



# From Features to Fluency: Predicting Perceived Speech Fluency of Preschool Children for Language Proficiency Assessments

Valentin Kany

<sup>1</sup>Language Science and Technology, Saarland University, Saarbrücken, Germany

valentin.kany@uni-saarland.de

## Abstract

This study investigates the effect of various fluency-related features on the perception of speech fluency in the speech of German-speaking preschool children. Speech data from 10 children (5 L1, 5 L2) were elicited by means of a serious-game-based method developed for Language Proficiency Assessments (LPAs). Listeners evaluated the perceived fluency in the stimuli on a 9-point Likert scale. The fluency assessment revealed a significant negative effect of the number of disfluent pauses and the number of other disfluencies (such as repairs, truncations, repetitions, and lengthenings) on the fluency rating. A significant positive effect was found for articulation rate. Overall, these results can serve as a basis for a unified evaluation of speech fluency of preschool children to extend LPAs.

**Index Terms:** speech fluency, fluency assessment, child speech, language proficiency assessment

## 1. Introduction

Language Proficiency Assessments (LPAs) of preschool children are ubiquitous in Germany and become obligatory in more and more federal states [1]. Most applied LPA methods test for the children’s vocabulary size, grammar skills [2, 3], and morphology [4] as indicators to check if the overall language competence is sufficient for the child to be able to cope with the tasks they will face in school. Despite the fact that several studies found speech fluency to be a prominent indicator of language proficiency ([5, 6, 7]), to the best of our knowledge, it seems not to be considered part of the standard procedure in LPAs in preschool practice. In previous studies, we attempted to develop semi-automatic methods to assess speech fluency in spontaneous preschool-child speech [8, 9]. While the individual speech fluency profiles from [9] offer an overview of the child’s speech with regard to a range of features assumed to be related to speech fluency (hereafter referred to as “fluency-related features”), they cannot be used to assess overall speech fluency, which would be the goal in an LPA. This is due to the fact that the profiles present the plain results extracted from manual and semi-automatic annotations and thus lack information on their actual influence on speech fluency.

Therefore, the present study conducts a perceived fluency assessment to investigate the influence the features have on perceived fluency, with the aim to make such profiles more meaningful and usable for an overall fluency assessment in LPA.

Previous studies found articulation rate (e.g. [10, 11, 12]), inter-pause intervals (e.g. [11, 12]), pause frequency [10], and pause duration [12] to be the best predictors of perceived fluency. However, these studies investigated adult speech and their fluency in a second language (L2). There are a few studies addressing fluency in child speech (e.g. [13, 14]), but their par-

ticipants are usually older and already attending school. At that age, the children’s speech data can already be collected through reading tasks [13]. For preschool children, in the past, the focus of studies was mainly on children with the speech disorder “stuttering” [15]. Studies addressing fluency in spontaneous child speech, especially at preschool age and for the German language, are hard to find.

## 2. Methodology

### 2.1. Data

The data and their annotations of fluency-related features used in this study were taken from [9]. The data were collected by means of a serious game that was developed as a tool for LPAs of preschool children [16]. The game tells a coherent story with 28 different scenes in which the child has to talk to a virtual character to help the character progress through the story. In order to do so, the child needs to answer two questions per scene that are posed by the character, which is secretly controlled by a human experimenter. The answers given by the children were recorded in German daycare centres in a separate room, with only the experimenter and one of the daycare centre’s staff, who functions as a confidant to the child, present in the room. For the recordings, the built-in microphone of the iPad (9th gen) on which the game is played, was used.

While there are other, larger child speech corpora available (e.g. [17, 18, 19]), we still decided to keep to this small-sized data since most of the corpora are either not in German language, cover a different or wider age range, or do not feature spontaneous speech. Apart from that, this work is part of a large-scale project that ultimately aims to integrate the fluency assessment into an (semi-)automated LPA with this game.

