Automatic Prediction of Child Speech Fluency with Game-Based Data from German Preschoolers

Valentin Kany[1,2], **Bernd Möbius**[1], **Jürgen Trouvain**[1]
[1]Language Science and Technology, [2]German Studies
Saarland University, Saarbrücken, Germany
valentin.kany@uni-saarland.de, {moebius, trouvain}@lst.uni-saarland.de

Abstract

This paper introduces an approach to automatically predict the speech fluency of preschool children as part of Language Proficiency Assessments. We use spontaneous speech data from children with German as native and second language aged 4–6 years, collected via a game–based elicitation method. The recordings were mainly annotated manually on various fluency-related phenomena. The resulting feature values were compared to human fluency ratings of the same data. The human ratings and the fluency-related acoustic features were used to build Cumulative Link Mixed Models (CLMMs) with and without splines to test their ability to predict the human ratings with multiple metrics (Spearman's ρ, MAE, quadratic weighted κ). Results show that a parsimonious linear model already reaches near-human agreement (quadratic weighted kappa $\kappa = 0.65$) and that incorporating non-linear spline effects does not improve predictive accuracy. These findings suggest that relatively simple CLMMs can substitute additional human raters in fine-grained fluency assessment of preschool children, which is a task that is already challenging for trained listeners.

Keywords: Speech Fluency, Language Proficiency Assessment, Automatic Speech Assessment, Child Speech, Disfluencies, Acquisition, Corpus (Creation, Annotation, etc.), Prosody, Speech Resource/Database, Tools, Systems, Applications

1. Background

Language Proficiency Assessments (LPAs) of preschool children are ubiquitous in many countries. For instance, in Germany, LPAs are mandatory in almost all federal states. However, the overall procedure and the practiced LPA method vary from state to state and make comparisons difficult (Faas et al., 2021). In general, between one and two years before entering primary school, either all children or a certain group of children take part in an LPA. They serve as a check to see whether they need language learning support to reach the required level to enter school. However, the classical methods applied in Germany come with some caveats. They induce an unfamiliar test situation for preschool children and their inflexible tasks hinder the child's production of spontaneous speech. This negatively affects the test's validity. In addition, the assessment is performed by just one human expert (with possibly different levels of quality), which makes the process 1) complex, 2) time consuming, and 3) the result rather subjective. These aspects lead to a loss of objectivity, reliability, and validity of the classical methods. Therefore, Roche et al. (2019) aimed to develop a standardised game-based method for LPAs. In this way, LPAs can be conducted without inducing a test setting that generates pressure for the child. The game introduces an immersive environment for the child and encourages the production of spontaneous speech. The possibility of automatic evaluation makes the method less time-consuming than other manual methods.

Conventional LPAs of preschool children in Germany usually test for vocabulary size, grammar skills (Schulz and Tracy, 2011; Gagarina et al., 2019), and morphology (Mayr and Ulich, 2003) as indicators for the child's overall language competence. However, speech fluency was also found to be a major predictor for language competence by several studies (e.g. Baker Smemoe et al. 2014; de Jong et al. 2021; Ginther et al. 2010; Iwashita et al. 2008), and has not yet received much attention in practice. Thus, we would like to use the proposed LPA method in Roche et al. (2019) to investigate speech fluency in spontaneous child speech and find a way to integrate fluency assessment into LPA.

Providing a uniform definition of "fluency" is almost impossible, as this term is used in a number of different sub-disciplines viewed from different angles with different research questions in mind. In speech pathology in the context of aphasia diagnosis tests for example, the term is used in non-conversational contexts. "Fluency" does not refer to the flow of words in connected speech, but to smoothness in a specific task in language production here (Lickley, 2015). In the field of second language acquisition, the term "fluency" is used to describe the ability to speak in a language in a native-like way. This entails producing speech "at the tempo of native speakers, unimpeded by silent pauses and hesitations" (Lennon, 1990). This abil-

ity is affected by a range of factors, such as accessibility of vocabulary, formulation of syntactically correct phrases, or preparation of motor commands for articulatory sequences (Lickley, 2015). Given this broad range of factors affecting fluency, this might be the reason why fluency was found to be a robust indicator of language proficiency in general (e.g. Baker Smemoe et al., 2014; de Jong et al., 2021; Ginther et al., 2010; Iwashita et al., 2008). Lickley (2015) further discriminates "fluency" in the context of second language acquisition from fluency in "typical speech", where the term is used to describe the smoothness of flow at several different levels described by several language production models (e.g. Levelt, 1993; Garrett, 1980). A basic distinction between the two levels of planning fluency (referring to smoothness of the internal processes) and surface fluency (referring to smoothness of overt speech) can be made (Lickley, 2015).

