Using TalkBank and CHILDES

May 22, 2012

Presenters:

Brian MacWhinney

Tutorial Programme/Overview

This tutorial will survey TalkBank public open-access corpora and the computational tools that have been developed for their analysis. These are the largest available corpora for spoken language data. Materials for this workshop can all be downloaded from:

http://childes.psy.cmu.edu and
http://talkbank.org

CHILDES, which is the largest single component of TalkBank, contains 60 million words of child-adult conversation across 26 languages; the adult segment of TalkBank includes 63 million words of adult-adult conversation with the bulk in English. All of the data are in a format specified by a detailed XML schema. As such, this is the largest consistently transcribed database of spoken language materials. Nearly all of the transcripts in TalkBank are linked on the utterance level to either audio or video. For CHILDES, about 25% is linked to media.

There are three basic goals for this tutorial:

1. The first goal is to learn to use the resources already on the web for corpus analysis. This involves a quick overview of the 384 corpora in the database across the web, and review of the corpus analysis facilities available in the CLAN programs.
2. The second goal is to explain the construction of the database, XML, and programs from a technical point of view.
3. The third goal is to learn how to produce new transcripts and annotations for inclusion in the database and analysis through CLAN. This involves learning to use the editor for speech and gesture transcription, and how to use the MOR and GRASP programs for adding morphosyntactic coding.

Because we will be installing programs and downloading data, we recommend that participants bring their own laptop computers with wireless access to the workshop, if possible.

Tutorial Description/Outline/Contents

Because we will be installing programs and downloading data, we recommend that participants bring their own laptop computers with wireless access to the workshop, if possible.

The specific topics to be covered in the workshop are, in order of presentation:

A. Learning the overall system and programs:

1. The relation of CHILDES to TalkBank.
2. Connecting to the web for downloading and browsing of TalkBank data.
3. Principles of TalkBank data sharing and IRB.
4. Downloading and using the CHAT and CLAN manuals and the database manuals.
5. Installing CLAN, setting up a working directory.
6. Running basic CLAN programs (MLU, FREQ, KWAL) on sample corpora from CHILDES and AphasiaBank.
7. Running package analyses through MORTABLE, FREQ, and EVAL.

B. Explaining the technical structure of the system.

1. The setup of the QuickTime Streaming server and the TalkBank Browser.
2. TalkBank XML and validation through Chatter.
3. Interoperability to Phon, ELAN, Praat, and other programs.
4. Metadata for OLAC and IMDI.

C. Understanding how to create new transcriptions and annotations.

1. POS Tagging of TalkBank corpora using the MOR program for 12 languages.
2. Tagging of bilingual corpora.
5. Development of new MOR taggers.
6. Dependency Parser tagging based on use of the MOR-coded POS tags.
7. Transcription in CLAN for Conversation Analysis, using Sonic CHAT and Sound Walker.
8. Gesture Analysis in CLAN.