My Science Tutor (MyST)–A Large Corpus of Children's Conversational Speech

Sameer S. Pradhan[1,2], **Ronald A. Cole**[3], **Wayne H. Ward**[4]

[1]cemantix.org, Cambridge MA, USA
[2]Linguistic Data Consortium, University of Pennsylvania, , Philadelphia PA, USA
[3]Boulder Learning Inc., Boulder CO, USA
[4]University of Colorado at Boulder, CO, USA
pradhan@cemantix.org

Abstract

This article describes the MYST corpus developed as part of the My Science Tutor project. To the best of our knowledge, this is one of the largest collections of children's conversational speech that is freely available for non-commercial use under the creative commons license (CC BY-NC-SA 4.0). It comprises approximately 400 hours of speech, spanning some 230K utterances spread across about 10,500 virtual tutor sessions. Roughly 1,300 third, fourth and fifth grade students contributed to this corpus. The current release contains a little over 100K transcribed utterances comprising close to 1.5M space separated transcribed tokens. It is our hope that the corpus can be used to improve automatic speech recognition models and algorithms. We report the word error rate achieved on the test set using a model trained on the training and development portion of the corpus. The git repository of the corpus contains the complete training and evaluation setup in order to facilitate a fair and consistent evaluation. It is our hope that this corpus will contribute to the creation and evaluation of conversational AI agents having a better understanding of children's speech, potentially opening doors to novel, effective, learning and therapeutic interventions.

Keywords: automatic speech recognition, speech corpora, children's conversational speech, virtual tutor, education

1. Introduction

According to the 2009 National Assessment of Educational Progress (NAEP, 2011), *only* 34 percent of fourth-graders, 30 percent of eighth-graders, and 21 percent of twelfth-graders in the U.S. performed at or above the *proficient level* in *science*. A more recent assessment, in 2019 (NAEP, 2021), reported a *statistically significant decrease* in the average score for fourth graders in science[1] compared with the most recent previous assessment, in 2015. Thus, approximately two thirds of students in the United States are not proficient in science[2].

This article describes a resource that was the result of a 13-year project conducted between 2007 and 2019. The project investigated improvements in students' learning proficiency in elementary school science using conversational multimedia virtual tutor, Marni. The operating principles for the tutor are grounded on research from education and cognitive science where it has been shown that eliciting self-explanations plays an important role (Chi et al., 1989, 1994, 2001; Hausmann and VanLehn, 2007a,b). Speech, language and character animation technologies play a central role because the focus of the system is on engagement and spoken explanations by students during spoken dialog with the virtual tutor. A series of studies conducted during this project demonstrated that students who interacted with the virtual tutor achieved substantial learning gains, equivalent to students who interacted with experienced human tutors, with moderate effect sizes (Ward et al., 2011, 2013) Surveys of participating teachers indicate that it is feasible to incorporate the intervention into their curriculum. Surveys given to students indicated that over 70% of students tutored by Marni were more excited about studying science in the future.

2. The MYST corpus

The MYST children's conversational speech corpus consists of spoken dialog between 3^{rd}, 4^{th} and 5^{th} grade students, and a virtual tutor in 8 areas of science. It consists of 393 hours of speech collected across 1,371 students. The collection comprises a total of 228,874 utterances across 10,496 sessions. It is freely available for research use[3] upon completion of a data use agreement.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B070008, and the National Science Foundation through Grants DRL 0733323 to Boulder Language Technologies and DRL 0733322 to the University of Colorado at Boulder. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education or the National Science Foundation.

[1] https://www.nationsreportcard.gov/highlights/science/2019/

[2] This does not consider the significant impact that the educational system experienced owing to the Covid-19 pandemic.

[3] https://myst.cemantix.org

LREC-COLING 2024, pages 12040–12045
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

2.1. Data Collection

As part of the study, students engaged in spoken dialog with a virtual science tutor—a lifelike computer character that produced accurate lip and tongue movement synchronized with speech produced by a voice talent. Analyses of the spoken dialog sessions indicated that, during a dialog of about 15 minutes, tutors and students produced about the same amount of speech, around 5 minutes each. This approach was used to develop over 100 tutorial dialog sessions, of about 15 minutes each. The students who participated in this study were enrolled in schools belonging to the Boulder Valley School District[4]. We did not record the gender, age or primary language of individual students. One could get a rough approximation of the age range based on the grades of the students in the study and information about the science modules. The MyST corpus was collected in two stages—Phase I and Phase II. In both phases, the scientific content covered is aligned to classroom science content of Full Option Science System (FOSS) modules, which typically last 8 weeks during the school year. FOSS is used by over 1 million children in over 100,000 classrooms in all 50 states in the U.S. FOSS modules are centered on science investigations. There are typically 4 Investigations in a module (e.g., in the Magnetism and Electricity module, the 4 investigations are Magnetism, Serial circuits, Parallel Circuits, and Electromagnetism). Each Investigation has 3 to 4 classroom "investigation parts" where groups of students work together to, for example, build a serial circuit to make a motor run, and record their observations in science notebooks. Shortly after conducting an "investigation part", students interact one-on-one with a virtual tutor for 15-20 minutes. The tutor asks the student questions about science presented in illustrations, animations or interactive simulations, with follow-up questions designed to stimulate reasoning and help students construct accurate explanations.

