Benchmarking NLP-supported Language Sample Analysis for Swiss Children's Speech

Anja Ryser, Yingqiang Gao, Sarah Ebling*
Department of Computational Linguistics
University of Zurich, Switzerland
{ryser, yingqiang.gao, ebling}@cl.uzh.ch

Abstract

Language sample analysis (LSA) is a process that complements standardized psychometric tests for diagnosing, for example, developmental language disorder (DLD) in children. However, its labour-intensive nature has limited its use in speech-language pathology practice. We introduce an approach that leverages natural language processing (NLP) methods that do not rely on commercial large language models (LLMs) applied to transcribed speech data from 119 children in the German-speaking part of Switzerland with typical and atypical language development. This preliminary study aims to identify optimal practices that support speech-language pathologists in diagnosing DLD more efficiently with active involvement of human specialists. Preliminary findings underscore the potential of integrating locally deployed NLP methods into the process of semi-automatic LSA.

Keywords: language sample analysis, developmental language disorder, automatic speech recognition.

1. Introduction

Developmental language disorder (DLD) is a neurodevelopmental condition, commonly diagnosed in children, that significantly affects an individual's ability to acquire and use spoken and written language, despite typically developed intelligence and no obvious sensory or neurological impairments or inadequate language exposure (Tomblin et al., 1996; Bishop, 2006; Lüke et al., 2023; van Wijngaarden et al., 2024).

As a recommended part of DLD diagnosis, language sample analysis (LSA) aims at evaluating the spontaneous[1] language production skills of children (Gallagher and Hoover, 2020; Ramos, 2024). It involves collecting and analysing samples of language during conversation, storytelling, play, or other activities. LSA provides detailed information about a person's linguistic abilities, including vocabulary, grammar, sentence structure, and pragmatic language use. These insights can then be used in diagnostics, setting therapeutic goals and monitoring progress.

Despite being an effective tool for practice, LSA is not often used by speech-language pathologists due to its time- and effort-intensive process (Klatte et al., 2022; Bawayan et al., 2022). Modern NLP methods and machine learning can help to alleviate some of these challenges with their time efficient approaches to big amounts of data

In this study, we evaluate the zero-shot capability of non-commercial NLP methods, namely ASR and part-of-speech (POS) tagging on data collected from 110 children living in Switzerland. These NLP methods are the foundation on which analyses of the speech and language samples, such as measurements for lexical density and diversity and many others, are based on. To ensure the quality and reliability of these analyses, it is crucial to have trustworthy NLP methods with no commercial LLMs involved, as described in this study.

Transcription	Variant	WER	CER	MER	WIL
Original	Swiss German	81.0	80.0	49.9	94.8
	Swiss Std. German	48.7	47.8	36.1	59.5
Normalized	Swiss German	57.2	55.7	38.6	73.8
	Swiss Std. German	45.6	44.8	35.2	54.7

Table 1: Average ASR results on Whisper transcriptions for original and normalized text.

With data collected from 119 children living in Switzerland, we present a case study demonstrating that NLP approaches that do not rely on commercial LLMs can effectively assist speech-language pathologists in identifying critical linguistic patterns essential for LSA. Results from analyses of Swiss German and Swiss Standard German speech transcriptions highlight the potential of automating LSA. To achieve the high quality performance needed in a clinical setting, our results also show the need for specific training and fine-tuning.

Our **main contributions** are three-fold: 1) We demonstrate a possible annotation process in semi-automated LSA for the elicitation of clinical linguistic features of DLD in Swiss German; 2) We release the first case study dataset in Swiss German and Swiss Standard German containing six speech transcriptions together with their annotations; 3)

*Corresponding author. Dataset on SWISSUBase

[1]We acknowledge that elicited language is never completely "spontaneous"; nevertheless, the term is common in connection with LSA.

Proceedings of the 6th RaPID@MENTAI.ai Workshop @ LREC 2026, pages 55–73

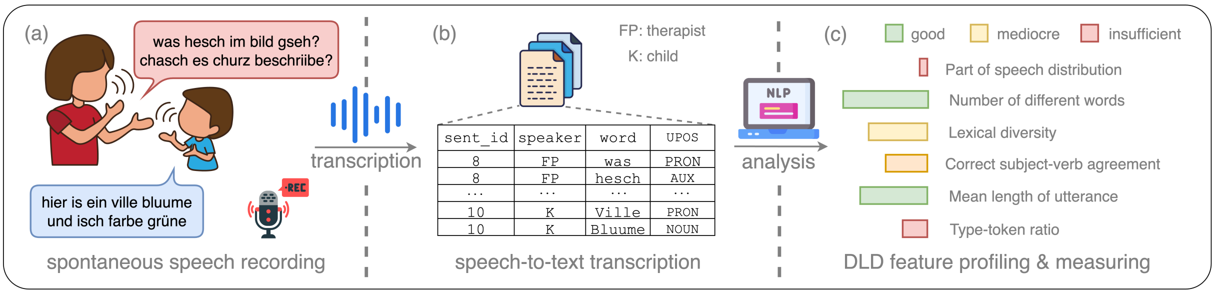


Figure 1: Our pipeline of LSA with NLP-supported approaches for diagnosis of DLD. **(a) Spontaneous speech recording**: a speech-language pathologist interacts with the child in a naturalistic setting and both of their speeches are recorded; **(b) Speech-to-text transcription**: the recordings are (automatically) converted into text and post-corrected by the speech-language pathologist and further (automatically) tokenized to words; **(c) DLD feature profiling and measuring**: approaches such as POS tagging, dependency parsing, stemming and lemmatization, etc., are applied to create the DLD feature profiles, where various linguistic measures are computed to evaluate the language abilities of children. The final diagnostic decision is made by the human expert (i.e., the speech-language pathologist), taking into account the output of the pipeline as well as other criteria. The speech utterances demonstrated are in Swiss German.

We provide an assessment of the effectiveness of NLP methods not based on commercial LLMs for LSA on both Swiss German and Swiss Standard German.

2. Related Work

2.1. Speech Transcription

Transcribing children's speech into text is a critical first step in LSA. However, advanced end-to-end approaches, such as ASR, still face substantial challenges due to the high inter- and intra-speaker variability in pronunciation, vocabulary, and speech rate among children across different ages as well as higher fundamental frequencies in children (Potamianos et al., 1997; Bhardwaj et al., 2022). Smith et al. (2017) trained a deep neural network on out-of-domain adult speech data, which was subsequently fine-tuned using speech data from children with DLD. Similarly, Rumberg et al. (2021) proposed a framework for age-invariant training, leveraging age-independent patterns derived from both adult and child speech. Jain et al. (2023) studied the adaption of the Whisper model (Radford et al., 2023) to child speech via self-supervised fine-tuning. Pokel et al. (2025) introduced a novel algorithm for ASR tailored to dysarthric German speech, which restructures word-level utterances into sentence-level sequences. This approach demonstrates promising results in improving the accessibility of speech transcriptions.

These approaches, often leveraging the power of transfer learning by utilizing adult speech data, have shown effectiveness in increasing performance on children's speech data. However, the fundamental lack of data collected from children with DLD still limits the applicability of ASR in LSA. In recent years, ASR systems for low-resource languages, such as Swiss German, have been developed (Kew et al., 2020; Nigmatulina et al., 2020; Arabskyy et al., 2021; Schraner et al., 2022; Timmel et al., 2024), supported by the availability of corpora of Swiss German speech data (Plüss et al., 2020; Dogan-Schönberger et al., 2021; Plüss et al., 2022, 2023; Stucki et al., 2025). However, due to licensing limitations, these models are not currently available and thus the ability of Swiss German models to generalize to speech data of children with DLD remains largely unexplored. Additionally, finetuning of commercial models running on commercial computation services with children's speech data is heavily restricted due to legal and ethical considerations and limitations. (Liu et al., 2024). As Whisper models perform well on Swiss German in a zero-shot setting (Dolev et al., 2024) and their availability, a Whisper model was used for this research for speech transcription.

2.2. Feature Analysis

Initial efforts to develop semi-automated LSA approaches have emerged in the past years, with a primary focus on English speech data. Gabani et al. (2011) analysed speech data collected from monolingual English-speaking children and proposed NLP methods to predict the presence of DLD. The study utilized eight categories of linguistic features (later expanded by Hassanali et al. (2012) to include syntactic and semantic features) to train language models as predictors, which provided important aspects for future research. Solorio (2013) provided a concise summary of the types of NLP-features employed in LSA for the diagnosis of DLD and highlighted questions for future research. Lüdtke et al. (2023) described the ideal hypothetical system capable of recording spontaneous speech of children while effectively separating background noise and speech from non-target individuals. This system can transcribe and segment recorded speech, offering a wide range of measurements, including environmental factors of recording, DLD profiles, and detailed analyses across various linguistic structures and elements.

These studies have laid the groundwork for advancing semi-automated LSA approaches, highlighting key linguistic features, methodological considerations, and future directions for improving DLD diagnosis. However, none of these works investigated LSA for children's speech in Swiss German.

