

Measuring Emotional Richness in Parental Speech: A Pilot Validation of Human vs. AI-Assisted Annotation

Weronika Araszkiewicz, University of Warsaw, weronika.araszkiewicz@psych.uw.edu.pl

Alicja Jeleń, University of Warsaw, alicja.jelen@psych.uw.edu.pl

Zofia Kordas, University of Warsaw, zofia.kordas@psych.uw.edu.pl

Magdalena Łuniewska, University of Warsaw, magdalena.luniewska@psych.uw.edu.pl

Parental emotional speech has been consistently linked to children's socioemotional and cognitive development (Ogren & Sandhofer, 2021). Research has demonstrated that the richness and variety of emotion-related language used by caregivers predicts children's emotional vocabulary, emotion understanding, and broader language outcomes. Despite the well-documented importance of this construct, there is currently no standardized, validated tool for assessing the emotional richness of parental speech in a systematic and scalable way. Existing approaches rely on labor-intensive manual annotation, which limits their applicability in both research and clinical settings. The development of reliable, efficient methods for evaluating parental emotional speech therefore remains an important methodological gap in the field.

In this pilot study, 15 Polish native speaking mothers of preschool children were asked to narrate two colourful and detailed picture scenes designed to elicit emotionally rich speech - A Day at the Lunapark and A Day at the Clinic. Participants were instructed to describe each picture as they would typically do while reading a picture book to their child.

The resulting narratives were audio-recorded and automatically transcribed. We evaluated the transcriptions to examine the emotional richness of the parental speech. Annotation of emotional words was done independently by a human researcher and by an AI chatbot. The final score was calculated by dividing the number of emotional words by the total word count of the transcription.

To validate the AI-assisted assessment, we compared the emotionality scores resulting from human vs. AI judgement. The scores were highly correlated (Pearson's $r = .72$), however, the analysis of means showed a significant difference between the two methods of evaluation. For the Lunapark picture scene, AI scores were significantly higher ($M = 2.64$) than the human assessment ($M = 1.20$), $p < .001$. The same pattern held for the Clinic scene, where AI scores ($M = 1.23$) again significantly exceeded human scores ($M = .59$), $p < .001$.

These findings suggest a promising but inconclusive trend: while the two methods show meaningful convergence, full comparability between human and AI-assisted emotionality assessment could not be established. The observed discrepancy may reflect the absence of a time limit for picture description - leading to greater variance in utterance length - or suboptimal AI prompting. Furthermore, we recognize the need to expand the concept of emotional richness to include additional measures beyond a simple proportion of emotion words. In the currently ongoing study, we address all of these issues.

References

Ogren, M., & Sandhofer, C. M. (2021). Emotion Words in Early Childhood: A Language Transcript Analysis. *Cognitive development*, 60, 101122. <https://doi.org/10.1016/j.cogdev.2021.101122>