ASHA American Speech-Language-Hearing Association

AJSLP

Research Note

Collaborative Commentary for Understanding Communication Disorders

Brian MacWhinney[a] and **Davida Fromm**[a]

[a] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA

ARTICLE INFO

Article History:
Received December 2, 2022
Revision received February 28, 2023
Accepted April 21, 2023

Editor-in-Chief: Katherine C. Hustad
Editor: Michael de Riesthal

https://doi.org/10.1044/2023_AJSLP-22-00385

ABSTRACT

Purpose: The goal of the Collaborative Commentary (CC) system is to make the TalkBank adult clinical databases—including AphasiaBank, DementiaBank, RHDBank, and TBIBank—open to commentary and analysis from the full community of researchers, instructors, students, and clinicians.
Method: CC allows a group leader to establish a commentary group and invite colleagues or students to join as members of the group. Members can then browse through the transcript database using the TalkBank Browser. When they wish to insert a comment, they click on the utterance line number or drag the cursor across a range of utterances and a window opens to receive the comment. The comment can include open text along with codes selected from a predefined set of codes created by that commentary group.
Results: CC was released for public use in August 2022. It is being used currently in five research projects and eight classes. An important feature of CC is its ability to evaluate the reliability of coding systems and to sharpen analytic categories. By familiarizing instructors and researchers with the capabilities of CC, we expect to see an increasing usage of CC for a variety of clinical and research applications.
Conclusions: CC can contribute to a better understanding of connected speech features in aphasia, dementia, right hemisphere disorder, and traumatic brain injury. CC represents an extreme innovation not only for the study of adult neurogenic communication disorders but also for the study of spoken language generally.

Collaborative Commentary (CC) is a tool that allows groups to collaboratively code and comment on transcripts in the TalkBank databases. TalkBank is the largest publicly available source of data on spoken language interactions. Of the 14 components databases in TalkBank, there are seven that include data relevant to the study of language disorders. These include AphasiaBank for aphasia, DementiaBank for dementia, FluencyBank for stuttering, RHDBank for right hemisphere disorder (RHD), TBIBank for traumatic brain injury (TBI), and the child clinical data in CHILDES (Child Language Data Exchange System) for child language disorders and PhonBank for child phonological disorders. TalkBank data have been used in over 9,500 publications. The corpora in TalkBank have been generously contributed from projects all around the world, from over 1,500 participants speaking any one of 34 languages. All data in TalkBank were contributed with informed consent. TalkBank data are coded in a consistent format, and the transcripts and media can be played back directly through the TalkBank Browser. This open accessibility makes TalkBank data an excellent target for collaborative commentary and analysis.

The CC system builds upon this TalkBank computational infrastructure to provide a new way of understanding communication in people with language disorders. Using CC, researchers, clinicians, and students can access video recordings of spoken language interactions in the browser, watch the video, follow along with the linked transcript, and enter codes or comments directly attached to utterances in the transcript. These codes or comments are then visible to everyone in the commentary group—which may be the class, the research group, or the clinical trainees. The rapid development and ubiquitous presence of broadband web

Correspondence to Brian MacWhinney: macw@cmu.edu. **Publisher Note:** This article is part of the Special Issue: Select Papers From the 51st Clinical Aphasiology Conference. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

connections offer enormous opportunities for collaborative learning and theory testing. No longer limited to working on isolated computer files, groups of people can now focus their mutual attention on a common set of interactions in a shared database to promote deeper and wider understandings, test theories, and elaborate analysis systems.

History

Nearly 2 decades ago, MacWhinney et al. (2004) proposed the design of a CC system for spoken language data. However, the web tools available for spoken language at that time made development difficult. During the same period, the W3C (WorldWide Web) Annotea project at http://www.w3.org/2001/Annotea developed tools for web-based commentary on written language materials. The development of commentary tools for written materials has continued with systems such as AlvisAE (Papazian et al., 2012), Cromer (Girardi et al., 2014), INCEpTION (Boullosa et al., 2018), Mimir at http://demos.gate.ac.uk/mimir/ (Tablan et al., 2015), MTAS (Multi Tier Annotation Search) at https://meertensinstituut.github.io/mtas/ (Brouwer et al., 2017), NeuroCurator (O'Reilly et al., 2017), and WebAnno at https://webanno.github.io/webanno/ (de Castilho et al., 2016). However, these tools only work on written documents, and most do not provide methods for group-based commentary. The Hypothesis project at https://web.hypothes.is provides commercial tools that can be adopted in classrooms. However, Hypothesis tools are still limited to commentary on written documents on the web.