DLD features can manifest in all linguistic categories. At the moment we are focusing on the grammatical level, which can be analysed based on transcripts of speech. As part of future research, phonetic and phonological features could also be analysed directly on the speech recordings.

In this work, we show that it is possible to perform LSA using NLP methods that are not based on commercial LLMs and are both ethical and effective. While we acknowledge LLMs are likely to dominate in LSA, we argue that other NLP methods still deliver good results without causing potential ethical issues.

3. Data

The dataset is scheduled to be published on SWISSUbase (DOI: `https://doi.org/10.48656/rqf2-sq76`) and will be released in the near future for public access.

3.1. Data Collection

Speech data collected from children are highly sensitive and require careful handling. In compliance with the regulations of the responsible research committee, we obtained the necessary ethical approval to collect speech utterances, accompanied by signed consent forms from the children's parents.[2] Speech utterances were recorded in both therapeutic and naturalistic settings (i.e., kindergarten), capturing spontaneous interactions between one therapist and one child. To ensure privacy, the data were stored in an anonymized format, with no metadata linked to identifiable codes.

We collected speech samples of duration between 10 and 20 minutes from 119 children with typical and atypical Swiss German and Swiss Standard German speech and 25 speech-language pathologists. Of these, we obtained permission to publish the data of 41 recordings. Although the dataset is relatively small, our objective is to initiate research into the application of semi-automated LSA for Swiss German speech utterances based on this dataset. The recordings were made with phones in standard quality and stored as mp3-files.

[2]For further details, please refer to the Ethics Statement section.

	Swiss German	Swiss Std. German
# recordings	91	19
# hours	17:57	3:35
# utterances	16,553	3,014
# words	126,733	21,356

Table 2: Statistics of the collected data. The table summarizes the total duration of recordings, the number of utterances and individual words for both speakers.

3.2. Data Statistics

Table 2 presents the overall statistics of the collected speech data. Speech recordings were obtained in both Swiss German and Swiss Standard German. The ages of the children range from four to eight, encompassing an important phase in the assessment of and intervention for DLD (Sansavini et al., 2021). The speech data were transcribed by students of speech and language therapy and subsequently verified by professionals native in both Swiss German and Swiss Standard German.

We engaged with children living across Switzerland. Therefore, the collected recordings are composed of different Swiss German dialects such as "*Züridüütsch*" (dialect spoken in the Canton of Zurich) and "*Baseldüütsch*" (dialect spoken in the Canton of Basel).

4. Methods

Starting from the collected recordings of spontaneous speech of Swiss children, we showcase our data processing methods specifically used for semi-automated LSA.

4.1. Speech Transcription

The first task is to transcribe speech into text based on the raw audio recordings. To do this, we apply two approaches:

Manual transcription by human experts. We recruited 13 students majoring in speech and language therapy, who had received training in the transcription method used for this study. For the transcriptions we use an adapted form of the *Dieth-Schreibung* (*Dieth spelling*) (Dieth, 1986) with added annotation of features important for language and speech pathology, such as annotation of stress, unintelligible sequences, pauses, mazes, and overlaps in turn taking. Each transcript was created by one student and checked by another. The manual transcriptions served as the ground-truth reference for our study. See Table 20 in Appendix E for dataset examples. Additionally, the authors normalized the transcript manually to compare the influence on non-standard orthography and an assimilation to Standard German. Where possible, orthographically correct versions of words were used while keeping the word order intact.

Transcription using ASR models. We deployed a Whisper model (Radford et al. (2023), Hugging Face checkpoint `openai/whisper-small`[3]) **locally** to transcribe speech recordings into spoken sentences. This lightweight Whisper model, with 244 million parameters, was selected for its ease of deployment on local computers to prevent data leakage and its multilingual transcription capabilities and before the data was collected.

Automatically transcribing the speech samples used in this study posed three challenges: (1) Most large ASR systems are not trained with Swiss German data; (2) Most ASR systems underperform in transcribing speech from children (Bhardwaj et al., 2022); (3) The speech transcribed contains non-standard, atypical or wrong grammar due to the atypical language development of the children. Combining these three difficulties is challenging, requesting highly specifically trained models to solve the task reliably.

Manual transcription of spontaneous speech demands substantial knowledge and expertise in LSA and considerable time investment. Therefore, we aimed to evaluate the performance of the state-of-the-art Whisper model on the speech transcription task, given its potential to significantly reduce the workload of speech-language pathologists in practice.

4.2. Part-of-speech (POS) Tagging

Since POS tagging delivers information for certain linguistic features used for the feature analysis as described in Section 2.2, and thus helps in identifying morphosyntactic errors in language samples, our second task is to perform POS tagging for transcriptions in Swiss German and Swiss Standard German. For an overview over both POS tag sets used in this work, see Table 3.

For Swiss German transcriptions, we applied two BERT-based models (Aepli and Sennrich, 2022), trained with two different POS tag sets: 1) `swiss_german_pos_model`, trained with the Universal POS tags (UPOS[4]), and 2)

UPOS Tags	Feature	UPOS Tags	Feature
ADJ	adjective	ADP	adposition
ADV	adverb	AUX	auxiliary
CCONJ	co. conjunction	DET	determiner
INTJ	interjection	NOUN	noun
NUM	numeral	PART	particle
PRON	pronoun	PROPN	proper noun
PUNCT	punctuation	SCONJ	sub. conjunction
SYM	symbol	VERB	verb
X	other		

STTS Tags	Feature
ADJA	Attributive adjectives
ADJD	Predicative or adverbial adjectives
APPO	Postpositions
APPR	Prepositions
APPRART	Prepositions with an article
APZR	Circumpositions (right part)
ADV	True adverbs
ART	Definite/indefinite articles
CARD	Cardinal numbers
KOKOM	Comparative particles
KON	Coordinating conjunctions
KOUI	Subordinating conjunctions with infinitive
KOUS	Subordinating conjunctions with clauses
ITJ	Interjections
NE	Proper nouns
NN	Common nouns
PTKA	Particles with adjectives or adverbs
PTKANT	Response particles
PTKNEG	Negation particles
PTKVZ	Separable verb prefixes
PTKZU	"zu" before infinitives
PAV	Pronominal adverbs
PDAT	Attributive demonstrative pronouns
PDS	Substituting demonstrative pronouns
PIAT	Attributive indefinite pronouns without determiners
PIDAT	Attributive indefinite pronouns with determiners
PIS	Substituting indefinite pronouns
PPER	Non-reflexive personal pronouns
PPOSAT	Attributive possessive pronouns
PPOSS	Substituting possessive pronouns
PRF	Reflexive personal pronouns
PRELAT	Attributive relative pronouns
PRELS	Substituting relative pronouns
PWAT	Attributive interrogative pronouns
PWAV	Adverbial interrogative pronouns
PWS	Substituting interrogative pronouns
TRUNC	Truncation
FM	Foreign language material
XY	Non-words
VAFIN	Auxiliary finite verbs
VMFIN	Modal finite verbs
VVFIN	Full finite verbs
VAIMP	Auxiliary imperative verbs
VVIMP	Full imperative verbs
VAINF	Auxiliary infinitives
VMINF	Modal infinitives (substitute infinitive)
VVINF	Full infinitives
VVIZU	Infinitives with "zu"
VAPP	Auxiliary past participles
VMPP	Modal past participles
VVPP	Non-inflected full past participles

Table 3: Overview of two POS tagging systems for German, UPOS (top) and STTS (bottom). The core difference is that STTS contains the level of detail required for LSA.

[3]Model available at https://huggingface.co/openai/whisper-small

[4]https://universaldependencies.org/u/pos/

swiss_german_stts_pos_model, trained with the Stuttgart-Tübingen-Tagset (STTS[5]). Both models were deployed **locally** to adhere to the data privacy rules.

For Swiss Standard German transcriptions, we tested the statistical model de_core_news_sm provided by spaCy[6] as a supplement to the BERT-based models.

The models were chosen due to their availability before the data was collected.

Inter-annotator Agreement We recruited three native Swiss German speakers and three speakers with profound knowledge of Swiss Standard German with strong background in computational linguistics to annotate the gold-standard UPOS and STTS tags for sentences in the manual speech transcriptions. These annotations serve as ground truth for evaluating the performance of the three models discussed in Section 4.2. Prior to initiating the annotation process, we conducted training sessions with all annotators, during which the task instructions were explained in detail. We attached the instruction sheet in Appendix A for reference.

	Swiss German	Swiss Std. German
A&B	0.804	0.861
B&C	0.850	0.844
A&C	0.802	0.900

(a) IAA for UPOS tagging.

	Swiss German	Swiss Std. German
A&B	0.910	0.926
B&C	0.939	0.921
A&C	0.926	0.945

(b) IAA for STTS tagging.

Table 4: Pairwise linearly weighted Cohen's Kappa (Cohen, 1968) among three human annotators who are native Swiss German speakers for UPOS and STTS tagging on Swiss German (100 sentences) and Swiss Standard (Std.) German (80 sentences) transcriptions.

