MyST Children's Conversational Speech Corpus

v0.1.0-ec3acab

The My Science Tutor (MyST) corpus consists of roughly 300 hours of children's speech. More than half of it has been transcribed at the word level.

Data Collection

The MyST corpus was collected in 2 stages, Phase I and Phase II

In both phases, the content covered is aligned to 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, Series 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 series circuit to make a motor run, and record their observations in science notebooks. Shortly after conducting an investigation part, students interact 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. 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, a session level transcript file is created for each audio file. No identifying information is stored with the data, only code numbers for schools and students. All students and their parents signed consent forms allowing Boulder Learning to enter and distribute their anonymous speech data.

The file structure for both corpora is

```
corpora/myst/data/<student_id>/<session_id>/<session_id>.<file-extension>
```

Phase I

The Phase I corpus contains speech from students in grades 3-5. All of the sessions have been transcribed using the transcription guidelines document in the `doc` folder of the release.

1. ME - Magnetism and Electricity
2. MS - Mixtures and Solutions

3. VB - Variables
4. WA - Water

There was no attempt to have any individual student cover all of the parts for a module. The focus of the collection was to get a wide variety of students rather than try to get complete coverage of material for individual students.

Phase II

The Phase II corpus contains sessions from students in grades 4-5. It uses 5 modules, with an average of 10 parts each. About half of the sessions have been transcribed using a simpler version of the same guidelines used for Phase I which excluded marking disfluencies.

1. EE - Energy and Electromagnetism
2. MX - Mixtures
3. SMP - Sun, Moon and Planets
4. SRL - Soil, Rocks and Landforms
5. LS - Living Systems

In this collection, teachers were asked to have students complete all parts for 2 modules, however, many teachers did not want to cover 2 modules and whatever data was collected was kept, even if students didn't complete the sequence.