In our case, we investigate speech of young children who are still in the process of learning German, either as L1 (native language) or L2 (second language). Thus, we assess a combination between fluency in the field of (second) language acquisition and fluency in "typical speech". With our LPA method, we can only observe speech at surface level and are thus interested in "surface fluency" and its link to "perceived fluency" of the children, i.e. if human listeners perceive the speech as fluent.

2. Aims

As mentioned above, aside from the typically assessed aspects in LPAs like vocabulary size, grammar skills, and morphology, we would like to use data collected by means of Roche et al. (2019)'s game-based method to develop a method to (automatically) assess the speech fluency of the children. Several studies suggest that fluency plays a considerable role in assessing the language competence level (e.g. Baker Smemoe et al. 2014; de Jong et al. 2021; Ginther et al. 2010; Iwashita et al. 2008), yet it is not considered in typical LPAs in practice.

In previous studies, we annotated parts of our collected data to extract several fluency-related features and used them, combined with biographic data from parent questionnaires, to create individual fluency profiles for the children (Kany and Trouvain, 2025). These profiles serve as a starting point in the development of an individual fluency score, as they provide an overview of all fluency-related features that are covered by our data and could possibly be considered in the final assessment.

Later, we conducted a human fluency assessment where raters evaluated the overall perceived fluency of the children in the recorded segments (Kany, 2025). With those ratings in combination with the feature values derived from our annotations and fluency profiles, we determined the individual effects the features have on the overall perceived fluency. In this study, we want to use the data and findings from the previous studies to develop and evaluate a model that is able to predict a human fluency rating of a recorded segment by the feature values derived from the annotations. This constitutes a major step on the road to develop an automatic speech fluency assessment that can be integrated into and enhance the proposed LPA method (Roche et al., 2019).

3. Data Acquisition

In order to develop a method to assess speech fluency of preschool children, we used the "Wuschel-App" (Roche et al., 2019) to collect data in German child daycare centres (kindergartens).

3.1. Principle Functionality of the Game

The game was developed to be played on a tablet. It tells a coherent story in which the main character, a dog named "Wuschel", is faced with a variety of obstacles and tasks, which need to be solved. To do so, he needs the help of the child playing the game.

The child needs to talk to the virtual character and answer his questions to progress through the story. The story is told in 28 scenarios with two questions each. The second question is a follow-up to the first one, offering the opportunity for the child to elaborate on his/her answer. The answers to those 56 questions are recorded and serve as the speech material in our data.

The game's process is controlled by a human conductor, as the whole system consists of two separate apps: the game app used by the child on a tablet and the conductor app used by the conductor on a smartphone to trigger the prompts. The task is based on the Wizard-of-Oz-Principle: the child thinks they interact with an autonomous system while the conductor decides when Wuschel's questions are played. This way, the conductor can control the timing of Wuschel's utterances and adjust them according to the child's behaviour. This helps to further enhance the dialog aspect of the game and supports the authenticity of the speech data.

3.2. Application in Practice

So far, we have collected data from 167 children in 30 kindergartens. The kindergartens provided us with a separate room to record our data. During a session, only the child, the conductor, and a member of the kindergarten staff were present in the room. The staff member served as a person of trust to the child.

Before the start of a session, the conductor instructed the kindergarten staff to not help the child with their task by using any words relevant to the solution. They were only allowed to support the child by motivating or comforting them. Then, the kindergarten staff brought the child into the room. The conductor greeted the child and talked to them for a minute before they introduced them to the game. They told the child about the basic principle of the game, i.e., that they have to help the dog "Wuschel" by talking to him and that they can only interact verbally with him. Then, the conductor withdrew from the child and sat down behind them to start and direct the game inconspicuously.

An average recording session lasted for about 30 minutes. This duration seems to be reasonable as we principally received very positive feedback from the participating children with a low dropout-rate of 2.5%. For our recordings we used the built-in microphone of an iPad (9th gen) as the device on which the game was played. While the quality of the recordings could certainly be better by using external microphones, this would create an uncomfortable situation for the child and affect the naturalness of our speech data. After the end of a session, the child was rewarded with colouring pictures of the game's characters and a medal for beating the game.

3.3. Data from Parent Questionnaires

For the interpretation of the speech data, additional meta data was elicited, e.g. biographical information from the participating children. These include the children's age, their first language, and their contact time with German, among other things. Therefore, the kindergarten staff handed out a questionnaire and a declaration of consent to the children's parents. The documents had to be filled out and returned to the kindergarten by the parents before their children were allowed to participate.

4. Speech Data from the App

So far, we have collected data from 167 children in 30 kindergartens in Germany (all located in the federal state Saarland). Our corpus comprises both, native speakers of German (L1) and second language (L2) speakers, the majority still being monolingual German L1 speakers (n=103). The children are between 4 and 6 years old and the L2 speaking children have a minimum of 1 year of contact time with the German language.