The tradition of qualitative data analysis has also led to the development of widely used commercial systems, such as ATLAS.ti at https://atlasti.com/, TransAna at https://www.transana.com/, Lexalytics at https://www.lexalytics.com/, MAXQDA (MAXimum Qualitative Data Analysis) at https://www.maxqda.com/, Dedoose at https://www.dedoose.com/, Quirkos at https://www.quirkos.com/, and NVivo at https://lumivero.com, for commenting on files on individual computers. However, none of these systems is linked to openly shared data, and most work mainly on individual files on the desktop. A review of these systems can be found at https://www.predictiveanalyticstoday.com/top-qualitative-data-analysis-software/.

To support responsible web-based commentary for openly shared data on spoken language interactions, a fully functional CC system must comply with a wide range of open science standards for data sharing, replicability, survivability, and consistency. These standards are embodied in the principles of the Core Trust Seal (CTS) at https://coretrustseal.org, which is promoted by the National Institutes of Health (NIH) Office of Data Science Strategy and the NIH/National Library of Medicine system of trusted NIH repositories, as well as the standards articulated in the FAIR (Findable, Accessible, Interoperable, Reusable; Wilkinson et al., 2016) and TRUST (Transparency, Responsibility, User Focus, Sustainability, Technology; Lin et al., 2020) principles promoted by NIH Office of Data Science Strategy and CTS. Based on adherence to all of these standards, TalkBank has received the CTS and has been designated as a trusted NIH repository at https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html. The features required for this type of responsible web-based commentary system are as follows.

1. There must be open web access to transcripts of spoken language interactions linked to media. The TalkBank CC system provides this by relying on the specially designed TalkBank Browser, which follows HTML5 (HyperText Markup Language 5) standards for web playback.

2. For consistent analysis, the transcripts must be in a common format that identifies the speaker and breaks up each turn into its component utterances or T-units (Foster et al., 2000). Within utterances, there must be consistent methods for transcribing features such as pauses, filled pauses, retracing, errors, and dialect forms. TalkBank CC achieves this by relying on the fact that all data in TalkBank are in the CHAT (Codes for Human Analysis of Transcripts) format, which provides all these features, as well as methods for Conversation Analysis notation and morphosyntactic analysis. The various codes and conventions of the CHAT format are described in the CHAT manual, which is available from https://talkbank.org/manuals/CHAT.pdf.

3. The transcripts being analyzed must be linked on at least the utterance level to either audio or video media. This is important for understanding conversational, proxemic, and gestural features of the interactions. TalkBank CC provides this facility through time links from the transcripts to the media. Broadband connections allow for smooth and direct playback from a highlighted utterance in the transcript to the corresponding segment of the media. Prior to 2022, these links were created by the transcriber. Our new batchalign system available at https://github.com/talkbank can insert these links automatically.

4. Data must be protected in accord with institutional review board requirements, General Data Protection Regulation regulations, and informed consent agreements. For clinical samples, this typically means that data must be password protected. The required levels of password protection will vary from corpus to corpus. TalkBank implements these protections through a system of user-based passwords and access permissions.

5. To provide flexibility, comments must be stored in a database separate from the main transcript database. TalkBank CC implements this by storing comments in a separate PostgreSQL (Post INGRES Structured Query Language) database organized by commentary groups. For each comment, the following features are stored: file PID (Permanent Identifier), begin and end time of the media segment, time of comment creation, text of the comment, ID of the user creating the comment, group membership of the user, group status of the user, identity of the coding system and the codes in that system, and links to other comments.

6. The system must support the control of commentary group structure. Commentary group owners can provide members with privileges to either read and write or just read. Additionally, group owners can choose to hide or share comments and codes from individual group members to allow for independent judgments by each group member and later collective review.

7. It must be possible to search through a group's comments on the basis of specific keywords, codes, transcripts, or group members. TalkBank CC provides this through searches controlled from the PostgreSQL database.

8. It must be possible to output the text of the transcript along with the newly attached comments to other programs for further analysis. These further analyses could involve checks for reliability between coders or statistical analysis of distributions of codes across transcripts. TalkBank Database facilitates this by allowing the user to save commented transcripts to the desktop for further analysis in the R statistical language (https://www.r-project.org) or TalkBank's CLAN (Computerized Language Analysis) programs (https://dali.talkbank.org/clan).