The corpus features data from about 150 children. However, only 10 children have been annotated up to this point and could be used in this study. The children were between 4;6 and 5;6 years old at the time of recording. Half of them speak German as their first language, the others speak German as their second language. Given the game’s structure, we received 56 recorded segments per child (28 scenes with 2 questions each). The uniqueness of the data (spontaneous speech of rather young children) and the acquisition method comes with a caveat that might be an important factor when it comes to the assessment of speech fluency: While other studies, especially those relying on read speech, can work with long coherent articulation phases, our data mostly contain rather short sequences of speech (average duration of 3.23 seconds).

Feedback from a pilot study revealed that not all of the 560 stimuli proved to be suitable for fluency assessment. Particularly short stimuli, containing only one or two words, were found to be unratable and always received the maximum fluency

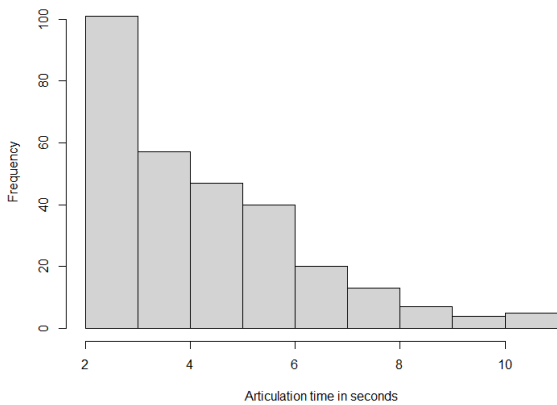


Figure 1: *Distribution of articulation time across all stimuli used in the fluency assessment. Bins are in 1-second-steps.*

rating by the participants. Thus, we decided to automatically filter out all stimuli that include less than 2 seconds of articulation time in a first step and manually excluded all remaining single-word-answers from the stimulus pool. In addition, the raters found the dialog aspect (caused by interactions with other persons in the room or the virtual character in the game) of the data confusing, as it caused the confounding of dialog fluency and utterance fluency. To avoid this, we decided to use only stimuli that represent single turns and ignore the dialog aspect for the moment. 500 ms silence was added at the start and end of each stimulus.

After the modification of the stimulus pool from the pilot study, we ended up with a total of 320 stimuli from 10 children to be featured in the main study. The articulation time of the stimuli is still noticeably skewed to the lower end (see Figure 1). Most of the stimuli include 2 to 4 seconds of articulation time which is due to the game-based data acquisition method.

## 2.2. Fluency Assessment

32 raters participated in the fluency assessment. They are German L1 speakers and between 18 and 62 years old. All of them have a background in linguistics, as they either studied or are still studying something linguistics-related. This way, they had similar prerequisites and at least a basic understanding of language, which ensured a higher consistency in the ratings. Nevertheless, they do not reflect the entirety of the target group of LPAs, since in practice these are conducted by preschool or elementary school teachers, doctors, psychologists, or speech therapists in Germany, depending on the federal state [1].

The fluency assessment took place online via LabVanced [20]. The raters were instructed to rate the fluency of the child in the presented audio on their overall perception of fluency. They were not given a definition of fluency, nor were they asked to pay attention to any of the aspects investigated in this study. For the assessment, we used a 9-point Likert scale with the negative pole labeled “not fluently at all” (1) and the positive pole “absolutely fluently” (9). We used this scale, as it is the most commonly used rating method among other fluency assessment studies [21]. Raters were allowed to listen to each presented stimulus twice and had the option to give some further feedback if they wanted to. The assessment started with a warm-up



Figure 2: *Screenshot of the experimental setup. Translations in parentheses were not visible during the actual task.*

phase of 5 stimuli excluded from the analysis to let them familiarise themselves with the task. For the main part of the assessment, the raters were then presented with 20 different stimuli in random order first. Subsequently, they got to rate the same 20 stimuli again in the exact same order for a second time to increase the robustness of the rating and check for consistency within a listener’s perception.

In the end, we received 2 separate ratings of 20 different stimuli per rater. There were 2 raters per set of 20 stimuli, which results in each stimulus being rated 4 times. The experimental setup of the main part of the assessment can be viewed in Figure 2.

