Automatic Detection of Disfluencies in L1 and L2 Child Speech

Martha Schubert[1], Valentin Kany[1 2]
Universität des Saarlandes
[1]Dept. Language Science and Technology,
[2]Dept. German Studies

BACKGROUND & PREVIOUS WORK

Background & Motivation:

- Common practice in Germany: Language Proficiency assessment (LPA) for preschool children [1]
- Most applied LPA methods: test for children's vocabulary size, grammar skills [2, 3] and morphology [4] but not fluency
- Speech fluency correlates with language proficiency, e.g. [5], [6], [7]
  → Need for an individual assessment of child's fluency
- Assessment by human raters is i) complex, ii) time consuming, iii) inconsistent
  → Aim: development of automatic annotation of fluency related phenomena

Previous Work:

- Most prior models based on adult speech, binary labels or single disfluency types → limited coverage of fine-grained disfluencies in spontaneous child speech
- Adaptation of BERT-based model [8] to German child speech

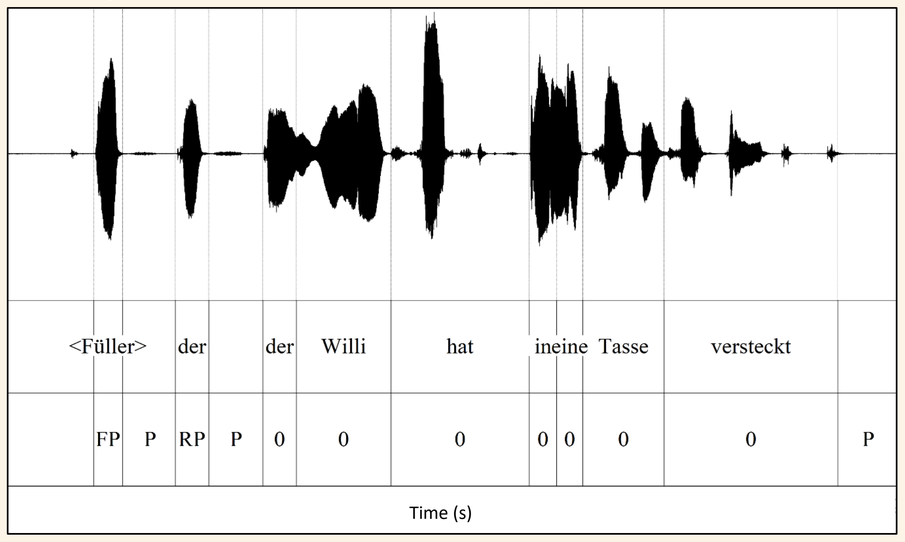
THE DATASET

Acquisition method:

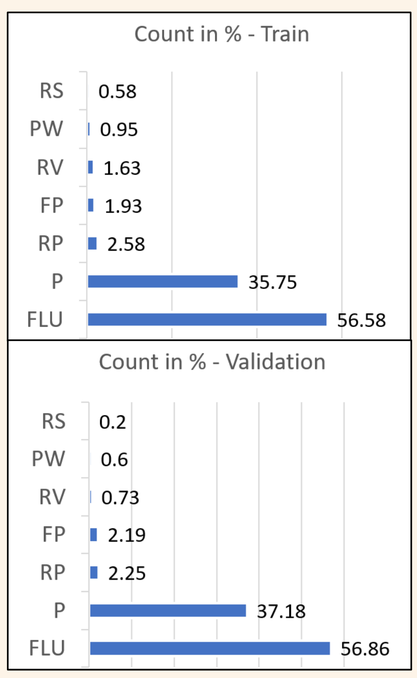
- Data acquisition with WUSCHEL [9], a game-based task in a custom-made app: children interact with virtual character, answer questions to progress through coherent scenes
- Recordings from 167 children aged 4-6 years

Data: 28 scenes, 2 answers each → 56 segments (∅ 7s duration) per child; muting non-child speech

Annotation: Manually annotated 14 children (748 utterances, 7204 tokens)



