Data paper

Analyzing Linguistic features and Interaction Patterns from a Guided Drawing Task: A Case Study of Hong Kong Kindergarten Children particularly PN, K2, K3

Kalen Chan, Anson Chui, Shunem Chung, Tom Lee

Author roles

Kalen Chan - Formal analysis, Data curation

Anson Chui - Methodology, Data curation

Shunem Chung - Resources, Data curation

Tom Lee - Conceptualization, Data curation

Abstract

This dataset provides a set of recordings and annotation about a guided 15-minute drawing task and a mini game between an instructor and one participant. The tasks are conducted in English, targeting PN, K2 and K3 children from Hong Kong. Audio data were collected from 6 kids, all studying in the same school Joyful World International Nursery & Kindergarten. One boy and one girl from each grade respectively.

While current existing datasets often focus on structured or classroom-based interactions, there is a room of improvement in Hong Kong for guiding and discovering individual youngsters to express themselves in creative, low-pressure settings. The project fills this gap by developing a linguistically dataset derived from interactive tasks, designed to empower parents and teachers with actionable insights into their children's language abilities.

Keywords: English, child education, art therapy, drawing, reframing problems

(1) Overview

Repository location

Zenodo: https://doi.org/10.5281/zenodo.15411387

GitHub: https://github.com/yauchunlee/Dataset_of_6_Kindergarten_Students

Context

This dataset was produced as part of a group research project for LIN3046 Language Information Management of the Education University of Hong Kong to fill the existing gap in the available datasets. The audio data collection from 6 kindergarten kids has been cleaned and transformed into transcription files for further linguistic analysis. Files derived from the transcription, which includes excerpts, sentence length, tokenization, linguistic features, and grammatical mistakes, are uploaded and stored properly to GitHub through the Visual Studio Code for public access. After receiving feedback from the instructors and classmates in the presentation, adjustments for the dataset were made and published on Github as version 1.2.

(2) Method

Steps

Participants

We targeted 10 random parents from one of the authors' workplace, an international kindergarten, interviewing them to rate their children's English proficiency. Next, we selected those children whose parents considered their English is not good.

Data collection

The tasks were fully conducted in English face-to-face and one-to-one. For the first drawing task, the instructor distributed a "Think-out-of-the-box"worksheet to the child with a big "rainbow" shape on it. Starting the task by asking them "This is not a rainbow, instead, this is a … ?" Next, she demonstrated how to draw outside the shape and turned the paper horizontally. Besides wooden colored pencils, additional selected decorations were offered, 3D stickers, glitter glue. And stuffed toys as their drawing accompany while encouraging their creativity. With guided questions including "Does the rainbow shape look like colourful bears' ears / Mr. Seal's hands?" The child then drew additional lines and sketches, jumping out of the box to illustrate what they intended.

After completing the first one, the second task was a coloured Mouse and Cat Mini-Game. The instructor and the player flipped the paper and chose a color. Then, they chased each others' coloured pencil on paper for around 20 seconds, and asked the player to describe what the result pattern looks like.

The tasks were inspired by design thinking, these activities allow children to create art while verbally explaining how they think.

Software

The raw audio files were denoised utilizing Myedit for maximizing audio clarity. The transcription files (.txt) from speech to text was conducted using the Azure AI Video Indexer. To ensure the conversion was performed accurately, our team members manually checked each of the text files using audio files as reference to add correspondent name tags, correct spelling and punctuation errors caused by misinterpretation, resolve context ambiguity, fix formatting (e.g., capitalization, spacing and indentation) to improve overall clarity.

After retrieving the excerpts of the interviewees' speech from the transcriptions using the Gemini-2.0-Flash, we calculated the statistics of the mean length of utterances and morphemes counts of each utterance through the Morpheme Counter (https://www.morphemecounter.com/). It is a web tool that utilizes the MorphyNet: a Large Multilingual Database of Derivational and Inflectional Morphology (+morpheme segmentation) dataset available on GitHub repository at

https://github.com/kbatsuren/MorphyNet. A total of 2820 morphemes across 625 utterances are collected from the 6 child participants.