## 3. Results

We received a total of 1280 fluency ratings for 320 different stimuli from 32 raters. Figure 3 provides an overview of the frequency of the different ratings across all stimuli. The given ratings are not distributed equally over the whole scale, nor are they normally distributed. They show a clear tendency towards higher ratings. Ratings of 8 and 9 were given the most often (210 and 215 times), while the lowest rating of 1 was only given 25 times. The neutral rating of 5 was also given relatively rarely compared to its surrounding rating options.

In a first step, we wanted to check if children with German as L1 are perceived to be more fluent than children with German as L2. Thus, we calculated a Cumulative Link Mixed Model for ordinal data (CLMM) [22] in R. We used L1 status (L1 vs. L2) as fixed effect, and random intercepts for stimulus, rater, and speaker. The results revealed a trend towards lower ratings for L2 speakers compared to L1 speakers, although the effect was not statistically significant ( $OR = 0.17$ , 95%-CI [0.02, 1.16],  $p = .070$ ). This tendency needs to be taken with a grain of salt since only 5 children of each category were analysed here.

A detailed analysis of the fluency-related features will be calculated and presented in the following section by a different model, independent from the speaker’s L1.

### 3.1. Effects of features on fluency rating

To investigate the effect the features have on the perceived fluency ratings, a CLMM [22] was calculated in R. Based upon the

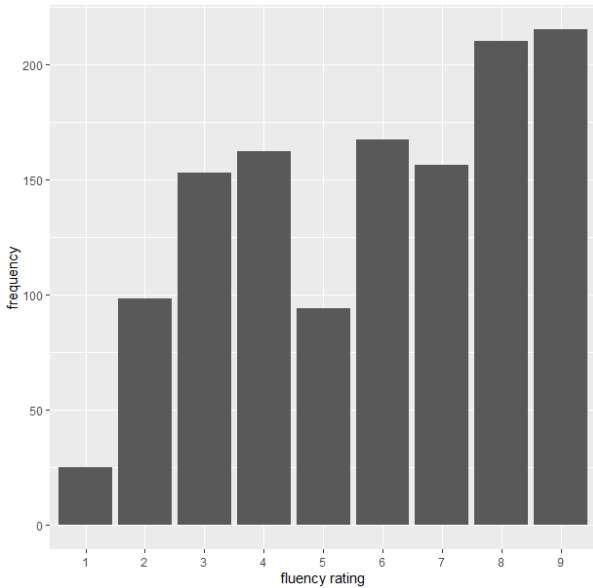


Figure 3: Frequency of the different fluency ratings given by the raters in the fluency assessment.

features from [9]’s fluency profiles, articulation time, articulation rate, longest articulation phase (duration of speech without any kind of disfluency), number of disfluent pauses (pauses that were perceived to interrupt the flow of speech in the annotation process), duration of disfluent pauses, number of filler particles (such as “uh” and “uhm”), and number of other disfluencies were included in the model. “Other disfluencies” comprise repairs (speech errors which are corrected shortly afterwards), truncations (abandonments of syllables, words or clauses at some point during the utterance), lengthenings (prolongations of speech sounds), and repetitions (reiterations of words) [9]. Articulation rate was measured automatically [5], whereas all other features were derived through manual annotation [9]. All predictors were z-standardised with the help of the scale function beforehand to avoid problems caused by the different scales of the measures and ensure comparability [23]. Rater, child, and stimulus were included as random effects to control for individual variances. The resulting model is shown in table 1.

It shows significant negative effects on the fluency rating for the number of disfluent pauses ( $OR = 0.45$ , 95%-CI [0.30, 0.69],  $p < .001$ ) and the number of other disfluencies ( $OR = 0.46$ , 95%-CI [0.34, 0.62],  $p < .001$ ). A higher articulation rate in the stimulus has a significant positive effect ( $OR = 1.32$ , 95%-CI [1.04, 1.69],  $p = .025$ ). All other predictors showed no significant effect. Nevertheless, there are slight tendencies for negative effects of the duration of disfluent pauses ( $OR = 0.75$ , 95%-CI [0.50, 1.14],  $p = .180$ ) and the number of filler particles ( $OR = 0.82$ , 95%-CI [0.61, 1.10],  $p = .179$ ).