The system is *strict turn-taking*; the tutor presents information, asks a question and waits for the student to respond. Students wear headsets with close-talking noise-canceling microphones. To respond, the student presses the spacebar on the laptop, holds it down while speaking, and releases it when done. Each student turn is recorded as a separate audio file. When transcribed, an utterance level transcript file is created for each audio file.

2.2. Transcription

Roughly 45% of all utterances have been transcribed at the word level. Phase I of the project used rich (slow, expensive) transcription guidelines[5]—the ones typically used by speech recognition researchers. However, for the purposes of this project, that level of detail was not required in the transcriptions, and during Phase II, a reduced (quick, cheaper) version of those guidelines[6] was used, allowing transcription of more data.

2.3. Data Composition

Some characteristics of the data collected in the two phases are described below. Phase I comprised sessions from students in grades 3-5 across four science modules. All the sessions from this phase have been transcribed using rich transcription guidelines. Phase II comprised sessions from students in grades 4-5. It included five modules, with an average of 10 parts each. Table 1 lists the modules included in each phase. Table 2 lists the size of the corpus based on a few different parameters.

Table 1: List of science modules in Phase I and II

Phase	Module	Description
I	**MS**	Mixtures and Solutions
	ME	Magnetism and Electricity
	VB	Variables
	WA	Water
II	**EE**	Energy and Electromagnetism
	LS	Living Systems
	MX	Mixtures
	SRL	Soil, Rocks and Landforms
	SMP	Sun, Moon and Planets

Table 2: Size of MyST corpus.

Description	**Quantity** Count (Hours)
Phase I	
Number of Students	421 —
Number of Sessions	1509 (102)
Transcribed Sessions	1509 (102)
Untranscribed Sessions	0 (0)
Phase II	
Number of Students	950 —
Number of Sessions	8987 (102)
Transcribed Sessions	1426 (102)
Untranscribed Sessions	3711 (0)

2.4. Corpus Structure

The directory structure for the corpus is as shown in Figure 1 below. Variables are enclosed in angle-

[4] https://www.bvsd.org/

[5] https://cemantix.org/myst/phase-i-guidelines.pdf

[6] https://cemantix.org/myst/phase-ii-guidelines.pdf

Figure 1: The MyST Corpus Structure.

```
corpora/
└──[name]/
    ├── data/
    |   ├── <partition>/
    |   |   ├── <student_id>/
    |   |   |   └── <session_id>/
    |   |   |       ├── <session_id>.<file-extension>
    |   |   |       ├── ...
    |   |   |       ├── ...
    |   |   |       ├── ...
    ...
    ...
    ...
```

brackets (`<variable>`) and can take values as described immediately after.

`<partition>` is one of *train*, *development* or *test*. `<student_id>` is a 6-digit ID with the first 3 digits representing the school code and the next 3 digits the student number. `<session_id>` is the ID for a particular session and is further represented as `<corpus>_<student_id>_<date>_<time>_<module>_<investigation>.<part>`. `<date>` is represented as `<YYYY>-<MM>-<DD>`. `<time>` is represented as `<hh>-<mm>-<ss>`, where `<hh>`, `<mm>` and `<ss>` represent two digit hour, minutes and seconds respectively[7]. `<module>` is a two- or three-character string enumerated in Table 1 earlier. `<investigation>` is a decimal number representing the respective investigation for a module. `<part>` is the utterance ID within a session. Numbers `001` onward represent the index of each utterance in a session[8]. `<file-extension>` is one of `.flac` or `.trn`. `.flac` is the compressed audio file and `.trn` is the transcription of the corresponding audio file.

3. Data Cleanup and Pre-processing

We did a pass over the corpus to clean up various types of errors that could be identified using statistics on the underlying audio and potentially erroneous data collection.

3.1. Session Quality

Bad—empty or corrupted sessions were removed using simple heuristics and based on missing data.

3.1.1. Session Length

Sessions that were less than a certain minimal threshold ($<$ 10 minutes), or longer than a certain maximum threshold ($>$ 1 hour) were inspected and corrected or removed.

3.1.2. Missing audio files

Sessions that were missing audio files for a significant number of utterances were deleted.

3.2. Audio Quality

All utterances were processed to identify all possible unacceptable recordings and were removed from the database. We performed the following checks for audio quality.

