

Peer Review and Communication History

Ms Title: Developing A Cross-Cultural Annotation System and MetaCorpus for Studying Infants' Real World Language Experience

Author Names: Melanie Soderstrom, Celia Rosemberg, Marisa Casillas, Erika Bergelson, John Bunce, Florencia Alam, Anne S. Warlaumont

Submitted: Sep 30, 2020

Editor First Decision—Revise & Resubmit

Nov 10, 2020

Dear Melanie Soderstrom,

I have now received all reviews of your manuscript, "Developing A Cross-Cultural Annotation System and MetaCorpus for Studying Infants' Real World Language Experience" from qualified researchers. I also independently read the manuscript before consulting these reviews. I agree that your manuscript has important strengths and also that there are some issues that need to be addressed. I therefore encourage you to submit a revised version for further consideration at Collabra: Psychology.

The reviewers did an outstanding job in their reviews. In your resubmission, please include a document with a point-by-point response to the reviewers' comments, outlining each change made in your manuscript or providing a suitable rebuttal.

The reviewers' comments are specific and helpful, so I will not repeat them here in detail. Both reviewers agree that an overview of the project right at the beginning would be helpful so that the reader could follow the structure of the paper more clearly.

I also noticed that some OSF links require permission. Please make sure that all relevant files are accessible. You can create a view-only-link, if you do not want to make the project page public yet.

In summary, I think this is a promising manuscript and, I hope you will revise it for further consideration at Collabra: Psychology. I look forward to receiving your revision.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This is the last opportunity for major editing, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please contact the editorial office at editorialoffice@collabra.org.

We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.

Sincerely,

Jan Philipp Röer

Reviewer 1

Open response questions

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

This manuscript describes the creation of a database and coding scheme for longform audio recordings of child language. The innovative cross-disciplinary team approach including language researchers as well as computer scientists capitalizes on trends towards open-science, 'Big Data,' and automated analysis. It is important work towards broadening the diversity of the participants we study. The manuscript is well thought out and appears to fully and clearly describe the authors' thought process during project development. The paper is very strong. I have only minor recommendations they may help with readability.

A flow chart may be helpful to illustrate each step of the project. The narrative describes each stage of this project from conceptual discussion to platform choice to coder training, etc. A visual may help the reader follow the structure of the paper more clearly.

Reliability (pg. 21). One of the objectives of this projects seems to be the need for a coding scheme that can be used across labs so that data can be standardized. Although a link is cited so that the reader can learn more about reliability, a sentence or two referring to the current reliability of the scheme here would be appreciated. Did the protocol and training work?

A sneak peek at tools...(pg. 21). Again, although a link is provided, elaboration on the progress being made by the 'tools' team would be appreciated, as it is one of the objectives of this progress.

Rating scale questions

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose "Neutral" if this is not an empirical manuscript)			✓		
The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose "Neutral" if this is not an empirical manuscript)			✓		
The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative			✓		

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
explanations are thoroughly considered, etc.). (Choose "Neutral" if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.)					
The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.)				✓	

Reviewer 2

Open response questions

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

Developing A Cross-Cultural Annotation System and MetaCorpus for Studying Infants' Real World Language Experience

General comments and summary of recommendation

This paper reports a collaborative project conceived by a group of child language researchers who sought to partner with a group of speech technology researchers, working to develop of A Cross-Cultural Annotation System and MetaCorpus for Studying Infants' Real World Language Experience.

The authors describe their project with two goals: First, a novel annotation system and datasets will be of interest to other child language researchers, and potentially researchers from other fields. Second, they hope to inspire other researchers to apply their collaborative and interdisciplinary approach (as opposed to current mainstream siloed approaches) to other research questions. They examine similarities and differences in children's early experiences with language.

This is a valuable collaborative project which brings together diverse, naturalistic datasets and builds state-of-the-art language processing tools to measure the range and types of variability in children's language experiences and relate this to variability in their language development. Building tools to automatically analyze datasets require an iterative process where smaller datasets are used to develop tools, these tools in turn allow faster annotation as these datasets are expanded.

Despite the importance of this collaborative effort to describe the ACLEW Annotation Scheme (AAS) and the datasets, I have some comments to further improve the manuscript.