9. To provide for replicability (Munafò et al., 2017), CC must store the state of the CHAT transcript database and the commentary PostgreSQL database for a given analysis at a given time point. TalkBank CC achieves this through version history in the TalkBank GitLab (Global Information Tracker Laboratory) repository.

10. To provide sustainability, the system must be built on widely used, open-source tools. TalkBank CC achieves this by relying on NodeJS, JavaScript, and PostgreSQL, along with current web standards.

The current version of CC satisfies these requirements. In the Applications section below, we discuss four further sets of CC features that we wish to create.

Learning CC

Eight short screencasts are available to help clinicians, instructors, and researchers to learn the CC features of TalkBank at https://talkbank.org/screencasts.

1. CC-overview. This screencast describes the location of CC and the CC manual on the web, the process of starting CC within the Browsable Database, ways of adding a comment to an utterance, and the process of searching for and creating tags.

2. CC-new_user. This screencast describes registering as a new user of CC.

3. CC-join_group. This screencast describes the process of joining a CC group.

4. CC-comments. This screencast describes the process of entering comments and codes into transcripts.

5. CC-search. This screencast describes the process of searching for comments with specific tags, words, creators, transcripts, or groups. Searches will return matches in a window, allowing the user to go directly to the location of the comment in the transcript to replay the media segment.

6. CC-contact_user. This facility describes how to send e-mail to the creator of a comment. The e-mail includes the exact location and text of the comment. This function is particularly useful for an instructor who wishes to provide feedback to a student.

7. CC-owner. This screencast describes functions controlled by the group owner, such as adding and removing members or setting specific types of data access.

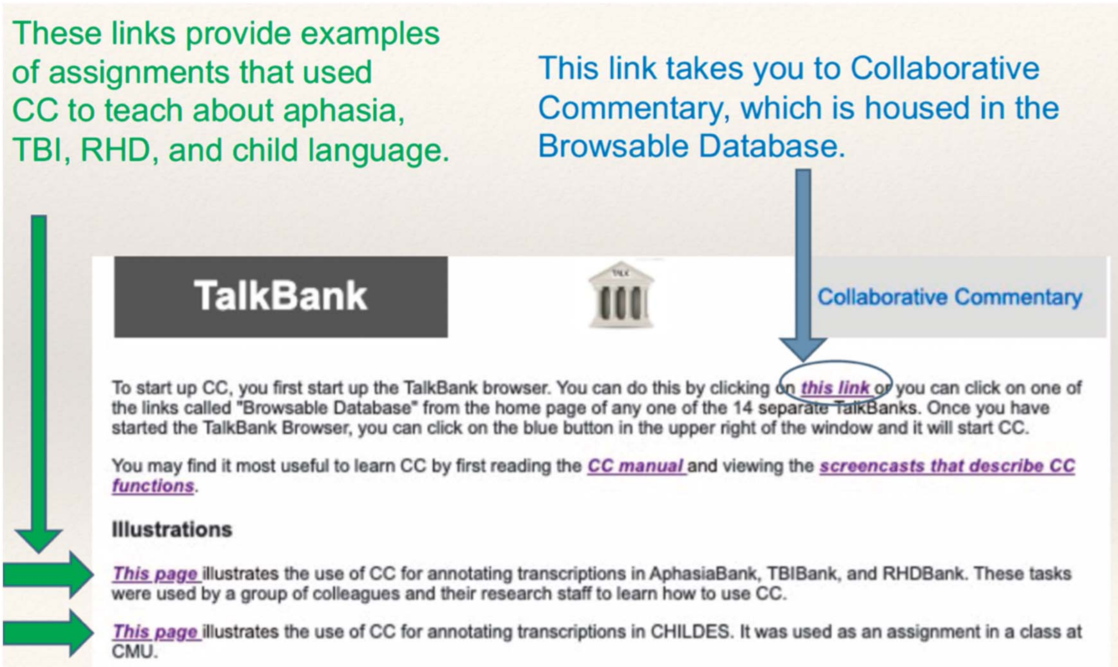
8. CC-manage. This screencast explains how the group owner can manage read and write permissions for users.

The manual at https://sla.talkbank.org/CCmanual/ outlines the steps for creating an account and the processes of tag set creation, comment insertion, and comment searching.