The game's structure of 28 scenarios with 2 questions each leads to a total of 56 recorded segments of speech per child. The mean duration of these segments in our corpus is 8.46 seconds (including pauses) and 3.23 seconds of articulation time (excluding pauses). This is markedly shorter compared to most other fluency assessment studies with adults (Derwing et al., 2009; Préfontaine et al., 2016; Suzuki and Kormos, 2019; Suzuki et al., 2021) but also to studies with children with usually read-aloud speech from older children to receive longer, continuous articulation phases (e.g. Gelin et al. 2025; Harmsen et al. 2025; Sappok 2023). In contrast, our data consists of spontaneous speech from preschool children gathered through a game-based elicitation method which simulates a dialog and thus, naturally causes the duration of coherent speech intervals to be shorter. This aspect needs to be kept in mind, as it might be a relevant factor for fluency assessment.

5. Processing of the Data

5.1. Cleaning

To reduce complexity in further processing steps, we muted all parts in the recordings that did not include any child speech. These parts usually consist of adult speech from other people in the room or background noise.

5.2. Annotation

The annotation of fluency-related phenomena such as filler particles, disfluent pauses, lengthenings, repetitions and repairs is a very complex task. It is highly dependent on research agenda, subdiscipline and specific research question (Trouvain et al., 2025). Thus, there is no common standard and we had to develop an annotation scheme that fits our purpose and our specific game-based data with its peculiarities.

The annotation of these phenomena should serve as a foundational basis for all further research, the development of automatic methods, and the assessment of the child's speech fluency. The manual annotations are done with Praat (Boersma and Weenink, 2024) in TextGrids. The TextGrids contain 6 tiers (see Figure 1).

The first tier contains a manually generated transcript that has been automatically aligned via WebMAUS (Schiel, 1999). It mainly serves as means of orientation for the annotators. The second tier deals with the annotation of pauses. The pauses were annotated with six possible attributes, taking into account if the pause is perceived to interrupt the flow of speech or not (disfluent (d) vs. fluent (f)), as well as their position in the dialog (i.e. if they are within or between turns or if they occur at the beginning or end of the recorded segment).

Tier 3 covers filler particles (FPs) produced by the child. FPs were assigned to one of four main types. We discriminate between monophthong vowel FPs ("äh"), diphthong vowel FPs ("ei", commonly found

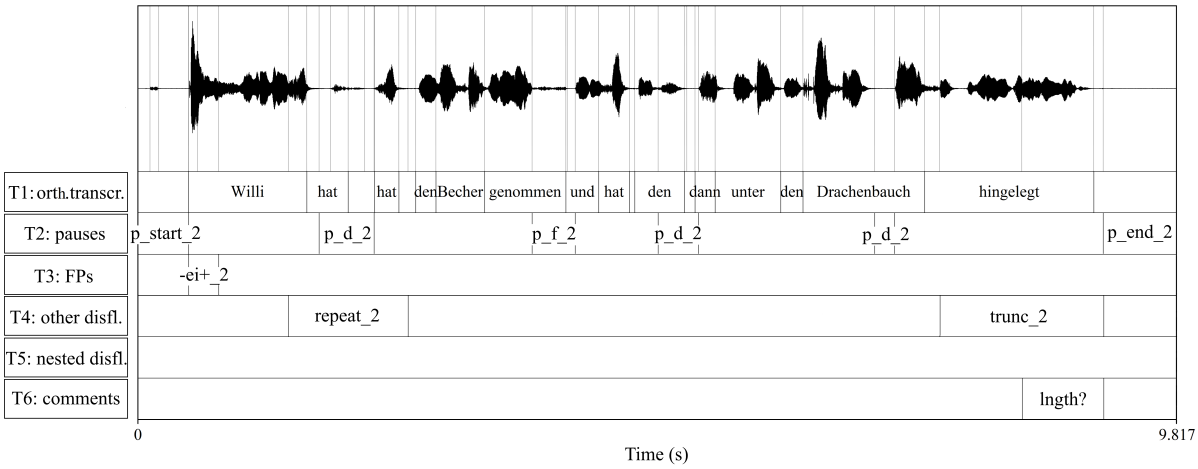

Figure 1: Example utterance with annotations on 6 tiers, reproduced from Kany and Trouvain (2025). See text for details.

in the local dialects), combinations of a monophthong plus a nasal consonant ("ähm"), and FPs consisting of only a nasal consonant ("hm"). They were further analysed on their position with respect to pauses. The label [-] is put in front of the label if the filler particle is preceded by a pause, while [+] describes that the filler particle is preceded by speech. The same is done at the end of the label to indicate if the filler particle is followed by speech or pause.

Tier 4 and 5 were used to annotate a group of phenomena that we called "other disfluencies", comprising 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) (Kany and Trouvain, 2025). The fifth tier is used in case nestings occur (e.g., a lengthening within the span of a repair). The sixth and last tier serves as a possibility to provide some comments or take notes in case anything interesting or unusual occurs that is not covered by the annotation scheme. All annotated labels on tier 2 to tier 5 include a number at the end to indicate if the utterance belongs to the first or second question of the scene. This might have an impact on the child's fluency because they might already be more familiar with the task of the scene and thus more fluent.

