Datasheet for Wordbank*

A Datasheet Used to Understand the Study on 'Pathways to Early Vocabulary Acquisition'

Yongqi Liu

2024-12-02

The Wordbank dataset is a comprehensive, publicly accessible resource designed to advance the study of early vocabulary acquisition in children aged 16–30 months. It aggregates anonymized data from the MacArthur-Bates Communicative Development Inventories (CDIs), spanning 29 languages and dialects, to support research on linguistic development, cross-linguistic comparisons, and developmental trends. Maintained by Stanford University's Language and Cognition Lab, the dataset features both norming and non-norming samples, offering robust analytical capabilities while capturing the diversity of linguistic environments. This datasheet provides an in-depth overview of the dataset's purpose, structure, collection methodology, distribution, and ethical considerations, highlighting its utility as a transparent and versatile tool for research and education.

Extract of the questions from Gebru et al. (2021) and this datasheet is based on dataset from Braginsky (2024) and CDI (2024).

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

    - The Wordbank dataset was created to provide a centralized platform for archiving, sharing, and analyzing anonymized data from the MacArthur-Bates Communicative Development Inventories (CDI). Its primary objective is to support researchers in studying the cognitive development processes of children during early childhood. By compiling data across multiple languages, the dataset serves as a comprehensive resource for both norming studies and diverse contributions from various research projects. It addresses the need for an accessible repository of CDI data, enabling

*Code and data are available at: Early Vocabulary Acquisition Pathway

researchers to generate norms for populations and individual vocabulary items interactively. It also bridges a critical gap by facilitating cross-linguistic comparisons of early lexical development and offering standardized benchmarks alongside diverse linguistic samples. Building on its predecessor, CLEX, Wordbank enhances functionality as both a research and teaching tool. It allows users to analyze developmental patterns, customize norms, and conduct interdisciplinary studies, making it an invaluable asset in linguistic, psychological, and developmental research.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

    - The Wordbank dataset was developed by the Language and Cognition Lab at Stanford University, under the direction of Dr. Michael C. Frank. This initiative is part of Stanford's broader commitment to advancing research in child language acquisition and cognitive development. The platform compiles data from the MacArthur-Bates Communicative Development Inventories (CDIs) across multiple languages, serving as a valuable resource for researchers worldwide.

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

    - The development of Wordbank was funded by a grant from the National Science Foundatio as well as generous support from the MB-CDI Advisory Board.

4. *Any other comments?*

    - The dataset is a crucial resource for advancing research on children's cognitive and linguistic development. It highlights critical windows of vocabulary growth and integrates both norming and non-norming samples, providing standardized benchmarks alongside diverse linguistic contexts. With its cross-linguistic breadth and detailed lexical categorization, it serves as a powerful tool for analyzing global patterns in early vocabulary acquisition and fostering a deeper understanding of developmental trajectories.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

    - The instances in the dataset are children aged 16–30 months and their parents' responses from the MacArthur-Bates Communicative Development Inventories (CDIs). Each instance is tied to a specific child and consists of detailed information about the words they understand (comprehension) and can use (production), categorized into lexical groups such as nouns, verbs, and adjectives.

2. *How many instances are there in total (of each type, if appropriate)?*

    - Wordbank now contains data from 92,771 children and 105,290 CDI administrations, across 42 languages and 89 instruments.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

    - No. The wordbank package collected a sample drawn from responses collected using the MacArthur-Bates Communicative Development Inventories (CDIs) across various languages and studies. It includes both norming samples, which are standardized and representative of general language acquisition trends, and non-norming samples, contributed by researchers studying specific subpopulations or contexts. While the norming data provide a validated baseline for comparison, the non-norming data add diversity but may not fully represent population-level trends.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.* Here is a single type of instance in the dataset, but each instance includes multiple dimensions, such as:

    - Instrument: A parent-report survey or questionnaire containing a specific set of items. For instance, the American English Words & Sentences form is one such instrument.
    - Item: An individual question within an instrument. This could be a specific word like dog (a canonical CDI item) or questions related to gestures, morphological and syntactic complexity, or other elements of early language and behavior.
    - Administration: A single instance where an instrument is completed for a child, accompanied by details like the child's age and the contributing lab.
    - Child: A unique individual for whom data is collected, along with associated demographic information.
    - Language: A specific language or language community for which a CDI instrument has been adapted. This includes distinctions between varieties, such as American and British English. While the term "language or dialect" is more precise, it is avoided here for simplicity in variable naming.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

    - Yes. Each instrument has a unique Data ID, and each sample individual has a unqiue Child ID, both of which can be used to identify each instance.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