Figure 4 shows the mean z-value of all stimuli rated with a certain fluency rating in the assessment. The graphs for the different features mirror the tendencies found in the CLMM. The mean number of filler particles in a stimulus as well as the duration of disfluent pauses decline with increasing fluency rating. In case of the number of filler particles, there is a particularly high decline from the rating of 1 to the rating of 2. On the other hand, the average value of the longest articulation phase only increases slightly with an increasing rating from 1 to 3 and re-

Predictors	Odds Ratios	CI	p
No. disfluent pauses	0.45	0.30 - 0.69	<0.001
Disfluent pause duration	0.75	0.50 - 1.14	0.180
No. other disfluencies	0.46	0.34 - 0.62	<0.001
Articulation time	1.09	0.70 - 1.70	0.711
No. filler particles	0.82	0.61 - 1.10	0.179
Articulation rate	1.32	1.04 - 1.69	0.025
Longest articulation phase	1.07	0.78 - 1.48	0.675

Table 1: The results of the Cumulative Link Mixed Model with fluency rating as dependent variable.

mains the same from then on.

### 3.2. Interrater Reliability

To ensure the reliability of the fluency ratings, the interrater reliability in the assessment was checked. Since each stimulus was rated by two separate raters on an ordinal Likert scale, weighted Cohen’s Kappa with squared weights was used to calculate the interrater reliability across the whole study. First, Cohen’s Kappa was calculated for only the first ratings of each rater, then for only the second ratings of each rater (both  $N = 320$  stimuli, 2 raters each). The resulting Kappa values are  $\kappa = 0.450$  for the first ratings and  $\kappa = 0.512$  for the second ratings, which indicates a moderate agreement among the raters [24].

Further, the intra-rater reliability within each rater between their first and second rating of the same stimuli was calculated. This was done to check the robustness of the raters’ perception of fluency. Again, weighted Cohen’s Kappa with squared weights was used to calculate the intra-rater reliability for each rater individually first (32 times  $N = 20$  stimuli, 2 ratings each). The mean Kappa value across all 32 raters is  $\kappa = 0.669$ , which indicates substantial agreement between the first and second rating of a stimulus [24].

## 4. Discussion

While there might be some differences and preferences in the perception of fluency between listeners, it is still possible to get a decent agreement, even on a rather fine-grained 9-point Likert scale. Also, raters seem to be consistent in their perception of fluency, as the substantial agreement between their first and second rating of a stimulus in the assessment suggests. These findings show that a unified overall assessment of a child’s speech fluency is not unrealistic and might be a valid addition to the grammar and vocabulary aspects that are already considered by the game-based LPA approach of [16].

The distribution of the ratings given by the raters shows a noticeable tendency towards higher ratings. The two highest ratings, 8 and 9, were given particularly frequently. Even though the shortest stimuli had already been excluded after the feedback from the pilot study, this might still have to do with the high amount of short stimuli in the dataset. Since they do not offer many opportunities for the speakers to become disfluent, they might get awarded with the highest fluency rating possible. The very low frequency of the lowest two ratings, 1 and 2, could also be explained by the “pardon young effect” mentioned by [13], where raters are tempted to rate less strictly when they perceive a young voice. The dent in frequency for the rating of 5 is probably due to it being the default value and the raters tend to rather make a decision in either direction than to just leave the default value.