6. Perception Experiment

In order to be able to build a model that predicts human perceived fluency ratings, we needed some baseline ratings first. Thus, we decided to conduct a human perceived fluency assessment in an earlier study (Kany, 2025) where people rate the recorded segments on the child's speech fluency. 32 raters participated with background in linguistics. This way, they all had similar preconditions and a basic understanding of language at minimum, which should support consistency in the ratings. In spite of that, they do not represent the entire main target group of LPAs which are conducted by preschool or elementary school teachers, doctors, psychologists, or speech therapists in Germany, depending on the federal state (Lisker, 2010).

320 different segments (from 10 children) were used as stimuli. Each presented segment was rated four times. The raters had to rate the overall fluency of the child per segment on a 9-point Likert scale. Each rater was presented with 20 segments in random order twice. In total, 1,280 fluency ratings were achieved (32 raters x 20 segments x 2 ratings per segment).

These ratings are not distributed equally over the whole Likert scale. They show a clear tendency towards higher ratings (see Figure 2) which might be caused by the short duration of most stimuli.

To find possible links between the ratings and the fluency-related features that can be derived from our annotations, we built a first Cumulative Link Mixed Model (CLMM) for ordinal data (Christensen, 2023) with the features from our fluency profiles (Kany and Trouvain, 2025) as fixed effects and rater, child, and stimulus as random effects in R. 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 (R Core Team, 2025). CLMMs are suitable for our usecase as they can deal with our ordinal dependent variable (ratings from 1 to 9) and allow a mixture between fixed and random effects, such as stimulus or child. The model only revealed three significant effects: the

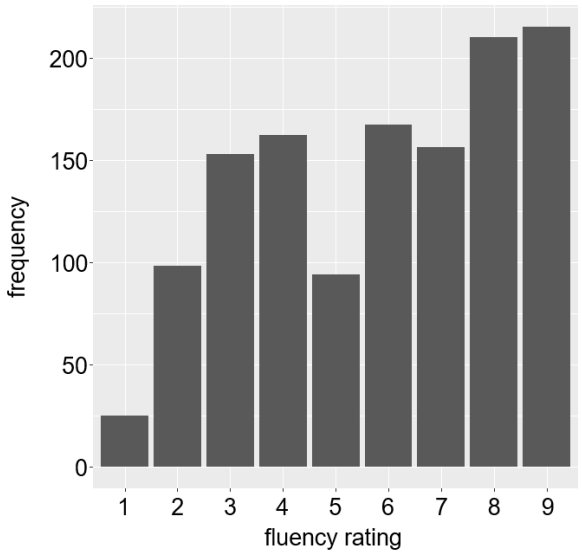

Figure 2: Frequency of the different fluency ratings given by the human raters in the fluency assessment (1 = not fluent at all, 9 = very fluent).

number of "disfluent pauses" in the stimulus, the number of "other disfluencies" in the stimulus, and the "articulation rate" in the stimulus. However, an additional graphical overview of all features with their mean value in relation to the different ratings suggested some non-linear relationships that could possibly not be modelled by the simple linear CLMM (see Figure 3).

7. Predicting Human Fluency Ratings

7.1. Modelling Non-Linear Effects

After identifying the crucial fluency features from the previous study and their link to the human fluency ratings, we aim to build a selection of new CLMMs that extend the baseline CLMM from the previous study by taking the non-linearity of certain features into account and test their ability to predict the human ratings.

First, we conducted a check for non-linearity for all z-standardised features by plotting the partial effect curves from Kany's (2025) CLMM for each feature in R. Then, we modelled all features that showed a considerably non-linear curve with penalised regression splines in our CLMM, as they provide a data-driven way to capture non-linear effects without the need to assume a specific functional form in advance (Wood, 2017). The plots support the first impression from the previous study, that the number of FPs in the audio segment has a non-linear effect on the fluency rating (see Figure 4): A low amount of FPs in the utterance does not affect the perceived fluency in a negative way. Only an extensive production of FPs seems to lead to a strong loss in fluency. The second feature that shows a noticeably non-linear effect curve is the "number of other disfluencies" (see Figure 4): The negative effect the production of more "other disfluencies" has on the fluency rating plateaus in the middle range. Only very high counts of disfluencies lead to a further decrease of the fluency rating.

