frontiers | Frontiers in Language Sciences

TYPE Original Research
PUBLISHED 23 April 2025
DOI 10.3389/flang.2025.1556481

OPEN ACCESS

EDITED BY
Lucia Colombo,
University of Padua, Italy

REVIEWED BY
Bernd J. Kröger,
RWTH Aachen University, Germany
Elizabeth Simmons,
Sacred Heart University, United States

*CORRESPONDENCE
Monami Nishio
nishio-mo@ncchd.go.jp

RECEIVED 07 January 2025
ACCEPTED 26 March 2025
PUBLISHED 23 April 2025

CITATION
Nishio M, Koyanagi A, Yakura H, Hanawa T and Shi S (2025) Language-specific development of noun bias beyond infancy.
Front. Lang. Sci. 4:1556481.
doi: 10.3389/flang.2025.1556481

Language-specific development of noun bias beyond infancy

Monami Nishio[1]*, Ayuha Koyanagi[2], Hiromu Yakura[3], Takaya Hanawa[4] and Shoi Shi[5]

[1]Department of Neonatology, National Center for Child Health and Development, Tokyo, Japan, [2]Fvital Inc., Tokyo, Japan, [3]Center for Humans and Machines, Max-Planck Institute for Human Development, Berlin, Germany, [4]Department of Pediatrics, The University of Tokyo Hospital, Tokyo, Japan, [5]International Institute for Integrative Sleep Medicine, University of Tsukuba, Tsukuba, Japan

Speech and language delays can significantly impact a child's learning, literacy, and social development, making early detection—particularly through vocabulary monitoring—essential. One well-established phenomenon in early language acquisition is the "noun bias," where infants acquire nouns more readily than verbs. However, the developmental trajectory of this bias beyond infancy remains unclear, especially across different languages. In this study, we analyzed spontaneous speech using AI-based voice analysis to examine vocabulary development in Japanese- and English-speaking children across a broad age range. We quantified changes in noun and verb use over time and found that noun growth plateaued earlier in English than in Japanese, resulting in a more pronounced and persistent noun bias in Japanese beyond infancy. These findings suggest that the early noun bias may gradually converge with adult-like noun-to-verb ratios, which differ substantially across languages (e.g., 23,800:7,921 in English vs. 71,460:7,886 in Japanese). This study demonstrates the utility of AI-based tools in advancing language development research and underscores their potential for clinical applications in identifying and assessing speech and language delays.

KEYWORDS

noun bias, artificial intelligence, voice, Japanese, English

Highlights

- Validated an AI-based voice analysis pipeline that accurately tracks vocabulary development in Japanese and English with high precision and recall (>0.7).
- Discovered noun growth plateaus around 40 months in English, while it continues to grow steadily between 10 to 50 months in Japanese.
- Revealed that noun bias diminishes rapidly in English but persists in Japanese, surpassing English by age two and persisting beyond infancy.
- Provided insights into how infancy noun bias transitions to adult noun-to-verb ratios, varying significantly between Japanese and English.

Introduction

Speech and language delays and disorders can pose significant challenges for children and their families (Barry et al., 2024). Research indicates that school-aged children with these delays are at an increased risk of developing learning and literacy difficulties, such as problems with reading and writing (Conti-Ramsden et al., 2012; Catts et al., 2008; Lewis et al., 2015). Observational cohort studies further suggest that these children may face

higher risks of social and behavioral issues, in addition to learning challenges, some of which can persist into adulthood (Dubois et al., 2020).

Tracking vocabulary development is critical for early detection and intervention of speech and language delays (Barry et al., 2024). Early language development in infancy often exhibits a "noun bias," where infants acquire nouns earlier and more rapidly than verbs (Gentner, 1982). This phenomenon, first observed in English, has sparked debate about its universality across languages (Frank et al., 2021). Studies across 16 languages confirm the presence of a noun bias across languages, but its strength varies—being more pronounced in noun-focused European languages and weaker in verb-centric non-European languages (Frank et al., 2021). It is uncertain whether the noun bias persists beyond infancy, though it is plausible that the noun-to-verb ratio gradually aligns with adult language patterns. For example, while English displays a stronger noun bias during early development compared to Japanese (Frank et al., 2021), the adult ratios suggest a reversal. The noun-to-verb ratio is approximately 23,800:7,921 in English (Chi, 2015) and 71,460:7,886 in Japanese (Kindaichi et al., 2022), with Japanese exhibiting a stronger noun bias. This may suggest a reversal in the noun-to-verb ratio at some point in development, most likely beyond infancy, although such changes have not been reported.

It has been challenging to assess vocabulary growth beyond infancy, as vocabulary expands rapidly and becomes difficult to track manually. Parental vocabulary checklists, such as the MacArthur-Bates Communicative Development Inventories (CDI) (Bates et al., 1994) and CDI-III, are widely used for screening. However, these are based on parental reports, therefore the quality of responses depends on the parent's awareness and observation of the child's language use. Additionally, the CDI is designed for children under 2 years old, while the CDI-III covers children up to 3 years old. As children grow beyond infancy, their vocabulary expands rapidly, making manual tracking or parental checklists less effective. Additionally, their linguistic abilities become more complex and nuanced, requiring a more advanced assessment tool. Consequently, standardized instruments such as the Peabody Picture Vocabulary Test (PPVT) (Olabarrieta-Landa et al., 2017) are better equipped to detect subtle variations or delays in vocabulary acquisition that may not be evident through parental observations alone. However, the PPVT requires administration by trained professionals who can ensure standardized procedures, accurate response interpretation, and strict adherence to scoring criteria, all of which are critical for maintaining reliability and validity. The global shortage of such qualified professionals severely limits the accessibility of those assessment tools (Olabarrieta-Landa et al., 2017).