    - Yes. When doing the survey, parents may skip certain items on the questionnaire, especially for abstract or less commonly used words, resulting in missing data for those items.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

    - No. Relationships between individual instances are not explicitly defined in the dataset. Each instance represents a single child's vocabulary data, and the dataset treats these instances independently. However, implicit relationships exist within subgroups, such as children belonging to the same norming sample or sharing similar demographic or linguistic attributes. These relationships are not directly encoded but can be inferred and analyzed using metadata like norming status, age, and language group.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - The dataset can be split into training, validation, and testing sets, stratified by norming status to maintain a balanced representation across norming and non-norming groups. This approach ensures that both standardized (norming) and diverse (non-norming) samples contribute proportionally to each split, supporting robust model training and evaluation while keeping the diversity characteristic in the dataset.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - A primary source of noise in the dataset stems from parental reporting biases, especially for abstract word categories such as those related to sensory and emotional vocabulary. These biases may arise from differences in interpretation or recall accuracy among respondents, potentially affecting the reliability of the data for these categories. Additionally, variability in reporting interpretations across linguistic and cultural contexts may also introduce inconsistencies.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there*

official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

- The dataset is self-contained but derived from the CDI and Wordbank repositories. External tools like Wordbank R package were used for data extraction and processing.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - No personally identifiable information is included. The dataset is anonymized and contains only aggregated linguistic data for an individual sample.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - Yes. During the initial data downloading part for conducting the paper, subpopulations are collected by only choosing the English language speaker and WS form word for further data handling convenience.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - Yes. From the Child ID, we can identify individuals.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - No. The dataset does not include data that might reveal sensitive personal attributes.

16. *Any other comments?*

- No.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data for each instance was reported by children's parents through the use of the CDI checklist and survey, which records the child's word comprehension and production. Parents provided this information either in electronic form or on paper-based surveys.
    - The Wordbank repository archives and organizes these responses, ensuring the data is accessible for research purposes. While the data relies on parental reporting, it is validated through the standardized CDI forms, which has been rigorously tested for reliability and consistency in capturing early vocabulary development. This ensures that the reported data aligns closely with established benchmarks for child language acquisition.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - Data collection utilized the CDI survey forms. It can be conducted by either electronically or via paper forms.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - The dataset is a sample drawn from responses using CDI across various languages. Norming samples were obtained through a probabilistic sampling strategy to ensure demographic representativeness, balancing factors such as age, gender, and region, and providing a standardized baseline for language acquisition. In contrast, non-norming samples were collected using convenience or purposive sampling, often targeting specific populations or linguistic environments to enhance diversity. While the norming samples support generalizations and cross-study comparisons, the non-norming samples offer investigations into underrepresented contexts, together creating a comprehensive resource for studying early vocabulary development.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - The data collection process involved researchers working in collaboration with parents, educators, and early childhood experts. Parents reported their children's vocabulary data through structured surveys, while educators and experts provided additional support in administering and validating the surveys. Compensation details are not explicitly provided, as the contributions were likely part of research studies where participation was voluntary or incentivized through non-monetary means such as feedback on child development or access to study results.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - Data collection occurred in 1990s, focusing on snapshots of vocabulary acquisition for children aged 16 to 30 months.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - Yes. Data collection adhered to ethical guidelines, with informed consent obtained from all participants. Some guidlines about the data usage can be found at: Wordbank Official Website.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - The data is obtained from third parties from the Wordbank (Braginsky 2024). Braginsky (2024) use the data from CDI (2024) website. The Wordbank team processes this data to create a standardized dataset and develops tools, the wordbankr package, for efficient access and analysis by researchers.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - Yes. Individuals (typically parents or guardians of the children) were notified about the data collection. Notification was not seen in the provided survey form from the MacArthur-Bates Communicative Development Inventories, but it may exist in other real-person communications and records. The consent process included

clear communication about the purpose of the study, how the data would be used, and assurances of anonymity. This information was typically conveyed through a consent form or verbal explanation, depending on the study protocol. The language of the notification included similar statements such as: "By completing this survey, you agree to participate in a study on early language development. Your responses will be anonymized and used solely for research purposes. Participation is voluntary, and you may withdraw at any time without penalty."

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - Yes. All individuals provided informed consent prior to participation. Consent was requested through the survey provided to parents or guardians. The consent form outlined the purpose of the study, the types of data being collected, and how the data would be used. An example of the consent language included: "By signing below, I agree to participate in this study on early language development. I understand that my child's vocabulary data will be anonymized and used solely for research purposes." The form also provided contact details for the research team in case participants had further questions or concerns.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - Yes. Participants were informed of their right to withdraw consent at any time. The consent form included instructions for revoking participation, such as contacting the research team via email or phone. Upon request, their data would be promptly removed from the study and any associated analyses. A specific clause stated: "Participants may withdraw their consent at any point without penalty. To do so, please contact the study coordinator at mbcdiboard@gmail.com."

