Assignment 9

Referee Report:
childes-db: A Flexible and Reproducible Interface to the Child Language Data Exchange System

Ellen Hsieh

The Child Language Data Exchange System (CHILDES) is an open dataset contains the transcripts and recordings relevant to the study of child language acquisition, which has played a critical role in research on child language development. However, the access to this database can be “complex for novices and difficult to automate for advanced users” (Sanchez et al., 2018, p.2). Therefore, in order to improve the accessibility and usability of the data, the paper introduces a new interface, childes-db (Sanchez et al., 2018), which allows the users to easily obtain and analyze the data they need. At the same time, it can also enhance the computational reproducibility that are not feasible by the former method, using the specialized format like CHAT and CLAN, when accessing the data.

In the introduction (Sanchez et al., 2018, p.3-6), the authors start with the introduction to CHILDES. Although giving some background for CHILDES could be helpful for the readers to comprehend the context of the paper, the coverage of CHILDES might be too long. For instance, the authors use half of the page to describe some inefficient methods that collect the data unavailable in in-lab experiments such as “audio recording”, “automatic speech recognition (ASR)” before the establishment of CHILDES.

Undeniably, the authors persuasively illustrate the need for creating a new system like childes-db by pointing out the shortcomings of CHILDES, and demonstrate how childes-db create a user-friendly interface for the users in a clear way. First, “childes-db treats CHILDES as a set of linked tables” (Sanchez et al., 2018, p.6). Therefore, it would be more intuitive for the users to analyze the data using certain operations such as filtering and grouping. This new interfaces “allow the API to take care of the formatting details” so the users can just access the data they need without concerning how to access the database. Secondly, the tables in the database are hierarchical collections. The data are stored in different level according to their attributes. For example, a token (a word) is the smallest unit in the database, and tokens comprise the utterances, then utterances are stored in a transcript labeled by the date. The highest level is the collection and the collections are classified by the languages (Sanchez et al., 2018, p.7). The users can easily access those data through childes-db interactive web apps or the childsr R package.

When elaborating the mechanism of childes-db, the authors emphasize on how they preserve the data integrity (Sanchez et al., 2018, p.9). To evaluate their success, they use the Pearson correlation coefficient to measure the comparison between the unigram count and CLAN. Besides, the authors talk about the versions and current annotation coverage in childes-db. It would be better if the authors can elaborate more on how they choose the information to cover since the information that are not covered seems important for studying the language acquisition as well such as tone direction and stress.

In the paper, the authors use two different examples to substantiate the effectiveness of childs-db. One is “the study case of the frequency with which children hear the color words

of English" (Sanchez et al., 2018, p.16). The investigation was conducted through using the web apps and childsr R package. The researchers only need to specify the ages of the children they want to include, the speaker, and the specific words they want to analyze on the web page. Then, childes-db will produce the plots of frequency of the words and the researchers can easily analyze it. Furthermore, to analyze the differences between gender, childsr R package is available for the researchers, which allow the users to get more details about the data in CHILDES. Another case which discusses how childs-db facilitates the course related to early language acquisition (Sanchez et al., 2018, p.23-25). Through the interactive web apps, the student can observe the differences of language learning based on different situations. Moreover, childes-db provides good opportunities for students to sharpen their programming skill while using childsr R package and empower students' research projects.

The authors do a great job in putting the paper in context of the broader literature and they do not miss any important citation or add any unnecessary citations. They review various literatures that are related to CHILDES, point out the disadvantages of it, and then provide a better way to access the data from CHILDES more effectively. The whole paper focus on the functionality of childes-db, which is clear and strong.

In the conclusion part, the authors mention the limitation about childes-db system such as it only operates on transcript data. Here comes the issue of how to "further computational and manual analyses of phonology, prosody, social interaction, and other phenomena by providing easy access to the video and audio data" (Sanchez et al., 2018, p.26). To achieve the goal, the media links that are not included previous should be included. Although the analysis of the audio or video would be harder than just analyzing the transcripts, nowadays

there are a lot of analysis tools for audio and video. The key points are to fully understand which tools and methods we should utilize to analyze the media files. Bases on that, we can also analyze the tone direction and stress, which are critical in communication. Language acquisition is not only about the syntax and word learning. The sound of the language is also important. Therefore, adding the access to media files in CHILDES and the analysis tools for audios and videos will improve the usability of cildes-db, thereby benefiting the studies on child language development.

References

Sanchez, Alessandro, Stephan C. Meylan, Mika Braginsky, Kyle E. Mac- Donald, Daniel Yurovsky, and Michael C. Frank, "childes-db: A Flexi- ble and Reproducible Interface to the Child Language Data Exchange System," under review, Communication and Learning Laboratory, University of Chicago, https://callab.uchicago.edu/papers/smbmyf-brm-underreview.pdf 2018.