1. The abstract should provide a clear summary of the three main objective(s) of the project: a. build a flexible and systematic annotation system b. implement sampling across language communities c. provide a well-designed corpus for the development of tools for automated analysis
2. In the introduction section, the following points should be clarified and reorganized: a. proposal for assessing and promoting replicability in infancy research: large-scale, multilaboratory replication efforts, b. comprehensive start-to-finish analysis pipeline that is publicly available for use, sensitive to cultural differences, and flexible to address a variety of research questions
3. Attention should be paid to document better this new collaborative ACLEW project showing different stages from DARCLE a. date of the ACLEW project b. guidelines for the ACLEW Annotation Scheme (AAS) from DARCLE (DAS) c. collaborative data collection around the world d. perspective and future
4. Figure 1, all tables and supporting material are appropriate and comprehensive for the reader a. Figure 1. Example annotations for a few seconds of audio b. Table 1. Details of the full corpora available in the ACLEW dataset. Some files (McDivitt, Warlaumont and some Bergelson and Casillas files) are available through the HomeBank datasharing system. The remaining files may be requested directly from the dataset owner. c. Table 2. Data sharing ground rules d. Table 3. Checklist for data-sharing and storage. e. Table 4. Descriptive data from random and high volubility sampling of the ACLEW dataset.
5. Check and remove some all the osf references which are preprints of this paper

p11 line 3 (<http://osf.io/4532e>) P15 line 17 (<http://osf.io/erkm8>) P20 line 5 (<http://osf.io/739g8>) p21 line 6 (<https://osf.io/vbpgf>)

6. One reference cited in the text p 4 lines 16-17 is missing in the bibliography Shneidman, L. A., & Goldin-Meadow, S. (2012). Language input and acquisition in a Mayan village: How important is directed speech? *Developmental science*, 15(5), 659-673

Rating scale questions

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose "Neutral" if this is not an empirical manuscript)				✓	
The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose "Neutral" if this is not an empirical manuscript)					✓
The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative					✓

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
<p>explanations are thoroughly considered, etc.). (Choose "Neutral" if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.)</p> <p>The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.)</p>					✓

Author Response

Feb 9, 2021

Dear Dr. Röer,

We thank you and the two reviewers for the very helpful feedback on our submission. I appreciate your patience as we worked through some issues with the data for Table 4, which has now been revised. In particular, one participant from the McDivitt corpus had to be replaced and we revised our data script to more accurately handle overlapping speech, which changed all the data slightly. Below we outline the changes made in response to this feedback. I trust this will address all of the concerns raised.

Response:

The reviewers' comments are specific and helpful, so I will not repeat them here in detail. Both reviewers agree that an overview of the project right at the beginning would be helpful so that the reader could follow the structure of the paper more clearly.

We have added a brief outline of the paper on page 2-3, and as suggested by Reviewer 1, a flowchart on page 8 that more explicitly outlines our workflow for the ACLEW project itself.

I also noticed that some OSF links require permission. Please make sure that all relevant files are accessible. You can create a view-only-link, if you do not want to make the project page public yet.

We have checked that all the links listed in the manuscript are set to public. Some links within those pages may be inaccessible if they reference specific participant files due to confidentiality. All files needed for the Annotation training should be public. Please let us know if there is anything you would expect to access that is not public.

Reviewer 1:

This manuscript describes the creation of a database and coding scheme for longform audio recordings of child language. The innovative cross-disciplinary team approach including language researchers as well as computer scientists capitalizes on trends towards open-science, 'Big Data,' and automated analysis. It is important work towards broadening the diversity of the participants we study. The manuscript is well thought out and appears to fully and clearly describe the authors' thought process during project development. The paper is very strong. I have only minor recommendations they may help with readability.

We thank the reviewer for this positive feedback.

A flow chart may be helpful to illustrate each step of the project. The narrative describes each stage of this project from conceptual discussion to platform choice to coder training, etc. A visual may help the reader follow the structure of the paper more clearly.

A flowchart has been added on page 8 to help visualize our process. We appreciate this suggestion.

Reliability (pg. 21). One of the objectives of this projects seems to be the need for a coding scheme that can be used across labs so that data can be standardized. Although a link is cited so that the reader can learn more about reliability, a sentence or two referring to the current reliability of the scheme here would be appreciated. Did the protocol and training work?

We have added a fuller description on page 23 addressing our reliability results.

A sneak peek at tools...(pg. 21). Again, although a link is provided, elaboration on the progress being made by the 'tools' team would be appreciated, as it is one of the objectives of this progress.

We have added some wording to better clarify the progress of the tools. We have avoided going into too much detail, however, in part due to the more technical nature of that component of the project.

Reviewer 2:

This paper reports a collaborative project conceived by a group of child language researchers who sought to partner with a group of speech technology researchers, working to develop of A Cross-Cultural Annotation System and MetaCorpus for Studying Infants' Real World Language Experience.

The authors describe their project with two goals: First, a novel annotation system and datasets will be of interest to other child language researchers, and potentially researchers from other fields. Second, they hope to inspire other researchers to apply their collaborative and interdisciplinary approach (as opposed to current mainstream siloed approaches) to other research questions. They examine similarities and differences in children's early experiences with language.

This is a valuable collaborative project which brings together diverse, naturalistic datasets and builds state-of-the-art language processing tools to measure the range and types of variability in children's

language experiences and relate this to variability in their language development. Building tools to automatically analyze datasets require an iterative process where smaller datasets are used to develop tools, these tools in turn allow faster annotation as these datasets are expanded.

We thank the reviewer for promoting the value of our work.

