Universität
Zürich UZH

Master's thesis
presented to the Faculty of Arts and Social Sciences
for the degree of
Master of Arts

Development of a Swiss German Language Sample Analysis Tool - Legal and Ethical Aspects of Data Collection

Author: Anja Ryser
Matriculation Number: 17-704-461

Supervisor: Dr. Sarah Ebling

Department of Computational Linguistics

Submission date: 01.06.2023

Abstract

This thesis is a contribution to the project DigiSpon whose goal is to research automated language sample analysis in Swiss German-speaking children. This thesis focuses on two aspects. First, the further development of the tool, including a graphical user interface and the extension of functionality in close collaboration with experts in speech therapy. Multiple iterations of the tool were improved integrating user feedback to adapt it to the specific needs of speech therapists. Different possibilities for development in the future were laid out as well. Second, different legal and ethical considerations concerning data collection in research and deployment of a language sample analysis tool used in speech therapy are highlighted. This includes the introduction of the most important legal regulations and ethical guidelines, followed by considerations for research in machine learning. In addition to general legal and ethical considerations, this thesis adapts the outcome specifically to the project at hand. The difficulty of adapting the general rules to practical ones used in a specific project highlights the importance of such considerations and shows that blindly following rules is not sufficient. This thesis stands at the start of the DigiSpon project and leaves many opportunities for further development in the future.

Zusammenfassung

Diese Arbeit ist ein Beitrag zum Projekt DigiSpon, dessen Ziel es ist, die automatisierte Spontansprachanalyse bei Schweizerdeutsch sprechenden Kindern zu erforschen. Diese Arbeit konzentriert sich auf zwei Aspekte. Erstens die Weiterentwicklung des Tools, einschließlich einer grafischen Benutzeroberfläche und der Erweiterung der Funktionalität in enger Zusammenarbeit mit Expert*innen der Logopädie. Mehrere Iterationen des Tools wurden durch Benutzerfeedback verbessert und auf die spezifischen Bedürfnisse von Logopäd*innen zugeschnitten. Es wurden verschiedene Entwicklungsmöglichkeiten für die Zukunft aufgezeigt. Als zweiter Aspekt wurden verschiedene ethische und rechtliche Überlegungen zur Datenerhebung in der Forschung und zur Veröffentlichung eines Spontansprachdiagnostiktools für die Sprachtherapie getätigt. Dazu gehört die Erläuterung der wichtigsten gesetzlichen Regelungen und ethischen Richtlinien, gefolgt von spezifischeren Überlegungen zur Forschung im Bereich Maschinelles Lernen. Zusätzlich zu allgemeinen rechtlichen und ethischen Überlegungen werden in dieser Arbeit die allgemeinen Überlegungen speziell an das hier beschriebene Projekt angepasst. Die Schwierigkeiten, die sich dabei zeigten, die generellen Regulierungen an ein spezifisches Projekt anzupassen, zeigt die Wichtigkeit solcher Überlegungen und zeigt auf, dass mehr benötigt wird als ein einfaches anwenden der Regulierungen. Diese Arbeit ist in der Anfangsphase des Projektes entstanden und lässt viele Möglichkeiten für die Weiterentwicklung in der Zukunft offen.

Acknowledgement

I want to thank first and foremost Dr. Sarah Ebling whose advice and care made this thesis possible. Thank you for your open ear and motivating words throughout this whole process. Thank you, Susanne Kempe, Julia Winkes and Pascale Schaller for your valuable feedback on my tool and your advice in the field of speech therapy. Thanks to my father for proofreading this thesis more than once. Thanks to my family for supporting me and for patiently listening to my rambles, even if you did not always fully understand what I was talking about.

Contents

List of Figures

List of Acronyms

ADHD	Attention Deficit/Hyperactivity Disorder
AI	Artificial Intelligence
AM	Acoustic Model
API	Application Programming Interface
ASD	Autism Spectrum Disorder
ASPA	Aachener SprachAnalyse
ASR	Automatic Speech Recognition
BESD	Behavioral, Emotional and Social Difficulties
BSD	Berufsverbandes Deutscher Soziologen
CAHAI	Committee on Artificial Intelligence from the Europe Council
CHILDES	Child Language Data Exchange System
CLAN	Computerized Language Analysis
CTC	Connectionist Temporal Classifier
DIRA	Data Integrated Reading Assessment
DLD	Developmental Language Disorder
DigiSpon	Digital unterstütze Spontansprachanalyse
DMP	Data Management Plan
DSG	Datenschutzgesetz
DSI	Digital Society Initiative
EEA	European Economic Area
GDPR	General Data Protection Directive
GUI	Graphical User Interface
HFH	Hochschule für Heilpädagogik Zurich
HRA	Human Research Act
IDG	Gesetz über die Information und den Datenschutz
IMS	Institut für Maschinelle Sprachverarbeitung (University of Stuttgart)
IQ	Intelligence Quotient
LENA	Language Environment Assessment
LSA	Language Sample Analysis

LSTM	Long Short-Term Memory recurrent network
ML	Machine Learning
MLU	Mean Utterance Length
MRI	Magnetic Resonance Imaging
NDW	Number of Different Words
NLP	Natural Language Processing
PER	Phoneme Error Rate
POS	Part of Speech
RADLD	Raising Awareness of Developmental Language Disorder
SALT	Systematic Analysis of Language Transcript
SNSF	Swiss National Science Foundation
SOV	Subject-Object-Verb
SST4SG	Speech-to-text for Swiss German
TALC	Tools for Analyzing Language and Communication
TDNN	Time Delay Neural Network
TNW	Total Number of Words
UZH	University of Zurich
WER	Word Error Rate
WFSA	Weighted Finite State Automaton
WMA	World Medical Association

1 Introduction

Children with Developmental Language Disorder (DLD) are impaired in their language development resulting in poorer communication, which has long-lasting linguistic as well as emotional and social consequences. Early intervention can mitigate the effects of DLD. The diagnostics of DLD, especially when using Language Sample Analysis (LSA), is a time-intensive process. LSA requires transcription of speech, annotation and coding of different phenomena and analysis of the transcript, taking multiple times longer than the recording took. Due to its time-consuming nature, the potential of LSA is not always used. Not all children with DLD are recognised early. A broad screening would be able to recognize more cases but is not possible to do with the already limited resources.

Supporting diagnostics with automated processes could improve efficiency and help save time in meticulous processes like transcribing spontaneous speech. Automating LSA can save precious time for speech therapists and can support the efficiency of the diagnostic process in children with DLD.

Whereas there are already multiple tools for automated LSA in English and German, similar projects have not been done in Switzerland focusing on Swiss German. My thesis is part of the newly launched project DigiSpon that tries to close this gap.

DigiSpon (short for *Digital unterstützte Spontansprachanalyse*, digitally supported LSA) is a project of the University of Teacher Education in Special Needs Zurich (HfH) (*Hochschule für Heilpädagogik*) in collaboration with the Department of Computational Linguistics at the University of Zurich, the Department for Special Education at the University of Fribourg and the University of Teacher Education in Berne. The project researches computer-aided transcription of spontaneous speech used in LSA and automated linguistic analysis supporting speech therapists in the diagnostic and assessment of Swiss German-speaking children. This can reduce the effort involved in a single session and save time when evaluating the data. Through this, therapy can be more individual and better adapted to the needs of the children.

The automatic transcription and diagnostic should be integrated into a user-friendly tool which can be used easily in the daily work of speech therapists.

The first steps in the project DigiSpon were done by Conrad [2021] and Carpentieri [2022], who developed a pipeline which transcribes an audio file and gives the possibility to correct the transcripts, annotate the data and analyse linguistic features. It is only usable from the command line and only works in German. This is a big obstacle for people not used to programming and the tool needs to be able to handle Swiss German speech for practical use in Switzerland.