Recent technological advancements have enabled recording in everyday settings and significantly reduced coding costs through partial automation using natural language models. For example, the Language ENvironment Analysis (LENA) system (Greenwood et al., 2011) is designed to capture and analyze the natural language environment of children by employing wearable audio recorders in real-life settings such as the home, childcare facilities, and other everyday environments. These recordings capture extended periods of naturalistic audio, which are then processed by an advanced computational framework. At its core, LENA uses a neural network-based architecture to automatically identify and categorize acoustic events relevant to language development, such as adult word count (AWC), child vocalization count (CVC), and conversational turn count (CTC). The system's neural networks are trained using large, annotated corpora of audio data, employing supervised learning techniques. LENA has advanced child language research over the past decade (Bergelson et al., 2023) but is less reliable for non-European languages due to its algorithm being trained on American English. For instance, the correlation between LENA's results and manual annotations for Adult Word Count (AWC) is 0.92 for English (Xu et al., 2009), but only 0.73 for Mandarin (Gilkerson et al., 2015) and 0.72 for Korean (Pae et al., 2016).

Meanwhile, the recent surge in publicly available general-purpose machine learning tools, such as OpenAI's GPT and Azure's whisper (Radford et al., 2023), provides means to instantly perform speech and text analysis in various languages. In a previous study, we demonstrated that a pipeline integrating readily available machine learning tools could accurately estimate the vocabulary of Japanese children based on voice recordings from nursery schools (Nishio et al., 2024). At the public nursery school, we positioned a smartphone on a tripod in front of the children while they engaged in activities such as origami, cutting paper with scissors, or coloring. Typically, there are two to three children at each table, and the teacher interacted with them simultaneously. Our analysis included seven children aged between 3 and 5 years old. Nursery school teachers, who regularly interact with the children, completed a developmental questionnaire called the Enjoji Scale of Infant Analytical Development (Enjoji and Yanai, 1961). The vocabulary size derived from these recordings showed a strong correlation with language skills as assessed by nursery school teachers.

In this study, we validate an AI-driven voice analysis pipeline to track vocabulary development in Japanese- and English-speaking children beyond infancy. Our analysis focuses on the growth of nouns and verbs to examine how the noun bias changes beyond infancy.

Methods

Datasets

We used two publicly available longitudinal voice recording datasets.

NTT INFANT dataset (ages 0–4, 300–700 recordings per child, totaling 2,415 h)

The NTT INFANT dataset (Amano et al., 2009) consists of longitudinal recordings of six monolingual Japanese-speaking children aged between 0 and 5 years. In total, the dataset comprises 541 h of recordings. Recordings took place in a room within the participants' houses using a SONY TCD-D10 digital audio recorder and a SONY ECM-959 stereo microphone, with 16-bit quantization and a 48 kHz sampling frequency. The microphone was either held by a parent or placed on a stand during recording. No specific tasks were assigned to the infants or their parents, allowing the

capture of natural daily life utterances. Although the majority of the recordings include utterances from the infant and parents, occasional contributions from siblings or relatives were also recorded due to the everyday setting. Trained transcribers listened to each utterance and produced transcriptions in both a kanji-hiragana-mixed form and a katakana-only form. One well-trained transcriber checked and corrected the other transcribers' work.

PhonBank English Providence Corpus (ages 1–3, 40–60 recordings per child, totaling 364 h)

The PhonBank English Providence Corpus (Evans and Demuth, 2012; Yung Song et al., 2013) consists of longitudinal recordings of six monolingual English-speaking children aged between 1 and 3 years. The entire corpus comprises 364 h of speech. Recordings were made in the children's houses during spontaneous interactions, primarily between the children and their mothers. During the recording sessions, both the child and the mother wore wireless Azden WLT/PRO VHF lavalier microphones attached to their collars while engaging in everyday activities. Recordings were made using a Panasonic PV-DV601D-K mini digital video recorder, with the audio subsequently extracted and digitized at a 44.1 kHz sampling rate. Both adult and child utterances were orthographically transcribed using Codes for the Human Analysis of Transcripts conventions (Macwhinney, 1992). The children's utterances were also transcribed by trained coders using International Phonetic Alphabet (IPA) transcription, to capture the phonetic representations of words. Overall reliability of IPA-transcribed segments ranged from 80–97% across files in terms of presence/absence of segments and place/manner of articulation.

Voice analysis pipeline

Our analysis employed a previously proposed voice analysis pipeline designed to function without the need for a large dataset or computationally intensive training (Nishio et al., 2024). This pipeline is immediately applicable not only to Japanese but also to various non-English languages. The details of each component of our pipeline are provided below.