Despite the importance of this collaborative effort to describe the ACLEW Annotation Scheme (AAS) and the datasets, I have some comments to further improve the manuscript.

1. *The abstract should provide a clear summary of the three main objective(s) of the project:*
 - a. *build a flexible and systematic annotation system*
 - b. *implement sampling across language communities*
 - c. *provide a well-designed corpus for the development of tools for automated analysis*

We have added additional wording to the abstract to address this suggestion.

2. In the introduction section, the following points should be clarified and reorganized:
 - a. proposal for assessing and promoting replicability in infancy research: large-scale, multilaboratory replication efforts,
 - b. comprehensive start-to-finish analysis pipeline that is publicly available for use, sensitive to cultural differences, and flexible to address a variety of research questions

We were not sure what the reviewer was asking for here. We could not find the text of “a” in our manuscript, “b” can be found in our abstract. We hope the above revisions to the manuscript in the abstract and the addition of the flowchart (see above), together with some additional wording edits on page 2-3 have addressed the reviewer’s concern.

3. *Attention should be paid to document better this new collaborative ACLEW project showing different stages from DARCLE*
 - a. *date of the ACLEW project*
 - b. *guidelines for the ACLEW Annotation Scheme (AAS) from DARCLE (DAS)*
 - c. *collaborative data collection around the world*
 - d. *perspective and future*

We thank the reviewer for the suggestion to take a little more time in describing these important relationships, particularly the emergence of ACLEW from DARCLE. We have added a paragraph on page 7 more explicitly describing the emergence of ACLEW from DARCLE, have added language on page 12 regarding the AAS vs. the DAS, and a sentence on page 25 acknowledging the importance of DARCLE for the emergence of the current project.

4. *Figure 1, all tables and supporting material are appropriate and comprehensive for the reader*
 - a. *Figure 1. Example annotations for a few seconds of audio*

- b. *Table 1. Details of the full corpora available in the ACLEW dataset. Some files (McDivitt, Warlaumont and some Bergelson and Casillas files) are available through the HomeBank datasharing system. The remaining files may be requested directly from the dataset owner.*
- c. *Table 2. Data sharing ground rules*
- d. *Table 3. Checklist for data-sharing and storage.*
- e. *Table 4. Descriptive data from random and high volatility sampling of the ACLEW dataset.*

- 5. *Check and remove some all the osf references which are preprints of this paper*

p11 line 3 (<http://osf.io/4532e>) P15 line 17 (<http://osf.io/erkm8>) P20 line 5 (<http://osf.io/739q8>) p21 line 6 (<https://osf.io/vbpqf>)

Thank you for pointing out this issue with some of our permissions. The relevant information should now be available.

- 6. *One reference cited in the text p 4 lines 16-17 is missing in the bibliography Shneidman, L. A., & Goldin-Meadow, S. (2012). Language input and acquisition in a Mayan village: How important is directed speech? *Developmental science*, 15(5), 659-673*

Apologies for the missing reference. We have updated the reference list accordingly.

Editor Final Decision—Accept

Mar 4, 2021

Dear Dr. Soderstrom,

I have now had a chance to read over your manuscript “Developing A Cross-Cultural Annotation System and MetaCorpus for Studying Infants’ Real World Language Experience”, along with the letter describing the changes you made. Thank you for your responsiveness to the concerns that the reviewers raised.

I am happy to say that your paper is now officially accepted for publication in *Collabra: Psychology*. Congratulations on this excellent work, I think it will make an important contribution to the literature and I look forward to seeing it published!

I hope your experiences with *Collabra: Psychology* have been positive and that you will continue to consider it as an outlet for your work.

As there are no further reviewer revisions to make, you do not have to complete any tasks at this point. Our managing editor will contact you in case there are any pre-production file related questions. You will

have an opportunity to check the page proofs before we publish your article. Thank you again for publishing in Collabra: Psychology.

Sincerely, Jan Philipp Röer

Reviewer 1

Open response questions

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

Thank you for the additions you made to your manuscript during review. You have clarified my questions. I look forward to future work from your team.

Rating scale questions

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose "Neutral" if this is not an empirical manuscript)				✓	
The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose "Neutral" if this is not an empirical manuscript)				✓	
The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose "Neutral" if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.)				✓	
The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.)				✓	

Reviewer 2

Open response questions

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

The corrections from the authors in the abstract and the new brief outline in the introduction make the paper more readable now. In particular, the addition of the flowchart on page 8 provides a good overview of the project so that the reader could follow the emergence and the history of the project more clearly. The emergence of ACLEW from DARCLE project is also now well documented in the new paragraph on page 7, on p 12 regarding the AAS vs. the DAS and on p 25 when the authors acknowledge the importance of DARCLE for the current project.

Rating scale questions

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose "Neutral" if this is not an empirical manuscript)					✓
The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose "Neutral" if this is not an empirical manuscript)					✓
The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose "Neutral" if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.)				✓	
The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.)				✓	