni-hannover.de

Abstract: In this paper we share experiences on collecting and annotating child speech data 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 (focus ASR) for linguistic and speech therapy purposes in an interdisciplinary team. We will reflect on the importance of collecting natural speech data for ASR model training and will summarize recommended methods for eliciting such spontaneous child speech at different ages. For annotating recorded data such as transcribing them and marking relevant parts for subsequent analysis, we will focus on possible ways to ensure communication between different researchers. Throughout, we will elaborate on the interdisciplinary collaboration in our project in order to ensure that requirements of model developers and end-users are met.

1 Background

Besides mainstream applications automatic speech recognition (ASR) of children has the potential to support child speech and language research, clinical assessment, intervention planning and advising parents. For example, the analysis of recorded child speech samples is a widely recognized and ecologically valid method in the diagnosis of developmental language disorders [1]. These samples can be analyzed with regard to all linguistic domains (phonology, morphology, syntax, semantics and pragmatics) creating a comprehensive picture of a child's speech and language abilities [2]. But the process of recording, manually transcribing and manually analyzing these samples is very time consuming. Therefore, despite its many advantages over the use of standardized speech and language tests, it is not used regularly by speech language therapists [3]. Automating the process via ASR would enable to record longer and more samples and thus to continuously monitor language development outside the therapy setting and plan interventions based on functional communicative abilities. Turning to research, automated support of language sample analysis would enable the collection of mass data to gain new and verify existing knowledge on language development.

Data also build 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 [4, 5]. In the case of automatic child speech recognition the latter is extremely variable, increasing the distance and to adult speech and its heterogeneity with decreasing age [6, 7]. At the same time collecting and manually processing representative child speech data for software development is a challenging and time consuming task as noted previously [8, 9]. This leads to a twofold sparseness: There is a lack of openly available child speech data for ASR training purposes,

especially natural speech [10], and at the same time more data is needed than for training ASR models of adult language [11]. Therefore, to collect and annotate robust and representative child speech data for model training and subsequently to develop the software to fit the intended purpose, an interdisciplinary collaboration of disciplines developing speech/language technology and disciplines with knowledge on speech/language development and child communication in addition to end-user involvement is beneficial [12, 13].

2 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 [2]. 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 these kind of data 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 illustrate this [10]. We compared modal performance in a test set of eight children across all four age groups of our kidsTALC corpus (AG1: 3;6 - 4;11; AG2: 5;0 - 6;11; AG3: 7;0 - 8;11; AG4: 9;0 - 10;11). Phone error rate was analyzed dependent on three types of elicitation contexts included in our kidsTALC corpus (narrative, picture description, conversational). Although all of these types represent connected speech the most unconstricted (natural) context is the conversational in free dialogue as marked by the highest average speech rate in all children except child K32 (Table 1). This context most representative of everyday communication also has the highest phone error rate (Table 2) thus emphasizing the need to collect these type of data to train ASR models for children.

Table 1 – Speech Rate in phones per seconds depending on the elicitation context for the children of the kidsTalc test set

	age	narrative	conversational	picture description
K32	3;9	-	6.7 (3.0)	6.7 (3.1)
K19	4;1	-	8.1 (3.0)	7.2 (3.2)
K4	5;8	-	9.3 (2.8)	7.6 (3.1)
K17	5;8	-	9.9 (3.9)	8.6 (3.5)
K12	8;1	8.1 (2.9)	-	7.4 (3.1)
K27	8;5	8.5 (2.4)	8.7 (5.6)	8.2 (4.6)
K30	10;7	9.1 (2.6)	10.7 (3.5)	8.8 (3.5)
K33	10;9	6.4 (1.6)	6.9 (4.1)	-

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 [14, 15, 16]. It is recommended that language samples of children should be collected in a location and in situations familiar to the children and that the activities and materials should be chosen with regard for the age, the fields of interest and the linguistic capacities of the children [17]. For example, interactions with the children with materials not necessitating fine motor activity are