Audio segmentation based on transcription intervals

Initially, we used Azure AI Speech (Microsoft Data Science Process Team, 2020) to identify audio segments containing speech. We then classified the speakers within these segments, assigning each a unique identifier. This process differentiates between speakers in the audio and lays the foundation for subsequent analyses.

Speaker identification

The recorded environment typically involves interactions between a teacher and a child, with the assumption that the teacher speaks more frequently and uses longer words compared to the child. Based on this premise, we identified the most frequently occurring speaker from the audio segmentation results as the "teacher" and the second most frequent as the "child." Other speakers were considered background noise or misidentifications and were excluded from further analysis.

Transcription of speech

After speaker identification, each segmented speech interval was transcribed into text. We applied three different model sizes (tiny, medium, and large) of both Whisper (Radford et al., 2023) and Azure Speech-to-Text. This step converts the audio data into text format, thereby enabling further linguistic analysis. Notably, Whisper has been trained on 680,000 h of multilingual and multitask data collected from the web, which enhances its robustness to diverse accents and grammatical structures across multiple languages.

Morphological analysis

From the transcription results, we extracted the child's sentences and performed a word-level morphological analysis using the Sudachi (Takaoka et al., 2018) dictionary. This analysis deepens our understanding of the child's language usage patterns by focusing specifically on four parts of speech commonly used by children: nouns, verbs, adjectives, and adverbs.

Evaluation of pipeline performance

To evaluate the accuracy of transcribing child speech segments, we calculated Character Error Rate (CER), Word Error Rate (WER), and BERT (Devlin et al., 2019) score. CER and WER are common metrics used to assess the performance of automatic speech recognition systems by comparing the generated output to the reference transcription. CER measures the proportion of incorrect characters in the output compared to the reference, while WER measures the proportion of incorrect words. For example, if a child's pronunciation leads to the word "shensuikan" instead of "sensuikan", despite the words conveying the same meaning, it would be counted as an error in both CER and WER, even though the intended meaning is preserved. These metrics focus on exact word or character matches, which can be problematic for speech recognition in children, where pronunciation variations are common. To address this limitation, we adopted the BERT score, which evaluates the quality of the transcription by comparing the contextual similarity between the predicted output and the reference sentence using the BERT model (Devlin et al., 2019). Unlike CER and WER, which rely on exact matches, the BERT score captures the semantic meaning and contextual relevance of the entire sentence, making it more robust to minor pronunciation differences and better suited for evaluating the accuracy of speech in natural settings, where context is crucial.

Developmental trajectory analysis

To flexibly model both linear and non-linear relationships between age and vocabulary development, we employed Generalized Additive Models (GAMs) (Wood et al., 2015;

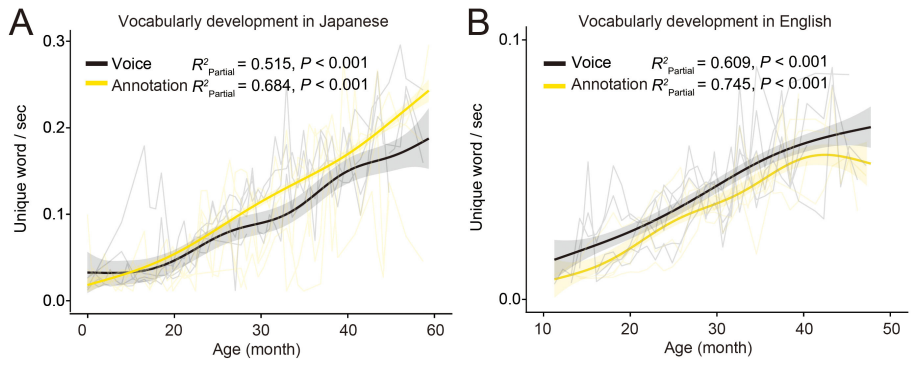

FIGURE 1
Transcription accuracy of the AI-based voice analysis pipeline. **(A, B)** Developmental trajectory of the total unique words in Japanese [**(A)** Voice: $R^2_{partial} = 0.515$ $P < 0.001$, Annotation: $R^2_{partial} = 0.684$ $P < 0.001$] and English [**(B)** Annotation: $R^2_{partial} = 0.745$ $P < 0.001$], based on manual annotation (yellow) and voice transcription (black).

Stasinopoulos, 2007) using the *mgcv* package in R (Stasinopoulos, 2007). GAMs extend traditional linear models by allowing for smooth, non-parametric relationships between predictor and outcome variables, making them well-suited for capturing complex developmental trends. Unlike linear models, which assume a strict linear relationship between predictors and the response variable, GAMs use smooth functions to model nonlinear patterns, providing greater flexibility in data analysis.

Our model was specified as follows:

$$Vocabulary_i = \beta_0 + f(Age_i) + \beta_1 Gender_i + \epsilon_i$$

where Vocabulary represents the vocabulary size of child *i*, Age is modeled as a smooth function *f(Age)* using thin plate regression splines, Gender is included as a linear covariate, and ε is the residual error term. The smooth function *f(Age)* allows us to estimate vocabulary development trajectories without assuming a specific parametric form.

To prevent overfitting while preserving model interpretability, we set the maximum basis complexity (*k*) to 3, ensuring a balance between flexibility and generalizability. The smoothing parameter was estimated using the restricted maximum likelihood (REML) approach, which optimally determines the degree of smoothness based on the data. This modeling approach enables us to capture both gradual and abrupt changes in vocabulary development, providing a more nuanced understanding of language growth over time.