4.3. Morphological Features

German is a morphologically rich language that utilizes all 17 UPOS tags. For instances where a single POS tag is insufficient to capture lexical distinctions — such as the straightforward example of nouns in German, which can exhibit three different genders: masculine, feminine, and neutral—, features that describe these linguistic differences become essential.

Key	Value
Case	Acc, Nom, Gen, Dat
Number	Sing, Plur
Gender	Fem, Masc, Neut
Person	1, 2, 3
PronType	Art, Dem, Ind, Int, Prs, Rel
Mood	Ind, Sub, Imp
Tense	Past, Pres
VerbForm	Fin, Inf, Part
Definite	Def, Ind
Degree	Cmp, Pos, Sup
Foreign	Yes
Poss	Yes
Reflex	Yes

Table 5: Values for morphological categories. Notice that morphological categories rely on the spaCy model de_core_news_sm and are therefore language- and model-dependent.

To broaden the scope of our study, we annotated the morphology of language samples by assigning specific values to identified linguistic features (i.e., keys in our annotations). Table 5 provides a summary of the morphological values associated with these keys. Additionally, Table 6 outlines the mappings between POS tags and linguistic features, presented as key-value pairs.

The automatic prediction of morphological key-value features can be achieved using traditional statistical models (Can, 2011; Silfverberg and Lindén, 2011) or deep learning approaches (Tkachenko and Sirts, 2018; Bohnet et al., 2018; Klemen et al., 2023), both of which require a substantial amount of Swiss German language samples. However, deep learning approaches require a large amount of training data, which is currently absent in the Swiss LSA research.

To address this requirement, we are working on collecting additional language samples to support the training of these models. A comprehensive investigation of morphology prediction, however, is beyond the scope of this study and will be comprehensively addressed in future work.

5. Results

5.1. ASR Transcriptions

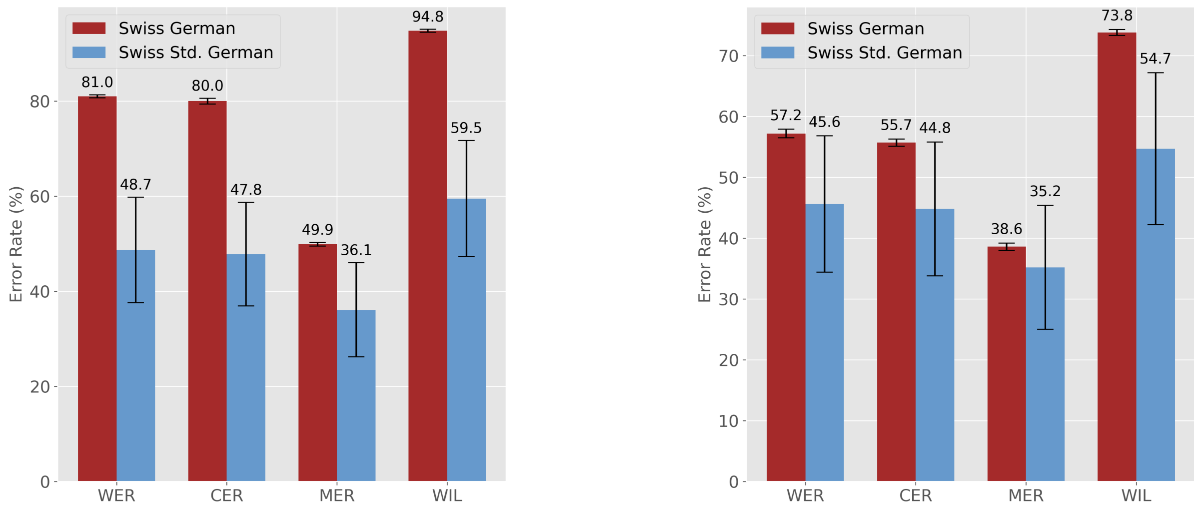
In Figure 2, we visualize the error rates of the Whisper-based ASR model transcribing the spontaneous speech recordings between one therapist and one child (Table 1). We report Word Error Rate (WER), Character Error Rate (CER), Match Error Rate (MER), and Word Information Lost (WIL) of Swiss German and Swiss Standard German tran-

[5]More details available at https://homepage.ruhr-uni-bochum.de/stephen.berman/Korpuslinguistik/Tagsets-STTS.html

[6]https://spacy.io/models/de, MIT licence

Tag	Key-Value Features
ADJA	{'Case': 'Acc', 'Degree': 'Pos', 'Gender': 'Fem', 'Number': 'Plur'}
ADJD	{'Degree': 'Pos'}
APPRART	{'Case': 'Dat', 'Gender': 'Neut', 'Number': 'Sing'}
ART	{'Case': 'Nom', 'Definite': 'Def', 'Gender': 'Neut', 'Number': 'Sing', 'PronType': 'Art'}
NN/NE	{'Case': 'Acc', 'Gender': 'Masc', 'Number': 'Sing'}
PDS/PDAT (plural)	{'Case': 'Nom', 'Number': 'Plur', 'PronType': 'Dem'}
PDS/PDAT (singular)	{'Case': 'Nom', 'Gender': 'Fem', 'Number': 'Sing', 'PronType': 'Dem'}
PIS	{'Gender': 'Neut', 'PronType': 'Ind'}
PIAT	{'Case': 'Nom', 'Gender': 'Fem', 'Number': 'Sing', 'PronType': 'Ind'}
PPER (other)	{'Case': 'Nom', 'Number': 'Sing', 'Person': '3', 'PronType': 'Prs'}
PPER (singular)	{'Case': 'Nom', 'Gender': 'Neut', 'Number': 'Sing', 'Person': '3', 'PronType': 'Prs'}
PPOSAT	{'Case': 'Acc', 'Gender': 'Fem', 'Number': 'Sing', 'Poss': 'Yes', 'PronType': 'Prs'}
PPOSS	{'Case': 'Nom', 'Gender': 'Masc', 'Number': 'Sing', 'Poss': 'Yes', 'PronType': 'Prs'}
PRF	{'Case': 'Acc', 'Number': 'Sing', 'Person': '3', 'PronType': 'Prs', 'Reflex': 'Yes'}
PRELS/PRELAT	{'Case': 'Acc', 'Gender': 'Neut', 'Number': 'Sing', 'PronType': 'Rel'}
PWS	{'Case': 'Nom', 'Gender': 'Masc', 'Number': 'Sing', 'PronType': 'Int'}
PWAV	{'PronType': 'Int'}
VERB	{'Mood': 'Ind', 'Number': 'Sing', 'Person': '3', 'Tense': 'Pres', 'VerbForm': 'Fin'}
VERB IMP	{'Number': 'Sing'}
VERB INF	{'VerbForm': 'Inf'}
VVPP	{'VerbForm': 'Part'}

Table 6: Morphological features for STTS tags (based on `spaCy`), sorted alphabetically. The UPOS tags are converted to STTS tags using a conversion look-up table.



(a) Original Transcriptions.

(b) Normalized Transcriptions.

Figure 2: Average ASR results on Whisper transcriptions with standard deviations.

scriptions.[7] The WER for Swiss German is 81%, and 48.7% for Standard German. After normalizing the transcripts, the WER is 57.2% for Swiss German and for Standard German 45.6%

5.2. Inter-annotator Agreement on Human Labelling of POS Tags

Inter-annotator agreement (IAA) scores were calculated based on 100 sentences for Swiss German and 80 sentences for Swiss Standard German, as one annotator for Swiss Standard German was unable to complete annotations for all 100 sentences. These sentences were randomly sampled from the corresponding transcriptions.

Overall, we achieved high IAA scores for both UPOS tagging (above 0.8) and STTS tagging (above 0.9), as measured using linearly weighted Cohen's Kappa (Cohen, 1968). The detailed scores are presented in Table 4.

5.3. Automatic POS Tagging

In Figure 3, we present the POS tagging performance of the BERT-based models and the `spaCy`

[7]We use the `JiWER` Python package for all error rate calculations, Apache-2.0 license.

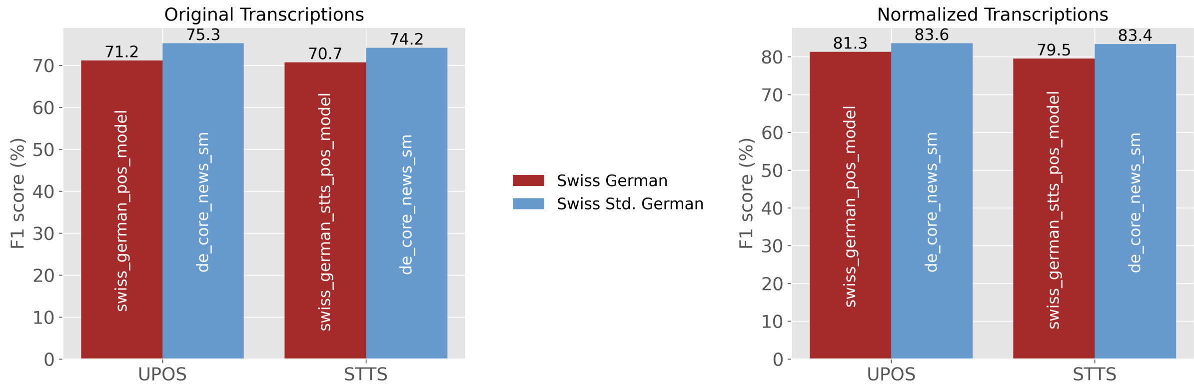


Figure 3: POS tagging results with BERT-based POS tagging model and `spaCy` model, measured on all transcription data for Swiss German and Swiss Standard German.

model, evaluated on the complete transcription datasets for both Swiss German and Swiss Standard German. All models achieved F1 scores above 70, with significantly higher results observed on normalized transcriptions (above or nearly 80) compared to original transcriptions.