Consequently, we build 3 different CLMMs that are compared against each other and the linear CLMM from the previous study:

- CLMM with "number of filler particles" as non-linear spline term
- CLMM with "number of other disfluencies" as non-linear spline term
- CLMM with "number of filler particles" + "number of other disfluencies" as non-linear spline terms

Based on the linear CLMM, all the models include the following features: 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 as fixed effects. Articulation rate was measured automatically (de Jong et al., 2021), whereas all other features were derived through our manual annotation introduced earlier. We further added "stimulus" as a random effect to control for individual variances. In contrast to our previous study, we had to exclude the random effects "rater" and "child" from our models, as we want to test their ability to rate new children. In this situation, we do not have this information.

We performed three likelihood-ratio tests (ANOVA in R) to test if these new non-linear models provide a better fit than the baseline linear model. Both the CLMM with "number of other disfluencies" modelled as a non-linear effect as well as the CLMM with "number of other disfluencies" and "number of filler particles" modelled as non-linear effects significantly improved model fit (LR $\chi^2(2) = 10.73$, $p < 0.001$; LR $\chi^2(4) = 13.04$, $p = 0.011$). The CLMM with only "number of filler particles" modeled as a non-linear effect did not significantly improve in model fit (LR $\chi^2(2) = 2.56$, $p = 0.279$), suggesting that only the non-linear effect of "number of other disfluencies" improves model fit significantly.

7.2. Evaluation

As our aim is to automatically predict human fluency ratings, we evaluated how well the models reproduced the ratings the audio segments received

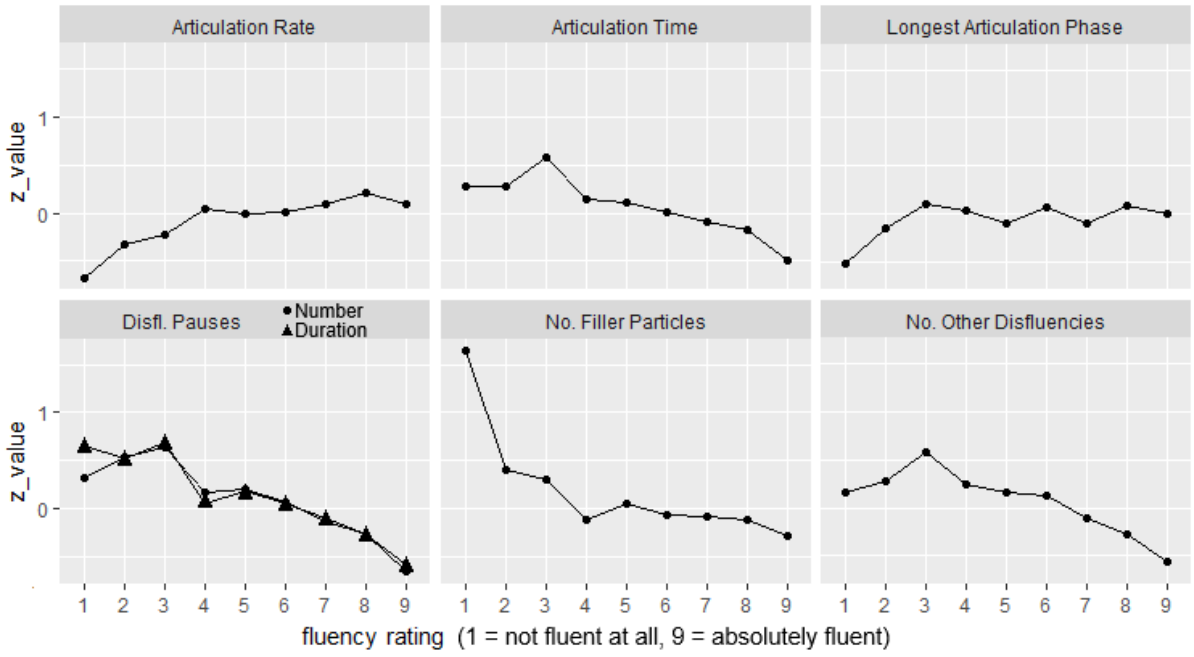

Figure 3: Overview of all features with their normalised mean z-value (y-axis) for stimuli awarded with a certain fluency rating (x-axis), reproduced from Kany (2025).

