Datasheet for Word Bank*

A Datasheet Used to Understand Children's Vocabulary Acquisition Pattern

Yongqi Liu

2024-11-29

This datasheet containes the motivation, composition, collection, preprocessing, uses, distribution and maintainence of the dataset from Word Bank and MacArthur-Bates Communicative Development Inventories.

Extract of the questions from Gebru et al. (2021) based on the data 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 serve as a platform for archiving, sharing, and exploring anonymized data from the MacArthur-Bates Communicative Development Inventories (CDI). Its primary goal is to help the study of early vocabulary acquisition across multiple languages by providing a centralized resource for both norming studies and contributed data from various research projects. The dataset solves the need for a comprehensive, accessible repository of CDI data, enabling researchers to generate norms for populations and individual vocabulary items interactively. It also fills a critical gap by supporting cross-linguistic comparisons of early lexical development and enhancing accessibility to standardized benchmarks and diverse linguistic samples. Wordbank builds on the foundation of its predecessor, CLEX, by offering improved functionality as both a teaching and research tool. It allows users to explore developmental patterns and generate customized norms for different populations, making it an invaluable resource for 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)?*

*Code and data are available at: https://github.com/Cassieliu77/Early_Vocabulary_Acquisition_Pathway.git.

- 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 Foundation (# 1451577) as well as generous support from the MB-CDI Advisory Board.

4. *Any other comments?* -The dataset is significant for its contribution to linguistic development research. It emphasizes critical windows of vocabulary growth and includes both norming and non-norming samples, offering standardized benchmarks while reflecting diverse linguistic environments. Its cross-linguistic scope makes it a valuable tool for understanding global patterns in early lexical development.

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 represent children aged 16–30 months and their responses from the MacArthur-Bates Communicative Development Inventories (CDIs). Each instance includes data on the child's vocabulary comprehension and production, focusing on early language acquisition. The dataset consists of multiple types of information for each child, such as:

    a. Vocabulary Comprehension: Words the child understands.
    b. Vocabulary Production: Words the child can produce.
    c. Lexical Categories: Words grouped by semantic or grammatical categories (e.g., Function Words, Living Things, Adjectives).
    d. Norming and Non-norming Status: Indicates whether the data come from a standardized norming sample or research projects contributed by individual researchers.

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

    - As of the latest data available, the Wordbank dataset comprises 92,771 children and 105,290 CDI administrations, encompassing 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).

- The dataset is 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.* Each instance consists of processed features rather than raw data. Specifically, the data includes: • Child Demographics: Age (in months) and norming status (norming or non-norming group). • Vocabulary Measures: • Word comprehension counts (words the child understands). • Word production counts (words the child can produce). • Lexical Categories: Words grouped by semantic or grammatical categories (e.g., Function Words, Living Things, Adjectives).
5. *Is there a label or target associated with each instance? If so, please provide a description.*

   - Yes, the target variable is the binary "High Vocabulary" status, indicating whether a child's combined comprehension and production exceeds 300 words.

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, the target variable is the binary "High Vocabulary" status, indicating whether a child's combined comprehension and production exceeds 300 words.

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.*

    - Data can be split into training, validation, and testing sets, stratified by norming status to ensure balanced representation across groups.

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

    - Potential biases in parental reporting are a source of noise, particularly for abstract word categories such as Sensory Words.

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 WordbankR were used for data extraction.

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.

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, the dataset focuses on vocabulary acquisition and does not include sensitive information such as ethnicity, religion, or health data.

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, subpopulations are identified by norming status and age groups, enabling analysis across different demographic and developmental categories.

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

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?*

    - The dataset offers valuable insights into vocabulary development while balancing standardization and diversity through its dual norming and non-norming sample structure.

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.*

    - Data was reported by parents using the CDI checklist, which captures word comprehension and production for their child.

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 tool, administered either electronically or via paper forms, validated against standardized linguistic benchmarks.

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

    - Norming samples were stratified to ensure demographic representativeness, while non-norming samples were more diverse but not strictly representative.