For each GAM, we assessed the significance of the association between the vocabulary and age using an analysis of variance (ANOVA), comparing the full GAM model to a nested model without the age term. A significant result indicates that including a smooth term for age significantly reduced the residual deviance, as determined by the chi-squared test statistic. For each GAM, we identified the specific age range(s) where the vocabulary significantly changed using the gratia package in R. Age windows of significant change were determined by examining the first derivative of the age smooth function (Δ Vocabulary/Δ age) and assessing when the simultaneous 95% confidence interval of this derivative did not include 0 (two-sided). To quantify the overall magnitude and direction of the association between the vocabulary and age, referred to as an overall age effect, we calculated the partial R^2 between the full GAM model and the reduced model for effect magnitude. We then signed the partial R^2 based on the average first derivative of the smooth function for effect direction.

Results

Voice analysis pipeline accurately tracks vocabulary development across languages

We first assessed the accuracy of the AI-based voice analysis pipeline using two publicly available longitudinal voice recording datasets. The NTT INFANT dataset (Amano et al., 2009) includes audio recordings from five monolingual Japanese children (ages 0–4, 300–700 recordings per child, totaling 2,415 h), while the PhonBank English Providence Corpus (Evans and Demuth, 2012; Yung Song et al., 2013) contains audio/video recordings from six monolingual English children (ages 1–3, 40–60 recordings per child, totaling 364 h), all collected during spontaneous interactions with parents at home. Native speakers transcribed the recordings. We quantified developmental changes using Generalized Additive Models (GAM) (Sydnor et al., 2023), calculating the differential in R^2 values between the full and reduced models excluding age effects.

We found a clear alignment between the developmental trajectory of overall vocabulary from manually annotated data and that derived from our voice analysis model, both for Japanese (Figure 1A, Voice: $R^2_{partial} = 0.515$ $P < 0.001$, Annotation: $R^2_{partial} = 0.684$ $P < 0.001$) and English (Figure 1B, Voice: $R^2_{partial} = 0.609$ $P < 0.001$, Annotation: $R^2_{partial} = 0.745$ $P < 0.001$). Our analysis showed that precision and recall, based on the BERT model (Devlin et al., 2019), exceeded 0.7 for both languages, with these scores improving as the children aged. In addition, Character Error Rate (CER) and Word Error Rate (WER) decreased through development (Table 1). These metrics reflect the child's growing proficiency in both the phonetic and syntactic aspects of language development. More specifically, a reduction in CER indicates that the accuracy of individual character recognition improved over time, suggesting that the

child's pronunciation became more accurate and easier for the AI to recognize. Similarly, the decrease in WER shows that overall word recognition improved, highlighting progress in the child's ability to produce and structure words correctly in speech.

Noun growth plateaus earlier in English compared to Japanese

Using this pipeline, we analyzed vocabulary development in nouns and verbs for both languages. Manually transcribed data revealed that nouns increased faster than verbs in both Japanese and English, consistent with previous studies (Gentner, 1982; Frank et al., 2021) (Figure 2A, Noun: $R^2_{partial} = 0.681$ $P < 0.001$, Verb: $R^2_{partial} = 0.606$ $P < 0.001$, Figure 2B, Noun: $R^2_{partial} = 0.663$ $P < 0.001$, Verb: $R^2_{partial} = 0.752$ $P < 0.001$). Across ages 10–50 months, nouns outnumbered verbs in vocabulary size for both languages (Figure 2C, Japanese: $t = -17.196$ $P < 0.001$, English: $t = -16.885$ $P < 0.001$). However, language-specific differences emerged: in English, noun growth plateaued around 40 months, while in Japanese, noun growth continued steadily throughout the 10–50-month period (Figures 2A, B). Additionally, Japanese showed a larger vocabulary size for both nouns and verbs compared to English, with a wider gap between noun and verb vocabulary sizes (Figure 2C Japanese: $t = -17.196$ $P < 0.001$, English: $t = -16.885$ $P < 0.001$). Using the voice analysis pipeline, we observed developmental trajectories for both languages that closely aligned with the manually annotated results (Figure 2D, Noun: $R^2_{partial} = 0.545$ $P < 0.001$, Verb: $R^2_{partial} = 0.368$ $P < 0.001$, Figure 2E, Noun: $R^2_{partial} = 0.563$ $P < 0.001$, Verb: $R^2_{partial} = 0.593$ $P < 0.001$). Consistent with manual annotations, nouns outnumbered verbs in both languages, but Japanese had a larger vocabulary size overall and a larger gap between noun and verb vocabulary sizes (Figure 2F, Japanese: $t = -12.874$ $P < 0.001$, English: $t = -10.934$ $P < 0.001$).