To prepare speech data for future reusability, we performed tokenization and part-of-speech (POS) tagging using the Natural Language Toolkit (NLTK) library in Python with Google Colab. We used the word_tokenize function to break down the speech into individual tokens and then applied the pos_tag function to assign grammatical tags to each token. After mounting the Google Drive in the Colab, the JSON files of the tokenization with POS tags are saved directly to the designated folder.

Quality control

To validate data quality, the audio was recorded using a Zoom H6 recorder. The meeting location was set in quiet and comfortable places such as clubhouses and coffee shops. As the author has been the interviewees' playgroup instructor for 2 - 3 years, she was appointed to be the communicative partner to ensure children feel comfortable expressing themselves and reduce performance anxiety.

All transcriptions were checked by at least 2 team members. The project also validated its methodology and document collection through consultations with the course instructors.

(3) Dataset Description

Repository name

Zenodo: https://doi.org/10.5281/zenodo.15411387

GitHub: https://github.com/yauchunlee/Dataset_of_6_Kindergarten_Students

Object name

Metadata, Transcriptions, Excerpts, Sentence_Length, Tokenization_with_POS_Tags, Linguistic_Features, Grammatical_Mistakes

Format names and versions

.csv, .txt, .md, and .json

Data Structure

The dataset is organized by having six individual files for each child participant, one in each of the following folders: 'Transcriptions', 'Excerpt's, 'Sentence_Length', 'Tokenization_with_POS_Tags', 'Linguistic_Features', and 'Grammatical_Mistakes'.

'Metadata' introduces background information for each child, 'Transcriptions' is the raw folder that contains the transcribed text that has been manually checked. 'Excerpts' compiles of utterances spoken by each child participant. 'Sentence_Length' includes the statistics of

the mean length of utterances and morphemes counts of each utterance. ‘Tokenization_with_POS_Tags’ contains the output of automatic tokenization and part-of-speech (POS) tags using the Natural Language Toolkit (NLTK), and ‘Linguistic_Features’ and ‘Grammatical_Mistakes’ highlights the characteristics in their speech.

Creation dates

2025-04-22 to 2025-05-14

Dataset creators

Kalen Chan (formal analysis, data curation, Education University of Hong Kong), Anson Chui Chi Cheuk (methodology, data curation, Education University of Hong Kong), Shunem Chung Shu Nim (resources, data curation, Education University of Hong Kong), Tom Lee Yau Chun (conceptualization, data curation, Education University of Hong Kong)

Language

English

License

CC BY 4.0

Publication date

2025-05-14

(4) Reuse Potential

The data collected from our study, which includes a guided drawing task along with recordings of six kindergarten children in Hong Kong, provides potential for reuse in academic research. The annotated transcripts of the children’s verbal responses and the metadata on their linguistic backgrounds offer a foundation for improving studies in language research, pedagogical strategies, and technology development.

4.1 Language Research

Firstly, the dataset is a valuable resource for investigating early English language development in bilingual contexts. The annotated transcripts and part-of-speech (POS) tags allow for analysis of vocabulary acquisition, grammatical accuracy, and syntactic range across these different age groups. For example, the frequent phonological simplification and error, specifically the consonant cluster reduction of “flip” for “sweet”, could inform computational models of child speech recognition, which it often struggles with developmental mispronunciations. Additionally, the task-based responses help researchers examine how structured activities affect speakers’ language output, which provides insights

for creating effective language assessment tools. Therefore, comparing the linguistic patterns of PN, K2, and K3 children could establish future developmental benchmarks for English proficiency in multilingual kindergartens.

4.2 Pedagogy

Secondly, the dataset of the recorded instructor-child interactions offer practical insights for educators and school curriculum developers. From the results of this project, it was found that K3 children produced more complex sentences during open-ended prompts, for example “I drew a house beside the tree”, and they were able to use connectives within their speech, which shows their communication skills. However, PN children mainly replied with single-word responses, such as “milk”, this signifies the importance of age-appropriate questioning. The data could serve as a training tool for teachers in adapting their language to students’ early development stages, such as using close-ended questions for young children and more descriptive prompts for older learners. Additionally, the connection between visual aids, such as the colour prompts and stickers, and increased vocabulary use suggests multisensory tasks improve language production. Hence, these insights could improve classroom activities by integrating visual and verbal components, optimizing language learning for young ESL children.