Regarding the effects of the features, the strongest effect on

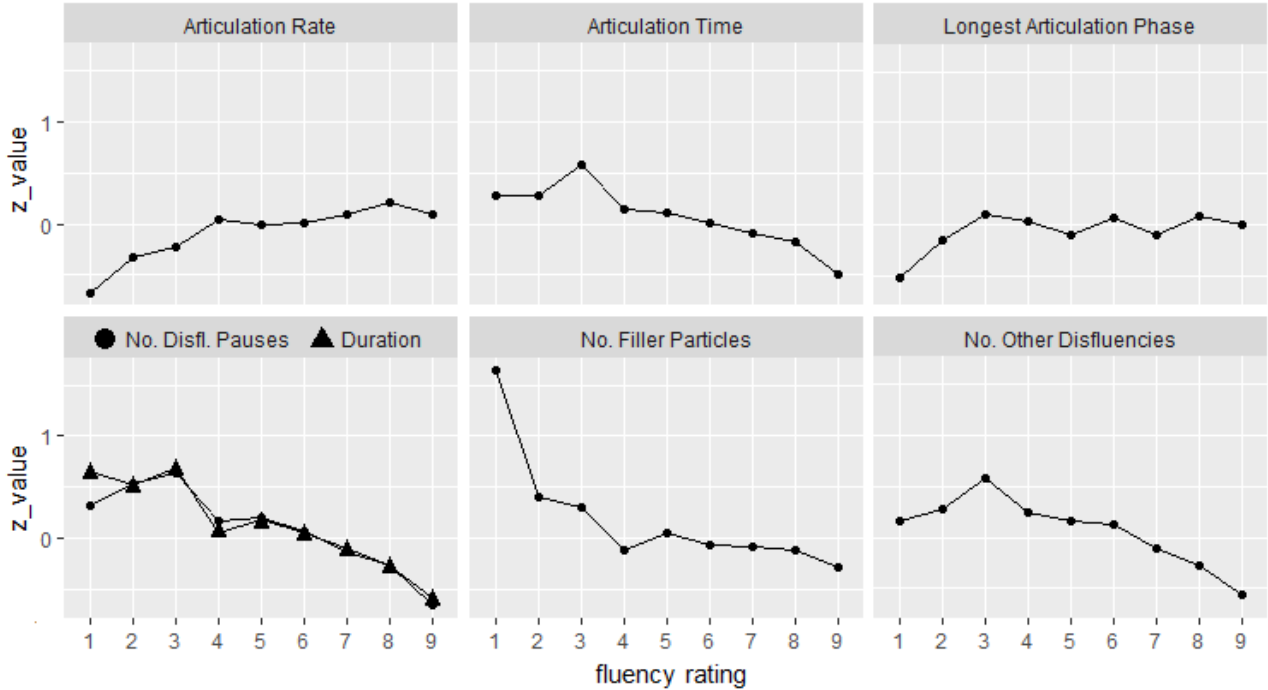


Figure 4: Overview of all features with their normalised mean z-value (y-axis) for stimuli awarded with a certain fluency rating (x-axis).

the rating was found for the number of disfluent pauses in the stimulus, whereas the duration of these pauses had no significant effect on the rating. This suggests that the sheer amount of pauses leads to the perception of disfluency more than their duration. In general, the strength of the effect might be explained by the fact that the “disfluent pauses” were already found to interrupt the flow of speech by the annotators during the annotation process of the data. The number of other disfluencies proved to have a similarly strong negative effect on the ratings as the number of disfluent pauses. Reparations and truncations are of syntactic nature and range across multiple linguistic units. This might make them more perceptible than other kinds of disfluencies. Since they have such a strong effect, breaking the rather broad category down to its individual types of disfluencies and analysing their effect on perceived fluency separately is of great interest. Then, you could check if the two “syntactic” disfluency types have a stronger influence on fluency perception than repetitions and lengthenings. Especially lengthenings might differ from the other types because they are usually limited to one linguistic unit only. However, such an analysis would require more data, as the frequencies of the individual disfluency types were too low to conduct an individual analysis. Articulation rate turned out to be the only significant positive effect on perceived fluency. This finding confirms the view of many other studies (e.g. [10, 11, 12]) that articulation rate is one major predictor of fluency in a language. Apparently, this seems to hold true for spontaneous speech of preschool children as well. Even though filler particles did not turn out to have a significant effect on perceived fluency, it is striking that the stimuli rated with the lowest rating of 1 contained a lot of them (see Figure 4). This observation might indicate that a moderate use of filler particles is tolerated by listeners, but if filler particles are produced excessively, it will lead to a complete loss of fluency in the listeners’ perception.