6. Discussion

6.1. Challenges of Manual and Automatic LSA

LSA poses many challenges despite being a gold standard tool in speech therapy (Klatte et al., 2022; Bawayan et al., 2022), which have to be taken into account in the manual use as well as in the development of a semi-automatic pipeline for LSA. Factoring them in from the first steps of development as well as searching for solutions is part of ongoing and future research.

Data collection. Speech and language samples of children are typically recorded in naturalistic environments, such as in free conversations, which include both relevant speech and background noise like TV soundtracks (Lüdtke et al., 2023), or, in our case, noise of other children playing and street noise in the background. Extracting meaningful speech utterances and transcribing them into text units which speech-language pathologists can directly work with, is often a challenging task.

Manual annotation. Linguistic features such as errors on the levels of syntax, morphology, semantics, or phonology are often manually annotated and analysed by speech-language pathologists, which due to its time-intensive nature makes the use in therapeutic assessment less common (Owens Jr et al., 2018).

Generalization difficulty. Language error patterns (such as wrong word order) vary from one child to another and from dialect to dialect, which limits the generalizability of LSA methods.

Use of LLMs In recent years, large language models (LLMs) have undergone significant advancements. However, querying commercial LLMs such as ChatGPT with research data, including language samples from children, is not compliant with Swiss data protection regulations. Furthermore, the use of commercial LLMs for language sample analysis (LSA) is neither ethical nor practical due to the highly sensitive nature of children's speech transcriptions and the risk of data contamination. As a result, progress in automating LSA for the diagnosis of developmental language disorder (DLD) remains limited and understudied.

Swiss German The linguistic landscape of Switzerland presents unique challenges for applying NLP methods (Parida et al., 2020). The prevalence of Swiss German dialects, which differ significantly from Standard German in their linguistic structure, complicates real-world NLP practice. Automatic speech recognition (ASR) on Swiss German typically produces Standard German transcriptions, thereby omitting or distorting important dialectal information or not being able to represent Swiss German faithfully, for example sentences such as “Ich gang go poste.” (*I'm going shopping.*) can be translated into “Ich gehe einkaufen.”, losing the “go”, which does not exist in German but is essential in Swiss German. Models trained on written Standard German typically perform poorly on Swiss German due to the lack of a standardized written form (Kew et al., 2020; Nigmatulina et al., 2020). Despite this, written Swiss German is increasingly used in digital communication, such as social media and online messaging, where individual writing styles introduce further variation

(Hollenstein and Aepli, 2014). In educational settings, children use Swiss Standard German. Swiss Standard German is a variety of Standard German but still contains words ("helvetisms") and grammar rules that render it different from Standard German of Germany or Austria. Many children who do not speak Swiss German as their native language primarily, or exclusively, communicate in Swiss Standard German. This highlights the necessity of NLP methods that can effectively process both Swiss Standard German and Swiss German while accommodating the inherent variation within Swiss German dialects.

Research in other languages such as English supports the effectiveness of LSA in diagnosing DLD (Ramos et al., 2022) and the effectiveness of using automated LSA (Miller et al., 1985; Pye, 1994). However, relevant studies are less present for German data and entirely missing for Swiss German. Our preliminary study investigates the potential of using NLP methods as a step closer to working with Swiss German data.

6.2. ASR Transcriptions

The majority of transcription errors committed by the Whisper model can be attributed to child speech. Specifically, Whisper often failed to recognize the children's utterances (but not the utterances of the adult therapists) or generated repeated words. After orthographically correcting the manual transcriptions (i.e., normalization; for examples, see columns **word** (original) and **normalized** in Appendix E), the error rates were significantly reduced (see Figure 2b). This highlights the continued need for normalization of speech transcriptions in practice to address the limitations of ASR models. This problem could also potentially be mitigated by fine-tuning the models with specific methods of transcription, which would allow keeping the important transcribed information as well as reaching a sufficient quality of transcription.

In our case study, indicated by the less prominent results of the Whisper transcriptions, fully relying on ASR transcriptions was not meaningful at this stage due to the limited availability of children's speech data in Swiss German, which precluded fine-tuning even the small variant of the Whisper model. This highlights the importance of both naturally expanding the dataset and exploring alternative approaches, such as data augmentation via speech synthesis using generative models (Ren et al., 2021; Ao et al., 2022; Toyin et al., 2024), to address the challenges of processing underrepresented and atypical languages in a speech-language pathology context.

6.3. Inter-annotator Agreement on Human Labelling of POS Tags

Our results indicate a strong level of agreement, demonstrating the reliability of our annotation process. Prior research on IAA for German and Latin POS tagging (Brants, 2000; Stüssi and Ströbel, 2024) has reported even higher IAA scores with professionally trained annotators, which suggests that the annotation quality could be further improved with expert training.

We argue that three primary factors contribute to the differences observed in our study: (1) While no major discrepancies exist, Swiss Standard German differs from the Standard German used in Germany, which can lead to disagreements among annotators; (2) Swiss German lacks standardized spelling, and our annotators originate from different Swiss German-speaking cantons, resulting in variations in their interpretation of syntactic functions; and (3) Due to the presence of erroneous, incomplete, or atypical sentence structures produced by children with DLD, some ambiguity remained, leading to variability in interpretation among annotators. In Appendix B, we provide examples of common discrepancies between our annotators.

6.4. Automatic POS Tagging

The results for the more general statistical `spaCy` model on Swiss Standard German data are consistently outperforming the two BERT-based POS tagging models specifically trained on UPOS and STTS for Swiss German annotations on the Swiss German data, further highlighting the difficulties specific to Swiss German in NLP methods. All in all, these findings highlight that the models used in this study can perform reasonably well on Swiss German and Swiss Standard German transcriptions while offering the advantages of being significantly less computationally expensive and more ethical in their deployment compared to commercial LLMs. More Swiss German training data in general, as well as finetuning the model to the specifics of our dataset (e.g. children's' speech and atypical speech) is expected to improve the performance.

6.5. Current Challenges

Despite achieving a reasonably good performance in automatic speech transcription and POS tagging, it is important to emphasize that the current results remain insufficient for therapeutic practice. This limitation is primarily due to the persistent need for manual correction of speech transcriptions. Since diagnostic applications demand highly precise speech transcription and linguistic analysis, there is strong motivation to further enhance NLP approaches not based on commercial LLMs, such as by fine-tuning existing BERT-based POS tagging models with additional data and developing more advanced ASR models for Swiss German speech.

In Appendix D, we present our prototype software developed specifically for speech-language pathologists. As we continue to gather user feedback from real-world practice, we adhere to human-in-the-loop design principles and plan to enhance the software with more rigorously validated features for better user experience.

7. Conclusion and Future Work

In this study, we have addressed the challenges of automating key steps in language sample analysis by employing non-commercial NLP models for ASR and POS tagging. We evaluated the zero-shot capabilities of these models on both tasks and reported the empirical performance. Our work underscores the feasibility of employing ethical NLP approaches not based on commercial LLMs in the setting of speech-language pathology, particularly when handling sensitive data such as children's speech. In future work, we aim to expand the dataset by incorporating more diverse samples of children's speech with and without DLD. Additionally, we plan to develop specialized ASR and POS tagging models tailored to Swiss German of children and evaluate whether analyses based on automatically created transcripts and automated POS tagging can reliably predict DLD in Swiss German-speaking children. We also plan to expand the linguistic analyses in broader contexts, especially on the syntactic level (such as dependency parsing, see the column **dependency** in Table 20 in Appendix E), ultimately facilitating the semi-automatic diagnosis of DLD for children in Switzerland.

8. Limitations

The primary limitations of our work are as follows: (1) Due to the data sparsity and difficulty of data acquisition, our sample size is relatively low compared to evaluations in contexts other than the speech and language disorder context; (2) We did not conduct further investigations into morphological features, as, to the best of our knowledge, no existing NLP approaches not based on commercial LLMs for morphological prediction in Swiss German are currently available; (3) We did not benchmark the evaluation with locally deployed LLMs of the newest generation as this will be part of a future study; (4) All Swiss German ASR models known to us transcribe the text directly into Standard German. However, for our application, it would be beneficial for the text to be in Swiss German. There is not yet any research showing that DLD approaches for other languages can be effectively applied to diagnosing DLD of Swiss German-speaking children. For this reason, the usefulness of the Standard German transcriptions obtained is limited; (5) Automatic tools for diagnosing DLD in children are not infallible; they should only be used in combination with other screening methods and by speech-language pathologists.

