Employing Machine Learning and Language Models to Differentiate Language Patterns in Mandarin-Speaking Preschoolers with Autism Spectrum Disorder

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder with significant challenges in communication and language. Mandarin-Speaking verbal preschoolers with ASD can exhibit difficulties in lexical and grammatical domains such as shorter Mean Length of Utterance in words (MLUw) compared to their typical developing (TD) peers (Su et al., 2018; Zhou et al., 2015). However, research on identifying ASD preschoolers using machine learning (ML) and language models (LM) based on naturalistic production data remains sparse. We report a pilot study that tests whether ML and LM can be used to distinguish between ASD and TD utterances, offering a potential for ecologically valid, cost-effective ASD screening that enables early intervention.

Primary data used utterances produced by Mandarin-speaking ASD (N = 7, $\text{Age}_{\text{Range}}$ = 4;0 - 6;0) and TD children (N = 43, $\text{Age}_{\text{Range}}$ = 3;0 - 6;0), extracted from adult-child interaction recordings from existing corpora. The ASD data were derived from the Mandarin Shanghai Corpus from the ASDBank (longitudinal, $N_{\text{ASD\_utterance}}$ = 2975), and the TD data were derived from the Beijing Child Mandarin Corpus (cross-sectional, Mai et al., in prep, $N_{\text{TD\_utterance}}$ = 3016). Data preprocessing involved manually filtering out utterances less than two words, non-verbal elements, and disfluencies, as they do not reveal rich syntactic structures (details in Table 1). Four ML models were employed: a Chinese BERT (Devlin et al., 2019), a multilingual XLM-RoBERTa, and ChatGPT to provide classification directly based on cleaned utterances and Logistic Regression to provide classification based on numerical data (e.g. MLUw) calculated from the cleaned transcripts. Except for ChatGPT which employs in-context learning, all models were trained on 78% of the data with a validation set consisting of 10% of the data, and tested on the remaining 12% data which avoids speaker overlap and ensures unbiased results. A trained speech therapist manually coded 20% of the test data ($N_{\text{utterance}}$ = 150) for whether they thought it was produced by a child with or without ADS for comparison.

Preliminary results revealed that ASD utterances have shorter MLUw and reduced lexical diversity which aligning with previous research. A significant difference in MLUw ($p < 0.05$) was observed between children with and without ASD, suggesting that an increase in MLUw corresponds to a decreased likelihood of ASD. As shown in Table 2, all models showed high accuracy and precision with strong reliability in classifying utterances produced by ASD children, except for ChatGPT. The ratings of the speech therapist exhibited lower accuracy, likely due to different criteria adopted (e.g., stereotyped and idiosyncratic use of words or phrases across utterances, behavioral gesture data, more disfluencies, which were missing from the utterances judged). The results suggest a potential for ML applications as an **easy, early and economical** screening tool for ASD before comprehensive human rating. On the other hand, the disparity in performance between the four models and human rating is attributable to differences in evaluative criteria. ChatGPT, like human raters, may incorporate broader contextual and pragmatic knowledge about the production and behavioral characteristics of ASD when assessing the given utterances.

Overall, this study highlights the effectiveness of ML methods in identifying ASD using readily available child speech data, paving the way for future development of practical, non-invasive early detection tools for Mandarin-speaking children in clinical settings.

Table 1: Data Preprocessing Examples

Nonwords	&-en .							
Gestures	&=laughs .	or	[=! contacts:toy]					
	Happened **between** utterances	or	Happened **within** utterances					
Code-switching	[- eng] chocolate .							
Retracing	shi	**<**wang	zuo	bian**>**	**[//]**	wang	you	bian
	yes	<to	left	side>	**Retracing**	to	right	side
	'yes, to the right side' (Jack, 4;0)							
Repetition	<tan	qing	tan	qing>	**[/]**	tan	qing .	
	<play	piano	play	piano>	**Repetition**	play	piano	
	'play piano' (Rebecca, 5;0)							
Utterances less than 2 words	baba	mama						
	father	mother						
	'father, mother' (Alice, 4;0)							

Note: All Data Preprocessing Examples are in standard CLAN format.

Table 2: Performance Comparison

Models	**Description**	**Data Type**	**Basic Unit**	**Test Set (Full, 12%)**		**Test Set (Partial, 2.4%)**	
				Accuracy	**Macro F1**	**Accuracy**	**Macro F1**
BERT	Encoder-only Language Model	Textual	Utterance	0.89	0.88	0.93	0.93
XLM-RoBERTa	Encoder-only Language Model	Textual	Utterance	0.88	0.87	0.92	0.92
ChatGPT	Decoder-only Language Model	Textual	Utterance	0.49	0.46	0.48	0.47
Speech Therapist	Human Rater	Textual	Utterance	NA	NA	0.62	0.61
Logistic Regression	Statistical Machine Learning	Numerical	Transcript	0.90	0.88	NA	NA

SELECTED REFERENCES

ASDBank Mandarin Shanghai Corpus. [Online]. Available: https://doi.org/10.21415/T5HW46

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Su, Y., Naigles, L. R., & Su, L. Y. (2018). Uneven expressive language development in Mandarin-exposed preschool children with ASD: Comparing vocabulary, grammar, and the decontextualized use of language via the PCDI-Toddler Form. Journal of autism and developmental disorders, 48(10), 3432-3448.

Zhou, P., Crain, S., Gao, L., Tang, Y., & Jia, M. (2015). The use of grammatical morphemes by Mandarin-speaking children with high functioning autism. Journal of Autism and Developmental Disorders, 45, 1428-1436.