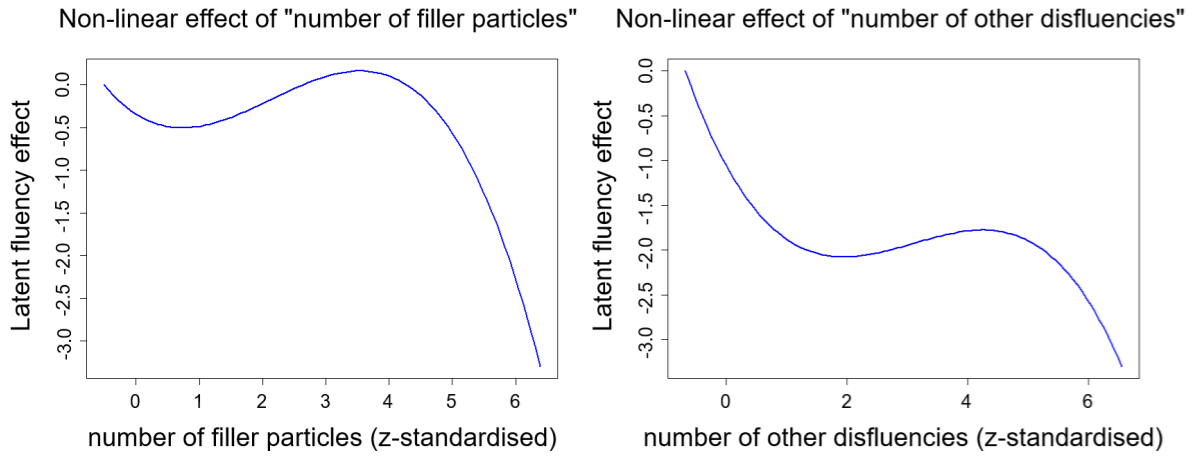

Figure 4: Partial effect curves of the features "number of filler particles" (left) and "number of other disfluencies" (right) in the baseline CLMM.

by human listeners in the previous assessment. To this end, we applied all four models to the entire dataset and compared their performance using three metrics: Spearman's Rho, Mean Absolute Error (MAE), and Quadratic Weighted Kappa (QWK). The results can be seen in Table 1. All models performed almost equally, most of the times values differed by only less than 0.01 (thus not visible in the table). This indicates that modelling the non-linearity of the effects does not help the model with the prediction of fluency ratings on a 9-point scale. However, the models provide reasonably accurate ratings overall. An MAE of 1.36 indicates that, on average, their ratings differ only by about 1.36 scale points from the human ratings on the 9-point scale. Furthermore, the QWA of 0.65 is slightly higher than the QWA between the human raters in the perceived fluency assessment (0.51), indicating that the agreement between the models and the human raters is higher than the agreement between the human raters themselves.

To get a better feel for the general performance of the models, we mapped the models' rating predictions and the actual human ratings onto confusion matrices. Since the models performed almost equally, we decided to only include the matrix of the

Model	Spearman's Rho	Mean Absolute Error	Quadratic Weighted Kappa
Baseline CLMM	0.72	1.36	0.65
Filler particles as spline	0.72	1.35	0.65
Other disfluencies as spline	0.72	1.36	0.65
Filler particles + other disfluencies as spline	0.72	1.36	0.65

Table 1: Performance of the baseline linear CLMM and the three CLMMs with non-linear spline terms evaluated using three different metrics.

model with "number of filler particles" as non-linear effect besides the baseline linear model here (see Figure 5), as it has a slightly lower MAE than the other models. Overall, the matrices underline the impression the evaluation metrics gave: The models are able to roughly predict the human ratings.

In most of the cases, the models stay within the range of +/-2 rating points from the human rating. It is conspicuous that the two displayed models have neither assigned a rating of 5 nor a rating of 1 to any of the audio segments. This has to do with the general distribution of the human ratings (see Figure 2). There are very few instances of a rating of 1 in general and thus, the models tend to prefer assigning a rating of 2 in those cases. However, other models not included in Figure 5 at least chose a rating of 1 in some rare cases. Contrary to that, the rating of 5 was never given by any of the models. This is due to the human rating distribution having a dent at the rating of 5. Human listeners tended to prefer making a decision towards any direction and chose either a rating of 4 or a rating of 6 in the assessment. As a consequence, the calculated probabilities for those ratings were always higher than the probability for the rating of 5 in our models.

8. Discussion and Conclusion

Aiming at a prediction of human fluency ratings, we included non-linear effects in our statistical models to account for the observed non-linearity of several fluency-related features. Incorporating the non-linear terms led to a significantly better model fit, yet the models failed to perform better than the baseline linear model in predicting the human fluency ratings. One possible explanation is that the non-linear terms refine the models' probabilities for choosing the correct human rating, which is recognised and rewarded by our likelihood-ratio tests. Still, these adjustments rarely change the most probable (hard) rating assignment and are thus barely visible in the final, rating-based evaluation metrics.

Modelling non-linear effects does not seem to yield any major benefits in predicting human perceived fluency ratings after all. The scarcity of low ratings (only 123 instances of ratings 1 and 2) in our data and already quite robust linear effects might be possible reasons for that. However, all in all, the models were mostly able to assign approximately the same rating to audio segments as human listeners. Between-rater agreement showed that they could further complement or even substitute human listeners in this task. This is a quite promising finding because fluency assessment of preschool children on a fine-grained 9-point scale is already challenging for trained listeners. In the final application-oriented task of assessing a child's speech fluency, the ratings will become even more robust, because 56 individually rated audio segments are aggregated into a single overall fluency score. If one is just interested in separating children's fluency into three categories (e.g., "not fluent", "average", "fluent") in a first step, all models discussed here already deliver solid results.