3.2.1. Clipping Rate

If there was a significant number of clipped frames, we removed or marked the audio file. We removed the entire session from the release this number affected a significant fraction of utterances in a session. If only a small number of files had large fraction of clipping, we tagged them as such as part of the session metadata, so that end users can determine whether to include or exclude that data from their study.

3.2.2. Silence

Sometimes there are significant amounts of leading and trailing silence in the audio files. We trimmed all such silence except for a small fraction at the beginning and end of the utterance. We did not, however, remove or compress silence that occurred within an utterance.

3.2.3. Background Noise

Utterances with a significant amount of noise or cross talk were removed. This was only possible for the cases that were transcribed or were part of a sample that we manually verified.

3.3. Transcription Quality

We fixed obvious spelling errors in the transcriptions. We tried to retain explicitly mispronounced words as much as possible.

3.4. Updated Pronunciation Dictionary

We also make available an updated pronunciation dictionary. We used CMU's pronunciation dictionary as a starting point and added words that were novel to this corpus. The additions made to the pronunciation dictionary are made available is part of the corpus release.

[7] In Phase I, we did not capture hour/minute/second for each session, so the corresponding fields for sessions in Phase I are set to `00`

[8] `000` is reserved to represent the entire session.

4. Evaluation

In order to promote reproducible, fair and balanced evaluation of automatic speech recognition (ASR) models using this corpus, we partitioned and structured the corpus upfront into *train*, *development* and *test* sets.

4.1. Identifying Partitions

These partitions were identified using stratified sampling strategy thus ensuring that they reasonably represent each of the science module in MyST , proportionately represent each phase, and each student is present in only one of the three partitions. We also included untranscribed data in all partitions in order to be able to allow limited semi-supervised training data augmentation using the untranscribed portions of the data, with an additional advantage of pseudo-unseen data—in the form of transcriptions that are as yet absent.

Table 3: Distribution of modules and speech across the experimental partitions.

Phase	**Science Module**	**Partition**			**Overall**
		Train	**Dev.**	**Test**	
		(Hrs.)	(Hrs.)	(Hrs.)	(Hrs.)
I	MS	31	5	5	41
	ME	30	4	4	38
	VB	14	2	2	18
	WA	4	1	1	6
II	EE	114	16	14	144
	LS	75	4	4	83
	MX	29	5	7	41
	SRL	16	2	1	19
	SMP	2	1	1	4
	Overall	315	40	39	393

4.2. Experimental Setup

We used SpeechBrain (Ravanelli et al., 2021) speech toolkit for our experiments. More specifically, we used an end-to-end transformer model. We fine-tuned the model pre-trained on LibreSpeech model, using the MyST training set. Owing to memory limitations, we were only able to use utterances less than 30 secs. during training.

4.3. Word Error Rate

We use the traditional evaluation metric of word error rate (WER) to report ASR performance. In spite of several quality checks, an initial release of the corpus through Linguistic Data Consortium (LDC2021S05) contained some transcription errors. We corrected the transcription errors occurring in the test and use that to report the baseline WER in this article. Given the improvements in ASR models in the recent past, we were able to use some heuristics to identify utterances with errorful transcriptions from the training and development set. The number of utterances with transcription errors was a small fraction of the total transcribed utterances. We trained the ASR model using the training and development set after filtering out those utterances. Given the small fraction of utterances with errorful transcriptions, we did not see a noticeable difference between the WER using models trained before and after filtering respectively. Table 4 shows the WER on the corrected test set transcriptions. We plan to make another quality control pass through the corpus to correct residual errors in the development and training set and release an updated version of the corpus in the near future.

4.4. Replicability

We understand the importance of ensuring that the research community can replicate and monitor the performance improvements on this data over time. In order to facilitate that, we are making available the data, the evaluation setup—the model architecture, WER evaluation program and all relevant configuration–in a `git` repository[9] of the corpus.

Table 4: Word Error Rate on the MyST test set using a model trained only on the training and development partitions.

MyST **Test Set**	WER (%)
WER	10.0
Insertions	2.9
Substitutions	5.1
Deletions	3.2

5. Related Work

Over the years researchers have created several speech corpora for the analysis of children's speech. Below are a few that are typically used for ASR evaluation. A thorough empirical evaluation of various end-to-end ASR systems specifically focused on children's speech was recently reported in Shivakumar and Narayanan (2022). They used a pre-release of this corpus in their study which did not contain the current experimental partitions, so their evaluation numbers are not directly comparable.

