LREC2012 Tutorial on Using the TalkBank and CHILDES Databases and Tools
Brian MacWhinney -- Carnegie Mellon University

Tuesday May 22, morning session
Lütfi Kirdar Istanbul Exhibition and Congress Centre

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. 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. The tutorial will review these issues:

1. The CHAT format for TalkBank data.
2. Access to TalkBank corpora through the web browser.
3. The conversion of CHAT data to XML through Chatter.
4. Analysis through the CLAN programs.
5. Transcription in CLAN and Conversation Analysis.
6. Interoperability between CHAT, ELAN, EXMaRALDA, Praat, and other systems.
7. POS Tagging of TalkBank corpora using the MOR program for 12 languages.
8. Tagging of bilingual corpora.
9. Development of new MOR taggers.
10. Dependency Parser tagging based on use of the MOR-coded POS tags.
11. Gesture Analysis in CLAN.