The main limitations of this study lie in the underlying data: For one thing, the major caveat of our game-based acquisition method is that we have to work with mostly short coherent intervals of speech. This offers less room to differentiate between fluent and disfluent utterances. For another, we have too few human rating data to develop our model on, especially on the lower end of the scale. Future research should therefore aim to stabilise the dataset with additional human ratings and automate the annotation process of fluency-related phenomena. In ongoing work, we fine-tuned a BERT-based token-level sequence labelling model (Devlin et al., 2018) on our data to automatically annotate the disfluency types relevant for the CLMMs presented in this study. Initial results look promising, as for some types (e.g. pauses, filler particles), detection accuracy already exceeded 90%. In a next step, we aim to incorporate an acoustic model to further improve performance and add those features that could not be detected by the BERT-based model.

Ultimately, our goal is to fully automate fluency assessment and integrate it into the WUSCHEL LPA framework (Roche et al., 2019) to make LPA in German kindergartens more objective, consistent, and efficient.

9. Acknowledgements

We are grateful to Julia Schu for her annotation work and to Diana Davidson for preparing the data. We would also like to thank all listeners for their participation in the rating task.

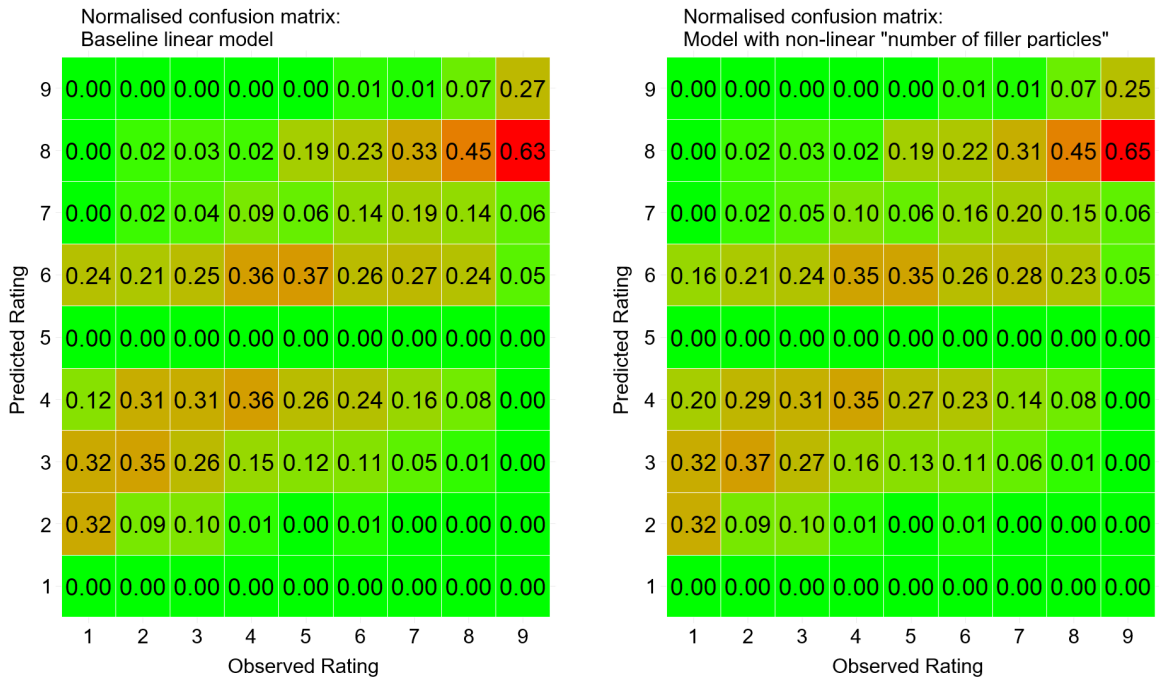

Figure 5: Confusion matrices of the baseline linear CLMM and the CLMM with feature "number of filler particles" modelled as non-linear spline. Values are normalised by the frequency of the rating in the human assessment (columns).

10. Bibliographical References

Wendy Baker Smemoe, Dan Dewey, Jennifer Bown, and Rob Martinsen. 2014. Does measuring L2 utterance fluency equal measuring overall L2 proficiency? Evidence from five languages. *Foreign Language Annals*, 47:707–728.

Paul Boersma and David Weenink. 2024. Praat: doing phonetics by computer (version 6.4.04).

Rune H. B. Christensen. 2023. *ordinal—Regression Models for Ordinal Data*. R package version 2023.12-4.1.

Nivja H. de Jong, Jos Pacilly, and Willemijn Heeren. 2021. Praat scripts to measure speed fluency and breakdown fluency in speech automatically. *Assessment in Education: Principles, Policy & Practice*, 28(4):456–476.

