EVALUATING PROSODIC ASPECTS OF ORAL READING PROFICIENCY IN SCHOOLCHILDREN: EFFECTS OF GENDER, GENRE AND GRADE

Simon Wehrle[1] & Christopher Sappok[2]

{[1]I*f*L-Phonetik, [2]Institute of German Language and Literature II}, University of Cologne, Germany
{simon.wehrle, csappok}@uni-koeln.de

ABSTRACT

Oral reading fluency is an essential component of school education. Nevertheless, research on the relevant prosodic correlates is rare and mostly limited to listener ratings or coarse acoustic measurements. We applied an innovative method for characterising intonation styles to a longitudinal data set of oral reading performances by German-speaking schoolchildren. We found that higher listener ratings of oral reading proficiency are correlated with more dynamic intonation styles. Further, both melodicity and listener ratings were found to increase over time for individual speakers. Additionally, we show that male children used a more melodic intonation style and received higher listener ratings than females, and that the type of text stimulus had a clear effect on intonational realisation. Our paper is the first to provide robust empirical evidence for an intonational correlate of oral reading proficiency, and simultaneously corroborates the validity of the method used for measuring intonation styles.

Keywords: oral reading, fluency, intonation style, Wiggliness, L1 acquisition

1. INTRODUCTION

Although research on oral reading fluency has received increased attention in recent years [1, 2], the role of prosody in this regard has not been sufficiently elucidated. Some early examples of research on prosody in read speech can be found in [3, 4] (on adults) and [5] (on children). Two recent meta-analytic studies [6, 7] give an overview of educational research on reading acquisition. The characteristics of and differences between read and spontaneous speech are rarely considered in detail in the studies surveyed (in contrast to e.g. [8]) and, relatedly, the identification of target models for oral reading in children might be questioned in many cases (for more details see [9]). Most importantly for our purposes, the characterisation of prosody in the studies surveyed is typically limited to listener ratings or coarse acoustic-phonetic measurements (e.g. [10]).

We investigated a longitudinal data set of oral reading by German-speaking schoolchildren, using an innovative method for capturing intonation styles [11–13] to investigate the contribution of intonation to perceived oral reading proficiency. We found a clear correlation between a more dynamic intonation style and higher perceptual ratings as well as effects of gender, text stimulus (genre) and age (grade).

2. DATA

2.1. LAUDIO corpus

The present study investigates a subset from a corpus of oral reading by schoolchildren that was recorded at a rural German school and spans four years of learning development. Recording sessions were conducted with one child at a time. Sessions lasted around 15 minutes on average and consisted of multiple readings of various texts. All children were invited to participate again in the following year(s). The final corpus contains recordings of children from German grades 3 to 7 (age range 8;3–13;9).

The final Longitudinal AUDIO (LAUDIO) [14] corpus consists of roughly 1000 recordings. A great variety of text stimuli was used, but two anchor texts were applied in every single recording session, which we will focus on here. These are coded as “DOL”, an entry in a children’s dictionary (53 words in length) and “SCHNEE”, a dialogue in the style of a fable (204 words in length).

2.2. Subsample of highly fluent readings

The subset of recordings investigated here was drawn by means of a large-scale screening procedure involving 270 DOL recordings. Each recording was rated by three university students (51 raters in total) using various scales, most importantly the so-called “NAEP-Fluency-Scale” [15]. This scale distinguishes four fluency levels through very detailed descriptions. The results obtained with this scale were sufficiently reliable (ICC inter-rater-reliability > .8) to allow the identification of a subset of particularly fluent performances.

The definition of “highly fluent” children used here is comparatively conservative, in two ways: 1) whereas for the NAEP scale, a silent familiarization with the text in question usually takes place, we only considered the more demanding *prima vista* recordings (first reading without rehearsal); 2) from

these, only recordings were selected that were judged to be at the highest NAEP level (4) by all raters.

The 26 recordings (by 16 children; 9 female, 7 male) for which this was the case make up the subsample of DOL readings investigated here (representing the uppermost quartile of all readings in terms of proficiency). We also included the corresponding SCHNEE readings recorded by the same speakers in the same session (n = 25; one of the recording was not suitable for prosodic analysis as the speaker suffered from a severe cold at the time).