THE MODEL

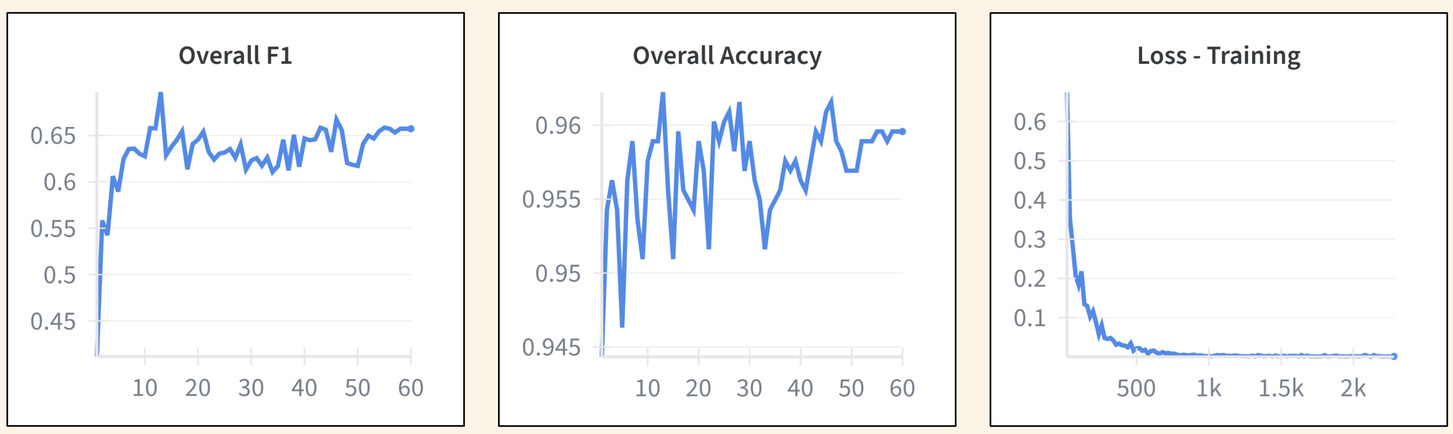


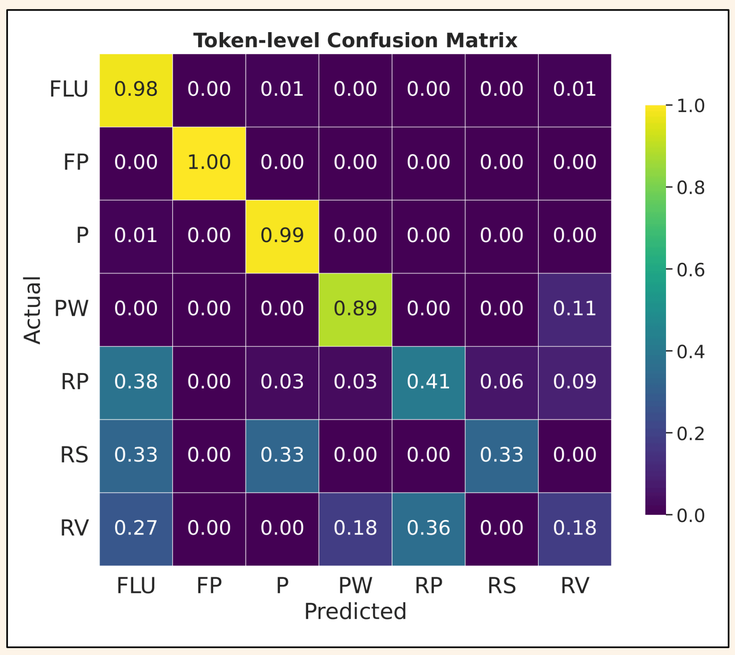
Preprocessing:
Consolidation of labels → final label set: filler particles (FP), partial words (PW), repetitions (RP), restarts (RS), revisions (RV), pauses (P), fluent tokens (FLU)

Training Process:

- German BERT fine-tuned for token-level sequence labeling
- Train/Val Split: 80/20
- 60 epochs, Learning rate: 3×10^{-5}

RESULTS & DISCUSSION





- **Final macro F1 score: 65.7%, final token-level accuracy: 96.0%**
  → outperforming majority baseline: 56.86%
- Performance is strongest for most frequent classes (FLU, FP, P)
- Rarest classes often misclassified as FLU
  → due to imbalanced dataset
- RVs: often misclassified as RPs
  → RVs (in contrast to RSs) often being minor corrections (e.g. function words)
  → repair = single word (similar linguistic properties to those of RPs)
- RSs: often misclassified as Ps
  → both appearing at utterance boundaries

REFERENCES

[1] Faas et al. (2021). Sprachstandsfeststellung, Sprachförderung und sprachliche Bildung. Pädquis Stiftung b.R., Berlin.
[2] Schulz & Tracy (2011). LiSe DaZ: Linguistische Sprachstandserhebung – Deutsch als Zweitsprache. Hogrefe, Göttingen.
[3] Gagarina et al. (2019). Main: Multilingual assessment instrument for narratives – revised. ZAS Papers in Linguistics, 63, 20.
[4] Mayr & Ulich (2003). Sismik – Sprachverhalten und Interesse an Sprache bei Migrantenkindern in Kindertageseinrichtungen.
[5] Iwashita et al. (2008). Assessed levels of second language speaking proficiency: How distinct? Applied Linguistics, 29, 24–49.
[6] De Jong et al. (2021). Praat scripts to measure speed fluency and breakdown fluency in speech automatically. Assessment in Education: Principles, Policy & Practice, 28, 456–476.
[7] Ginther et al. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. Language Testing, 27, 379–399.
[8] Romana et al. (2024). Automatic disfluency detection from untranscribed speech. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32, 4727–4740.
[9] Weidinger et al. (in press). Assessing multilingual children from a usage-based perspective: The WUSCHEL approach. Usage-based approaches to multilingualism: Language acquisition, language contact, multilingual language use.

SUMMARY AND FUTURE WORK

- Transformer model achieves 96% token accuracy and 65.7% macro F1; rare disfluencies remain challenging
- Ongoing annotation with more data will reduce class imbalance and enable finer-grained disfluency labels
- Adding acoustic features via WavLM aims to capture prosody and improve detection of rare disfluencies as well as disfluencies not included in text-based detection