Tracey Derwing, Murray Munro, Ron Thomson, and Marian Rossiter. 2009. The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, 31:533–557.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805.

Stefan Faas, Alicia Götz, and Christiane Müller. 2021. *Sprachstandsfeststellung, Sprachförderung und sprachliche Bildung*. Pädquis Stiftung b.R., Berlin.

Natalia Gagarina, Daleen Klop, Sari Kunnari, Koula Tantele, Taina Välimaa, Ute Bohnacker, and Joel Walters. 2019. MAIN: Multilingual Assessment Instrument for Narratives – Revised. *ZAS Papers in Linguistics*, 63:20.

Merrill Garrett. 1980. Levels of processing in sentence production. In Brian Butterworth, editor, *Language Production Vol. 1: Speech and Talk*, pages 177–220. Academic Press, London.

Lucile Gelin, Lucas Block Medin, Alexandre Cruel, and Alice Liu. 2025. Combining word and phoneme speech recognition for fluency assessment of young children's oral reading. In *10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 16–20.

April Ginther, Slobodanka Dimova, and Rui Yang. 2010. Conceptual and empirical relationships between temporal measures of fluency and oral english proficiency with implications for automated scoring. *Language Testing*, 27(3):379–399.

Wieke Harmsen, Max van der Velde, Roeland van Hout, Catia Cucchiarini, and Helmer Strik. 2025. Unraveling the relationship between objective and subjective measures of oral reading fluency. In *10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 71–75.

Noriko Iwashita, Annie Brown, Tim McNamara, and Sally O'Hagan. 2008. Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29(1):24–49.

Valentin Kany. 2025. From Features to Fluency: Predicting Perceived Speech Fluency of Preschool Children for Language Proficiency Assessments. In *10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 118–122.

Valentin Kany and Jürgen Trouvain. 2025. Annotation of disfluencies in child speech. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2025*, pages 247–254. TUDpress, Dresden.

Paul Lennon. 1990. Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40(3):387–417.

Willem J.M. Levelt. 1993. *Speaking: From Intention to Articulation*. MIT Press.

Robin J. Lickley. 2015. Fluency and disfluency. In Melissa Redford, editor, *The Handbook of Speech Production*, pages 445–474. Wiley Online Library.

Andrea Lisker. 2010. *Sprachstandsfeststellung und Sprachförderung im Kindergarten sowie beim Übergang in die Schule. Expertise im Auftrag des Deutschen Jugendinstituts*. Deutsches Jugendinstitut, München.

Toni Mayr and Michaela Ulich. 2003. *SISMIK. Sprachverhalten und Interesse an Sprache bei Migrantenkindern in Kindertageseinrichtungen*. Herder, Freiburg.

Yvonne Préfontaine, Judit Kormos, and Daniel Ezra Johnson. 2016. How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language Testing*, 33(1):53–73.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Jörg Roche, Stefanie Haberzettl, Giulio Pagonis, Moiken Jessen, and Nicole Weidinger. 2019. Serious Games in der Sprachstandsermittlung. In Jörg Roche, editor, *Propädeutikum wissenschaftliches Arbeiten: Schwerpunkt DaF/DaZ und Sprachlehr-/Spracherwerbsforschung*, pages 340–358. Narr Francke Attempto Verlag.

Christopher Sappok. 2023. Oral reading proficiency and prosody–a perceptual pilot study on especially fluent german students (grade 3 to 7). In *Proc. of the 20th International Congress of Phonetic Sciences (ICPhS)*, pages 1538–1542, Prague.

Florian Schiel. 1999. Automatic Phonetic Transcription of Non-Prompted Speech. In *Proc. 14th International Congress of Phonetic Sciences (ICPhS)*, pages 607–610, San Francisco.

Petra Schulz and Rosemarie Tracy. 2011. *LiSe-DaZ: Linguistische Sprachstandserhebung – Deutsch als Zweitsprache*. Hogrefe, Göttingen.

Shungo Suzuki and Judit Kormos. 2019. Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, 42:435–463.

Shungo Suzuki, Judit Kormos, and Takumi Uchihara. 2021. The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *Modern Language Journal*, 105:143–167.

Jürgen Trouvain, Ludivine Crible, Malte Belz, Simon Betz, Štefan Beňuš, Lorraine Baqué, Marina Cantarutti, Jessica Di Napoli, Ivana Didirková, Maria Machuca, Lucia Mareková, Oana Niculescu, Pauliina Peltonen, Aurelie Pistono, Loredana Schettino, Vered Silber-Varod, and Simon Williams. 2025. On variability in the identification and labelling of disfluencies — preliminary results from 23 annotations of the same data. In *12th edition of the Disfluency in Spontaneous Speech Workshop (DiSS 2025)*, pages 57–61, Lisbon.

Simon N. Wood. 2017. *Generalized Additive Models: An Introduction with R, Second Edition (2nd ed.)*. Chapman & Hall/CRC.