For Swiss German-speaking children with DLD data is not available, and there are no trained models for Automatic Speech Recognition (ASR) that perform sufficiently on this task. The next step of the project is to collect spontaneous speech data from Swiss German-speaking children. Data collection for research and later in the project during deployment poses its very own challenges when navigating regulations and ethical guidelines.

1.1 Research Questions

In my thesis, I work further towards the goal of the DigiSpon project. The questions I want to answer are:

1. How can the pipeline of Conrad [2021] and Carpentieri [2022] be further developed to expand its practical use for the DigiSpon project?

2. What needs to be considered legally and ethically in data collection for machine learning regarding the context of DigiSpon?

3. How could the insights of research question 2 be practically applied in the DigiSpon project?

1.2 Thesis Structure

In Chapter 1 I introduce the DigiSpon project and my research questions and give an overview of the work done for this thesis.
In Chapter 2 I summarize important research for my thesis, including developmental language disorder, automatic language sample analysis, automatic speech recogni-

tion and the first results of DigiSpon.

In Chapter 3 I describe the further development of the tool.

Chapter 4 describes the legal and ethical aspects of data collection for machine learning and transfers the theoretical knowledge to the data collection of DigiSpon. I summarize my findings and conclude this thesis with future research in Chapters 5 and 6.

2 Related Research

2.1 Developmental Language Disorder

2.1.1 Definition and Demographics

Language development is an important part of the development of children. It is genetically determined and heavily dependent on environmental factors. Because of that language development is very heterogeneous and children show big individual differences. Children with typical language development have learned to express themselves in correct grammatically ordered structure in an understandable way in their first language and can communicate appropriately to the situation when reaching the age of four [AWMF, 2010]. Disordered language development deviates from the norm in either its time frame or the content of the language learning [AWMF, 2010]. If a child's language development deviates 1.5 standard deviations from the norm value of their age group and is significantly below the appropriate level of their mental age, they meet the criteria for a DLD diagnosis [Zauke and Neumann, 2019]. Disordered language development can impact productive and receptive skills in one or more areas in a broad range of severity [AWMF, 2010]. If in addition to the disordered language development no other developmental disorders or physical impairments are found which can cause language disorders as a secondary symptom, a child is diagnosed with DLD. A child with DLD has a typical non-verbal Intelligence Quotient (IQ), is not on the autism spectrum, has no sensory development disorder, has a typical hearing capacity and shows typical motor skills in the body and the speech apparatus [AWMF, 2010]. Later evolving disorders such as attention difficulties, atypical social behaviour and mental illnesses are caused by impaired communication skills and count as secondary disorders [AWMF, 2010].
In the last years, there was a broad range of terminology used to describe DLD. In the English world terms such as specific language impairment, language delay, developmental language disorder and developmental dysphasia were used [Bishop et al., 2016]. In a DELPHI consensus study experts in the field recommended the terminology "developmental language disorder" and "language disorder" when comorbidities

are present [Bishop et al., 2017]. A similar DELPHI procedure was conducted by a consortium consisting of experts from Germany, Switzerland and Austria, which recommends using the terminology *Sprachentwicklungsstörung* and *Sprachentwicklungsstörung assoziiert mit...* correspondingly. This terminology replaces other used terms such as *spezifische Sprachentwicklungsstörung* or *umschriebene Sprachentwicklungsstörung* [Kauschke et al., 2023].

American studies show the prevalence of DLD is around 7.4% [Tomblin et al., 1997] and boys are two to three times more likely to be affected than girls [Kiese-Himmel, 2022a]. Similar studies are still lacking for the German-speaking area, but experts expect similar numbers [AWMF, 2010]

2.1.2 Course of the Disorder

There is some evidence that DLD signs can be observed even in pre-verbal linguistic development [AWMF, 2010]. Other studies describe the first signs of atypical language development in two-year-old children [[Berufsverband Deutscher Psychologinnen und Psychologen, 2011], [AWMF, 2010]]. At this age, the development shows substantial variability between individuals. Children with atypical language development at this age are classified as late talkers [AWMF, 2010]. One-third of them show typical development of language at the age of three. The other two-thirds still show delays in development at three years old. Half of them are diagnosed with DLD. 16% of late talkers still show significant impairment at school age [Berufsverband Deutscher Psychologinnen und Psychologen, 2011]. 40-80% of children showing symptoms in preschool age have symptoms four to five years later. 40-75% of them also show difficulties in the development of written language [AWMF, 2010]. Treated as well as untreated DLD shows residual symptoms 28 years later [Berufsverband Deutscher Psychologinnen und Psychologen, 2011].

In adolescence, difficulties in written language, on a semantic and lexical level are at the forefront of DLD. Difficulties in comprehension, atypical communication behaviour and socio-emotional development emerge in 75% of adolescents with DLD. 40% develop mental health problems [Kolonko and Seglias, 2004]. As most content in school is taught verbally, children with lower skills in linguistic reception are disadvantaged in all subjects and have trouble understanding and explaining more complex issues. The phenomenon of sinking IQ was observed in children with DLD, where they started with typical IQ values that sank below average when they got

older. This is probably caused by the difficulties in learning caused by impaired language understanding [Till et al., 2017].

St Clair et al. [2011] looked at longitudinal studies from 7 to 16-year-old children with a DLD diagnosis. They studied behavioural, emotional and social difficulties (BESD) such as problems with hyperactivity and conduct, as well as peer relationships. The severity decreased insignificantly in all areas when they got older, but difficulties in relationships with peers increased with age. They show more difficulties than their typically developing peers but do not fall in the clinical range of behavioural disorders. The severity and impact of symptoms vary wildly between individuals. As these symptoms are not visible, and not always traced to their origin in DLD, this can leave a vulnerable group without support for their difficulties. The heaviness of the impairment of pragmatic skills seems to be the biggest factor when predicting difficulties in relationships with peers.

Similar results were found in Conti-Ramsden et al. [2013] which looked at self-reported difficulties. The findings of St Clair et al. [2011] were confirmed in the self-reporting as well. The participants reported the most difficulties in maintaining relationships with peers. It was shown that children with impairments in receptive language development had more difficulties in maintaining peer relationships, as well as higher emotional and behavioural difficulties.

Through atypical communication and difficulties maintaining relationships, children with DLD develop aggression or retreat as coping strategies. Through that, children with DLD have a higher risk of developing comorbid mental illnesses, mostly depression and anxiety [Clegg et al., 2005].

Clegg et al. [2005] conducted a follow-up study with men in their thirties with a DLD diagnosis with receptive impairments. They showed severe symptoms of DLD and deficits in theory of mind (cognitive capacity to understand and attribute mental states to others) and phonological processing. They reached lower skills in literacy. A higher childhood IQ correlated with higher success in adult life. Men with DLD had significantly worse outcomes in social adaptation (unemployment, less close friendships and romantic relationships) than the control group. They showed higher rates of schizotypal features and general mental health problems.

2.1.3 Etiology

Twin studies showed a significant genetic factor in DLD. Environmental factors such as language input from caregivers are negligible. DLD occurs clustered in families, which supports the theory of genetic causes. Typically developing children can compensate for missing input in their families over time, which is not the case in children with DLD. Through this, impaired input from parents with DLD does not impact typically developing children but can be an added negative factor in children with DLD, which is another explanation for the clustering in families. The genetic factor seems to be polygenic and multi-factorial with a sex-specific threshold which explains the enormously heterogeneous appearance of DLD as well as the higher risk in boys [Monaco, 2007]. There is no known biomarker, but slight differences were found in the size of language-related brain areas, as well as a slightly reduced amount of grey matter in the brains of children with DLD [Raising awareness of developmental language disorders, 2020].