Noun bias persists longer in Japanese beyond infancy

To investigate whether noun bias persists beyond infancy, we calculated the noun-to-verb ratio based on manual annotations across the age range of 10–50 months for both Japanese and English (Figure 3A, Japanese: $R^2_{partial} = 0.375$ $P < 0.001$, English: $R^2_{partial} = 0.770$ $P < 0.001$). The noun bias was strongest in early childhood, gradually weakening as development progressed in both languages, consistent with previous findings (Frank et al., 2021). However, the decline of noun bias is slower in Japanese compared to English. As a result, after ~20 months of age, the noun bias became more pronounced in Japanese compared to English. Therefore, across the age range of 10–50 months, Japanese exhibited a significantly stronger overall noun bias than English (Figure 3B, $t = 4.970$, $P < 0.001$). This pattern was also observed in the results from the voice analysis pipeline, confirming its reliability in tracking cross-language vocabulary development (Figure 3C, $t = 6.686$, $P < 0.001$).

TABLE 1 Accuracy of AI-based voice analysis pipeline.

	CER	WER	BERT precision	BERT recall
Japanese				
Age 0	2.347	1.002	0.643	0.679
Age 1	1.546	0.992	0.699	0.712
Age 2	1.343	1.016	0.709	0.724
Age 3	1.278	0.978	0.736	0.758
Age 4	1.400	0.972	0.732	0.759
Total	1.374	0.992	0.718	0.738
English				
Age 1	0.928	1.355	0.702	0.713
Age 2	0.885	1.270	0.723	0.739
Age 3	0.852	1.188	0.728	0.738
Total	0.884	1.264	0.720	0.733

Character Error Rate (CER), Word Error Rate (WER), and BERT-based precision and recall scores for transcription results in Japanese and English. Metrics are reported by age group and for all participants combined (Total). CER and WER represent transcription accuracy, while BERT precision and recall scores assess semantic alignment with ground truth. Lower CER/WER values and higher BERT scores indicate better performance.

Discussion

In this study, we validated the use of an AI-based voice analysis tool to track vocabulary development across a wide age range beyond infancy. While existing tools like the LENA system (Greenwood et al., 2011) are primarily trained on English, limiting their applicability to other languages, our approach utilizes widely adopted machine learning tools, enabling broader cross-linguistic compatibility. By applying these tools to densely annotated voice recordings of infants and toddlers in English and Japanese, we found that the growth of nouns plateaus around 40 months in English, earlier than in Japanese. As a result, while noun bias is similarly strong during infancy in both languages, it diminishes more rapidly in English but persists in Japanese beyond infancy. This leads to noun bias in Japanese surpassing that of English around age two. These findings offer insight into how the noun bias observed in infancy evolves to align with the noun-to-verb ratio seen in adult language.

Measuring the genuine language competence of children presents challenges for both researchers and clinicians (Siu, 2015). Despite the small sample size, we have demonstrated the validity of our pipeline in reliably measuring children's language development. This approach is innovative not only for its time efficiency and low physical cost but also for its usability in natural education and childcare settings. It allows for data collection over much longer periods compared to clinical assessments conducted in time-constrained hospital environments, enhancing assessment reliability. Additionally, since children are more familiar with their everyday environments, their performance is likely to more accurately reflect their true language abilities compared to unfamiliar clinical settings.

Noun bias is a well-established feature of early vocabulary development, though its universality is still debated (Frank