2.3. Listener ratings

Thirteen students attending a seminar on reading fluency took part in a perception experiment on 156 randomised stimuli from the LAUDIO corpus [9, 14]. We focus here on the ratings of perceived *overall quality* of a given oral reading performance (judged on a 10-point scale), which were found to be sufficiently reliable across listeners (ICC > .8).

3. METHOD

3.1. Intonation style: Wiggliness and Spaciousness

While previous work has examined the LAUDIO corpus regarding the influence on listener ratings of factors such as articulation rate (weak positive correlation) and number of errors (strong negative correlation) [16], our focus here is on an analysis of intonational realisation and its relation to ratings.

Specifically, we follow [11, 12, 17] in using a two-dimensional characterisation of intonation style (Wiggliness and Spaciousness). Wiggliness measures the time-varying dynamics of pitch in the form of slope changes per second (range in this data set 0.5–4.6). Spaciousness captures pitch excursions in the form of the largest f0 rises and falls, measured in semitones (ST; range 1.5–15.6). Spaciousness is closely related to more conventional operationalisations of pitch range, while there is no direct analogue to Wiggliness in other approaches (although there is a close resemblance to the concept of macro-rhythm [18, 19]).

We divided each recording into 6 intervals, the pitch contours of which were carefully manually corrected and smoothed by a previously trained annotator before undergoing automatic processing to extract Wiggliness and Spaciousness values [13]. In total, we were able to analyse 300 intervals (from 26 readings by 16 speakers across 2 stories).

Listener ratings are available for all stories at the recording level; separate listener ratings at the interval level had been collected only for the DOL text at the time of analysis (n = 156 intervals).

3.2 Statistical analysis

We used Bayesian modelling for statistical analysis [20–23]. All code, scripts, model specifications and data frames are available at https://osf.io/axqv9/.

4. RESULTS

4.1 Effect of intonation style on listener ratings

We first examined the correlation between Wiggliness/Spaciousness and listener ratings at the recording (rather than interval) level. Across text material (DOL or SCHNEE), we found a clear correlation between Wiggliness and listener ratings. Bayesian modelling of Wiggliness by rating, with speaker as a random effect, shows this to be a robust effect (δ = 0.2, 95% CI [0.09, 0.3], P (δ > 0) = 1). There is also a strong tendency for a positive correlation between Spaciousness and listener rating, but this effect cannot be assumed to be entirely robust (δ = 0.32, 95% CI [-0.06, 0.72], P (δ > 0) = 0.92). Fig. 1 plots the interplay of listener ratings with Wiggliness and Spaciousness.

A speaker-specific analysis revealed that the positive correlation between Wiggliness (and, to a lesser extent, Spaciousness) and listener ratings held true for all speakers, but also that ratings were strongly influenced by idiosyncratic characteristics. Some speakers always received high ratings (8 or 9) and some speakers always received relatively low ratings (5 or 6), but the correlation with intonation style was still evident within those individual ranges.

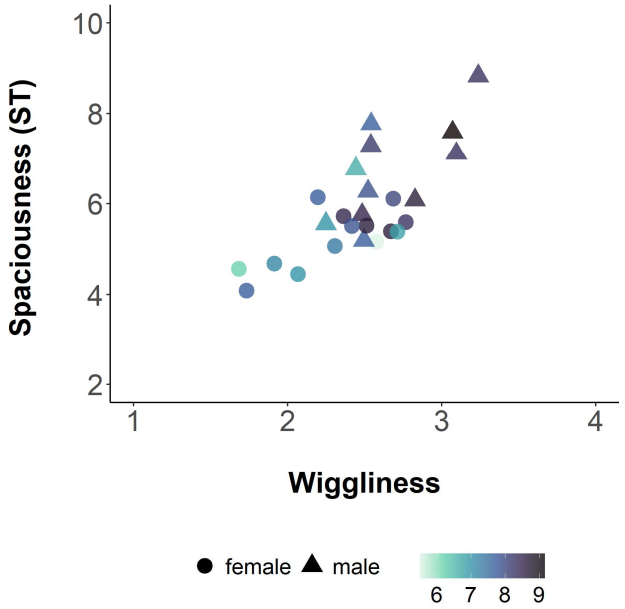

Figure 1: Plot of intonation style and corresponding listener ratings. Spaciousness (in ST) on the y-axis, Wiggliness on the x-axis. Darker shades of blue represent higher listener ratings. Circles represent female speakers, triangles male speakers.