Interface Illustrations

Here we provide illustrations of a few of the screens used by CC. Figure 1 shows the top-level CC page (https://talkbank.org/CC/) with links to the TalkBank Browser, the manual, screencasts, and applications of CC. Figure 2 displays the page at https://sla.talkbank.org/TBB, which you can use to start the TalkBank Browser and the CC process. The two screenshots in Figure 3 show the dialogs that help you join a group.

Figure 1. Starting page for information about CC: https://talkbank.org/CC/. CC = Collaborative Commentary; TBI = traumatic brain injury; RHD = right hemisphere disorder.



Applications

In this section, we describe three current applications of CC along with four that are planned for future work. The three current applications are for *Teaching Clinical Analysis*, *Teaching Clinical Practices*, and *Research Projects*. The four applications proposed for future work are *Automatic Tagging, Coding Untranscribed Data*, *Citizen Science,* and *ASR and Online Transcription*.

Teaching Clinical Analyses

In this section, we present four ways in which CC is being used currently at several universities to teach the analysis of clinical samples.

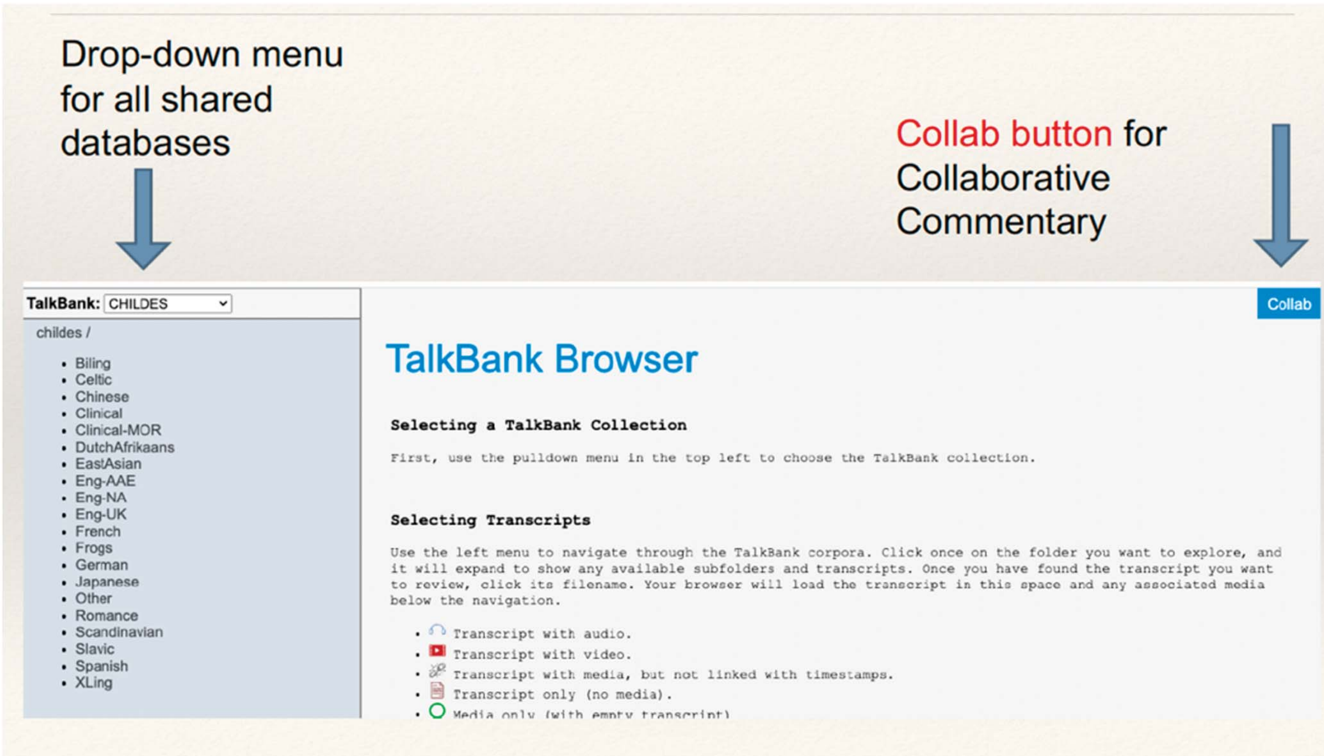
1. Students learning about aphasia are being asked to identify and code behaviors such as paraphasia (semantic, phonemic, mixed), circumlocution, agrammatism, paragrammatism, jargon, anomia, empty speech, conduit d'approche, stereotypy, and perseveration. These features are coded in a tag set created by the instructor, which students can use to mark each case consistently. Along with the tags, students add comments about utterances. The instructor then provides feedback on the tagging and annotation by adding comments to the transcripts or through the e-mail facility.