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - Yes. A data protection impact analysis was conducted to ensure compliance with ethical and privacy standards. This analysis evaluated risks such as unintended re-identification of individuals, data misuse, and potential biases in reporting. Outcomes of the analysis included ensuring data was anonymized before analysis to prevent identification, storing data on secure servers with restricted access and regularly reviewing datasets to remove personally identifiable information (PII).

12. *Any other comments?*

    - This study adhered to the highest ethical standards in data collection, processing, and usage. It aligns with guidelines for working with human subjects, ensuring privacy and respect for participants throughout the research process. Future iterations of this dataset will continue to prioritize transparency, participant rights, and ethical integrity.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - Yes. In this study, data cleaning involved removing incomplete responses, grouping lexical items into broad categories, and standardizing age and vocabulary scores.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - The raw data is archived for future research purposes and stored alongside cleaned versions. The complete reproducible project can be found at Early Vocabulary Acquisition Pathway.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - The software used to preprocess, clean, and label the data is available through the wordbankr R package. This package facilitates access to the Wordbank database, an open repository of developmental vocabulary data. You can find the wordbankr package and its documentation on GitHub. Additionally, all subsequent data analysis tasks were conducted using R Studio, ensuring a robust and reproducible workflow for processing and analyzing the dataset.

4. *Any other comments?*

    - No.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

    - Yes. The dataset has been utilized to investigate vocabulary acquisition patterns in children aged 16–30 months. Specifically, it has been used to: i). Predicting high vocabulary proficiency based on variables like age, norming status, and lexical

categories. ii). Examining patterns in the development of different word categories, such as Function Words and Sensory Words, to understand their relative acquisition timelines. iii). Investigating the impact of norming and non-norming environments on language development to identify contextual factors influencing vocabulary growth.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

   - The dataset is linked to studies and resources available on the CDI Wordbank platform, as well as to related papers focused on early language development. Future research using this dataset will reference these repositories as part of their methodology. The Wordbank repository, which includes tools for accessing and analyzing the dataset, is accessible on GitHub at: https://github.com/langcog/wordbankr.

3. *What (other) tasks could the dataset be used for?*

   - Exploration of Language Developmental Milestones: Studying the progression of specific word categories over time.
   - Cross-Linguistic Comparisons: Comparing vocabulary acquisition patterns across different languages by aligning this dataset with similar CDI datasets from other languages.
   - Bilingual Studies: Examining how bilingual environments influence vocabulary development using norming and non-norming comparisons.
   - Intervention Design: Identifying underrepresented word categories and norming group's language environment to inform targeted educational programs or speech therapy practices.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

   - Bias in Reporting: The reliance on parental reports may introduce inaccuracies, especially for abstract or infrequent word categories. Dataset consumers should interpret these results with caution and consider complementing them with observational or experimental data.
   - Underrepresentation of Certain Groups: The lack of norming sample may not fully capture linguistic and socio-economic diversity. Care should be taken to avoid overgeneralizing findings to underrepresented populations.

- Lexical Category Distribution Imbalance: The raw data reveals a significant imbalance in the distribution of lexical categories, with certain categories being underrepresented. This uneven category colelction can introduce bias during data analysis, as conclusions drawn from overrepresented categories may not generalize well to those with insufficient samples. Addressing this issue may require careful statistical adjustments or targeted data collection to ensure a more balanced representation of lexical categories.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

   - No.

6. *Any other comments?*

   - The dataset provides a rich resource for studying early language development but requires careful consideration of its limitations. Researchers are encouraged to combine it with other data sources and methodologies to enhance robustness and applicability. Transparency in reporting and responsible use are critical to maximizing the dataset's value while minimizing potential harm.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

   - Yes, the dataset is distributed to third parties as a public-access data tool through the Wordbank platform. Researchers, educators, and other users can freely access the data under appropriate licensing terms, which require that the dataset remains anonymized and is used ethically and responsibly. This open-access approach facilitates transparency, reproducibility, and collaborative research in early childhood language development, allowing third parties to leverage the data for various academic and educational purposes.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

   - The dataset is distributed via the Wordbank.
   - It can be accessed by library(wordbankr) in the R Studio, which provides programmatic access to the data. Users can explore and download data directly from the platform.
   - The direct access of raw data from CDI needs to get the authorized. Once authorized, adaptation developers are also free to determine how the instrument will be distributed.