An examination of perceptual ratings and intonation style at the level of intervals broadly confirms the findings from the recording-level analysis, but emphasises the special role of the metric of Wiggliness (capturing pitch dynamics). Only higher Wiggliness values, and not higher Spaciousness values (capturing pitch excursions) were found to be clearly correlated with higher listener ratings. Bayesian modelling confirms a robust positive effect of Wiggliness (δ = 0.15, 95% CI [0.09, 0.22], P ($\delta > 0$) = 1), but not of Spaciousness (δ = 0.06, 95% CI [-0.06, 0.19], P ($\delta > 0$) = 0.81).

4.2 Effects of gender and genre (text material)

The above analysis and the plot in Fig. 1 also point to an effect of gender. We will discuss this here together with the effect of text stimulus, or genre—the SCHNEE dialogue compared with the DOL monologue—on intonation style. Fig. 2 shows Wiggliness and Spaciousness values by gender (male/female) and text type (DOL/SCHNEE).

Turning first to effects of gender, it is clear that both Wiggliness and Spaciousness values tended to be higher for male compared to female speakers. Bayesian modelling confirms this, but also shows that the gender difference was clearer in terms of Spaciousness (δ = -1.2, 95% CI [-1.91, -0.48], P ($\delta > 0$) = 1) compared to Wiggliness (δ = -0.32, 95% CI [-0.65, 0.02], P ($\delta > 0$) = 0.94).

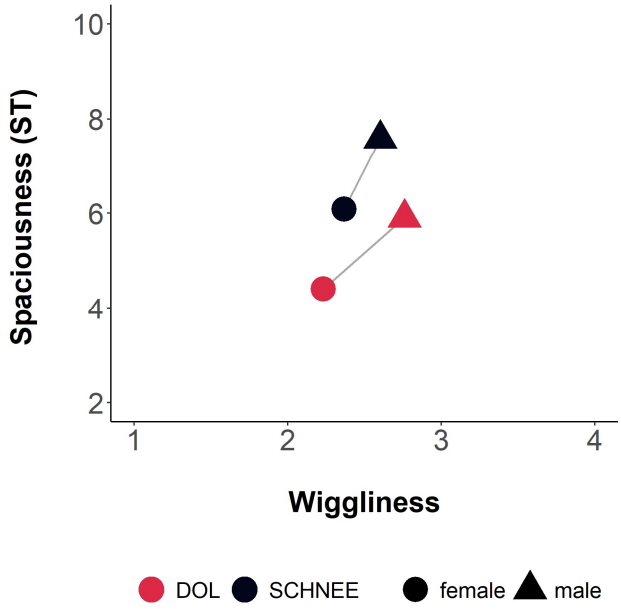

Figure 2: Intonation style by speaker gender and text material. Spaciousness (in ST) on the y-axis, Wiggliness on the x-axis. Circles represent female speakers, triangles male speakers. Red icons represent readings of the DOL monologue, black icons readings of the SCHNEE dialogue.

In contrast, when we focus on a comparison of text materials, it is clear that the intonational difference between readings of the DOL monologue and the SCHNEE dialogue is found in the dimension of Spaciousness only, with Wiggliness (pitch dynamics) not seeming to play any clear role. Bayesian modelling confirms this in unambiguously showing a robust effect for Spaciousness (δ = 1.53, 95% CI [1.1, 1.97], P ($\delta > 0$) = 1) but no effect for Wiggliness (δ = 0.06, 95% CI [-0.11, 0.24], P ($\delta > 0$) = 0.7).

A speaker-specific analysis further confirms this pattern: *all* 15 speakers who produced readings of both stories produced higher Spaciousness values in the SCHNEE dialogue compared to the DOL monologue (with no clear pattern for Wiggliness).

4.3 Effects of grade (individual development by age)

To investigate developmental changes by age, we need to consider data at the level of the individual. Of the 16 individuals in the data set, 7 produced readings at different age grades, with 5 subjects producing readings in two (consecutive) school years, 1 subject in three consecutive school years (aig-03), and 1 subject in four consecutive school years (aig-17).