2. Students in classes on acquired adult communication disorders are being asked to look for and comment on a participant's ability to comprehend, self-monitor, and self-correct—all important considerations for treatment planning. They are to identify which cases may demonstrate coexisting apraxia of speech and to code the relevant behaviors that led them to that diagnosis. Students practice coding correct information units, a frequently used outcome measure in aphasia assessment (Bryant et al., 2017). At the macrostructural level, students code features of discourse such as global coherence, cohesion, and story grammar.

3. Students are being asked to examine Case #2 in the RHDBank Grand Rounds tutorial at https://rhd.talkbank.org/education/class-rhd and to respond to these questions: *How would you judge Phil's stroke story? Did it embody some of the characteristics of right hemisphere discourse? If so, which ones—information content, organization, coherence, prosody, etc.? Would you agree that Phil's Cinderella story could be described with these terms—verbose, fluent, intelligible, confabulatory, tangential? What other terms might you use instead of or in addition to those terms?*

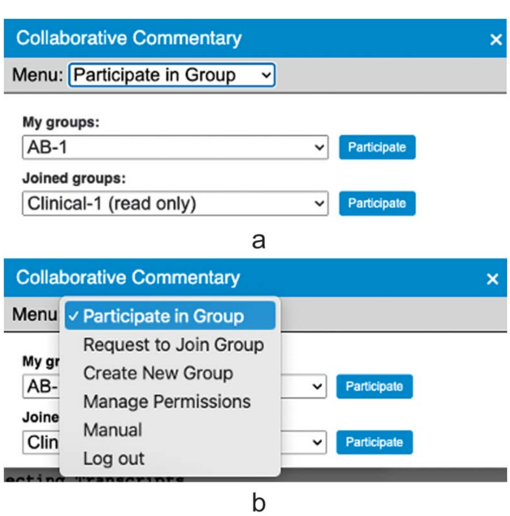
4. Students are being asked to analyze video 6a in TBIBank from a 56-year-old woman who presents

Figure 2. Starting page for the TalkBank Browser with the Collab button.



with aphasia in addition to having a cognitive communication disorder resulting from a motor vehicle accident. The TBI Grand Rounds section provides a brief description of the case and poses the following question: *What features of her spoken discourse are more consistent with aphasia versus cognitive-communication disorder?* Figure 4 shows in red the illustrative comments and codes for this exercise. The tags identify instances of anomia and semantic paraphasias, providing evidence consistent with aphasia. Other tags mark instances where the speaker demonstrated difficulty understanding the main idea of the story and made an irrelevant comment, providing evidence of a cognitive communication disorder.

Figure 3. Dialogs used to participate in a group, join a group, or manage permissions.



Teaching Clinical Practices

CC can also be used for teaching and learning clinical practices. Group members can observe the proper administration and scoring of formal and informal tests. For example, AphasiaBank corpora include recorded and transcribed material from the short form of the Boston Naming Test (Kaplan et al., 2001), the Verb Naming Test from the Northwestern Assessment of Verbs and Sentences (Cho-Reyes & Thompson, 2012), the Quick Aphasia Battery (Wilson et al., 2018), the Famous People Protocol (Holland et al., 2014), and the picture description from the spontaneous speech segment of the Western Aphasia Battery–Revised (Kertesz, 2007).

To learn clinical interaction skills, students can observe the transcript and linked video from a skilled clinician and then use tags to identify specific behaviors, such as

Figure 4. Collaborative Commentary comments and tags for the TBIBank Grand Rounds exercise.

comm-conv (conversational comment), comm-impr (comment on improvement), cues (provides cues), hum (uses humor), interp (interprets, restates), ques-c (closed-ended question), ques-o (open-ended question), reinf (reinforce), sugg (suggest), supp (support, sympathize), and time (slow pace).

Research Projects

Research projects that involve coding or scoring behaviors are using CC in several ways. CC allows research groups to work together asynchronously on shared materials. For example, CC makes it possible for several research assistants to be engaged in parallel at the work of coding CC transcripts. People can work at home or in the office on their own schedule, and the results will be stored in a single consistent form for later analysis.