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)?*

    - Data was collected by researchers, often in collaboration with parents, educators, and early childhood experts.

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 over several years, focusing on snapshots of vocabulary acquisition for children aged 16–30 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.

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 Word Bank (Braginsky 2024). Braginsky (2024) use the data from CDI (2024) website to construct their package for use.

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.*

    - Data was collected by researchers, often in collaboration with parents, educators, and early childhood experts.

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 a standardized form provided to parents or guardians, either digitally or in print. 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 [contact email/phone]."

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. • 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, 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.

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

- Preprocessing scripts written in R are available for replicating the data preparation process.

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). Develop logistic regression models to predict high vocabulary proficiency based on age, norming status, and lexical categories. ii). Analyze trends in the acquisition of different word categories, such as Function Words and Sensory Words. iii). Evaluate the impact of linguistic environments (norming vs. non-norming groups) on language development.
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 published on the CDI Wordbank platform and related papers on early language development. Future papers using this dataset will reference these repositories as part of their methodology. The Wordbank repository is accessible at: https://wordbank.stanford.edu/.
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 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 norming sample may not fully capture linguistic and socio-economic diversity. Care should be taken to avoid overgeneralizing findings to underrepresented populations.
- Data Transformation: Standardization of age and vocabulary scores may obscure fine-grained variations, which could be relevant for specific applications.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - TBD
6. *Any other comments?*
    - The dataset offers 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 publicly available for use by third parties. It is hosted on the Wordbank platform, managed by Stanford University's Language and Cognition Lab, and is intended for educational, research, and analytical purposes. Researchers, educators, and other users can access the data under appropriate licensing terms, with the stipulation that it remains anonymized and is used ethically and responsibly. The platform ensures the data is freely accessible to promote transparency, reproducibility, and collaborative research in early language development.
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 website (wordbank.stanford.edu) and can be accessed using tools like the WordbankR R package, which provides programmatic access to the data. Users can explore and download data directly from the platform.
3. *When will the dataset be distributed?*

- TBD

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 dataset is distributed under applicable terms of use that ensure ethical usage and compliance with data privacy standards. Licensing details and terms of use can be found on the Wordbank platform. There are no fees for accessing the data.

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 third-party IP-based restrictions are explicitly mentioned for this dataset. The data is anonymized and made available for educational and research purposes without additional restrictions.

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 regulatory restrictions apply to the dataset. It is publicly accessible and complies with ethical standards for anonymized data sharing.

7. *Any other comments?*

    - The Wordbank dataset is an invaluable 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?*

    - TThe 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 (wordbank.stanford.edu), which is regularly updated and maintained to ensure accessibility and usability for the research community. The lab actively supports users by providing tools like the WordbankR package and responding to queries related to the dataset.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - The dataset manager can be contacted via email at wordbank-contact@stanford.edu. Additional inquiries can be directed to the institution's research office at [https://wordbank.stanford.edu/about/] & [https://mb-cdi.stanford.edu/board.html] for the orignal 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 we are always interested in adding more datasets to our site! To see which researchers have already contributed, click here: http://wordbank.stanford.edu/contributors. If you are interested in contributing your own data, please contact us via email at 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, retention is guided by ethical standards and legal requirements. Anonymized data will be retained indefinitely for research purposes, but raw data containing any identifiable information will be deleted after a predetermined period (e.g., 5 years) to comply with data protection regulations. Participants were informed about these retention limits during the consent process.

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 the dataset repository. However, they will not be actively updated. Users will be notified of new versions and encouraged to transition to the most recent iteration.

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, contributions are welcome through a collaborative platform such as GitHub. Potential contributors will submit proposed changes or extensions, which will undergo validation by the research team. Accepted contributions will be documented in a changelog and credited to contributors. Clear guidelines for extensions will be provided to ensure consistency and quality.

8. *Any other comments?*

   - maintaining transparency and open communication with dataset users is a priority. Regular updates, user feedback mechanisms, and adherence to data management best practices will ensure the dataset remains a valuable resource for researchers and educators in the field of 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.