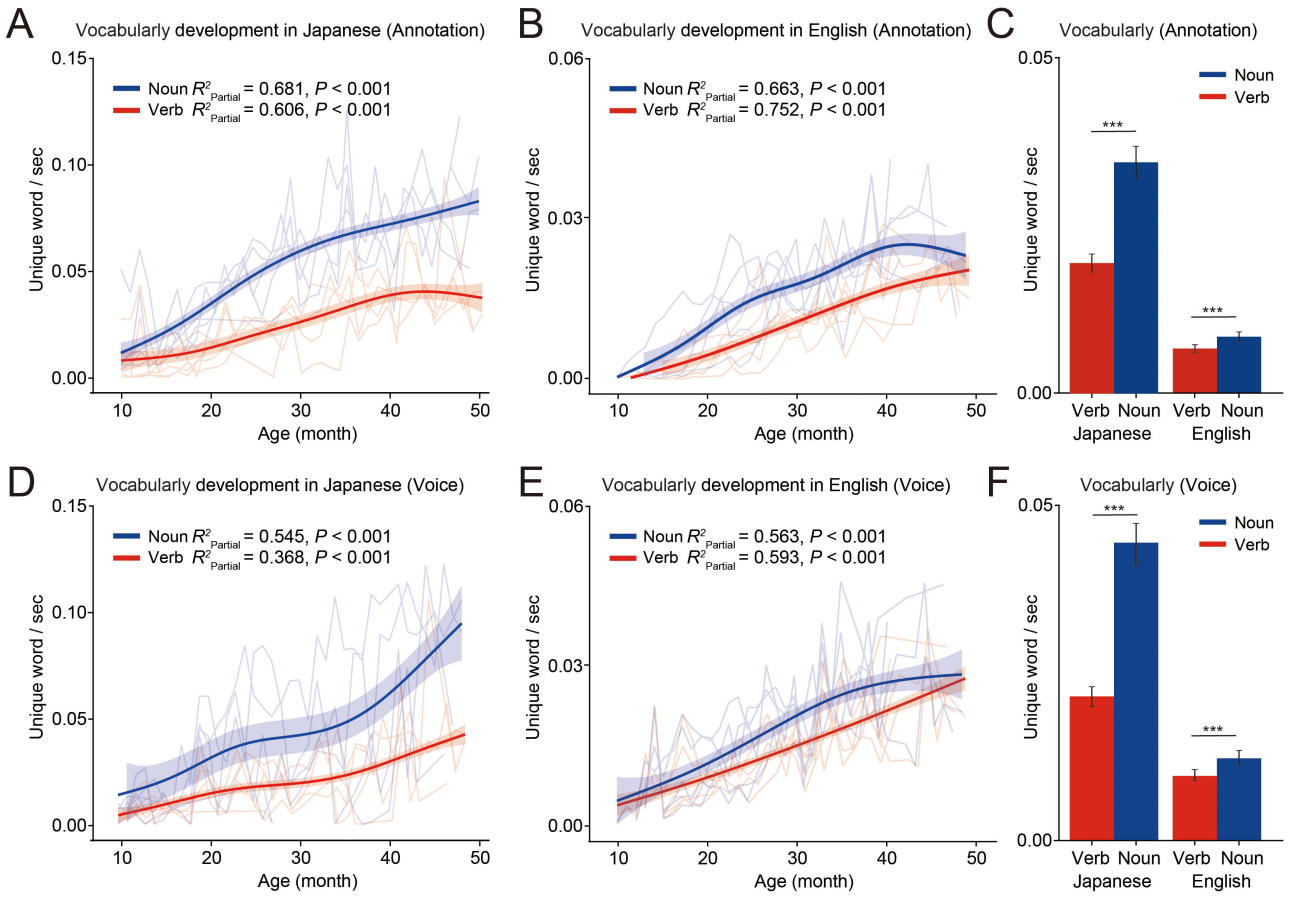

FIGURE 2
Vocabulary development in Japanese and English. **(A, B)** Vocabulary development in Japanese [**(A)** Noun: $R^2_{partial} = 0.681$ $P < 0.001$, Verb: $R^2_{partial} = 0.606$ $P < 0.001$] and English [**(B)** Noun: $R^2_{partial} = 0.663$ $P < 0.001$, Verb: $R^2_{partial} = 0.752$ $P < 0.001$] based on manual annotation. The mixed-effects model predicted the developmental trajectories of noun (blue), verb (red), adjective (green) and adverb (purple), with the 95% confidence interval shown as shaded areas. The individual trajectories are plotted as line plots ($N = 5$ for Japanese, $N = 6$ for English). **(C)** Comparison of total vocabulary for nouns and verbs based on manual annotations (Japanese: $t = -17.196$ $P < 0.001$, English: $t = -16.885$ $P < 0.001$). $^{***}P < 0.001$, Paired t-test. **(D, E)** Vocabulary development in Japanese [**(D)** Noun: $R^2_{partial} = 0.545$ $P < 0.001$, Verb: $R^2_{partial} = 0.368$ $P < 0.001$] and English [**(E)** Voice: $R^2_{partial} = 0.563$ $P < 0.001$, Verb: $R^2_{partial} = 0.593$ $P < 0.001$] based on voice analysis pipeline. **(F)** Comparison of total vocabulary for nouns and verbs based on voice analysis (Japanese: $t = -12.874$ $P < 0.001$, English: $t = -10.934$ $P < 0.001$). $^{***}P < 0.001$, Paired t-test.

et al., 2021). Discrepancies between studies are partly due to variability in methods for vocabulary counting, including parental questionnaires (Bates et al., 1994), clinical screening tools in lab environments (Olabarrieta-Landa et al., 2017), and annotations of naturalistic environmental recordings (Greenwood et al., 2011). Using annotations of household interaction recordings between children and parents, we showed that noun bias is equally strong in both Japanese and English in children under 2 years old.

Additionally, the trajectory of noun bias beyond infancy has remained unclear. Tools like the MacArthur-Bates Communicative Development Inventories (Bates et al., 1994) focus primarily on infancy. Using an AI-based analysis tool applicable across ages, we found that the noun bias remains strong in Japanese beyond infancy, while it diminishes more rapidly in English. This may be linked to the higher noun-to-verb ratio in Japanese (71,460:7,886) (Kindaichi et al., 2022) compared to English (23,800:7,921) (Chi, 2015). This is the first study to clearly demonstrate how noun bias in infancy transitions to the adult language-based noun-to-verb ratio throughout development. However, further research is needed to test this hypothesis in languages with varying noun-to-verb ratios, as well as in datasets with larger sample sizes.

Although this study demonstrates the promising potential of AI-based voice analysis, there are several limitations to note. First, the datasets used in this study are relatively small, including only five Japanese-speaking and six English-speaking children. This limited sample size may affect the generalizability of our findings to larger populations. The recording environments also differ slightly between the two datasets, which may influence the differences in vocabulary size between the languages. Secondly, while the AI-based pipeline achieved high precision and recall, it still requires improvement, especially for the speech of younger children. Errors in speech recognition, particularly with children's speech, could introduce biases in the analysis. Although the pipeline aligns well with manually annotated data, further validation through human experiments and cross-validation with independent datasets is needed to confirm the robustness of the findings.