In a study testing narrative skills, it was seen that children with focal brain damage (damage to a specific brain region through accidents or strokes) were able to catch up with typical language development over time, but children with DLD could not. This suggests that there is no brain damage involved in the causes of DLD, but that the problem lies at a deeper level of language processing. [Reilly et al., 2004]

AWMF [2010] proposes two main theories of involved impaired mechanisms causing DLD. First, the input process deficit hypothesis assumes that the impairment is in the auditive processing of language in combination with a reduced capacity of the phonological working memory. Through the reduced capacity in processing, children cannot use the given input efficiently which leads to slower or impaired language development [Bishop, 1994]. The second hypothesis is the grammar-specific deficit. It assumes the cause of DLD is an impairment in grammar production which hinders children from producing grammatically correct sentences and would explain the heightened difficulties in more complex structures and verbs [Clahsen et al., 1997].

In a picture naming task, it was noticed that both groups, control and children with DLD made many semantic errors. But where the control group produced semantically similar errors (such as "yarn" instead of "rope"), the errors of children with DLD consisted more of words related to the object on the picture ("foot" instead of "shoe"). This speaks for a less robust semantic-lexical representation of words in

children with DLD [Lahey and Edwards, 1999].

Redmond [2004] compared conversational profiles of children with DLD, ADHD (Attention Deficit/Hyperactivity Disorder) and ASD (Autism Spectrum Disorder) and found similarities in the language of all three groups. Even though the differences were big enough to separate the three groups, the similarities could speak for a similar processing impairment. One hypothesis is that similar to ADHD and ASD, the cause of DLD could be a "defective time parsing mechanism". This theory is supported by the evidence of Deevy and Leonard [2004], where children were tested on WH-question structures (structures following question words such as "who", "where", "what" etc.). In simple short structures as well as in short but complex or long but simple structures, no difference was found between typical children and children with DLD; however, in the long complex structures, the performance of the children with DLD collapsed. This supports the hypothesis that children with DLD have an impairment in their processing capability. Through this limitation, important information from the beginning of the sentence is fading, which leads to incomplete processing. It is assumed that this originates in a limited ability to parse and cluster the language input, which leads children with DLD to process an unstructured and unanalysed representation of language which exceeds their processing capacity.

2.1.4 Symptoms

Symptoms in children with DLD vary with their age. Even though their symptoms are heterogeneous, there are symptoms found typically in each age [Von Suchodoletz, 2013]:

- Years 1-3: Milestones in language development are delayed or missing. This can affect all milestones, such as cooing, babbling, first words and multi-word phrase forming as well as limited lexical vocabulary. Pre-verbal linguistic skills can also be affected such as delayed word imitation, reduced building of consonants in babbling and a delay in gesture production [Kiese-Himmel, 2022b].

- Kindergarten and Preschool: Omission of whole words or beginning and ending syllables of words. Sentence structures are not formed grammatically. The inflexion of words such as tense in verbs or plural forming in nouns is erroneous.

- School age: Children can form correct short and simple sentences. They learned to avoid more complex structures through short answers without elaborations and through reformulation to simpler structures to hide their difficulties. Their difficulties show when they need to follow narratives or communicate more complex issues. They show a small vocabulary compared to their typically developing peers. These children show difficulties in the acquisition of written language. Their difficulties lead to other symptoms such as attention deficit, erratic, restless or fidgety behaviour and anxious or aggressive reactions when interacting with peers. These symptoms can mask the underlying language disorder.

In multilingual children symptoms of DLD are found in all languages. Their language development is delayed compared to their monolingual peers with DLD. This difference can also be seen in typically developing bilingual children, as through the multiple languages the input for each of them is reduced and can be caught up in the first few years in both groups. There is some evidence that bilingualism is beneficial in children with DLD. The specific structures affected by DLD change depending on language[Paradis et al., 2003].

Different categorisation schemes were developed to describe symptoms of DLD. One categorizes omission and comission errors [AWMF [2010], Grimm [2003]], (examples from Grimm [2003]):

- Omission mostly affects: Affixes, mostly ones with no lexical meaning which are used for grammatical functions (such as "*ge-*" or "*-t*" in verbs), particles such as prepositions, conjunctions, adverbs, interjections and negation, articles, pronouns and auxiliary verbs. These words as well as unstressed syllables are left out when speaking.

- Comission errors are mostly found in over generalisations ("*Ich bin gekommt.*"), case errors ("*Den Tante Besuch kommt*") and errors in word order ("*Heute morgen kein schön Wetter is*").

Another categorisation separates errors in linguistic areas [Siegmüller et al., 2022]:

- Phonetic and Phonological difficulties: Children with DLD often have combined phonetic and phonological impairments. Their speech motor skill to produce phones occurring in their first language and use them adequately in natural contexts is impaired. They cannot reliably recognize and distinguish meaning-differentiating phonemes in words. Because of this, they do not use

the right phonemes in typical positions or omit phonemes in words. They add, delete, double or mix syllables in words.

- Lexical difficulties: The development of the lexicon is delayed. They either have a reduced vocabulary and difficulties in learning new rare words or reach a typical lexicon size but are not able to use it in natural speech. Verbs cause major problems for children with DLD due to their complexity. Children can show difficulties in finding the right word, which they show through gestural or facial expressions. Children may use many neologisms, paraphrases or lists of synonyms instead of the word that they cannot remember. This leads to the use of stereotypical phrases and frequent repetitions as well as heightened use of pause fillers to fill the place of the missing word.

- Semantic difficulties: Children with DLD have difficulties building taxonomic structured semantic categories. They do not use semantic features to structure their semantic knowledge and rely on general associations. This leads to impaired use of abstraction as the rigid semantic representation is not able to represent more complex issues. This leads to difficulties in assigning concepts to categories and under- or over-generalisations.

- Grammatical difficulties: Children with DLD show difficulties in longer and more complex grammatical structures. Utterances are reduced in length and complexity and often they use fixed structures to compensate for their deficits. Children with DLD often do not use the second position placement but instead prefer SOV-structure (subject-object-verb such as *"da ich wohn"*). They produce errors in case, tense and number in their sentence structures (for example using the infinite form of a verb, *"Dann verwandeln hab eine Nuss"*). The phase of placing finite verbs at the end of a sentence is prolonged compared to typical development (*"du auto sehen?"*).

This thesis looks at DLD in children speaking Swiss German as their first language. Research in DLD in Swiss German is still lacking. Through the linguistic similarity to German, it can be assumed that most German findings can be transferred to Swiss German. However, there are differences between the two languages that need to be considered when looking at the symptoms in Swiss German-speaking children with DLD [Till et al., 2017].

2.1.5 Diagnostics

Studies have shown that early intervention is the most effective way to improve children's language development significantly. To guarantee that affected children are reached with interventions, an early recognition of DLD is essential [Penner, 2002].

Some federal states in Germany carry out state-wide screenings for all children in kindergarten. If children have not reached all important milestones of their age they are referred to clinical diagnostics. However, the benefit of this technique is controversial as not all states use evidence-based, evaluated and normed tests, which makes results insufficiently reliable [AWMF, 2010].

Kiese-Himmel [2022b] argues that screening in kindergarten is too late as the neuroplasticity is already reduced. Early screening can be done as soon as the children are 18-30 months old with questionnaires for parents, in which the size of vocabulary is tested. They argue that even this is too late for early intervention and look at possible diagnostic features in younger children whose neuroplasticity is highest. To prevent treating late-talkers who do not profit as much from the already limited resources, it needs features that differentiate between late-talkers and children with DLD. They suggest evaluating mother-child interaction and haptic processing in 12-18 months old children. Even earlier diagnostics could make use of phonetic analysis of prelinguistic protophones and neuroimaging techniques such as MRI, in which studies have shown differences in the symmetry of the cerebellum, which could predict disordered language development. Whether these diagnostic features are reliable and whether intervention this early shows benefits is not yet evidence-based.

The most important partners in the early recognition of DLD are parents and caregivers, as well as daycare supervisors and teachers. They are the ones referring children to a specialist if the development of a child is atypical. For this, they need to be educated about symptoms and what to do when a child's language development starts to deviate from the norm. The initiative "Raising Awareness of Developmental Language Disorder" (RADLD) tries to educate these groups through video materials and interviews with affected people and experts in the field to raise awareness and expertise, helping to recognise these children as early as possible [Raising awareness of developmental language disorders [2020], Bishop et al. [2012]].