3. *When will the dataset be distributed?*

- The Wordbank dataset is already being distributed and is publicly available. Users can access it at any time through the Wordbank platform or associated tools like the wordbankr R package. The data is made accessible under licensing terms that ensure ethical and anonymized use, promoting its role as a collaborative and open-access resource for research and education. There is no future distribution delay, as the dataset is actively maintained and regularly updated to include new CDI data. The update can be tracked on the Github repository.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

    - Yes. The Wordbank dataset is distributed under applicable terms of use (ToU) to ensure ethical and responsible utilization. The data is provided as a public-access resource under licensing terms that require users to:
        1. Use the dataset only for educational, research, or non-commercial purposes.
        2. Ensure the data remains anonymized to protect participants' privacy.
        3. Acknowledge the Wordbank platform in any derived works.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

    - No. No third parties have imposed intellectual property (IP)-based or other restrictions on the data associated with the instances. The dataset is made freely available under the management of the Wordbank platform and adheres to open-access principles. The data must remain anonymized and be used only for ethical and non-commercial purposes, as outlined in the platform's terms of use (ToU). These conditions ensure privacy protection and responsible data use. There are no fees associated with accessing or using the dataset, further supporting its open-access mission.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

    - No export controls or other regulatory restrictions apply to the Wordbank dataset or its individual instances. Its design ensures compliance with global data privacy standards, making it accessible to researchers and educators worldwide without regulatory limitations.

7. *Any other comments?*

    - The Wordbank dataset is an invaluable and public resource for research on child language acquisition. Its cross-linguistic breadth and interactive tools make it an

essential asset for linguistic and developmental studies. Continued updates ensure its relevance and usability for global research communities.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The Language and Cognition Lab at Stanford University, under the leadership of Dr. Michael C. Frank, is responsible for hosting, supporting, and maintaining the Wordbank dataset. The dataset is available on the Wordbank platform, and it is regularly updated and maintained to ensure accessibility and usability for the research community.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - The dataset manager can be contacted via email: wordbank-contact@stanford.edu. Additional inquiries can be directed to the institution's Wordbank Development Team or CDI Advisory Board for the original survey framework from The MacArthur-Bates Communicative Development Inventories.
3. *Is there an erratum? If so, please provide a link or other access point.*
    - No.
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
    - Yes. The dataset will be updated periodically. The number of CDIs stored in Wordbank is continually growing, and the development team are always interested in adding more datasets to the site. To see which researchers have already contributed, click here: Wordbank contributors. If you are interested in contributing your own data, please contact via: wordbank-contact@stanford.edu.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
    - Yes. The dataset includes data related to children but obey strict anonymization protocols, ensuring that no personally identifiable information (PII) is retained. While the anonymized data is retained indefinitely for research purposes, raw data containing any identifiable information is subject to retention limits based on the original study's ethical guidelines. Parents are informed during the consent process about how their data will be handled, including the retention policy. Anonymized data is retained indefinitely to support transparency, reproducibility, and ongoing

research. Raw data containing identifiable information is retained only for a limited period, as specified by the original study's ethical guidelines.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - Older versions will remain archived for reproducibility purposes and will be accessible via wordbankr repository.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

    - Yes. Wordbank provides a mechanism for researchers to extend, augment, or contribute to the dataset. By contributing, researchers join a collaborative consortium dedicated to advancing the understanding of children's early language development through pooled data. Contributions are validated to ensure consistency and quality. The Wordbank team reviews and integrates new data, maintaining alignment with existing standards for CDI data. Researchers can submit their datasets through established channels, and accepted contributions are documented, credited, and made available to the broader community.
    - Wordbank follows the collaborative ethos of resources like CHILDES, Databrary, and CLEX. By participating, contributors gain visibility for their work and enable global researchers to conduct cross-linguistic analyses of CDI data. This collaborative framework not only enhances the dataset's scope but also fosters a shared mission to advance research in early language development.

8. *Any other comments?*

    - Wordbank maintains regular updates, implements user feedback mechanisms, and adheres strictly to data management to ensure the dataset remains a reliable, valuable and up-to-date resource for researchers and educators studying early childhood language development.

References

Braginsky, Mika. 2024. *Wordbankr: Accessing the Wordbank Database*. https://github.com/langcog/wordbankr.

CDI, MacArthur-Bates Communicative Development Inventories. 2024. "MacArthur-Bates Communicative Development Inventories." https://mb-cdi.stanford.edu/.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for Datasets." *arXiv Preprint arXiv:1803.09010*. https://arxiv.org/abs/1803.09010.