- CID children's speech corpus (American English, read speech, 436 children aged between 5 and 17 years) (Lee et al., 1999)

[9] `https://myst.cemantix.org`

- CMU Kid's speech corpus (American English, read speech, 76 children, aged between 6 and 11 years) (Eskenazi, 1996)
- CU Kid's Prompted and Read Speech corpus (American English, read speech, 663 children, aged between 4 and 11 years) (Cole et al., 2006),
- CU Kid's Read and Summarized Story corpus (American English, spontaneous speech, 326 children, aged between 6 and 11 years) (Cole and Pellom, 2006),
- OGI Kid's speech corpus (English, read speech, 1100 children, aged between 5 and 15 years) (Shobaki et al., 2000).
- BIRMINGHAM corpus (British English, 159 children, aged between 4 and 14 years, part of corpus PF-STAR) (D'Arcy and Russell, 2005)

6. Conclusion and Future Work

In this work we describe a large corpus of conversational children's speech and present a baseline WER using a state of the art ASR system. We are making this corpus freely available for research through a `git` repository. This should make it easier for users to identify and propose corrections to any residual transcription errors. With the help of sponsors and volunteers from the larger research community, we hope to manually transcribe the untranscribed utterances. We recommend that future users use this data repository as the definitive version of the corpus and the relevant documentation.

In spite of an exponentially large collection of data at our finger tips, it is difficult to get access to a reasonably large collection of specific kinds of data needed to train accurate end-to-end machine learning models. In this case that is *children's conversations*. One of the larger models for reporting performance on children's speech (Liao et al., 2015) used roughly 20K training utterances. However, the data underlying for that study is not generally available and the work cannot be replicated. Our hope is that the large MYST corpus of children's conversational speech will allow researchers to improve upon a consistent evaluation benchmark. Improvements in automatic transcription of children's conversations can open doors to transformational applications in various domains. Improved applications for use in education and in healthcare have the potential of making a significant global impact.

7. Ethics Statement

The The University of Colorado's Institutional Review Board reviewed and approved all components of the My Science Tutor (MyST)project to assure student privacy. The review board approved the Parental Consent forms and the Student Assent forms. The final Parental Consent and Student Assent forms approved by the IRB explicitly provide permission for the distribution of anonymous student speech data and transcriptions. We manually verified that we had parental consent and student assent for every student in the released corpus. No identifying information is stored with the data. All school codes and student IDs were anonymized.

8. Bibliographical References

M. Chi, M. Bassok, M. Lewis, P. Reimann, R. Glaser, and Alexander. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2).

M. Chi, N. De Leeuw, M. Chiu, and C. LaVancher. 1994. Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3):439–477.

M. T. H. Chi, S. A. Siler, H. Jeong, T. Yamauchi, and R. G. Hausmann. 2001. Learning from human tutoring. *Cognitive Science*, 25(4):471–533.

R. Cole, J. P. Hosom, and B. Pellom. 2006. University of colorado prompted and read children's speech corpus. Technical report.

R. Cole and B. Pellom. 2006. University of colorado read and summarized stories corpus. Technical report.

Shona D'Arcy and Martin Russell. 2005. A comparison of human and computer recognition accuracy for children's speech. In *Proc. Interspeech 2005*, pages 2197–2200.

Maxine S Eskenazi. 1996. *Kids: a database of children's speech*. Ph.D. thesis, Acoustical Society of America.

R. G. M. Hausmann and K. VanLehn. 2007a. Explaining self-explaining: A contrast between content and generation. *Artificial Intelligence in Education*, pages 417–424.

R. G. M. Hausmann and K. VanLehn. 2007b. Self-explaining in the classroom: Learning curve evidence. In *29th Annual Conference of the Cognitive Science Society*, Mahwah, NJ.

Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468.

Hank Liao, Golan Pundak, Olivier Siohan, Melissa Carroll, Noah Coccaro, Qi-Ming Jiang, Tara N Sainath, Andrew Senior, Françoise Beaufays, and Michiel Bacchiani. 2015. Large vocabulary automatic speech recognition for children.

NAEP. 2011. *The Nation's Report Card: Science 2009*. Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

NAEP. 2021. *The Nation's Report Card: Science 2019*. Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. 2021. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Prashanth Gurunath Shivakumar and Shrikanth Narayanan. 2022. End-to-end neural systems for automatic children speech recognition: An empirical study. *Computer Speech & Language*, 72:101289.

Khaldoun Shobaki, John-Paul Hosom, and Ronald Cole. 2000. The ogi kids' speech corpus and recognizers. In *Proc. of ICSLP*, pages 564–567. Citeseer.

W. Ward, R. Cole, D. Bolanos, C. Buchenroth-Martin, E. Svirsky, S. V. Vuuren, and L. Becker. 2011. My science tutor: A conversational multimedia virtual tutor for elementary school science. *ACM Trans. Speech Lang. Process.*, 7(4).

Wayne Ward, Ron Cole, Daniel Bolanos, C. Buchenroth-Martin, E. Svirsky, and Tim Weston. 2013. My science tutor: A conversational multimedia virtual tutor. *Journal of Educational Psychology*, 105(4):1115–1125.