For all these 7 speakers, listener ratings improved over time (or stayed equivalent, in one case). Concurrently, Wiggliness values increased, representing a more dynamic and varied intonation style. Spaciousness also tended to increase over time, but this correlation was less robust (see section 4.1). Fig. 3 (on the following page) plots listener ratings and Wiggliness values by school year/grade for all 7 speakers that produced longitudinal data. There is a clear overall trend for higher subjective listener ratings and more dynamic intonation styles over time, although some individuals deviated from this pattern. For instance, speaker aig-55 produced near-identical intonation and received near-identical ratings in grades 6 and 7, while speaker aig-17 received higher listener ratings in grade 6 than in grade 5 despite a slight reduction in pitch dynamics.

5. DISCUSSION

We used a two-dimensional characterisation of intonation style to investigate prosodic aspects of oral reading proficiency in schoolchildren. We found that both dimensions are positively correlated with listener ratings, but that a higher degree of Wiggliness (pitch dynamics), in particular, is the most robust prosodic correlate of more highly rated oral reading performances. This is a significant finding for at least two reasons. Our results provide the first suggestion of a reliable intonational correlate of oral reading proficiency (that is supported by thorough prosodic analysis). At the same time, these results corroborate

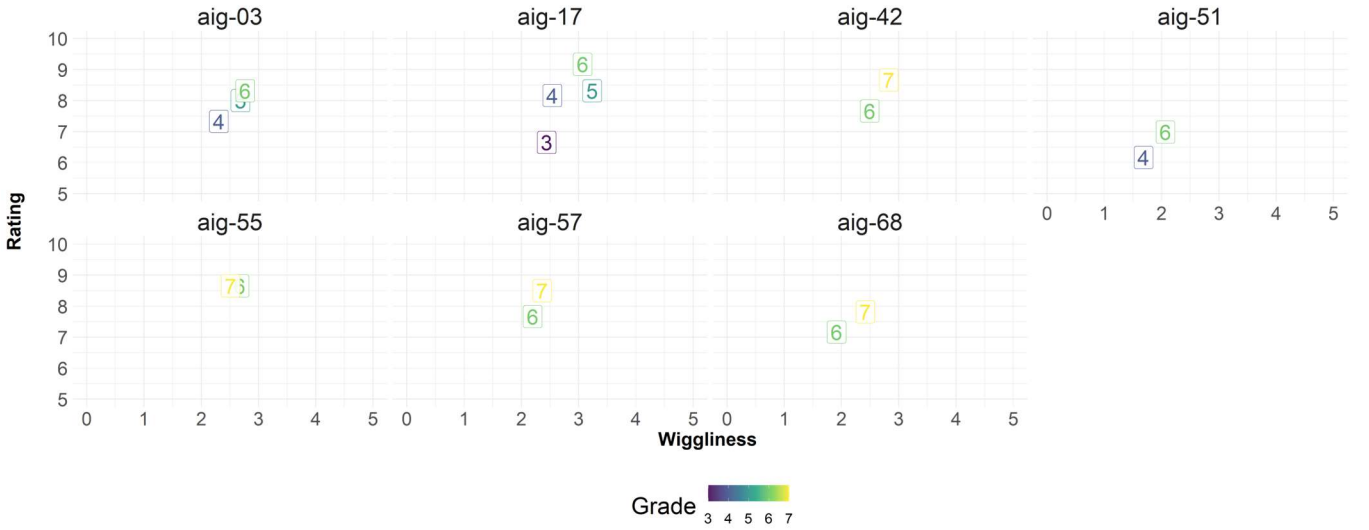

Figure 3: Wiggliness values (on the x-axis) and listener ratings (on the y-axis) by speaker and grade (/school year; higher grades represented with brighter colours). Only speakers that were recorded at more than one time point are included.

the validity of measuring intonation styles along the two dimensions of Wiggliness and Spaciousness: the positive correlations with listener ratings strongly suggest that the metrics used accurately capture acoustic features that guide perceptual judgements.

From a longitudinal perspective, we concordantly observed that both listener ratings and Wiggliness values increased with age. This signifies that schoolchildren are clearly able to develop a more dynamic intonation style, and produce oral readings that are rated to be of a higher quality, in the course of their individual development. All children for whom longitudinal data are available in the current data set showed increases (or equivalent values) in terms of both listener rating and intonation style.