Productive disorders are more probable to be recognized by parents and teachers

when compared to receptive disorders, as they have more obvious and observable symptoms. It is important to recognize receptive disorders as well, as children with receptive disorders are affected more severely and show significant symptoms in adulthood. The receptive type of DLD is additionally less likely to benefit from interventions. Receptive types need screenings of the physical and psychological state of development which are more complex and time-consuming [Berufsverband Deutscher Psychologinnen und Psychologen, 2011].

Children with DLD growing up bilingually or with a foreign first language are less likely to be recognized as their delay in language development through other languages cover up the impairments caused by DLD. Atypical language development through their multilingualism does not classify as DLD. DLD is a deeper laying impairment of language development and affects all languages a child is speaking [Kolonko and Seglias, 2004].

DLD is a differential diagnosis, which means it is diagnosed by exclusion of other possible diagnoses. After referral through parents, teachers or others the diagnosis starts with collecting the patient's history and physical and psychological examinations to rule out other factors causing secondary language disorders [AWMF, 2010]. In the second step, the areas affected and the severity of impairments are evaluated. This happens through questionnaires and conversations with parents regarding specific language skills as well as analysis of interactions through standardized tests of different language skills and through LSA in which elicited spontaneous speech of the child is evaluated for language performance (the ability to use language adequately and promptly in real-world situations) [AWMF [2010], Berufsverband Deutscher Psychologinnen und Psychologen [2011]].
Berufsverband Deutscher Psychologinnen und Psychologen [2011] published a possible diagnostics algorithm as a guideline for practitioners. It can be seen in Figure 1.

Standardized tests are heavily language-dependent. For Swiss German-speaking children, there are not many standardised tests that are normed and evaluated. Most Swiss speech therapists use standardized tests from Germany [Till et al., 2017]. This works in structures that are sufficiently similar in both languages. Speech therapists translate the used items ad-hoc (for example *"Ds Meitschi isst e Öpfu"* to *"Das Mädchen isst einen Apfel"*) [Dubagunta et al., 2022]. This does not always work as the differences in both languages are too big in pronunciation, vocabulary and grammar (such as plural forms, *"Chuchi – Chuchine"* to *"Küche – Küchen"* and
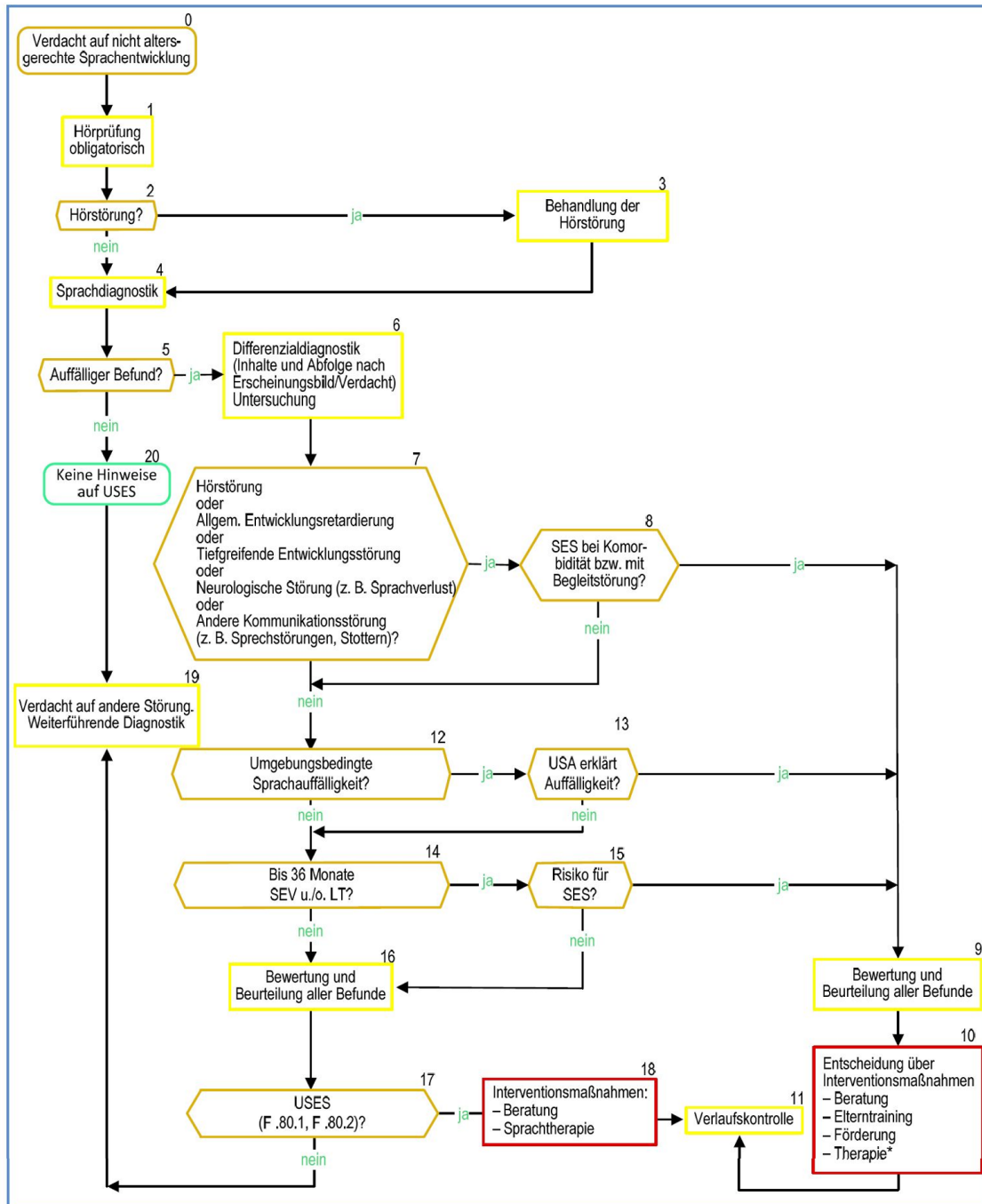
Figure 1: Diagnostic algorithm as suggested in Berufsverband Deutscher Psychologinnen und Psychologen [2011].
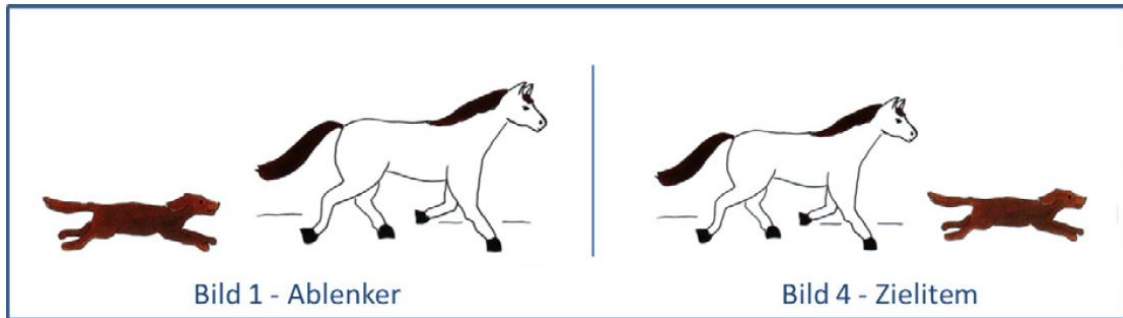
Figure 2: An example card from the TROG-D test [Dubagunta et al., 2022]

relative clauses "*Dr Tisch, wo d Tasse drufsteit*" to "*Der Tisch, auf dem die Tasse steht*") [Dubagunta et al., 2022]. These differences make it difficult to use German tests. Additionally, as these tests are mostly translated ad-hoc, translations can differ and the tests are not normed any more [Dubagunta et al., 2022]. For most children, German is a foreign language in their first years of contact when they show big differences in skill in the two languages. As German is spoken in school and has gained importance over the years, children have more and more contact with it and are proficient enough to be tested in German when they are older. Speech therapists in Berne tested 65% of children in Swiss German in kindergarten, but only 34% were tested in Swiss German in school [Dubagunta et al., 2022]. As early recognition is important and most Swiss German-speaking children are not skilled enough in German in kindergarten, it is important to have standardized and normed Swiss tests [Dubagunta et al., 2022].