9. Ethical Statement

Our study received ethical approval from the responsible university's research committee. Informed consent was obtained from the parents of all participating children and the speech-language pathologists, allowing for the recording, processing, storage, and controlled sharing of data. To ensure accessibility, the consent form was provided in simplified language to facilitate understanding for parents. Children were informed about the study and provided their oral consent to participate. None of the collected data have been processed through any commercial LLMs. All data processing was conducted locally. Permission was granted to publish the data from 41 recordings in a public repository. In addition to the small subset used for this paper, the whole dataset will be released in the near future.

10. Acknowledgement

This work was supported by the Foundation Special Education Centre Fribourg (Stiftung Heilpädagogisches Zentrum Fribourg) and the Foundation for Speech Therapy in the Canton of Zurich (Förderstiftung für das Sprachheilwesen im Kanton Zürich). We thank Susanne Kempe-Preti, Pascale Schaller, Julia Winkes, Sonja Schäli, Lena Graf, Ramona Rüegg, Lukas Fischer, Anne Göhring, Michelle Wastl and Angela Heldstab for their valuable input to the study.

11. Bibliographical References

Noëmi Aepli and Rico Sennrich. 2022. Improving Zero-shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-level Noise. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4074–4083.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing

Li, Yu Zhang, et al. 2022. SpeechT5: Unified-modal Encoder-decoder Pre-training for Spoken Language Processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738.

Yuriy Arabskyy, Aashish Agarwal, Subhadeep Dey, and Oscar Koller. 2021. Dialectal Speech Recognition and Translation of Swiss German Speech to Standard German Text: Microsoft's Submission to SwissText 2021. *arXiv preprint arXiv:2106.08126*.

Rebecca Bawayan, Jennifer A Brown, Rebecca Bawayan, and Jennifer A Brown. 2022. Language Sample Analysis Consideration and Use: A Survey of School-based Speech Language Pathologists. *Clinical Archives of Communication Disorders*, 7(1):15–28.

Vivek Bhardwaj, Mohamed Tahar Ben Othman, Vinay Kukreja, Youcef Belkhier, Mohit Bajaj, B Srikanth Goud, Ateeq Ur Rehman, Muhammad Shafiq, and Habib Hamam. 2022. Automatic Speech Recognition (ASR) Systems for Children: A Systematic Literature Review. *Applied Sciences*, 12(9):4419.

Dorothy VM Bishop. 2006. What Causes Specific Language Impairment in Children? *Current directions in psychological science*, 15(5):217–221.

Bernd Bohnet, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, and Joshua Maynez. 2018. Morphosyntactic Tagging with a Meta-BiLSTM Model over Context Sensitive Token Encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2642–2652.

Thorsten Brants. 2000. Inter-annotator Agreement for a German Newspaper Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*.

Burcu Can. 2011. *Statistical Models for Unsupervised Learning of Morphology and POS Tagging*. Ph.D. thesis, University of York.

Jacob Cohen. 1968. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological bulletin*, 70(4):213.

Eugen Dieth. 1986. *Schwyzertütschi Dialäktschrift*. Cornelsen Schweiz AG.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. Swissdial: Parallel Multidialectal Corpus of Spoken Swiss German. *arXiv preprint arXiv:2103.11401*.

Eyal Liron Dolev, Clemens Fidel Lutz, and Noëmi Aepli. 2024. Does Whisper Understand Swiss German? An Automatic, Qualitative, and Human Evaluation. *arXiv preprint arXiv:2404.19310*.

Keyur Gabani, Thamar Solorio, Yang Liu, Khairun-nisa Hassanali, and Christine A Dollaghan. 2011. Exploring A Corpus-based Approach for Detecting Language Impairment in Monolingual English-speaking Children. *Artificial Intelligence in Medicine*, 53(3):161–170.

John F Gallagher and Jill R Hoover. 2020. Measure What You Treat: Using Language Sample Analysis for Grammatical Outcome Measures in Children with Developmental Language Disorder. *Perspectives of the ASHA Special Interest Groups*, 5(2):350–363.

Khairun-nisa Hassanali, Yang Liu, and Thamar Solorio. 2012. Evaluating NLP Features for Automatic Prediction of Language Impairment Using Child Speech Transcripts. In *INTERSPEECH*, pages 1339–1342.

Nora Hollenstein and Noëmi Aepli. 2014. Compilation of a Swiss German Dialect Corpus and Its Application to POS Tagging. In *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects*, pages 85–94.

Rishabh Jain, Andrei Barcovschi, Mariam Yiwere, Peter Corcoran, and Horia Cucu. 2023. Adaptation of Whisper Models to Child Speech Recognition. *arXiv preprint arXiv:2307.13008*.

Tannon Kew, Iuliia Nigmatulina, Lorenz Nagele, and Tanja Samardzic. 2020. UZH TILT: A Kaldi Recipe for Swiss German Speech to Standard German Text. In *SwissText/KONVENS*.

Inge S Klatte, Vera Van Heugten, Rob Zwitserlood, and Ellen Gerrits. 2022. Language Sample Analysis in Clinical Practice: Speech-language Pathologists' Barriers, Facilitators, and Needs. *Language, Speech, and Hearing Services in Schools*, 53(1):1–16.

Matej Klemen, Luka Krsnik, and Marko Robnik-Šikonja. 2023. Enhancing Deep Neural Networks with Morphological Information. *Natural Language Engineering*, 29(2):360–385.

Dancheng Liu, Jason Yang, Ishan Albrecht-Buehler, Helen Qin, Sophie Li, Yuting Hu, Amir Nassereldine, and Jinjun Xiong. 2024. Automatic

Screening for Children with Speech Disorder using Automatic Speech Recognition: Opportunities and Challenges. In *Proceedings of the AAAI Symposium Series*, volume 4, pages 308–313.

Ulrike Lüdtke, Juan Bornman, Febe De Wet, Ulrich Heid, Jörn Ostermann, Lars Rumberg, Jeannie Van der Linde, and Hanna Ehlert. 2023. Multidisciplinary Perspectives on Automatic Analysis of Children's Language Samples: Where Do We Go From Here? *Folia Phoniatrica et Logopaedica*, 75(1):1–12.

Carina Lüke, Christina Kauschke, Andrea Dohmen, Andrea Haid, Christina Leitinger, Claudia Männel, Tanja Penz, Steffi Sachse, Wiebke Scharff Rethfeldt, Julia Spranger, et al. 2023. Definition and Terminology of Developmental Language Disorders—interdisciplinary Consensus Across German-speaking Countries. *Plos one*, 18(11):e0293736.

Jon Miller, Robin Chapman, et al. 1985. Systematic Analysis of Language Transcripts. *Madison, WI: Language Analysis Laboratory*.

Iuliia Nigmatulina, Tannon Kew, and Tanja Samardzic. 2020. ASR for Non-standardised Languages with Dialectal Variation: the Case of Swiss German. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 15–24.

Robert E Owens Jr, Stacey L Pavelko, and Dennis Bambinelli. 2018. Moving Beyond Mean Length of Utterance: Analyzing Language Samples to Identify Intervention Targets. *Perspectives of the ASHA Special Interest Groups*, 3(1):5–22.

Shantipriya Parida, Esaú Villatoro-Tello, Sajit Kumar, Petr Motlicek, and Qingran Zhan. 2020. Idiap Submission to Swiss-German Language Detection Shared Task. In *SwissText/KONVENS*, page 7.

Michel Plüss, Jan Milan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Huerlimann, Tanja Samardzic, Manfred Vogel, et al. 2023. STT4SG-350: A Speech Corpus for All Swiss German Dialect Regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772.

Michel Plüss, Manuela Huerlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, et al. 2022. DS-200: A Swiss German Speech to Standard German Text Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256.

Michel Plüss, Lukas Neukom, Christian Scheller, and Manfred Vogel. 2020. Swiss Parliaments Corpus, An Automatically Aligned Swiss German Speech to Standard German Text Corpus. *arXiv preprint arXiv:2010.02810*.

Niclas Pokel, Pehuén Moure, Roman Boehringer, and Yingqiang Gao. 2025. Adapting Foundation Speech Recognition Models to Impaired Speech: A Semantic Re-chaining Approach for Personalization of German Speech. In *12th edition of the Disfluency in Spontaneous Speech Workshop (DiSS 2025)*, pages 82–86.

Alexandros Potamianos, Shrikanth S Narayanan, and Sungbok Lee. 1997. Automatic Speech Recognition for Children. In *Eurospeech*, volume 97, pages 2371–2374.

Clifton Pye. 1994. The CHILDES Project: Tools for Analyzing Talk.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-scale Weak Supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Michelle N Ramos, Penelope Collins, and Elizabeth D Peña. 2022. Sharpening Our Tools: A Systematic Review to Identify Diagnostically Accurate Language Sample Measures. *Journal of Speech, Language, and Hearing Research*, 65(10):3890–3907.