While Wiggliness is thus the most relevant intonational correlate of perceived oral reading proficiency, the complementary measure of Spaciousness was shown to be decisive for effects of gender and genre (text material). Across year groups, males produced more melodic intonation styles than females and concurrently received higher listener ratings. The underlying factors are not clear. Voice change in puberty is not a sufficient explanation, as males differed from females for all year groups (starting at age 8). One speculative interpretation is that we might find males lying at *either* extreme end of the distribution in the full corpus (rather than this subset of particularly fluent speakers), i.e., that males may have produced both the most and the least melodic intonation styles (in terms of Spaciousness).

Differences as a function of genre, or text stimulus, are more easily explained. Spaciousness values were far higher for the SCHNEE compared to the DOL stimulus. The DOL text is not only monologic, it is also, by objective measures, a very difficult text for this age group (even though it was taken from a children's dictionary) [16]. In contrast, the SCHNEE text features a lively dialogue between two characters, a big burly snow man and a cute little hare. Most children clearly distinguished these two characters in terms of Spaciousness. Although this is not entirely surprising [24–26], it does have important implications, as it 1) underlines the importance of elicitation methods and text materials in studies on the production of prosody [8, 27–29] and 2) further corroborates the distinctiveness and complementarity of the Wiggliness and Spaciousness metrics.

Overall, the current work is thus an important contribution to 1) our understanding of intonational aspects of oral reading proficiency and 2) the characterisation and measurement of intonation styles more generally. The most important limitation of the current study is the restricted subset of data, featuring only the most fluent readers. We are in the process of expanding the scope from the current 26 to over 300 recordings in order to verify whether intonation style can also serve as a reliable correlate of oral reading proficiency in the context of less fluent speech.

6. ACKNOWLEDGEMENTS

We would like to give big thanks to Kim Stroeks for her extensive work on data processing.

7. REFERENCES

[1] K. Holle, 'Flüssiges und phrasiertes Lesen (fluency). Lesetheoretische Grundlagen und unterrichtspraktische Hinweise', *Schriftspracherwerb empirisch. Konzepte–Diagnostik–Entwicklung*, pp. 87–119, 2006.

[2] C. Rosebrock and D. Nix, 'Forschungsüberblick: Leseflüssigkeit (Fluency) in der amerikanischen

Leseforschung und-didaktik', *Didaktik Deutsch*, no. 20, 2006.

[3] G. Fant and A. Kruckenberg, 'Preliminaries to the study of Swedish prose reading and reading style', *STL-QPSR*, vol. 2, no. 1989, pp. 1–83, 1989.

[4] E. Blaauw, 'The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech', *Speech Communication*, vol. 14, no. 4, pp. 359–375, Sep. 1994, doi: 10.1016/0167-6393(94)90028-0.

[5] R. Cowie, E. Douglas-Cowie, and A. Wichmann, 'Prosodic characteristics of skilled reading: Fluency and expressiveness in 8-10-year-old readers', *Language and Speech*, vol. 45, pp. 47–82, 2002, doi: 10.1177/00238309020450010301.

[6] A. P. Wolters, Y.-S. G. Kim, and J. W. Szura, 'Is Reading Prosody Related to Reading Comprehension? A Meta-analysis', *Scientific Studies of Reading*, vol. 26, no. 1, pp. 1–20, Jan. 2022, doi: 10.1080/10888438.2020.1850733.

[7] E. Godde, M.-L. Bosse, and G. Bailly, 'A review of reading prosody acquisition and development', *Read Writ*, vol. 33, no. 2, pp. 399–426, Feb. 2020, doi: 10.1007/s11145-019-09968-1.

[8] L. E. De Ruiter, 'Information status marking in spontaneous vs. read speech in story-telling tasks–Evidence from intonation analysis using GToBI', *Journal of Phonetics*, vol. 48, pp. 29–44, 2015.

[9] C. Sappok, 'Oral reading proficiency and prosody – a perceptual pilot study on especially fluent German students (grade 3 to 7)', in *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Repulic, 2023.