Dubagunta et al. [2022] looked at the TROG-D test and whether it is possible to directly adapt it into a Swiss German version (TROG-CH). The TROG test is used to test grammatical comprehension. It contains items consisting of four cards from which the child picks the correct picture according to the sentence the speech therapist is giving. The wrong cards are minimally different in their lexicon or grammar. Each item tests a grammatical structure. The test contains 18 grammatical test items as well as three control items that ensure that the child knows the vocabulary required for the test.

Through the similarity in the cards, sentences can differ only in their case such as "*Der Hund jagt das Pferd*" vs "*Den Hund jagt das Pferd*" (this example can be seen in Figure 2). In German this item is unambiguously solvable. However, in Swiss German it is not ("*Dr brun Hung jagt ds Ross*" can have both meanings). In

the test, the second phrase is uttered (*"Den Hund jagt das Pferd"*). Swiss children tend to interpret this sentence as a nominative structure, as the correct accusative structure is not commonly used in Swiss German, which leads to 95% of the children choosing the "wrong" item. Speech therapists sometimes translate the sentence using a more natural Swiss German accusative structure, but as these translations are made spontaneously, they are not normed. Allowing both pictures as a correct answer alters the test severely and the chance to randomly pick the right card raises from 25 to 50%. This example shows the need for an officially translated and normed test specifically for Swiss German [Dubagunta et al., 2022]. Other items that are unambiguously solvable in both languages do not show these problems and can be used without further adaptation.

Dubagunta et al. [2022] compared the test results of children from kindergarten to third grade with and without DLD in the TROG-D test and a newly created, standardized TROG-CH test. Their goal was to find out whether there are differences in the two tests in the different groups and whether Swiss speech therapists need adapted tests. For this, they left out the items of the test which could not be solved unambiguously. They found that children with typical language development solved more items correctly in the TROG-CH test than children with DLD. In the TROG-D test typically developed children show more difficulties solving the items as in Swiss German. Through this, the two groups cannot be separated reliably any more. Children with DLD show similar results in both languages. They can transfer gained knowledge from Swiss German to German e.g. if they master a sentence structure in Swiss German they are also able to use it correctly in German, which was not the case for all typically developing children. It seems that in further diagnostics the original test can be used. However, as some children with typical development showed below-average performance in TROG-D, a Swiss German test would be beneficial in the first step of diagnostics [Dubagunta et al., 2022]. The tested group is more likely to be tested in German in practice, which leaves the question after the need for Swiss German tests in younger children open.

2.1.5.1 Language Sample Analysis

In addition to standardized tests, LSA is an important diagnostic tool in which children are encouraged to use speech spontaneously and freely, similar to their everyday lives. Speech is elicited through different techniques such as interviewing (talking about themselves, their families, interests or activities), retelling (listening to a story once, then retelling it with the support of pictures), storytelling (inventing a story to pictures) or role-playing (pretending to be on a phone call with their

mother) [Schlett et al., 2013].

The advantage of LSA is that it can test many different aspects of language development simultaneously in a natural setting. This includes aspects which cannot be tested with standardized tests such as word-finding difficulties in a conversation setting and difficulties in conveying meaning in conversations. LSA can be used repeatedly without influencing the test results, which is not the case for most standardized tests. Studies in younger children have shown that LSA has a high sensitivity in separating children with DLD from typically developing children, as well as from children with ADHD or ASD [Redmond, 2004].

Measurements used in the diagnostics of children with DLD with the help of LSA are: speaking rate, utterance formulation, lexical diversity, average utterance length, morphosyntactic complexity as well as many more.

Measurements that have shown to be reliable in discriminating children with DLD from their peers are [Redmond, 2004]:

- Number of unique words in 100 utterances. This measurement loses its reliability in older children.

- A composite measure of tense marking, which includes the correct production of regular simple past tense ("-ed"), third person singular verb inflexion ("-s") and the use of auxiliary "be"-forms.

- Errors in the production of composite tenses, such as present continuous, past continuous and past perfect. These tenses use auxiliary verbs in formation, sometimes more than one. This measurement reliably separates DLD from other language disorders which makes it a unique marker of DLD.

- Mean length of utterance. Children with DLD tend to use shorter utterances than typically developing children.

This study as well as most others was done with English-speaking children. Not all measurements can be transferred to children speaking German or Swiss German and if it is possible, evidence has to be found that they reliably differentiate German-speaking children with DLD from typically developing children.

Schlett et al. [2013] show that some elicitation methods may be better suited for different types of analysis. For analysing phonetic and phonological features, all elicitation methods are suitable. When analysing grammar, retelling and storytelling

are better suited as they elicit a bigger range of different grammatical structures which is needed for a complete analysis. In role-playing children use more answer fragments and incomplete phrases which makes analysis more difficult. Similar results were found in the automatic processing of language samples [Gabani et al., 2011].

2.1.6 Treatments

Many studies have shown that children with DLD benefit from language therapy. Interventions are beneficial even if residual symptoms were found in adulthood [AWMF [2010], Von Suchodoletz [2013], Berufsverband Deutscher Psychologinnen und Psychologen [2011], Fricke et al. [2013]]. To achieve the best possible development in writing and reading skills, therapy should be concluded before starting school. Some studies even showed that children with DLD profited from intervention in the pre-verbal phase, for example, training in triangular eye contact which is strongly linked with joint attention, an important precursor to language [Fricke et al. [2013], AWMF [2010]]. Early interventions can prevent later social, emotional and psychological difficulties and can generally strengthen linguistic skills and general development in children with DLD [AWMF, 2010].

If a child is recognized as a late-talker, intervention can lower the risk of developing a language disorder. The support for their language development can happen at home. Parents receive support through counselling and through training on ways to more actively elicit speech in the child. This can happen for example through dialogic book reading, where parents read out picture books and intentionally involve the child in the telling of the story and through eliciting conversations at the dining table. If this is not sufficient, the child can be integrated into a language support program, where children are supported specifically in their language development in a group through play. At this age, interventions happen in day-to-day life and therapy settings are only used in severe cases. During the intervention, the child's progress is closely monitored and the intervention is adapted to the child's specific needs and impairment [AWMF, 2010].

Fricke et al. [2013] monitored children with "poorly developed" language skills in nursery school and reception (English school grade from age 4 to 5) while receiving language intervention. In this study, teaching assistants were trained to carry out the intervention in a group setting. They used multi-sensory teaching techniques

to improve the children's vocabulary, narrative skills, active listening, letter-sound knowledge and phoneme awareness. Children were supported in gaining confidence in independent speaking. Interventions were first held three times a week at 15 minutes and were later increased to three times a week at 30 minutes in a group as well as twice a week a 15-minute individual sessions. This intervention was ended before the children started school. It improved their oral language and spoken narrative skills immediately. The improvements lasted for 6 months. Children receiving this intervention showed benefits in developing reading and writing skills, where they performed similarly to their typically developed peers.

If a child is diagnosed with DLD, they normally receive language therapy. There is no general treatment for DLD. Therapies are planned individually for each child and focus on the specific deficits a child has. Before therapy, an assessment of all language-related skills and the severity of the impairment needs to be conducted. The primary goal is to raise the level of language skills as close as possible to the age-appropriate level and to address secondary difficulties (such as attentional or social problems) [Kannengieser, 2013].

Individual goals in therapy could for example be [Kannengieser, 2013]:

- Facilitate language acquisition, language processing and communication.