1. Coding systems are being developed for gesture, fluency, agrammatism, paragrammatism, main concepts, global coherence, conversational sequencing, and more. Research teams can then compute the interrater reliability of code use to determine if the coding system needs further development. If coders did not establish adequate reliability on the first pass, the system can be refined by modifying the training or the codes themselves. Additional trials can be used to establish coding reliability scores within CC.
2. One research group is using a set of macrostructure codes to evaluate discourse on a specific task in a subgroup of the TBIBank Togher protocol database.
3. In another study, group members come from perspectives that yield contrasting and conflicting interpretations of conversational and linguistic features. These differences are being studied by placing their comments and analyses in parallel on specific utterances.

Automatic Tagging

The next three applications of CC rely on ongoing programming developments that are not yet complete. However, we have received funding that will allow us to complete each of these three projects. The first project, described in the current section, involves elaborating TalkBank transcripts with codes entered from automatic tagging systems. One example of this is CLAN's MOR (MORphosyntax) system for automatic tagging of morphological and syntactic dependency structure. Because morphosyntactic codes are so important for computing clinical profile measures such as IPSyn (Index of Productive Syntax; Scarborough, 1990), C-NNLA (Computerized–Northwestern Narrative Analysis; Thompson et al., 1995), Developmental Sentence Score (Lee, 1966), C-QPA (Computerized–Quantitative Production Analysis; Rochon et al., 2000), EVAL (EVALuation; Forbes et al., 2012), and FluCalc (FluencyCalculation; Bernstein-Ratner & MacWhinney, 2018), the codes produced by MOR are integrated directly into the main transcript database on the %mor and %gra lines. It is also possible to use the batchalign automatic speech recognition (ASR) diarization system we have created (https://github.com/talkbank) to add begin and end time codes to both utterances and words.

Some automatically generated codes are best kept separate from the main database. For example, codes for emotion terms using the Linguistic Inquiry and Word Count system (Pennebaker et al., 2001) or codes for conversational moves using rhetorical structure theory (Mann & Thompson, 1992) can be stored in segments of the CC database. Similarly, the utterance-level tags for DementiaBank transcripts created by participants in the INTERSPEECH ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) challenges for 2020 and 2021 (Haider et al., 2019; Luz et al., 2021) can be stored in the CC database.

A related method for adding codes to TalkBank files involves reformatting codes contributed by users that were not entered directly into the CC database. For example, we are now in the process of incorporating contributed codes for conversational moves in the DementiaBank database (Farzana & Parde, 2022; Farzana et al., 2020). Finally, we are working to implement a system of semiautomatic coding in which every utterance in a transcript is assigned a tag from a prespecified tag set for things like speech acts, main concepts, and rhetorical features.

Coding Untranscribed Media

A second extension of CC involves coding untranscribed media. Currently, data in the clinical banks have all been fully transcribed. Typically, these involve recordings of sessions lasting less than 1 hr. However, researchers are also interested in working with recordings taken across much longer timescales. For example, the audio recordings in HomeBank at https://homebank.talkbank.org each last for a full 16-hr day. Audio and video recordings of this type will become more common in the future. To use CC with this type of daylong media, a CLAN command can create a CHAT file composed of regular time segments for display in CC on the web. Researchers can then listen through the media and enter codes for specific segments without having to create a full transcript. In this way, CC provides functionality much like that found in systems such as The Observer XT at https://noldus.com.

Citizen Science

A third extension of the CC system will be to provide support for "citizen science." As an example of such applications, the Zooniverse system at https://www.zooniverse.org has organized nearly 500 projects in which citizens use web interfaces to register observational, experimental, and analytic data. Using CC, we can open up carefully deidentified, edited, and vetted data for crowd-sourced transcription, commentary, and analysis. If citizen volunteers could go through segments of these materials to identify speakers, note distracting noises, spot code-switching, or point out periods of silence, their annotations would provide important training data for further automatic analysis of daylong transcripts. Apart from Zooniverse, citizen science applications can also be routed through systems such as Amazon Mechanical Turk, Qualtrics, Prolific, Pushkin, or Cloudresearch for participant recruitment, gamification, and payment.

ASR and Online Transcription

The major barrier limiting more widespread use of language sample analysis is the time it takes to transcribe interactions (Overton & Wren, 2014). Advances in speech recognition and web technology now make it possible to lower this barrier. Using our batchalign system (https://github.com/talkbank), a clinician or researcher can send a 30-min audio or video recording to the web for ASR processing and forced alignment and receive back a fairly accurate transcription within a few minutes. We are working to also make it possible to automatically contribute the result to the TalkBank database. Once in the database, the researcher and assistants will be able to play through the transcript on the web to add comments and correct residual errors.