[10] M. R. Kuhn, P. J. Schwanenflugel, and E. B. Meisinger, 'Aligning Theory and Assessment of Reading Fluency: Automaticity, Prosody, and Definitions of Fluency', *Reading Research Quarterly*, vol. 45, no. 2, pp. 230–251, 2010, doi: 10.1598/RRQ.45.2.4.

[11] S. Wehrle, F. Cangemi, H. Hanekamp, K. Vogeley, and M. Grice, 'Assessing the Intonation Style of Speakers with Autism Spectrum Disorder', in *Proc. 10th International Conference on Speech Prosody 2020*, 2020, pp. 809–813.

[12] S. Wehrle, F. Cangemi, M. Krüger, and M. Grice, 'Somewhere over the spectrum: Between robotic and singsongy intonation', *Il parlato nel contesto naturale. Speech in the natural context*, no. 4, pp. 179–194, Dec. 2018, doi: 10.17469/O2104AISV000010.

[13] S. Wehrle, 'A brief tutorial for using Wiggliness and Spaciousness to measure intonation styles', Jun. 2022, Accessed: Dec. 14, 2022. [Online]. Available: https://osf.io/5e7fd/

[14] C. Sappok and J. Fay, 'Prosodische Aspekte von Leseflüssigkeit messen. Evaluation einer Ratingprozedur mit Audioaufnahmen von DrittklässlerInnen', *Didaktik Deutsch: Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur*, vol. 23, no. 44, pp. 61–83, 2018.

[15] G. S. Pinnell, J. J. Pikulski, K. K. Wixson, J. R. Campbell, P. B. Gough, and A. S. Beatty, *Listening to children read aloud: Data from NAEP's integrated reading performance record (IRPR) at grade 4*. Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education, 1995.

[16] C. Sappok, 'Exploring Advanced Prosody–eine Best-Practice-Untersuchung zum lauten Lesen in der weiterführenden Schule', *Weiterführende Grundlagenforschung in Lesedidaktik und Leseförderung*, pp. 68–97.

[17] S. Wehrle, 'A Multi-Dimensional Analysis of Conversation and Intonation in Autism Spectrum Disorder', PhD Thesis, University of Cologne, Cologne, Germany, 2021.

[18] S.-A. Jun, 'Prosodic typology revisited: Adding macro-rhythm', in *Speech Prosody 2012*, Shanghai, China, 2012.

[19] S.-A. Jun, 'Prosodic typology: By prominence type, word prosody, and macro-rhythm', in *Prosodic Typology II*, Oxford University Press, 2014, pp. 520–539.

[20] P.-C. Bürkner, 'brms: An R package for Bayesian multilevel models using Stan', *Journal of statistical software*, vol. 80, no. 1, pp. 1–28, 2017.

[21] R Core Team, 'R: A language and environment for statistical computing'. R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[22] RStudio Team, 'RStudio: Integrated Development Environment for R'. RStudio, PBC, Boston, MA, 2021. [Online]. Available: http://www.rstudio.com/

[23] Stan Development Team, 'Stan Modeling Language Users Guide and Reference Manual, Version 2.29.' 2022. [Online]. Available: https://mc-stan.org

[24] J. J. Ohala, 'Cross-language use of pitch: an ethological view', *Phonetica*, vol. 40, no. 1, pp. 1–18, 1983.

[25] L. Hinton, J. Nichols, and J. J. Ohala, *Sound symbolism*. Cambridge University Press, 2006.

[26] C. Gussenhoven, 'Foundations of intonational meaning: Anatomical and physiological factors', *Topics in Cognitive Science*, vol. 8, no. 2, pp. 425–434, 2016.

[27] A. Janz, 'Navigating Common Ground Using Feedback in Conversation- A Phonetic Analysis', MA thesis, University of Cologne, Cologne, Germany, 2022.

[28] M. Grice, M. Savino, and M. Refice, 'The intonation of questions in Bari Italian: Do speakers replicate their spontaneous speech when reading', *Phonus*, vol. 3, pp. 1–7, 1997.

[29] C. Dideriksen, R. Fusaroli, K. Tylén, M. Dingemanse, and M. H. Christiansen, 'Contextualizing conversational strategies: backchannel, repair and linguistic alignment in spontaneous and task-oriented conversations', in *CogSci'19*, Cognitive Science Society, 2019, pp. 261–267.