## 5. Conclusion

In an earlier study, we developed individual speech fluency profiles for preschool children to assess their speech fluency as part of an LPA [9]. However, these profiles only include the raw values of various fluency-related features. Thus, the aim of the present study was to test the influence these features have on perceived fluency, which is necessary to make an overall, unified assessment for the children.

The study revealed that the features’ effect on perceived fluency varies from feature to feature. The number of disfluent pauses and the number of other disfluencies turned out to have the strongest significant effects on fluency. Regarding the pauses, a further analysis of their location and its influence on perceived fluency would be interesting, since Kahng [25] suggests it to be a significant factor in adult speech. In terms of the feature “number of other disfluencies”, a larger study including more speech data would be interesting to differentiate between syntax-related disfluencies and disfluencies on the single-unit level. The same holds true for the number of filler particles, which showed no significant effect in this study but still suggested an influence of an excessive use on perceived fluency. Perhaps some particles have a stronger effect than others. Articulation rate turned out to be the only significant positive effect.

With these findings, this study contributes to the further development of individual speech fluency profiles for children, as they can be used to add weights to the plain measurements. This way, they can eventually be used to derive an overall fluency score that could be included in an LPA to contribute to a more precise evaluation of preschool children’s language proficiency. In general, a more differentiated view on fluency-related aspects of child speech can be helpful in diagnostic or didactic contexts. Future research could also benefit from involving LPA stakeholders so their practical expertise is reflected in the fluency score.

## 6. Acknowledgements

I am grateful to Julia Schu for the annotation work and to Diana Davidson for preparing the data. I would also like to thank Bernd Möbius and Jürgen Trouvain for their valuable feedback during the writing process, as well as all listeners for their participation in the rating task.