Conclusions

CC can contribute to a better understanding of connected speech features in aphasia, dementia, stuttering, RHD, and TBI, as well as child language speech disorders. CC represents an extreme innovation not only for the study of adult neurogenic communication disorders but also for the study of spoken language generally. Codes can be tracked across interactions and participants. The reliability of coding systems can be evaluated. Students can learn how to analyze clinical types and interactions. CC allows researchers to sharpen their coding and interpretation of language behaviors and the challenges people face during conversational interaction. The interpretation of the scope and causes of these difficulties can then inform assessment, classification, treatment, and treatment evaluation. Most importantly, CC makes both the raw data and the results of our comments and analyses fully public and replicable.

Author Contributions

Brian MacWhinney: Conceptualization (Lead), Funding acquisition (Lead), Writing – original draft (Lead), Writing – review & editing (Lead). **Davida Fromm:** Resources (Lead), Validation (Lead), Writing – original draft (Supporting), Writing – review & editing (Supporting).

Data Availability Statement

To obtain the username and password for access to CC and the adult clinical databases (AphasiaBank, DementiaBank, FluencyBank, RHDBank, and TBIBank), researchers and licensed clinicians should first read and accept the TalkBank Ground Rules at https://talkbank.org/share and then send an e-mail request for membership to macw@cmu.edu indicating their contact information, affiliation, and which banks interest them. Membership is usually granted quite quickly. Graduate and undergraduate students should obtain access by asking their faculty advisors to join as members.

Acknowledgments

This work was supported by National Science Foundation Human Networks and Data Science–Infrastructure Grant 53286.1.1122907 for the development of TalkBank Collaborative Annotation. John Kowalski created CC, the TalkBank Browser, and TalkBankDB. However, he declined being included as an author. The authors wish to acknowledge the contributions of the many colleagues who have contributed their recordings and transcripts to the clinical components of TalkBank, including AphasiaBank, DementiaBank, and TBIBank. Without their collegial involvement in data sharing, this type of open analysis of adult language disorders would not have been possible.

References

Bernstein-Ratner, N., & MacWhinney, B. (2018). Fluency Bank: A new resource for fluency research and practice. *Journal of Fluency Disorders, 56,* 69–80. https://doi.org/10.1016/j.jfludis.2018.03.002

Boullosa, B., de Castilho, R. E., Laskari, N. K., Klie, J.-C., & Gurevych, I. (2018). Integrating knowledge-supported search into the ital annotation platform. In E. Blanco & W. Lu (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 127–132). The Association for Computational Linguistics.

Brouwer, M., Brugman, H., & Kemps-Snijders, M. (2017). MTAS: A Solr/Lucene based multi tier annotation search solution. *Selected Papers From the CLARIN Annual Conference 2016: Linköping Electronic Conference Proceedings*, *136*(2), 19–37.

Bryant, L., Spencer, E., & Ferguson, A. (2017). Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology, 31*(10), 1105–1126. https://doi.org/10.1080/02687038.2016.1239013

Cho-Reyes, S., & Thompson, C. K. (2012). Verb and sentence production and comprehension in aphasia: Northwestern Assessment of Verbs and Sentences (NAVS). *Aphasiology, 26*(10), 1250–1277. https://doi.org/10.1080/02687038.2012.693584

de Castilho, R. E., Mújdricza-Maydt, E., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., & Biemann, C. (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. In E. Hinrichs, M. Hinrichs, & T. Trippel (Eds.), *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)* (pp. 76–84). The COLING 2016 Organizing Committee.

Farzana, S., & Parde, N. (2022). Are interaction patterns helpful for task-agnostic dementia detection? An empirical exploration. In O. Lemon, D. Hakkani-Tur, J. J. Li, A. Ashrafzadeh, D. H. Garcia, M. Alikhani, D. Vandyke, & O. Dušek (Eds.), *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 172–182). Association for Computational Linguistics.

Farzana, S., Valizadeh, M., & Parde, N. (2020). Modeling dialogue in conversational cognitive health screening interviews. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 1167–1177). European Language Resources Association.