In conclusion, our findings highlight the potential of AI-based voice analysis for quantitatively assessing vocabulary development

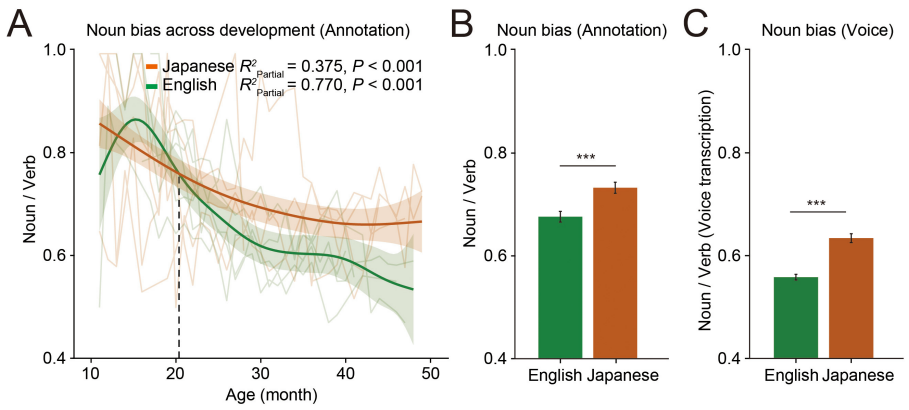

FIGURE 3
Noun bias in Japanese and English. **(A)** Developmental changes in noun bias for Japanese (orange) and English (green), based on manual annotation (Japanese: $R^2_{partial} = 0.375$ $P < 0.001$, English: $R^2_{partial} = 0.770$ $P < 0.001$). The mixed-effects model predicted the developmental trajectories of noun bias, with the 95% confidence interval shown as shaded areas. The individual trajectories are plotted as line plots ($N = 5$ for Japanese, $N = 6$ for English). **(B)** Comparison of overall noun bias based on manual annotation between English and Japanese ($t = 4.970$, $P < 0.001$). $^{***}P < 0.001$, Independent t-test. **(C)** Comparison of overall noun bias based on voice transcription between English and Japanese ($t = 6.686$, $P < 0.001$). $^{***}P < 0.001$, Independent t-test.

across languages and beyond infancy. This approach offers promising possibilities not only for advancing research but also for improving clinical screening for speech and language delays or disorders in the future.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

Author contributions

MN: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. AK: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. HY: Funding acquisition, Supervision, Writing – original draft, Writing – review & editing. TH: Writing – original draft, Writing – review & editing. SS: Funding acquisition, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. All phases of this study were supported by the research funding from Japan Science and Technology Agency (JST-Mirai Program).

Acknowledgments

We extend our sincere gratitude to the Corporate Planning Team at Fvital Inc., led by Ms. Masumi Ajima, for their invaluable support in providing resources. We dedicate this work to Dr. Yoichi Sakakihara, with deep respect and gratitude.

Conflict of interest

AK was employed by Fvital Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. We used ChatGPT (version 4.0) by OpenAI for grammar and language refinement from 11/1/2024 to 12/20/2024. The author(s) take full responsibility for the integrity of the content.

Publisher's note

References

Amano, S., Kondo, T., Kato, K., and Nakatani, T. (2009). Development of Japanese infant speech database from longitudinal recordings. *Speech Commun*. 51, 510–520. doi: 10.1016/j.specom.2009.01.009

Barry, M. J., Nicholson, W. K., Silverstein, M., Chelmow, D., Coker, T. R., Davis, E., et al. (2024). Screening for speech and language delay and disorders in children: US preventive services task force recommendation statement. *JAMA*. 331, 329–334. doi: 10.1001/jama.2023.26952

Bates, E., Hartung, J., Marchman, V., Fenson, L., Dale, P., Reznick, J. S., et al. (1994). Developmental and stylistic variation in the composition of early vocabulary. *J. Child Lang*. 21, 85–123. doi: 10.1017/S0305000900008680

Bergelson, E., Soderstrom, M., Schwarz, I. C., Rowland, C. F., Ramírez-Esparza, N., Hamrick, L., et al. (2023). Everyday language input and production in 1,001 children from six continents. *Proc. Natl. Acad. Sci. USA*. 120:e2300671120. doi: 10.1073/pnas.2300671120

Catts, H. W., Bridges, M. S., Little, T. D., and Tomblin, J. B. (2008). Reading achievement growth in children with language impairments. *J. Speech Lang. Hear. Res*. 51, 1569–1579. doi: 10.1044/1092-4388(2008/07-0259)

Chi, A. (2015). A review of Longman dictionary of contemporary English (6th edition). *Lexicography*. 2, 179–186. doi: 10.1007/s40607-015-0023-6

Conti-Ramsden, G., St. Clair, M. C., Pickles, A., and Durkin, K. (2012). Developmental trajectories of verbal and nonverbal skills in individuals with a history of specific language impairment: from childhood to adolescence. *J. Speech Lang. Hear. Res*. 55, 1716–1735. doi: 10.1044/1092-4388(2012/10-0182)

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (Minneapolis, MN: Association for Computational Linguistics).