## 7. References

- [1] A. Lisker, “Sprachstandsfeststellung und Sprachförderung im Kindergarten sowie beim Übergang in die Schule,” *Expertise im Auftrag des Deutschen Jugendinstituts*, 2010. [Online]. Available: [http://www.dji.de/bibs/Expertise\\_Sprachstandserhebung-Lisker\\_2010.pdf](http://www.dji.de/bibs/Expertise_Sprachstandserhebung-Lisker_2010.pdf)
- [2] P. Schulz and R. Tracy, *Linguistische Sprachstandserhebung - Deutsch als Zweitsprache (LiSe-DaZ): Language Test for Children with German as a Second Language*, 2011.
- [3] N. Gagarina, D. Klop, S. Kunnari, K. Tantele, T. Välimaa, U. Bohnacker, and J. Walters, “Main: Multilingual assessment instrument for narratives – revised,” *ZAS Papers in Linguistics*, vol. 63, p. 20, 2019.
- [4] T. Mayr and M. Ulich, “Sismik-sprachverhalten und interesse an sprache bei migrantenkindern in kindertageseinrichtungen,” *Ein Instrument zur systematischen Beobachtung der Sprachentwicklung. Freiburg*, 2003.
- [5] N. H. de Jong, J. Pacilly, and W. Heeren, “Praat scripts to measure speed fluency and breakdown fluency in speech automatically,” *Assessment in Education: Principles, Policy & Practice*, vol. 28, no. 4, pp. 456–476, 2021. [Online]. Available: <https://doi.org/10.1080/0969594X.2021.1951162>
- [6] A. Ginther, S. Dimova, and R. Yang, “Conceptual and empirical relationships between temporal measures of fluency and oral english proficiency with implications for automated scoring,” *Language Testing*, vol. 27, no. 3, pp. 379–399, 2010. [Online]. Available: <https://doi.org/10.1177/0265532210364407>
- [7] N. Iwashita, A. Brown, T. McNamara, and S. O’Hagan, “Assessed levels of second language speaking proficiency: How distinct?” *Applied Linguistics*, vol. 29, no. 1, pp. 24–49, 03 2008. [Online]. Available: <https://doi.org/10.1093/applin/amm017>
- [8] V. Kany and J. Trouvain, “Computergestützte Bestimmung des Sprechflusses bei Vorschulkindern,” in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2024*, T. Baumann, Ed. TUDpress, Dresden, 2024, pp. 62–69. [Online]. Available: [https://www.essv.de/pdf/2024\\_62\\_69.pdf](https://www.essv.de/pdf/2024_62_69.pdf)
- [9] —, “Annotation of disfluencies in child speech,” in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2025*, S. Grawunder, Ed. TUDpress, Dresden, 2025, pp. 247–254. [Online]. Available: [https://www.essv.de/pdf/2025\\_247\\_254.pdf](https://www.essv.de/pdf/2025_247_254.pdf)
- [10] C. Cucchiari, H. Strik, and L. Boves, “Quantitative assessment of second language learners’ fluency by means of automatic speech recognition technology,” *The Journal of the Acoustical Society of America*, vol. 107, pp. 989–99, 2000.
- [11] Y. Préfontaine, J. Kormos, and D. E. Johnson, “How do utterance measures predict raters’ perceptions of fluency in french as a second language?” *Language Testing*, vol. 33, no. 1, pp. 53–73, 2016. [Online]. Available: <https://doi.org/10.1177/0265532215579530>
- [12] P. Tavakoli, “Fluency in monologic and dialogic task performance: Challenges in defining and measuring l2 fluency,” *International Review of Applied Linguistics in Language Teaching*, vol. 54, 2016.
- [13] 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 Republic*, 2023.
- [14] 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, no. 1, pp. 47–82, 2002, pMID: 12375819. [Online]. Available: <https://doi.org/10.1177/00238309020450010301>
- [15] V. Tumanova, E. Conture, E. Lambert, and T. Walden, “Speech disfluencies of preschool-age children who do and do not stutter,” *Journal of Communication Disorders*, vol. 49, 2014.
- [16] J. Roche, S. Haberzettl, G. Pagonis, M. Jessen, and N. Weidinger, *Serious Games in der Sprachstandsermittlung*. Narr Francke Attempto Verlag, 2019, pp. 340–358.
- [17] C. D. Hernandez Mena, D. E. Mollberg, M. Borský, and J. Gunason, “Samrómur children: An Icelandic speech corpus,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odiijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 995–1002. [Online]. Available: <https://aclanthology.org/2022.lrec-1.105/>
- [18] K. Radha, M. Bansal, and R. B. Pachori, “Automatic speaker and age identification of children from raw speech using sinenet over erb scale,” *Speech Communication*, vol. 159, p. 103069, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167639324000414>
- [19] P. Escudero, G. Escobar, M. Hernandez Gallego, C. Diskin-Holdaway, and J. Hajek, “A database of multilingual child speech with recordings from a longitudinal project for multilingual education,” 12 2024.
- [20] C. Goeke, H. Finger, D. Diekamp, K. Standvoss, and P. König, “Labvanced: A unified javascript framework for online studies,” 2017.
- [21] S. Suzuki, J. Kormos, and T. Uchihara, “The relationship between utterance and perceived fluency: A meta-analysis of correlational studies,” *Modern Language Journal*, vol. 105, 2021.
- [22] R. H. B. Christensen, *ordinal—Regression Models for Ordinal Data*, 2023, r package version 2023.12-4.1. [Online]. Available: <https://CRAN.R-project.org/package=ordinal>
- [23] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2025. [Online]. Available: <https://www.R-project.org/>
- [24] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977. [Online]. Available: <http://www.jstor.org/stable/2529310>
- [25] J. Kahng, “The effect of pause location on perceived fluency,” *Applied Psycholinguistics*, vol. 39, no. 3, p. 569–591, 2018.