The following information was communicated by Prof. Brian MacWhinney on the info-childes listserv. It is reproduced here with his permission.

Dear Info-CHILDES,

Here is a quite lengthy update on several major advances in the CHILDES and TalkBank projects over the last year. Apologies for multiple postings to mailing lists. Here we call attention to each of six major developments.

1. Database Additions: Over the last year we have added several new child language corpora:
• From Mits Ota, the Edinburgh corpus of 20 UK children followed across two years,
• From Marilyn Nippold, a corpus of older children discussing how they learned chess,
• From Martine Sekali, a longitudinal video case study of two British children,
• For African-American English, we added corpora from Claire Cameron, DSLT (the DELV Project), and Isabel Barriere.
• Kathy Tamis-Lemonda contributed the password-protected Lego video corpus.
• From Leslie Rescorla, a longitudinal corpus from 38 late talkers and 22 controls.
• From Elena Tribushinina, a corpus from nearly 1000 Russian-Dutch bilingual children.
• From Liesbeth Schlichting and Jacqueline Van Kampen, a longitudinal corpus from four Dutch children
• From Rob Zwitserlood, a corpus of 150 DLD and TD Dutch-speaking children ages 4-8.
• For German, Nikolas Koch's longitudinal study of four children.
• From Gisela Szagun, a revised version of her longitudinal German corpus.
• From Velka Popova, additional subjects and material for the longitudinal corpus for Bulgarian.
• From Erika Hoff, a corpus of Frog Stories from Spanish-English bilingual children in Miami.
• From Liliana Tolchinsky, the GRERLI Spanish corpus from older children and teenagers.
• From Gordan Hrzica and Jay Roch, a corpus of MAIN story descriptions in Croatian.
• From Nino Tsintsadze and colleagues in Tbilisi, a longitudinal study of four children learning Georgian
• From Claire Cameron at UBuffalo, conversations and Frog Story descriptions
• apologies if I forgot any

2. Automatic Speech Recognition (ASR): In the summer of 2022, we worked with Houjun Liu to apply the Rev-AI ASR system and the Montreal Forced Aligner (MFA) to CHILDES and other TalkBank data using a Python script. This system has been remarkably successful, reducing transcription time to about 4 times recording time. We are now using it to automatically transcribe new data, transcribe untranscribed audio, and time-align older TalkBank data. An open-access article describing this "Batchalign" system is now available at https://doi.org/10.1044/2023_JSLHR-22-00642 and we have made the Batchalign system publicly available at https://github.com/talkbank . We are happy to provide email and Zoom support for users who want to explore use of this system.

3. Universal Dependency Tagging: Using other facilities in a development version of the same Batchalign pipeline, we applied grammatical dependency taggers from the Universal Dependencies (UD) project to several additional languages in CHILDES. Specifically, we used UD taggers available through the Stanza interface to tag almost all of the data in CHILDES for Danish, Dutch, Afrikaans, Turkish, Swedish, Norwegian, and Icelandic. We invite speakers of these languages to evaluate the accuracy of these taggers. The tagging produces %mor and %gra lines that conform with CHAT, but the tags themselves

are slightly different from those assigned by the MOR grammars in CHILDES. This work is very exciting, because it allows us to integrate work in CHILDES with work in the wider Computational Linguistic community that relies increasingly on crosslinguistic comparison based on UD tagging. UD taggers are available for over 100 languages, including these languages in CHILDES for which MOR taggers are not available: Irish, Welsh, Thai, Indonesian, Korean, Bulgarian, Croatian, Czech, Polish, Hungarian, Russian, Serbian, Slovenian, Arabic, Basque, Estonian, Farsi, Greek, Quechua, and Tamil. Being able to apply UD to these additional languages opens up CHILDES to much broader and more powerful crosslinguistic comparisons. We can make this system available from github to interested users. Although UD does a great job in terms of tagging grammatical relations, its analysis on the morphological level is less complete than that produced by MOR and we will keep both MOR and UD systems available for users.

4. KIDEVAL, IPSyn, and DSS: Last year, working with Jenny Roberts and Evelyn Altenberg, we published a version of automatic IPSyn that corrects for a variety of errors. This year, Ji Seung Yang and Nan Bernstein pursued further psychometric analysis of IPSyn items and found that two of the scales led to negative correlations with age and should therefore be eliminated. We are working on a similar analysis now for DSS. We are also receiving testing input from Pam Hadley at Illinois to correct residual errors in IPSyn and to automatically compute Subject-Verb diversity.

5. TalkBankDB: Last year we completed initial development of the TalkBankDB data base search engine system at https://talkbank.org/DB . This year, we found that use of this system was so heavy that it put a strain on our servers' capacities. To deal with this, we rewrote the server access code to offload computations to the client machine and to avoid reduplicative queries. We also installed a larger amount of memory on the server and now we are no longer experiencing any crashes. Going forward, we will be adding additional methods for cross-domain analyses for phonology, lexicon, syntax, and discourse. We are also continuing expansion of support for direct analysis of TalkBankDB data from R and Python.

6. Collaborative Commentary: Based on funding from NSF, we are continuing development of the Collaborative Commentary (CC) system at https://talkbank.org/CC . CC allows project groups to create a set of tags for language behaviors and locate instances of those tags in CHILDES data available directly through the TalkBankBrowser in the web. Eight research groups are using the alpha version of this system for teaching and research. Three are using CC for data from AphasiaBank, one for ClassBank data, one for DementiaBank data, and three for CHILDES data. Four instructors have used the system with students to teach coding and commentary for Conversation Analysis, discourse analysis, error analysis, and developmental milestone tracking.

Finally, we have plans to implement by the end of the summer a more standard and comprehensive authentication system for TalkBank.

Best regards,

— Brian MacWhinney
Teresa Heinz Professor of Cognitive Psychology,
Language Technologies and Modern Languages, CMU