Michelle Nichols Ramos. 2024. *Using Language Sample Analysis to Identify Developmental Language Disorder in Bilingual Children*. Ph.D. thesis, University of California, Irvine.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. FastSpeech 2: Fast and High-quality End-to-end Text to Speech. In *International Conference on Learning Representations*.

Lars Rumberg, Hanna Ehlert, Ulrike Lüdtke, and Jörn Ostermann. 2021. Age-Invariant Training for End-to-end Child Speech Recognition Using Adversarial Multi-task Learning. In *Interspeech*, pages 3850–3854.

Alessandra Sansavini, Maria Elena Favilla, Maria Teresa Guasti, Andrea Marini, Stefania Millepiedi, Maria Valeria Di Martino, Simona Vecchi, Nadia Battajon, Laura Bertolo, Olga Capirci, Barbara Carretti, Maria Paola Colatei, Cristina Frioni, Luigi Marotta, Sara Massa, Letizia Michelazzo, Chiara Pecini, Silvia Piazzalunga, Manuela Pieretti, Pasquale Rinaldi, Renata Salvadorini, Cristiano Termine, Mariagrazia Zuccarini, Simonetta D'Amico, Anna Giulia De Cagno, Maria Chiara Levorato, Tiziana Rossetto, and Maria Luisa Lorusso. 2021. Developmental Language Disorder: Early Predictors, Age for the Diagnosis, and Diagnostic Tools. A Scoping Review. *Brain Sciences*, 11(5).

Yanick Schraner, Christian Scheller, Michel Plüss, and Manfred Vogel. 2022. Swiss German Speech to Text System Evaluation. *arXiv preprint arXiv:2207.00412*.

Miikka Silfverberg and Krister Lindén. 2011. Combining Statistical Models for POS Tagging using Finite-state Calculus. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 183–190.

Daniel V Smith, Alex Sneddon, Lauren Ward, Andreas Duenser, Jill Freyne, David Silvera-Tawil, and Angela Morgan. 2017. Improving Child Speech Disorder Assessment by Incorporating Out-of-domain Adult Speech. In *Interspeech*, pages 2690–2694.

Thamar Solorio. 2013. Survey on Emerging Research on the Use of Natural Language Processing in Clinical Language Assessment of Children. *Language and Linguistics Compass*, 7(12):633–646.

Samuel Stucki, Mark Cieliebak, and Jan Deriu. 2025. SwissGPC v1. 0–The Swiss German Podcasts Corpus. *arXiv preprint arXiv:2509.19866*.

Elina Stüssi and Phillip Ströbel. 2024. Part-of-speech Tagging of 16th-Century Latin with GPT. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 196–206.

Vincenzo Timmel, Claudio Paonessa, Reza Kakooee, Manfred Vogel, and Daniel Perruchoud. 2024. Fine-tuning Whisper on Low-Resource Languages for Real-World Applications. *arXiv preprint arXiv:2412.15726*.

Alexander Tkachenko and Kairit Sirts. 2018. Modeling Composite Labels for Neural Morphological Tagging. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 368–379.

J Bruce Tomblin, Nancy L Records, and Xuyang Zhang. 1996. A System for the Diagnosis of Specific Language Impairment in Kindergarten Children. *Journal of Speech, Language, and Hearing Research*, 39(6):1284–1294.

Hawau Toyin, Hao Li, and Hanan Aldarmaki. 2024. STTATTS: Unified Speech-to-text And Text-to-speech Model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6853–6863.

Vivian van Wijngaarden, Hester de Wilde, Dieuwke Mink van der Molen, Jildo Petter, Inge Stegeman, Ellen Gerrits, Adriana L Smit, and Marie-José van den Boogaard. 2024. Genetic Outcomes in Children with Developmental Language Disorder: A Systematic Review. *Frontiers in pediatrics*, 12:1315229.

A. Instructions for Human Annotation of Part-of-speech Tagging on Gold Transcriptions (translated from the original German file)

Task Description. Your task is to annotate a small dataset of spontaneous speech sentences of children with typical and atypical language development in Swiss German and Swiss Standard German. You received an Excel sheet with around 100 sentences from different children (K as *Kind* (child) in German) and professionals (FP as *Fachperson* (specialist) in German) interviewing the children in either Swiss German or Swiss Standard German. These sentences are randomly selected from different gold transcription files. In the first column you find the human transcriptions, while all other columns are generated automatically and need to be corrected. Please correct the annotations of:

- Part-of-speech tagging with the UPOS tags (more information here[1]).
- Part-of-speech tagging with the STTS tags (more information here[2]).
- Morphology for certain part-of-speech tags.
- Subject-verb agreement (SVA).

Annotation Process. Please adhere to the following steps:

1. Read this manual carefully.
2. Read the documentation of the tag sets.
3. Perform trial annotations using the control sentences constructed individually and send them back.
4. If you pass the trial annotations, you will receive real annotation sentences.
5. During the annotation, please keep in mind that:
    - For the annotation you can use whatever assistance you want (such as look up tables, Duden, the internet, etc.);
    - Please **do not use the `spaCy` model `de_core_news_sm` as well as the `swiss_german_stts_pos_model` and the `swiss_german_pos_model` models from Huggingface**, as they are baselines of our study;
    - Please **do not copy and paste the sentences into commercial LLMs**, as we must respect the data protection policy of Switzerland;
    - The sentences are random samples of different transcriptions. Please use only the current sentence as context;
    - You can annotate in whatever order you like.
6. For morphology, annotate as much as you can as long as it is determinable. If something is not determinable, please leave it out blank.
7. For subject-verb agreement:
    - Please mark all conjugated verbs with 'v';
    - Please mark the main part that determines the verb form (subjects) as 'sb';
    - For the contracted forms of Swiss German (such as "gehen wir" → "gömmer", "kann er sie entsorgen" → "chanerse entsorge"), use sb_v or v_sb, depending on the order in the contraction.
8. Leave `<sentence>`, the sentence separator, as empty.
9. Tag all `UNK` as X/XY.
10. Tag all `NAME` (anonymized name) as PROPN/NE.
11. Use the tag PROAV instead of PAV (adhering to `spaCy`).

[1] https://universaldependencies.org/u/pos/

[2] https://homepage.ruhr-uni-bochum.de/stephen.berman/Korpuslinguistik/Tagsets-STTS.html

12. In STTS, the verb `sein` is always tagged as VAxxx, even when used as a full verb. However, UPOS distinguishes between the functions of the verb: when a verb is used as a full verb, can can thus be replaced by `sich befinden`, it is tagged as VERB.

13. The STTS tag set labels all occurrences of auxiliary verbs (`sein, werden, haben`) as well as copula with VAxxx, modal verbs (`müssen, dürfen, wollen, mögen, können, sollen`) are labelled with VMxxx.

14. Ja/Nein: Please annotate as PTKANT if

    - Standing alone;
    - Is part of an answer;
    - Used as a query (Rückfrage).

    Except when used as modal particle (in German "Abtönungspartikel", for instance, "Das ist ja wirklich schön"), annotate as ADV.

15. Interjections encompass:

    - All signals of understanding (in German "Verständigungssignale", for instances, "oh', "ah", "aha", "wow", "hmm");
    - All onomatopoeias (for instances, "brummbrumm", "gluglug", "miau", "wuff").

16. Corrections, word fragments, errors (which are then corrected, for instance, "eine Bri-Brille"), please annotate as X/XY.

17. If a child made a grammatical error and used the wrong form, **please annotate what was exactly said and not what it should be**, as we want to identify these errors later. If not all information can be clearly determined from the used word itself, assume the correct form.

18. For cases that are hard to tag (e.g., the child used irregular forms, wrong or fantasy words, etc.), please do your best.

Rules of Swiss German. Most rules are identical between Swiss German and Standard German. However, there are still some linguistic differences. Please read the chapter *Differences to German UD Guidelines* here[3] as a reference.

Conversion between UPOS tag set and STTS tag set. See more information here[4]. Please be aware that the conversion does not apply to all cases.