- Raising awareness of impaired linguistic structures

- Developing self-confidence through awareness of the disorder and strategies to compensate it

- Preventing or correcting unfavourable adaptation of speech in parents

Before starting school, the children are examined again and it is decided, whether they can attend regular school. If a child does not catch up with their peers and language development is significantly delayed or impaired, other institutions such as language kindergartens (*Sprachheilkindergarten*) or a special school (*Sonderschule*) are options. If a child can attend regular school, therapy accompanying the school is usually the most effective solution [AWMF, 2010].

Therapies are most effective in children with productive impairments (phonological, phonetic or lexical) and less effective in receptive disorders. In general, both groups benefit from interventions [AWMF, 2010].

Whereas therapy in younger children is a well-researched topic, studies for interventions in older children (age 13 and up) are rare. Most therapy techniques are not as efficient any more for this group, as they are heavily based on play, which is not suited for this age group. [Kolonko and Seglias, 2004] interviewed speech therapists on this topic, which was met with enthusiasm, which shows the need for this research. Therapy in this group should be led by the goal of terminating the therapy soon, which includes the disconnection process from the therapist, the management of residual symptoms and the use of aiding tools, including assistive technology [Kolonko and Seglias, 2004].

2.2 Automatic Diagnostics of Language Impairments

Even though LSA is an integral part of the diagnostics of DLD, it is not as often used as it could be because of its labour-intensive process [Lüdtke et al., 2023]. The meticulous process of preparing a child's speech for LSA includes elicitation and recording, manual transcription, manual coding for specific occurrences (such as coding errors or actions a child is making (e.g. throwing a toy etc.)), manual annotation (such as grammatical structures or part-of-speech (POS) tags) and finally the analysis of the sample [Lüdtke et al. [2023], Solorio [2013]]. This process can take several hours of work on a single sample of a child. Just transcribing a language sample can take 5 to 50 times the duration of the audio recording, depending on the depth of transcription [Lüdtke et al., 2023]. Ideally, this analysis is made in children repetitively to monitor the progress in development [Lüdtke et al., 2023]. Developments in Natural Language Processing (NLP) offer the possibility to use automated processes to reduce the workload of LSA, which can make it more efficient and thus more accessible analytic tool for practitioners [Lüdtke et al., 2023]. Results derived from automated computing processes are more consistent and better reproducible than manual analyses [Solorio, 2013].

Lüdtke et al. [2023] describe their ideal LSA pipeline. Their ideal system can assess mono- and multilingual children and reliably process recordings even when children are code-switching. The system records children in their natural environment (at home) over several hours. It can differentiate between relevant speech and background noises, including background speech (from a TV). The system codes different languages and speakers and can assign each utterance to the right speaker. The system analysis works in two channels: it analyses the audio recording (for background noise, frequency of digital media use, sounds from actions (such as throwing a toy

etc.)), as well as the transcripts. The recording is transcribed phonetically as well as orthographically and the transcripts are annotated in segments and single words. After transcribing, the files are coded and annotated with different important features such as POS tagging, linguistic phenomenons and grammatical structure. This system outputs a wide range of measurements covering environmental factors, developmental language profiles and detailed linguistic analysis for multiple linguistic structures and elements.

This ideal system does not fully correspond to the system we plan to develop. The goal of this project is to develop a tool for a clinical setting of speech therapists, which excludes some of the named features (such as media exposure or background noise). It shows many features that our tool would benefit from, such as speaker recognition, automatic transcription, coding and annotation and automatic analysis. It shows the wide range of NLP techniques used in such a project. Research is conducted in many of these areas, but complete pipelines for LSA are still rare.

Gabani et al. [2011] classified children with DLD and typically developing children with the help of manual transcriptions from recordings of six-year-old and 13 to 16-year-old English-speaking children. They tackled this experiment as a text classification task using shallow, surface-level features.
They used features separated into 8 different categories [Gabani et al., 2011]:

- Language productivity: Total Number of Words (TNW) and degree of conversational report (counting the number of utterances of the investigators, assuming that children with DLD need more help when speaking, leading to more utterances of the investigator)

- Morphosyntactic skills: measured as the ratio of the number of raw verbs to the total number of verbs, subject-verb agreement and number of different POS-tags.

- Vocabulary knowledge: measured in Number of Different Words (NDW)

- Speech fluency: measured through numbers of repetitions, revisions, partial words and filler words

- Probabilities from language models: using two language models, one trained on transcripts of typically developed children, one with transcripts of children with DLD, the probabilities from both language models for a transcript are calculated and derived values from this (such as difference of probability).

- Standard scores: distance measurements between the score of a transcript and a pre-calculated expected score were used as features for both scores of children with DLD and typically developed children. Standard scores were calculated in mean length of utterance (MLU), NDW and total number of utterances

- Sentence complexity: measured in MLU, number of words, average number of syllables per word, average number of clauses per sentence and the Flesch-Kincaid score (score of readability of a text)

- Error patterns: A few specific markers were identified for DLD in English, such as the inconsistent use of verb tense and determiner-noun number disagreement (for example "these book"). POS tags and bi-gram patterns were used to recognize these patterns

A feature analysis showed that the system performed best with all eight categories. Some features may be redundant, but they do not worsen the performance. The system scored 85% in f-measures [Gabani et al., 2011].

This corpus-based approach using text classification is adaptable to bilingual children speaking Spanish and English and has the potential to be adapted to other languages [Gabani et al. [2009], Solorio and Liu [2008]].

Hassanali et al. [2012] added deeper NLP features to the system of Gabani et al. [2011]. These are readability, situational model (measurement of the micro world a text is about), general word and text, syntax-based, referential and semantic, entity grid model (measuring local coherence), narrative structure and quality of text features. Testing in a dataset of transcripts of 14-year-old adolescents, adding these features improved the baseline system. Better results were achieved in narrative storytelling than in spontaneous speech, furthering the evidence that the elicitation method is an important performance factor.

All of the features described above proved to be valuable in predicting DLD with automated LSA. The studies testing these features were conducted on small sample sizes on different populations and language tasks, bigger studies are needed to compare them directly and to estimate the efficiency of individual features. For now, they are important pointers for the direction in which this field can advance. Performances generally are slightly better on more structured elicitation methods such as storytelling compared to more free forms, such as spontaneous speech or free play [Solorio, 2013].

2.2.1 Tools in Use

There are approaches to an automatic pipeline of LSA already used in practice. In this chapter, I will present the most common and advanced ones. All of the systems I describe are primarily for English analysis, with some tools supporting a few other languages in reduced functionality. The tools I describe are: SALT, CLAN and LENA.

The Systematic Analysis of Language Transcript (SALT) [Miller et al., 1985] software is probably the most used LSA software by speech therapists and other practitioners. It provides an editor to manually transcribe recordings according to an extended annotation scheme. The annotation scheme is very elaborate and able to encode a wide range of aspects including meta-information about the file, utterance boundaries, unintelligible parts of the recording, divergences of child forms from adult standards as well as code-switching [Lüdtke et al., 2023]. Because of the sheer size of the annotation scheme, it takes a prolonged time to learn the annotation and coding system [Snider]. This process is well supported by guides to many different topics provided by SALT on their website[1]. These annotations are used for automatic analysis, focusing on "the assessment of speech and language, [...] including measures of syntax, lexicon, discourse, fluency and speaking rate" (p. 5, Lüdtke et al. [2023]).
SALT offers multiple reference databases with data about different age groups, elicitation methods and languages of children with and without language disorders, including DLD, to which users can compare their results to [Heilmann et al., 2010]. SALT provides, additionally to their software, elicitation materials such as picture books, reference books and training on all their products, as well as an online story-elicitation tool, in which a cartoon character guides through the process and children can be recorded remotely [SALT].

Computerized Language Analysis (CLAN) [MacWhinney, 2000] is an open-source software and was originally developed to process language data in the "Child Language Data Exchange System" (CHILDES) project, including searching, manipulating and analysing language samples [Lüdtke et al., 2023]. Through its automatic analysis and annotation, it is also used in LSA. It offers semi-automatic coding such as morphological parsing, POS-tagging and grammatical dependency parsing in different languages, but mainly in English Lüdtke et al. [2023]. Of all the tools

---

[1]https://www.saltsoftware.com/resources

described, it offers the most options for automatic annotation and analysis of conversational interactions.