Table 2 – PER of an ASR system trained only on kidsTalc, depending on the elicitation context for the children of the kidsTalc test set

	age	narrative	conversational	picture description
K32	3;9	-	33.4	30.5
K19	4;1	-	30.9	25.9
K4	5;8	-	34.5	27.8
K17	5;8	-	31.7	25.7
K12	8;1	29.0	-	34.1
K27	8;5	17.5	30.3	20.5
K30	10;7	31.9	51.2	37.7
K33	10;9	22.1	30.4	-

recommended because such demanding activities engender shorter utterances [17]. The context in which the language sample is gained was also found to have an influence on the amount of utterances and on the complexity of the language produced by children [15]. For example, while play-based activities may be appropriate to elicit a greater amount and more complex speech from younger children, in older children story telling may be the favorable context [8]. Table 3 provides an overview of recommended sampling contexts to collect continuous speech at different ages.

Table 3 – Recommended language sample contexts by age [2, 18].

	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

Furthermore, Evan and Craig [19] found, that the use of questions (in an interview context) resulted in more and more complex utterances than during free play with children. Different question types induce different answers and vary as to the syntax and the semantic relations of the expected replies [20]. Closed prompts have only one correct answer and lead to short responses (e. g. yes/no answers) while open-ended questions offer many possibilities of answers which are generally more than one- or two-word responses [21]. The latter are therefore more suitable to elicit longer and more complex utterances. In addition, the use of questions by the examiner was found to elicit more non-imitative utterances and more different words in children than commenting [22]. Finally, the reaction of the examiner after the production of child's utterances or after asking a question might have an impact on the child's language production. Examiners should leave children time, as wait time is important to enable children to produce (complete), more complex answers or produce extended talk [23]. The combination of an overall engaging communication (e.g. friendly and inviting tone of voice, eye contact, getting down to the child's level, use of gestures and vivid mimic), following the child's lead of interest, using open-ended and follow-up questions or substantive feedback with enough wait time, can be summarized as the recommended conversational style to make the child feel comfortable talking to you as an examiner and generate natural language production by the child [21, 17].

Additionally, considerations from an information science perspective should complement

guidelines for collecting child speech data. These may also address the location, the materials used, the elicitation method, and the conversational style of the person interacting with the child. Collecting language samples in a familiar environment of a young child, for example in the daycare center, where spontaneous speech data are often recorded, might present challenges for the development of an ASR software because of the background noise. The choice of elicitation materials must also undergo constrains as to its noisiness (e.g. soft toys would be preferable to favored plastic ones). Elicitation methods should be chosen with regard for the quality of the recordings (e.g. language samples with many children interacting may be too challenging at first) and for their ability to elicit longer and complete utterances. Finally, the elicitation method as well as the conversational style of the examiner and the feedback elements used should minimize 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, for example by the use of non verbal feedback in order to keep the conversation going.

Regarding sample length, it should be kept in mind that the younger the children the more recording time is needed to collect a sufficient amount of child speech. Per minute of child speech in the kidsTALC corpus on average 2.44 min of audio recording are required in the youngest age group (3;6 - 4;11), while only 1.33 min of audio recording are required in the oldest age group (9;0 - 10;11) almost doubling the amount of time it takes to record the the same amount of child speech from the youngest to the oldest children in our corpus.

3 Annotation Child Speech Data

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). This agreement should balance the needs of end-users, here the required detail of phonetic transcription for child speech assessment purposes, and the challenge of model training, the more detailed the more demanding. Additional aspects that need to be kept in mind when deciding on a certain level of detail in transcription are the effort to compile very detailed transcriptions and the decreasing inter-transcriber agreement with increasing detail leading to inconsistent transcriptions. 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, material used, speaker roles).

Ensuring communication between researchers collecting the data, those annotating (e.g. transcribing) and those training the ASR model is central for further processing child data. In our TALC project we use several tools and methods to foster communication between the various persons involved. These range from regular interdisciplinary meetings, end-user involvement in the project, shared manuals of individual process steps, to project management software tools such as version control systems. 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 may sometimes be missed without notification).