[3] `https://universaldependencies.org/gsw/`

[4] `https://universaldependencies.org/tagset-conversion/de-stts-uposf.html`

B. Canonical Examples of POS Tagging Disagreements between Human Annotators

B.1. Examples in Swiss German

Disagreement 1 Whether a word is a proper noun (PROPN), noun (NOUN), or foreign word (X).

	ebe	genau	ja	de	het	der	paw	patrol	gfalle	?
A	ADV	ADV	PART	ADV	AUX	PRON	*X*	*X*	VERB	PUNCT
B	ADV	ADV	PART	ADV	AUX	PRON	*PROPN*	*PROPN*	VERB	PUNCT
C	ADV	ADV	PART	ADV	AUX	PRON	*PROPN*	*PROPN*	VERB	PUNCT

Table 7: Example sentence corresponding to the English translation *so exactly then you liked paw patrol?*

Disagreement 2 Whether a word is an adverb (ADV) or, e.g., an adjective (ADJD) or interjection (ITJ).

	ja	hm	genau	ah	schpannend
A	PTKANT	ITJ	*ADJD*	ITJ	ADJD
B	PTKANT	ITJ	*ADV*	ITJ	ADJD
C	PTKANT	ITJ	*ADV*	ITJ	ADJD

Table 8: Example sentence corresponding to the English translation *yes hm exactly ah interesting*.

	gäll	was	dänksch	was	isch	das
A	*ITJ*	PWS	VVFIN+	PWS	VAFIN	PDS
B	*ADV*	PWS	VVFIN	PWS	VAFIN	PDS
C	*ITJ*	PWS	VVFIN+	PWS	VAFIN	PDS

Table 9: Example sentence corresponding to the English translation *what do you think what this is?*

Disagreement 3 Distinguishing different types of pronouns (e.g., PDS, PDAT).

	gend	recht	gas	uf	dene	trotti
A	VVFIN	ADV	NN	APPR	*PDS*	NN
B	VVFIN	ADV	PTKVZ	APPR	*PDAT*	NN
C	VVFIN	ADV	NN	APPR	*PDS*	NN

Table 10: Example sentence corresponding to the English translation *they really step on the gas on these scooters*

Disagreement 4 Whether to tag attributive pronouns (PPOSAT in STTS) as determiner (DET) or pronoun (PRON) in UPOS.

	oh	verzell	mal	wy	fyrsch	dù[5]	din	gebùrtstag	?
A	INTJ	VERB	ADV	ADV	VERB	PRON	*DET*	NOUN	PUNCT
B	INTJ	VERB	ADV	CCONJ	VERB	PRON	*PRON*	NOUN	PUNCT
C	INTJ	VERB	ADV	ADV	VERB	PRON	*DET*	NOUN	PUNCT
STTS	ITJ	VVIMP	ADV	PWAV	VVFIN	PPER	*PPOSAT*	NN	$.

Table 11: Example sentence corresponding to the English translation *oh tell me how do you celebrate your birthday?*

[5] Adding an accent symbol to a vowel describes its quality adapted from Dieth (1986).

Disagreement 5 Whether a word is a concatenation of multiple words (e.g., VVFIN+) or not (e.g., VVFIN) (especially in the case of a verb in second person singular, as it can stand without PPER in Swiss German).

	gäll	was	dänksch	was	isch	das
A	ITJ	PWS	*VVFIN+*	PWS	VAFIN	PDS
B	ADV	PWS	*VVFIN*	PWS	VAFIN	PDS
C	ITJ	PWS	*VVFIN+*	PWS	VAFIN	PDS

Table 12: Example sentence corresponding to the English translation *what do you think this is?*

B.2. Examples in Swiss Standard German

Disagreement 1 Whether a word is nominalized (used as a noun) or not.

	aber	du	hast	recht	das	macht	man	doch	eigentlich	mit	einem	stock
A	CCONJ	PRON	AUX	*NOUN*	PRON	VERB	PRON	ADV	ADV	ADP	DET	NOUN
B	CCONJ	PRON	AUX	*NOUN*	PRON	VERB	PRON	ADV	ADV	ADP	DET	NOUN
C	CCONJ	PRON	AUX	*ADV*	PRON	VERB	PRON	ADV	ADV	ADP	DET	NOUN

Table 13: Example sentence corresponding to the English translation *but you are right actually you do this with a stick*.

Disagreement 2 For incomplete or erroneous words, what is the right way to interpret it (the correct word in the first example should be "meinte" ("meant") as one word, and in the second example, "warte" ("wait").

	mein	te	an
A	*PPOSAT*	*NN*	PTKVZ
B	*PPOSAT*	*XY*	XY
C	*VVFIN*	*VVFIN*	APPR

(a) English translation of the example sentence: *mean-ed on*.

	wate	an	freitag
A	*NN*	APPR	NN
B	*XY*	APPR	NN
C	*VVIMP*	APPR	NN

(b) Example sentence corresponding to the English translation *wait on friday*.

Table 14: Comparative morphological annotation examples from two German utterances.

Disagreement 3 Whether a verb is used as auxiliary (VA) or full verb (VV) (differences exist between the two tag sets).

	ja	diese	zeitung	ich	hab	das	gerne	aber	ich	hat	ein	zeichnung
A (UPOS)	PART	DET	NOUN	PRON	VERB	PRON	ADV	CCONJ	PRON	*AUX*	DET	NOUN
A (STTS)	PTKANT	PDAT	NN	PPER	VAFIN	PDS	ADV	KON	PPER	*VAFIN*	ART	NN
B (UPOS)	PART	DET	NOUN	PRON	VERB	PRON	ADV	CCONJ	PRON	*AUX*	DET	NOUN
B (STTS)	PTKANT	PDAT	NN	PPER	VAFIN	PDS	ADJD	KON	PPER	*VAFIN*	ART	NN
C (UPOS)	PART	DET	NOUN	PRON	VERB	PRON	ADV	CCONJ	PRON	*VERB*	DET	NOUN
C (STTS)	PTKANT	PDAT	NN	PPER	VAFIN	PDS	ADV	KON	PPER	*VAFIN*	ART	NN

Table 15: Example sentence corresponding to the English translation *yes this newspaper I like this but I has a drawing*.

C. Per-speaker-group ASR and POS Tagging Results

Observation 1 Whisper has significantly less ASR errors on transcriptions of speech-language pathologists than of children.

Original Transcripts	**WER** $\downarrow$		**CER** $\downarrow$		**MER** $\downarrow$		**WIL** $\downarrow$	
	FP	K	FP	K	FP	K	FP	K
Swiss German	0.772	0.860	0.469	0.556	0.761	0.850	0.927	0.970
Swiss Std. German	0.377	0.600	0.278	0.445	0.370	0.591	0.463	0.722

Table 16: ASR results evaluated on **original** speech transcriptions of speech-language pathologists (FP) and children (K).

Normalized Transcripts	**WER** $\downarrow$		**CER** $\downarrow$		**MER** $\downarrow$		**WIL** $\downarrow$	
	FP	K	FP	K	FP	K	FP	K
Swiss German	0.506	0.681	0.337	0.466	0.494	0.664	0.666	0.840
Swiss Std. German	0.365	0.550	0.275	0.430	0.359	0.541	0.444	0.651

Table 17: ASR results evaluated on **normalized** speech transcriptions of speech-language pathologists (FP) and children (K).

Observation 2 POS tagging models have higher F1 scores on transcriptions of speech-language pathologists (i.e. adults) than of children.

Original Transcriptions	**UPOS** (**F1 Score** $\uparrow$)		**STTS** (**F1 Score** $\uparrow$)	
	FP	K	FP	K
Swiss German	0.735	0.673	0.728	0.675
Swiss Std. German	0.844	0.659	0.831	0.654

Table 18: POS tagging results on **original** speech transcriptions of speech-language pathologists (FP) and children (K).

Observation 3 Swiss Standard German transcriptions achieved generally higher evaluation results than Swiss German transcriptions.

Normalized Transcriptions	**UPOS** (**F1 Score** $\uparrow$)		**STTS** (**F1 Score** $\uparrow$)	
	FP	K	FP	K
Swiss German	0.835	0.787	0.823	0.762
Swiss Std. German	0.878	0.796	0.852	0.807

Table 19: POS tagging results on **normalized** speech transcriptions of speech-language pathologists (FP) and children (K).

Observation 4 Orthographic normalization of speech transcriptions helps in boosting the performance of both ASR model and POS tagging model.

D. The LSA Software for Therapeutic Practice of DLD Diagnosis

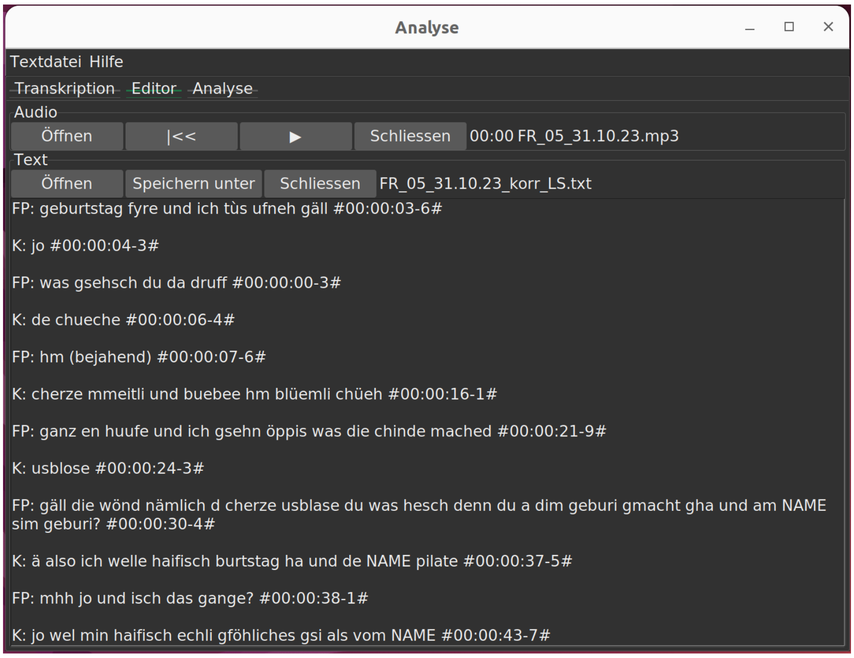


Figure 4: Editor of the software. After automatically transcribing the recordings, the transcript is opened in the editor. Here, the recordings can be played and the transcript can be corrected.