Language Environment Assessment (LENA) [Gilkerson and Richards, 2008] was originally developed to give automatic feedback to parents on their linguistic interactions with their children at home to encourage caregiver-child interaction [Lüdtke et al., 2023]. Through its original purpose, LENA includes recording with automatic recognition of speech and environmental sounds and can record speech for several hours. LENA's analysis happens directly on the audio recording and includes language recognition, speaker segmentation, annotating overlapping speech, electronic media, environmental noise coding and segmentation of age-specific children's vocalisations, such as distinguishing babbling from crying in toddlers [Lüdtke et al., 2023]. As the analysis is done directly on the audio, LENA is not able to compute standard measures such as MLU, as it skips the transcription process. LENA is one of the most advanced tools in audio analysis and offers interesting possibilities for LSA.

SALT and CLAN are reliant on manual transcriptions and work only in their respectable, manually annotated, coding scheme. All three tools lack automatic transcription using ASR and complete, automatic annotation and coding. This is because these processes are highly complex and dependent on language, as well as very sensitive to the data they are trained on [Lüdtke et al., 2023].

2.2.2 TALC Project

The "Tools for Analysing Language and Communication" (TALC) project started in 2019 and researches machine learning approaches in linguistic, speech therapy and pedagogical settings. Simultaneously, it is the goal to establish ethical and legal frameworks for the application of machine learning in this area. TALC consists of two project branches; one in Germany with a focus on German and one in South Africa, focusing on Afrikaans and Lesotho. It is planned that the South African branch extends its research to isiXhosa in 2024 [TALC, d]. In May 2023, three sub-projects were launched:

- The kidsTALK database [Ehlert et al., 2023] acts as the foundation for all subsequent research and development of tools. It consists of audio files and the corresponding transcripts from monolingual and multilingual children with typical language development in all languages which are part of this project. The first part of the dataset was published in 2022 [TALC, c].

- TALC-LSA focuses on the development of a hardware- and software-system to record, semi-automatically transcribe and linguistically analyse spontaneous language samples of children. To ensure easy and practical use in the field, the system will be launched on a small, portable recorder, which can record multiple hours of speech [TALC, a]. In May 2023, data collection was in progress and the first papers regarding ASR were published [Gebauer et al. [2023], Rumberg et al. [2021], Rumberg et al. [2022]].

- TALC-DIRA (Data Integrated Reading Assessment) researches and develops software for assessing reading skills in school children supported by machine learning. The goal is to develop a tool, which can do all assessments at once. Until now, these processes were separate and needed to be done linearly. The software will be launched as an application for tablets that can be used for automated screening in schools in multiple languages. The first pilot phase is planned for the school year of 2023/24 [TALC, b].

2.2.3 Automatic LSA in German and Swiss German

Most of the described research was done in English-speaking children. DigiSpon focuses on Swiss German-speaking children.

CLAN, as described in chapter 2.2.1, provides limited analysis in German, such as morphological parsing and a few others [Lüdtke et al., 2023]. This is not sufficient for a complete diagnosis in German children.

Barthel et al. [2006] developed the "Aachener Sprachanalyse" (ASPA) tool for the analysis of aphasic patients. It provides computer-assisted analysis of quantitative methods of linguistics, such as the percentage of words, type-token ratio and MLU. This tool was developed in 2006 and is not able to compete with more recent techniques.

Apart from the TALC project [TALC, d] currently in progress, no similar German projects were found. No projects on this topic were found in Swiss German.

2.3 Automatic Speech Recognition

2.3.1 ASR of Children's Speech

ASR of children's speech bears its challenges. Children have higher inter- and intra-speaker variability. They show higher variability in pronunciation, as well as vo-

cabulary and grammatical structures. Because of this, ASR systems trained on children's speech need more training data than systems trained on adult speech. Children show big variability in speech depending on age. ASR proved to be most difficult on children in kindergarten or younger. The second big problem in ASR of children's speech is the sparsity of data [Lüdtke et al., 2023]. Children's speech has higher fundamental and formant frequencies due to the smaller vocal tract. These frequencies are not in the range most ASR systems perform well. Children's speech is slower than typical adult speech [Liao et al., 2015]. When ASR is needed in LSA and other diagnostic contexts, the focus is on the precise representation of what the child said word by word, including repetitions, filler words and errors. Depending on the analysis, the precise phonemes are used in the transcription [Lüdtke et al., 2023]. Most ASR systems are trained for voice assistants, where the meaning of an utterance is more important than the exact formulation. This makes most pre-trained ASR systems useless for this task [Lüdtke et al., 2023]. Through its high variability and its sparse data, ASR of children's speech is similar to low-resource languages and applying techniques used in low-resource languages helped to improve performance in ASR with children's speech [Lüdtke et al., 2023].

Liao et al. [2015] built a large vocabulary ASR system for YouTube Kids (A YouTube application specifically designed for children which only shows restricted content), which was trained on large amounts of children's speech data. They filtered Voice Search traffic from Google for utterances from children using a neural network classifier. The utterances were transcribed manually and offensive content was filtered out. The system was trained on 2.6 million utterances from adults and 1,9 million utterances from children. Liao et al. [2015] tested techniques which were shown to improve ASR of children's speech, such as pitch features, spectral smoothing and vocal tract length normalization. The improvement observed in other studies was not found in this system. This system was trained on a larger training dataset than most other ASR systems for children's speech, which raises the performance of the system to a point where these techniques are not able to improve performance any more. The system reached lower performances than ASR of adults' speech trained on similar amounts of data, which confirms the assumption that ASR of children's speech needs more data than ASR of adults' speech.

Most projects do not have the amount of data used by Liao et al. [2015] to train a complete system. Approaches are needed that achieve sufficient performance with the limited data available.

Smith et al. [2017] combined transfer learning with constrained decoding for a screening system for children with language impairment. A deep neural network was trained on out-of-domain adult speech and fine-tuned with speech data of children with DLD. Using transfer learning lowered the Phoneme Error Rate (PER) by two per cent to 14.2% compared to a system that was trained only on children's data (16.3%). The constrained decoding, a hierarchical neural network, was constructed with the help of experts in speech pathology and represents typical errors found in the pronunciation of children with DLD.

Rumberg et al. [2021] propose a framework for age-invariant training, leveraging age-invariant knowledge from adults' and children's speech. For this, they use adversarial multi-task learning. They train the system on child and adult speech, using age as a feature. They simultaneously train the acoustic model, as well as a discrimination model to estimate the speaker's age at the last layer of the encoder. An adversarial loss punishes the system if the discriminator estimates the age correctly, thus forcing the system to learn age-invariant features only. The best performance was reached when combining the age-invariant training with the more traditional methods of feature space adaptation, such as fundamental frequency normalization.

Adult speech data can be used to produce synthesized data similar to children's speech by adapting fundamental and formant frequencies and prolonging vowels. This synthetic data can be used to extend the limited amount of available children's speech data [Lüdtke et al., 2023].

Most ASR systems output an orthographic transcription. This works fine for systems such as voice assistants, where the main goal is the semantic representation. If tasks further down the pipeline need utterances transcribed word by word, this is not always sufficient. Orthographic systems are limited, idealized representations of speech. Many different phonemes are represented by the same letter. Younger children are to be assumed to deviate heavily from standard pronunciation. When training a system with orthographic transcripts, it ends up biased towards the standard transcription and can idealize variations, which would be important to capture for diagnostics. Better representation of speech is achieved with phonetic transcripts. These are much more expensive to produce than orthographic transcripts and even fewer are available for training an ASR system [Rumberg et al., 2022].