Forbes, M., Fromm, D., & MacWhinney, B. (2012). AphasiaBank: A resource for clinicians. *Seminars in Speech and Language, 33*(3), 217–222. https://doi.org/10.1055/s-0032-1320041

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*(3), 354–375. https://doi.org/10.1093/applin/21.3.354

Girardi, C., Speranza, M., Sprugnoli, R., & Tonelli, S. (2014). Cromer: A tool for cross-document event and entity coreference. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 3204–3208). European Language Resources Association (ELRA).

Haider, F., de La Fuente, S., & Luz, S. (2019). An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing, 14*(2), 272–281. https://doi.org/10.1109/JSTSP.2019.2955022

Holland, A., Fromm, D., Forbes, M., & MacWhinney, B. (2014). *The Famous People Protocol* [Paper presentation]. International Aphasia Rehabilitation Conference, The Hague.

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston Naming Test–Second Edition*. Lippincott, Williams & Wilkins.

Kertesz, A. (2007). *Western Aphasia Battery–Revised (WAB-R)*. The Psychological Corporation.

Lee, L. L. (1966). Developmental sentence types: A method for comparing normal and deviant syntactic development. *Journal of Speech and Hearing Disorders, 31*(4), 311–330. https://doi.org/10.1044/jshd.3104.311

Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M. E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D. V., Stockhause, M., & Westbrook, J. (2020). The TRUST Principles for digital repositories. *Scientific Data, 7*(1), Article 144. https://doi.org/10.1038/s41597-020-0486-7

Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D., & MacWhinney, B. (2021). Editorial: Alzheimer's dementia recognition through spontaneous speech. *Frontiers in Computer Science, 3,* Article 780169. https://doi.org/10.3389/fcomp.2021.780169

MacWhinney, B., Martell, C., Schmidt, T., Wagner, J., Wittenburg, P., Brugman, H., Broeder, D., & Hoffert, E. (2004). Collaborative commentary: Opening up spoken language databases. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, & R. Silva (Eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)* (pp. 11–15). European Language Resources Association (ELRA). https://psyling.talkbank.org/years/2004/LREC-cc.pdf

Mann, W. C., & Thompson, S. A. (1992). *Discourse description: Diverse linguistic analyses of a fund-raising text*. John Benjamins. https://doi.org/10.1075/pbns.16

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1,* 0021. https://doi.org/10.1038/s41562-016-0021

O'Reilly, C., Iavarone, E., & Hill, S. L. (2017). A framework for collaborative curation of neuroscientific literature. *Frontiers in Neuroinformatics, 11,* Article 27. https://doi.org/10.3389/fninf.2017.00027

Overton, S., & Wren, Y. (2014). Outcome measurement using naturalistic language samples: A feasibility pilot study using language transcription software and speech and language therapy assistants. *Child Language Teaching and Therapy, 30,* 221–229. https://doi.org/10.1177/0265659013519251

Papazian, F., Bossy, R., & Nédellec, C. (2012). AlvisAE: A collaborative web text annotation editor for knowledge acquisition. In N. Ide & F. Xia (Eds.), *Proceedings of the Sixth Linguistic Annotation Workshop* (pp. 149–152). Association for Computational Linguistics.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count (LIWC): A computerized text analysis program*. Erlbaum.

Rochon, E., Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (2000). Quantitative analysis of aphasic sentence production: Further development and new data. *Brain and Language, 72*(3), 193–218. https://doi.org/10.1006/brln.1999.2285

Scarborough, H. (1990). Index of Productive Syntax. *Applied Psycholinguistics, 11*(1), 1–22. https://doi.org/10.1017/S0142716400008262

Tablan, V., Bontcheva, K., Roberts, I., & Cunningham, H. (2015). Mímir: An open-source semantic search framework for interactive information seeking and discovery. *Journal of Web Semantics, 30,* 52–68. https://doi.org/10.1016/j.websem.2014.10.002

Thompson, C. K., Shapiro, L. P., Tait, M. E., Jacobs, B. J., Schneider, S. L., & Ballard, K. J. (1995). A system for the linguistic analysis of agrammatic language production. *Brain and Language, 51,* 124–129.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., . . . Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*(1), Article 160018. https://doi.org/10.1038/sdata.2016.18

Wilson, S. M., Eriksson, D. K., Schneck, S. M., & Lucanie, J. M. (2018). A quick aphasia battery for efficient, reliable, and multidimensional assessment of language function. *PLOS ONE, 13*(2), Article e0192773. https://doi.org/10.1371/journal.pone.0199469