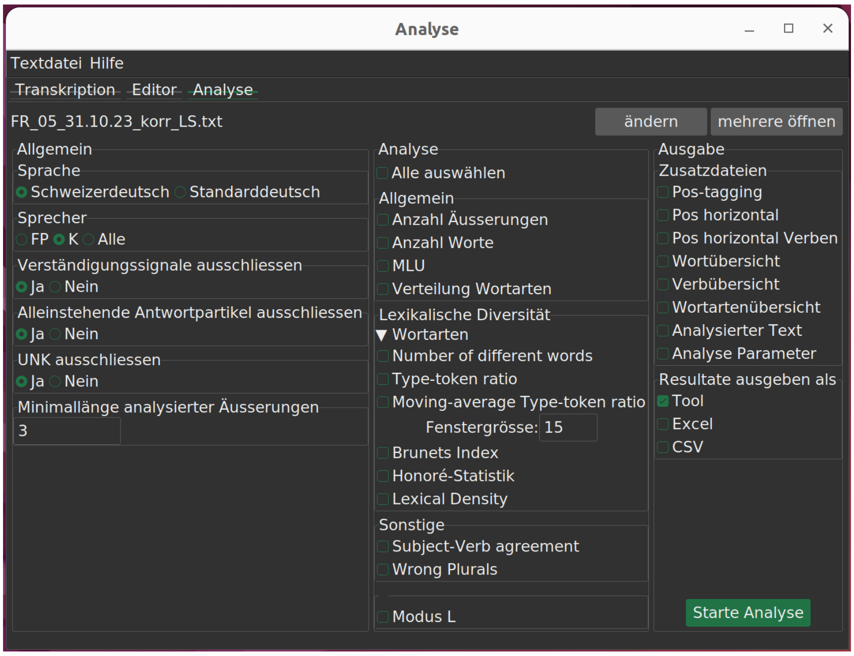


Figure 5: Analysis: After correcting the automatic transcription manually, the analysis can be started. Providing different options for the analysis, such as which speakers should be analyzed and what should be filtered out, a personalized analysis can be executed, containing values such as mean length of utterance, distribution of POS tags, and subject verb agreement as well as additional files with overviews of all verbs and POS tagging.

E. Annotation Examples

send_id	speaker	word_id	word	normalized	lemma	UPOS tag	STTS tag	morphology	SVA	dependency
62	FP	0	warst	Warst	sein	VERB	VAFIN	{‘Mood’: ‘Ind’, ‘Number’: ‘Sing’, ‘Person’: ‘2’, ‘Tense’: ‘Past’, ‘VerForm’: ‘Fin’}	v	ROOT
62	FP	1	du	du	du	PRON	PPER	{‘Case’: ‘Norm’, ‘Number’: ‘Sing’, ‘Person’: ‘2’, ‘PronType’: ‘Prs’}	sb	sb
62	FP	2	diesen	diesen	dieser	DET	PDAT	{‘Case’: ‘Acc’, ‘Gender’: ‘Masc’, ‘Number’: ‘Sing’, ‘PronType’: ‘Dem’}		nk
62	FP	3	sommer	Sommer	Sommer	NOUN	NN	{‘Case’: ‘Acc’, ‘Gender’: ‘Masc’, ‘Number’: ‘Sing’}		oa
62	FP	4	auch	auch	auch	ADV	ADV			mo
62	FP	5	schon	schon	schon	ADV	ADV			mo
62	FP	6	in	in	in	ADP	APPR			mo
62	FP	7	der	der	der	DET	ART	{‘Case’: ‘Dat’, ‘Definite’: ‘Def’, ‘Gender’: ‘Fem’, ‘Number’: ‘Sing’, ‘PronType’: ‘Art’}		nk
62	FP	8	badi	Badi	Badi	NOUN	NN	{‘Case’: ‘Dat’, ‘Gender’: ‘Fem’, ‘Number’: ‘Sing’}		oa
62	FP	9	?	?	?	PUNCT	$.			punct
63	K	0	ähm	Ähm	ähm	INTJ	ITJ			mo
63	K	1	ja	ja	ja	PART	PTKANT			mo
63	K	2	abeR	aber	aber	CCONJ	KON			mo
63	K	3	is	ist	sein	AUX	VAFIN	{‘Mood’: ‘Ind’, ‘Number’: ‘Sing’, ‘Person’: ‘3’, ‘Tense’: ‘Pres’, ‘VerbForm’: ‘Fin’}	v	sb
63	K	4	de	der	der	PRON	PDS	{‘Case’: ‘Nom’, ‘Gender’: ‘Masc’, ‘Number’: ‘Sing’, ‘PronType’: ‘Dem’}	sb	uc
63	K	5	mine	meine	mein	PRON	PPOSAT	{‘Case’: ‘Nom’, ‘Gender’: ‘Neut’, ‘Number’: ‘Sing’, ‘Poss’: ‘Yes’, ‘PronType’: ‘Prs’}		nk
63	K	6	haus	Haus	Haus	NOUN	NN	{‘Case’: ‘Nom’, ‘Gender’: ‘Neut’, ‘Number’: ‘Sing’}		ams
63	K	7	zün	zün	zün	X	XY			pd
63	K	8	ist	ist	sein	VERB	VAFIN	{‘Mood’: ‘Ind’, ‘Number’: ‘Sing’, ‘Person’: ‘3’, ‘Tense’: ‘Pres’, ‘VerbForm’: ‘Fin’}	v	ROOT
63	K	9	da	da	da	ADV	ADV			mo
63	K	10	drauf	drauf	drauf	ADV	PROAV			mo
63	K	11	badi	Badi	Badi	NOUN	NN	{‘Case’: ‘Nom’, ‘Gender’: ‘Fem’, ‘Number’: ‘Sing’}	sb	oc
63	K	12	kann	kann	können	AUX	VMFIN	{‘Mood’: ‘Ind’, ‘Number’: ‘Sing’, ‘Person’: ‘3’, ‘Tense’: ‘Pres’, ‘VerbForm’: ‘Fin’}	v	sb
63	K	13	kann	kann	können	AUX	VMFIN	{‘Mood’: ‘Ind’, ‘Number’: ‘Sing’, ‘Person’: ‘3’, ‘Tense’: ‘Pres’, ‘VerbForm’: ‘Fin’}	v	sb
63	K	14	türe	Türe	Tür	NOUN	NN	{‘Case’: ‘Acc’, ‘Gender’: ‘Fem’, ‘Number’: ‘Sing’}		sb
63	K	15	offen	offen	offen	ADV	ADJD	{‘Degree’: ‘Pos’}		oc
63	K	16	und	und	und	CCONJ	KON			cd
63	K	17	da	da	da	ADV	ADV			mo
63	K	18	ist	ist	sein	VERB	VAFIN	{‘Mood’: ‘Ind’, ‘Number’: ‘Sing’, ‘Person’: ‘3’, ‘Tense’: ‘Pres’, ‘VerbForm’: ‘’}	v	cj
63	K	19	so	so	so	ADV	ADV			mo
63	K	20	bile	viele	vieler	PRON	PIAT	{‘Case’: ‘Nom’, ‘Gender’: ‘Fem’, ‘Number’: ‘Plur’, ‘PronType’: ‘Ind’}		sb
63	K	21	badi	Badi	Badi	NOUN	NN	{‘Case’: ‘Nom’, ‘Gender’: ‘Fem’, ‘Number’: ‘Sing’}	sb	pd
63	K	22	und	und	und	CCONJ	KON			cd
63	K	23	kann	kann	können	AUX	VMFIN	{‘Mood’: ‘Ind’, ‘Number’: ‘Sing’, ‘Person’: ‘3’, ‘Tense’: ‘Pres’, ‘VerbForm’: ‘Fin’}	v	cj
63	K	24	dach	Dach	Dach	NOUN	NN	{‘Case’: ‘Nom’, ‘Gender’: ‘Neut’, ‘Number’: ‘Sing’}		mo
63	K	25	da	da	da	ADV	ADV			mnr
63	K	26	badi	Badi	Badi	NOUN	NN	{‘Case’: ‘Nom’, ‘Gender’: ‘Fem’, ‘Number’: ‘Sing’}		mo
63	K	27	lauffen	laufen	laufen	VERB	VVINF	{‘VerbForm’: ‘Inf’}		oc

Table 20: Swiss Standard German language sample annotated. Words like “abeR” with R in upper case indicate the emphasizing tone.