TalkBank within CLARIN

Brian MacWhinney
Department of Psychology
Carnegie Mellon University, Pittsburgh USA
macw@cmu.edu

Abstract

TalkBank promotes the use of corpora, web-based access, multimedia linkage, and human language technology (HLT) for the study of spoken language interactions in a wide variety of discourse types across many languages, involving children, second language learners, bilinguals, people with language disorders, and classroom learners. Integration of these materials within CLARIN provides open access to access a large amount of data to support researchers as well as a good test bed for the development of new computational methods.

1 Introduction

The TalkBank system (http://talkbank.org) provides web-based access to spoken language corpora and resources for a variety of research areas in Psychology, Linguistics, Education, and Speech Pathology. There are currently seven funded TalkBank projects. The National Institutes of Health (NIH) funds the development of the CHILDES database (http:childes.talkbank.org) for the study of child language development (MacWhinney, 2000), PhonBank (phonbank.talkbank.org) for the study of phonological development (Rose & MacWhinney, 2014), AphasiaBank (aphasia.talkbank.org) for the study of language in aphasia (MacWhinney & Fromm, 2015), and FluencyBank (fluency.talkbank.org) for the study of the development of fluency and disfluency in children and language learners. The National Science Foundation (NSF) provides additional funding for FluencyBank, as well as funding for HomeBank (http://homebank.talkbank.org) with daylong audio recordings in the home (VanDam et al., 2016). The National Endowment for the Humanities (NEH) and the Deutsche Forschungs Gesellschaft (DFG) provide funding for web-based access to materials from Classical Latin and Historical German (http://sla.talkbank.org). In addition to these seven funded projects, TalkBank has developed resources for TBIBank (traumatic brain injury), RHDBank (right hemisphere damage, ASDBank (autism), DementiaBank (dementia), CABank (conversation analysis), SamtaleBank (Danish), GestureBank (gesture), SLABank (second language acquisition) (MacWhinney, 2015), BilingBank (bilingualism), ClassBank (classroom interactions), and TutorBank (tutoring interactions). All of these resources use a common transcript format called CHAT which is linked to the CLAN analysis programs and other open-access resources. Except for some of the corpora from clinical areas and the daylong recordings from the home, these resources are available without passwords.

TalkBank includes 328 corpora contributed by researchers across all of these fields. In many cases, after corpora have been contributed, they undergo additional reformatting, curation, indexing, annotation, and linkage to media. The result is a unified open-access database with a fully consistent system of transcription and annotation across all corpora. We believe that this type of data integration is important for maximizing the value of the corpora contributed to TalkBank and that this method can serve as a model for other CLARIN data sites.

In 2014, the TalkBank center at Carnegie Mellon University in Pittsburgh became a CLARIN-B site, and in 2016 it became a CLARIN-K site. This is the first CLARIN site outside of Europe. This paper will summarize the ways in which TalkBank has implemented CLARIN standards and how it can provide resources for the larger CLARIN community.

2 Operation and use of the CLARIN infrastructure

The various TalkBank databases have each generated large amounts of scholarly productivity. The CHILDES database, which began in 1984, has generated over 6500 published papers. PhonBank, which began in 2007 has generated 480 papers, and AphasiaBank has generated 256 papers. We have not conducted output monitoring for the databases that are not grant supported, such as CABank, but these are also widely used in many ongoing research projects. Since 2003, the childes.talkbankorg website has received over 4 million hits since 2003 and the talkbank.org website has received over 1.2 million hits.

Apart from figures on web traffic, our monitoring of usage of the databases relies primarily on the requirement in the Ground Rules (http://talkbank.org/share) that published papers cite specific articles describing parts of the TalkBank system. We then conduct searches in scholar.google.com for articles that cite these sources, thereby allowing us to track usage through published papers. This method does not pick up the use of the data in student papers, talks, and non-English materials, but it is an accurate overall indicator of impact.

TalkBank's approach to sustainability focuses on integrating our corpora and tools with the basic research agenda of each of our participating language research communities. To the degree that we achieve such integration, funding for our work is tied to ongoing funding for basic research. For example, when developing tools for the study of child language development, we focus on methods for automatic morphosyntactic coding, because of the importance of grammatical analysis in language acquisition theory. For aphasia, we focus on both morphosyntax, lexical access, error analysis, and aspects of fluency. For the projects on disfluency and stuttering, we work on the application of tools for automatic speech recognition (ASR), including diarization and word-level alignment in order to characterize the linguistic environment and distribution of disfluencies. We also seek to achieve sustainability and survivability by using open-source software tools with full documentation and by linking to tool chains in the CLARIN infrastructure.

3 Design and construction of the CLARIN infrastructure

Our goals for TalkBank-CLARIN integration are to make TalkBank materials fully accessible and discoverable for CLARIN users, and to integrate CLARIN tools into the TalkBank analysis chains.