4.3 Collaboration and Technology

Lastly, the dataset’s annotations of turn-taking, response latency, and nonverbal cues provide an opportunity to enhance technologies designed for children. For instance, digital voice assistants like Siri or Alexa could benefit from this data to refine speech recognition algorithms for child users, specifically for interruptions and repetitive phrases, or parental supervision. Notably, the transcripts reveal how rephrasing elicits longer responses, which is an applicable analysis to developing conversational AI that inputs children’s linguistic needs. The metadata could assist developers in personalizing educational software for bilingual learners, which ensures culturally and linguistically appropriate content.

4.4 Limitations

Despite its potential, the dataset does have limitations that may prevent its reuse. The small sample size limits the generalizability, especially for studying broader and diverse bilingual contexts. The absence of video recordings may affect analysis of nonverbal cues, which is important in child communication studies. It is worth noting that some SEN symptoms were observed from a few interviewees which caused limited language output. Additionally, the dataset’s specific scope of utilizing guided drawing tasks may not fully capture the range of children’s speech abilities and requirements. It likely reduces chances of in-depth detailed analysis for sociolinguistic research. Therefore, these utility barriers could possibly minimum diverse reuse across different domains.

Overall, the dataset is curated for reusability, especially as a versatile tool for both academic and applied research. By sharing the resource under the CC BY 4.0 license, it is available for

collaborations across disciplines that are in need for refining language assessment models, teacher training programs, and advancing child-friendly AI technologies. This dataset could be expanded for diverse language pairs or longitudinal studies as well.

Acknowledgements

This research would not have been possible without the generous contributions of the participating parents and children. We are deeply grateful for their willingness to share and give consent to their authentic interactions, which form the foundation of this dataset. We also thank Chaak Ming Lau, Sam Tak Sum Wong, and Haoran Ho Yin Cheung for their guidance on the dataset construction.

Competing interests

The author(s) has/have no competing interests to declare.