To reach an acceptable inter-transcriber agreement, which is generally lower in transcribing child data and additionally in phonetic transcriptions [24] training of transcribers and communication between transcribers is of utmost importance. 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 (oral) child speech and 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 transcription manual.

In the context of ARS software development, it is particularly important to pay attention to transcription accuracy. For example, overlapping portions should be accurately marked and separated from non-overlapping portions. Additionally, it is necessary that timestamps and speaker changes are aligned. Again, accurate work is necessary. In order to be able to guarantee a high standard, it has proven useful that each transcription and annotation step is controlled by another (trained) person. If possible, different transcribers should work on one transcript, in order for errors to be mutually controlled and quality to be assured in this way. Another option is the independent creation of a transcript by several persons in parallel, in order to be able to compare the results afterwards and to discuss a common best possible version. However, this is not very economical and often not feasible due to limited financial and human resources. Considerations for increasing quality should be discussed based on a cost-benefit analysis.

4 Conclusion

The interwoven sometimes opposing demands of those developing and those using machine learning software especially in human applications call for an interdisciplinary approach to model design as well as collecting and annotating data for model training. Regarding ASR for the assessment and linguistic analysis of child speech and language, a background in child language development and/or speech and language therapy should guide data collection and annotation in order to obtain robust and representative data. Communication between researchers of different disciplines and working on different aspects of a project is essential to address the challenges of automating child speech recognition as well as to develop software for end-user needs. Over the years of our project duration we have discovered that the most powerful tools of interdisciplinary collaboration are shared goals, mutually beneficial outcomes for all involved disciplines, openness to see the world through the eyes of the others and to learn each others "language" all reached in continuous cooperative dialogue.

References

[1] VONIATI, L., S. ARMOSTIS, and D. TAFIADIS: *Language sampling practices: A review for clinicians*. *Evidence-Based Communication Assessment and Intervention*, 15(1), pp. 24–45, 2021.

[2] NIPPOLD, M. A.: *Language Sampling With Children and Adolescents: Implications for Intervention, Third Edition*. Plural Publishing, 2020.

[3] PAVELKO, S. L., R. E. OWENS, M. IRELAND, and V. D. L. HAHS: *Use of Language Sample Analysis by School-Based SLPs: Results of a Nationwide Survey*. *Language, Speech, and Hearing Services in Schools*, 47(3), pp. 246–258, 2016.

[4] FENSON, L., P. S. DALE, J. S. REZNICK, E. BATES, D. J. THAL, S. J. PETHICK, M. TOMASELLO, C. B. MERVIS, and J. STILES: *Variability in Early Communicative Development*. *Monographs of the Society for Research in Child Development*, 59(5), pp. i–185, 1994.

[5] POTAMIANOS, A. and S. NARAYANAN: *Robust recognition of children's speech*. *IEEE Transactions on Speech and Audio Processing*, 11(6), pp. 603–616, 2003.

[6] HAZEN, T. J.: *Automatic Alignment and Error Correction of Human Generated Transcripts for Long Speech Recordings*. 2006.

[7] GEROSA, M., D. GIULIANI, and F. BRUGNARA: *Acoustic variability and automatic recognition of children's speech*. *Speech Communication*, 49(10), pp. 847–860, 2007.

[8] KNIGHT, R.-A., C. BANDALI, C. WOODHEAD, and P. VANSADIA: *Clinicians' views of the training, use and maintenance of phonetic transcription in speech and language therapy*. *International Journal of Language & Communication Disorders*, 53(4), pp. 776–787, 2018.

[9] ROY, B. C. and D. K. ROY: *Fast transcription of unstructured audio recordings*. *MIT web domain*, 2009.

[10] RUMBERG, L., C. GEBAUER, H. EHLERT, M. WALLBAUM, L. BORNHOLT, J. OSTERMANN, and U. LÜDTKE: *kidsTALC: A Corpus of 3- to 11-year-old German Children's Connected Natural Speech*. In *Proceedings INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, pp. 5160–5164. ISCA, 2022.