In the context of achieving CLARIN-B Centre status and Data Seal of Approval (DSA) recognition, we implemented a method for automatic generation of CMDI metadata for the 4562 items (transcripts and media) in the TalkBank repositories. This script uses Arbil to create metadata which then is further configured by a Java program for OAI-PMH server publication to VLO. We create permanent IDs (PIDs) through both Handle Server and the EZ-Cite/Datacite facility for creating Digital Object Identifiers (DOIs). Both DOI and ISBN numbers are included on HTML pages for each corpus in TalkBank, and the accuracy and consistency of the PIDs, DOIs, and ISBNs is verified through a local Java database system.

We have implemented CLARIN single sign on through Shibboleth and the InCommon system used in the United States. However, the compatibility of this system with current CLARIN SSO systems has not yet been fully tested and verified. One potential problem here is that several of the databases require special community-specific passwords and that method has not yet been established through CLARIN. In addition, many TalkBank materials can be accessed without a password, making it relatively less important to have CLARIN login functional. However, it will be helpful to have this system working, because we are interested in limiting access to some materials to verified academics and the CLARIN system would allow for this. If we only implement this through InCommon access, we might be blocking access from Europe.

Our current software for word-level search relies on running the CLAN programs on data on the web. Although these are not database programs, they run fairly rapidly. However, to provide fuller database and corpus facilities, we have created a Pepper importer from CHAT data to ANNIS (http://corpus-tools.org) as well as a local ANNIS server (http://gandalf.talkbank.org:8080/annis-gui-3.4.4/). Hopefully, CLARIN will be able to integrate ANNIS or similar tools into its resource framework.

For alignment, we experimented with the use of WebMAUS, but we have achieved better results with SpeechKitchen (http://speechkitchen.org) and we will pursue that approach (Metze, Riebling,

Warlaumont, & Bergelson, 2016). It may be possible for CLARIN to make use of the SpeechKitchen framework, once we have tested it and developed it for more general use.

Our current approach to interoperability has focused on the writing of bi-directional converters between the CHAT format and ELAN, EXMARaLDA, Transcriber, LENA, Pepper, SALT, Phon, CONLL, and Praat.

4 CLARIN Knowledge infrastructure and dissemination

In June, TalkBank was recognized as a CLARIN-K Centre for Knowledge distribution. We seek to provide users with knowledge and help in the analysis of spoken language texts using either CLAN or any of the other software analysis system with which CLAN and CHAT are compatible. We can offer support through email, mailing lists, and phone with extremely quick turnaround. We have been creating online screencasts demonstrating the use of TalkBank tools, and we welcome suggestions for the creation of additional methods. These resources can become particularly important if CLARIN seeks to provide a higher level of support for the study of spoken language interactions.

5 Relations to Other Infrastructure

In the United States, the major alternative method for the archiving of language corpora is the Linguistic Data Consortium at the University of Pennsylvania. We have collaborated with LDC on various projects and shared several corpora. All TalkBank data are open for analysis and inclusion in LDC, as they are included in the MPI corpus collection within CLARIN. Unlike LDC, TalkBank emphasizes the use of a common data format and corpus structures across all corpora, thereby increasing interoperability and the possibility for comparison across corpora.

TalkBank has also contributed to the construction of segments of the SamtaleBank corpus for spoken Danish and we hope to continue the construction of this resource.

6 Challenges

The major challenge currently facing TalkBank integration into CLARIN is a fiscal challenge. Because the United States is not a member of the European Union, it has no clear mechanism for provided financial support for CLARIN membership. We hope that, by creating a CLARIN Infrastructure with multiple research sites in the United States, such as Brandeis, UPenn, Illinois, or Columbia, in addition to CMU, we can provide additional motivation and justification for NSF support for CLARIN.

We also seek to coordinate our activities with the LAPPS project (Ide et al., 2015) which may be able to achieve forms of tool integration that include both TalkBank and other CLARIN use cases. We are currently using new data from FluencyBank as a test use case for the LAPPS tool chain.

7 Conclusion

TalkBank plays an important role within the larger CLARIN infrastructure in terms of providing resources for the analysis of spoken language interactions. Unlike many other resources in this area, TalkBank resources are available through completely open access and rely on a consistent data format. These features may serve as a model for other resources in CLARIN.

References

Ide, N., Pustejovsky, J., Cieri, C., Nyberg, E., DiPersio, D., Shi, C., . . . Wright, J. (2015). *The language application grid.* Paper presented at the International Workshop on Worldwide Language Service Infrastructure.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. 3rd Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2015). Multidimensional SLA. In S. Eskilde & T. Cadierno (Eds.), *Usage-based perspectives on second language learning* (pp. 22-45). New York, NY: Oxford University Press.

MacWhinney, B., & Fromm, D. (2015). AphasiaBank as Big Data. *Seminars in Speech and Language, 37*, 10-22.

Metze, F., Riebling, E., Warlaumont, A. S., & Bergelson, E. (2016). *Virtual machines and containers as a platform for experimentation*. Paper presented at the Cognitive Science, Philadelphia, PA.

Rose, Y., & MacWhinney, B. (2014). The PhonBank Project: Data and software-assisted methods for the study of phonology and phonological development. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford handbook of corpus phonology* (pp. 380-401). Oxford: Oxford University Press.

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., Palma, P. D., & MacWhinney, B. (2016). HomeBank: An online repository of daylong child-centered audio recordings. *Seminars in Speech and Language*.