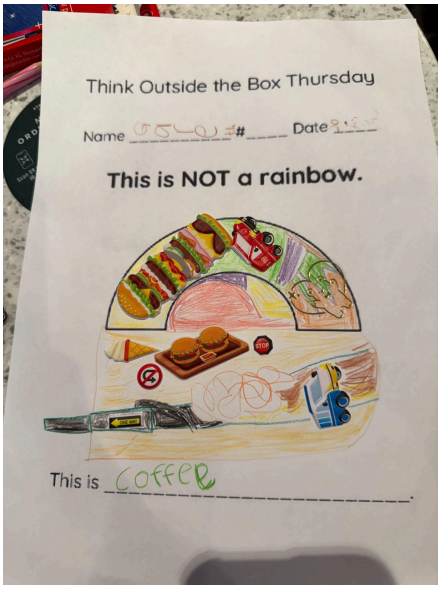
Appendices

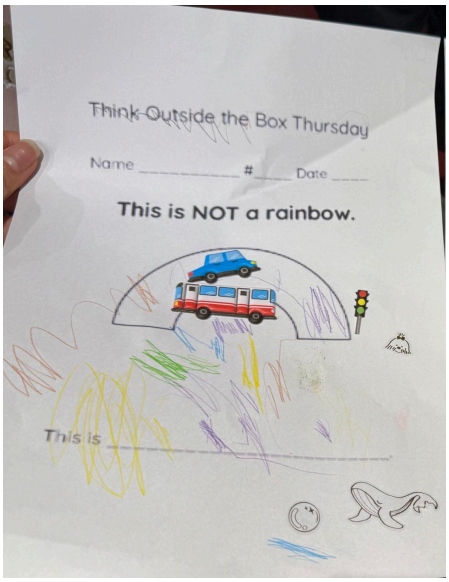
Think Outside the Box Thursday

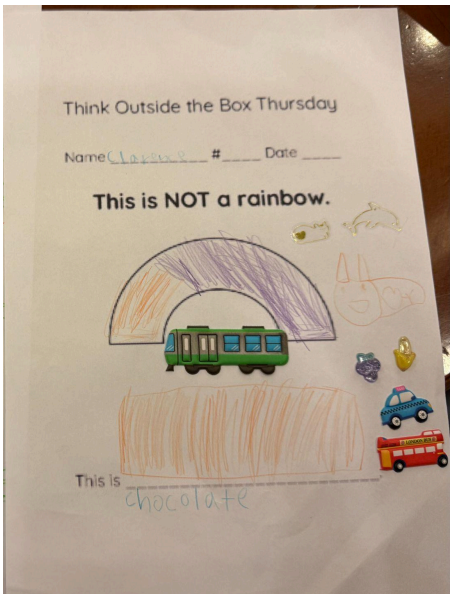
Name _____ #____ Date ____

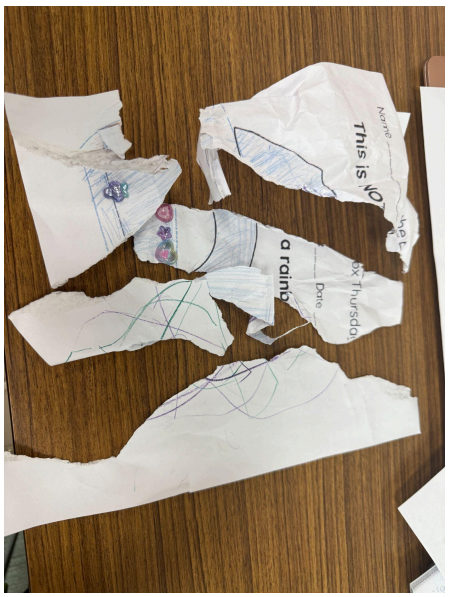
This is NOT a rainbow.

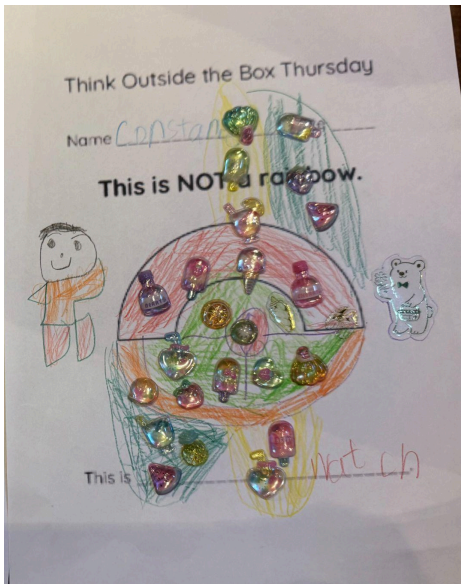
This is _____.

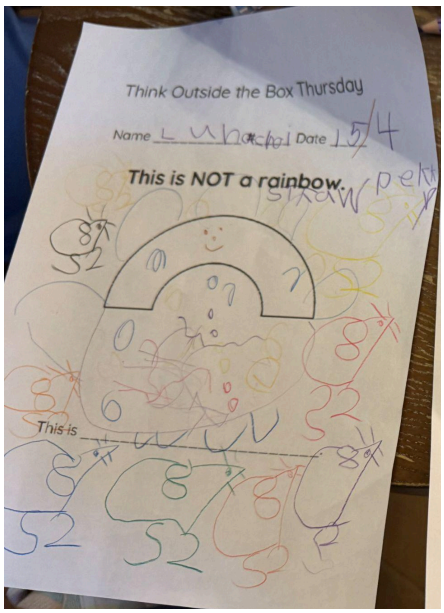









Think Outside the Box Thursday
Name
This is NOT a rainbow.
This is


Think Outside the Box Thursday
Name
Date
This is NOT a rainbow.
This is


12
Luna
Watercolour pencils
Classic
FREE