Dubois, P., St-Pierre, M. C., Desmarais, C., and Guay, F. (2020). Young adults with developmental language disorder: a systematic review of education, employment, and independent living outcomes. *J. Speech Lang. Hear. Res*. 63, 3786–3800. doi: 10.1044/2020_JSLHR-20-00127

Enjoji, M., and Yanai, N. (1961). Analytic test for development in infancy and childhood. *Pediat. Int*. 4:7. doi: 10.1111/j.1442-200X.1961.tb01032.x

Evans, K. E., and Demuth, K. (2012). Individual differences in pronoun reversal: Evidence from two longitudinal case studies. *J. Child Lang*. 39, 162–191. doi: 10.1017/S0305000911000043

Frank, C., Braginsky, M. M., Yurovsky, D., and Marchman, A. V. (2021). *Variability and Consistency in Early Language Learning: The Wordbank Project*. Cambridge, MA: MIT Press.

Gentner, D. (1982). *Why Nouns Are Learned before Verbs: Linguistic Relativity Versus Natural Partitioning*. Technical Report No. 257, Jun. 1982, [online] Available online at: http://eric.ed.gov/ERICWebPortalldetail?accno=ED219724

Gilkerson, J., Zhang, Y., Xu, D., Richards, J. A., Xu, X., Jiang, F., et al. (2015). Evaluating language environment analysis system performance for Chinese: a pilot study in Shanghai. J. Speech Lang. Hear. Res. 58, 445–452. doi: 10.1044/2015_JSLHR-L-14-0014

Greenwood, C. R., Thiemann-Bourque, K., Walker, D., Buzhardt, J., and Gilkerson, J. (2011). Assessing children's home language environments using automatic speech recognition technology. *Commun. Disord. Q*. 32:1525740110367826. doi: 10.1177/1525740110367826

Kindaichi, K., Saeki, U., Oishi, H., Nomura, M., and Kimura, Y. (2022). *Shinsen Kokugo Jiten*. 10th ed. Tokyo: Syogakukan.

Lewis, B. A., Freebairn, L., Tag, J., Ciesla, A. A., Iyengar, S. K., Stein, C. M., et al. (2015). Adolescent outcomes of children with early speech sound disorders with and without language impairment. *Am. J. Speech Lang. Pathol*. 24, 150–163. doi: 10.1044/2014_AJSLP-14-0075

Macwhinney, B. (1992). The CHILDES project: tools for analyzing talk. *Child Lang. Teach. Ther*. 8:211. doi: 10.1177/026565909200800211

Microsoft Data Science Process Team (2020). *Azure AI guide for predictive maintenance solutions - Team Data Science Process | Microsoft Docs*. Redmond, WA.

Nishio, M., Koyanagi, A., Takamori, A., Hasuike, K., Shimoura, Y., and Yakura, H. (2024). "Automatic assessment of language developmental disorders in non-english contexts," in *Proceedings of the 2024 IEEE International Conference on E-Health Networking, Application and Services* (Nara: IEEE).

Olabarrieta-Landa, L., Rivera, D., Ibáñez-Alfonso, J. A., Albaladejo-Blázquez, N., Martín-Lobo, P., Delgado-Mejía, I. D., et al. (2017). Peabody picture vocabulary test-III: normative data for Spanish-speaking pediatric population. *NeuroRehabilitation*. 41, 687–694. doi: 10.3233/NRE-172239

Pae, S., Yoon, H., Seol, A., Gilkerson, J., Richard, J., Ma, L., et al. (2016). Effects of feedback on parent-child language with infants and toddlers in Korea. *First Lang*. 36:6. doi: 10.1177/0142723716649273

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., et al. (2023). "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning in Proceedings of Machine Learning Research 202:28492-28518*. Available online at: https://proceedings.mlr.press/v202/radford23a.html

Siu, A. L. (2015). Screening for speech and language delay and disorders in children aged 5 years or younger: us preventive services task force recommendation statement. *Pediatrics* 136, e474–e481. doi: 10.1542/peds.2015-1711

Stasinopoulos, M. (2007). Generalized additive models: an introduction with R. by S. N. WOOD. *Biometrics* 63:4. doi: 10.1111/j.1541-0420.2007.00905_3.x

Sydnor, V. J., Larsen, B., Seidlitz, J., Adebimpe, A., Alexander-Bloch, A. F., Bassett, D. S., et al. (2023). Intrinsic activity development unfolds along a sensorimotor–association cortical axis in youth. *Nat. Neurosci*. 26, 638–649. doi: 10.1038/s41593-023-01282-y

Takaoka, K., Hisamoto, S., Kawahara, N., Sakamoto, M., Uchida, Y., Matsumoto, Y., et al. (2018). "Sudachi: a Japanese Tokenizer for Business," in *11th International Conference on Language Resources and Evaluation* (Miyazaki: European Language Resources Association (ELRA)).

Wood, S. N., Goude, Y., and Shaw, S. (2015). Generalized additive models for large data sets. *J. R. Stat. Soc. Ser. C Appl. Stat*. 64:12068. doi: 10.1111/rssc.12068

Xu, D., Yapanel, U., and Gray, S. (2009). "Reliability of the LENA^{TM} language environment analysis system in young children's natural home environment," in. *LENA Technical Report* (Boulder, CO: LENA Foundation), 5.

Yung Song, J., Demuth, K., Evans, K., and Shattuck-Hufnagel, S. (2013). Durational cues to fricative codas in 2-year-olds' American English: voicing and morphemic factors. *J. Acoust. Soc. Am*. 133, 2931–2946. doi: 10.1121/1.4795772