[11] LIAO, H., G. PUNDAK, O. SIOHAN, M. K. CARROLL, N. COCCARO, Q.-M. JIANG, T. N. SAINATH, A. SENIOR, F. BEAUFAYS, and M. BACCHIANI: *Large vocabulary automatic speech recognition for children*. In *Proceedings INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*, pp. 1611–1615. ISCA, 2015.

[12] KUSTERS, R., D. MISEVIC, H. BERRY, A. CULLY, Y. LE CUNFF, L. DANDOY, N. DÍAZ-RODRÍGUEZ, M. FICHER, J. GRIZOU, A. OTHMANI, T. PALPANAS, M. KOMOROWSKI, P. LOISEAU, C. MOULIN FRIER, S. NANINI, D. QUERCIA, M. SEBAG, F. SOULIÉ FOGELMAN, S. TALEB, L. TUPIKINA, V. SAHU, J.-J. VIE, and F. WEHBI: *Interdisciplinary Research in Artificial Intelligence: Challenges and Opportunities*. *Frontiers in Big Data*, 3, 2020.

[13] SARAVANAKUMAR, R. and S. PARIJA: *Artificial Intelligence and Interdisciplinary Research*. *Journal of Scientific Dentistry*, 11(2), pp. 43–43, 2021.

[14] KLEIN, H. B., N. MOSES, and B. R. JEAN: *Influence of Context on the Production of Complex Sentences by Typically Developing Children*. *Language, Speech, and Hearing Services in Schools*, 41(3), pp. 289–302, 2010.

[15] SOUTHWOOD, F. and A. F. RUSSELL: *Comparison of Conversation, Freeplay, and Story Generation as Methods of Language Sample Elicitation*. *Journal of Speech, Language, and Hearing Research*, 47(2), pp. 366–376, 2004.

[16] MILLER, J., F.: *Assessing Language Production in Children: Experimental Procedures*. University Park Press, 1981.

[17] KROECKER, J., K. LYLE, K. ALLEN, E. FILIPPINI, M. GALVIN, M. JOHNSON, A. KANUCK, S. LOCCISANO, C. MANONI, J. NIETO, L. FEENAUGHTY, C. SLIGAR, S. STAROWICZ, L. SZPAKOWSKI, K. WIND, S. YOUNG, and R. OWENS: *Effect of Student Training on the Quality of Children's Language Samples*. *Contemporary Issues in Communication Science and Disorders*, 37(Spring), pp. 4–13, 2010.

[18] PETZOLD, M. J., C. M. IMGRUND, and H. L. STORKEL: *Using Computer Programs for Language Sample Analysis*. *Language, Speech, and Hearing Services in Schools*, 51, 2020.

[19] EVANS, J. L. and H. K. CRAIG: *Language Sample Collection and Analysis*. *Journal of Speech, Language, and Hearing Research*, 35(2), pp. 343–353, 1992.

[20] JEAN-BAPTISTE, R., H. B. KLEIN, D. BRATES, and N. MOSES: *What's happening? And other questions obligating complete sentences as responses*. *Child Language Teaching and Therapy*, 34(2), pp. 191–202, 2018.

[21] HINDMAN, A. H., B. A. WASIK, and D. E. BRADLEY: *How Classroom Conversations Unfold: Exploring Teacher–Child Exchanges During Shared Book Reading*. *Early Education and Development*, 30(4), pp. 478–495, 2019.

[22] YODER, P. J., B. DAVIES, and K. BISHOP: *Getting children with developmental disabilities to talk to adults*. In S. F. WARREN and J. REICHLE (eds.), *Causes and Effects in Communication and Language Intervention*, no. 1 in Communication and Language Intervention Series, pp. 255–276. P.H. Brookes, 1992.

[23] WASIK, B. A. and A. H. HINDMAN: *Why Wait? The Importance of Wait Time in Developing Young Students' Language and Vocabulary Skills*. *The Reading Teacher*, 72(3), pp. 369–378, 2018.

[24] RAMSDELL, H. L., D. KIMBROUGH OLLER, and C. A. ETHINGTON: *Predicting phonetic transcription agreement: Insights from research in infant vocalizations*. *Clinical Linguistics & Phonetics*, 21(10), pp. 793–831, 2007.