Rumberg et al. [2022] improve the quality of a system trained on orthographic

script on a phonetic level with a Connectionist Temporal Classifier (CTC) trained on a small dataset of phonetic transcriptions. The CTC is constrained through a Weighted Finite State Automaton (WFSA) function as a pronunciation dictionary, constraining possible phone sequences. As there is variation in how different phones are pronounced, the WFSA was extended with possible common alternative pronunciation found in a corpus. Using the constrained CTC improved the system by a relative 14% compared to the baseline trained only on the orthographic transcripts. The created transcripts showed effectiveness in improving the training of a new system [Rumberg et al., 2022]. This system achieved similar performance to the German MAUS, a phonetic segmentation tool [Schiel et al., 2011]. MAUS performs slightly better in phone-level segmentation but is not able to represent phonemes deviating from the standard pronunciation. The system described by Rumberg et al. [2022] is more robust in out-of-domain speech and deviation from the standard [Gebauer et al., 2023].

2.3.2 Swiss German ASR

ASR in a few selected high-resource languages reaches Word Error Rates (WER) of less than 5% [Schraner et al., 2022]. German ASR models reach these values. Even though Swiss German is similar to German to some extent, it differs in phonetics, vocabulary, morphology and syntax enough to require specifically trained models in ASR. Swiss German qualifies as a low-resource language [Schraner et al., 2022]. In low-resource languages, WER is significantly higher (uni-dialectal Swiss German systems reach WER of 20% [Nigmatulina et al., 2020]). Since 2021, large Swiss German datasets have been available, which made the training of end-to-end systems with high performances possible [Schraner et al., 2022].
Swiss German is a low-resource, non-standardised language with high dialectal variation. The non-standardised writing of Swiss German, used in informal settings such as text messages leads to difficulties when producing transcripts due to spelling ambiguities, its high variability and its huge vocabulary size [Schraner et al., 2022]. All of this makes ASR in Swiss German a particularly challenging task [Nigmatulina et al., 2020].

Arabskyy et al. [2021] propose a hybrid ASR system that recognises Swiss German speech and outputs translated German transcripts, which needs an inherent translation step in the system. Their models consist of four parts: a lexicon containing German to Swiss German translations using parallel corpora, a 1st pass language

model specialized in Swiss German particularities (such as rare compound words and clitics), a transfer-learned acoustic model trained on German data and a strong neural language model for 2nd pass to produce accurate German predictions. This system won the Shared Task 3 at SwissText 2021 on Swiss German Speech to Standard German Text with a WER of 39%, out-competing the second-best model by 12%.

Kew et al. [2020] present a model submitted to the GermEval 2020 Task 4 on low-resource speech-to-text, whose goal it was to translate spoken Swiss German to written German. They implemented a time delay neural network (TDNN) acoustic model (AM) using the Kaldi Speech Recognition Tool kit [Povey et al., 2011] extended with a pronunciation lexicon (using an 11,000-word dictionary mapping German words to their Swiss pronunciation) and a German language model. This system reached a WER of 45.45% on the test set of the task. The transcripts are comprehensible. The most common error was missing numbers, which is explainable as the lexicon does not contain numbers. Another frequent error was missing words. The system had difficulties handling the cut-off beginning and end of recordings, which could explain some of the missing words at the end and beginning of the sample.

Nigmatulina et al. [2020] compared the two main types of possible output text of a Swiss German ASR system, namely dialectal transcripts, which approach a representation on the phonemic level, capturing differences in dialects and normalised transcripts, where the writing resembles standard German. The systems were trained on 14 different dialects.
Looking at WER, the system trained on standardised text achieved better results (29.39%), but is not able to perform well at the character level. Using FlexWER, (which counts words as correct when the gold standard and the predicted word share a normalised version) the system trained on dialectal transcripts performs better. Systems trained on normalized transcripts are more robust against noise, but systems trained on dialectal transcripts are a good alternative if dialectal variations are important in the downstream task. In addition to the findings of different methods, Nigmatulina et al. [2020] established a new benchmark system in multi-dialect ASR for Swiss German, with the best system reaching a WER of 29.39%

2.3.3 Combining the Challenges

In the DigiSpon project, we combine the challenges of ASR of children's speech (especially disordered speech), with the challenges of Swiss German ASR. This project looks at a very small population of a low-resource language.

As transfer learning improves performance in both areas, our ASR system could profit from it as well. Using German children's speech data, as well as/ or adult Swiss German data (such as the Swiss Parliament Corpus [Plüss et al., 2020]) to train the system and using Swiss German Speech of children with DLD to fine-tune it could achieve good results.

Swiss German ASR profits from the use of a lexicon, which could be implemented in the tool. When integrating typical errors of children with DLD, as well as Swiss German words, this could improve performance. However, as no project such as this was done before, all of this is speculation and needs to be tested.

One important decision to make is whether the output of the system is Swiss German or German. The transcripts, which will be collected in the next phase of the project, are written in Swiss German and this, as well as the importance of phonemic correctness for this pipeline, speaks for Swiss German output. Using German output would make it possible to use German NLP tools further down the pipeline, which are more widely distributed than NLP tools for Swiss German.

Datasets similar to the ones that will be created for DigiSpon could enhance performance. This includes the kidsTALC database [TALC, c], consisting of German recordings for ASR of children's speech or data collected in the project EmTik under the coordination of Dieter Isler, currently running at the *Pädagogische Hochschule Thurgau*, collecting data from Swiss German-speaking children in kindergarten and school [Isler, 2022]

2.4 DigiSpon

My contribution to the DigiSpon project is directly based on the bachelor's theses of Sophia Conrad and Sofia Carpentieri. They did the first steps towards developing a tool to support automatic LSA which first helps to collect and prepare data for machine learning, then provides linguistic analyses.
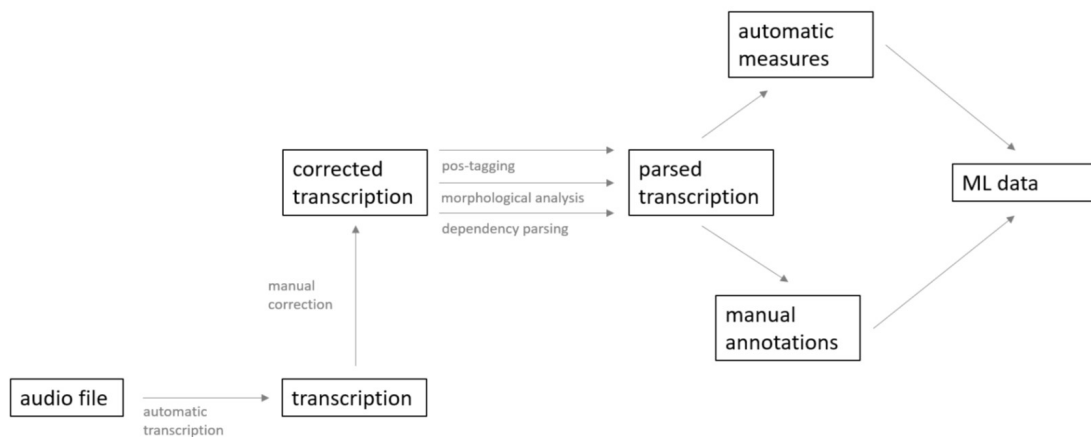
Figure 3: An overview of the first pipeline by Conrad [2021]

The first rendition of the pipeline for DigiSpon was developed in the work of Conrad [2021]. The pipeline takes an MP3 file as input and creates linguistic measurement as well as a transcript as output. The transcripts can later be used for machine learning. It consists of automatic transcription using IMS-Speech [Denisov and Vu, 2019], manual correction of the transcript and manual annotation with the web tool Prodigy[2] and automated linguistic analysis.

The implemented measurements are:

- Type-token ratio

- Moving-average type-token ratio

- Brunét's index

- Honoré's statistic

- Lexical density

For further information about the single measurements the reader is referred to Conrad [2021].

The pipeline is used linearly from start to end using the command line. A graphical

---

[2]https://prodi.gy/
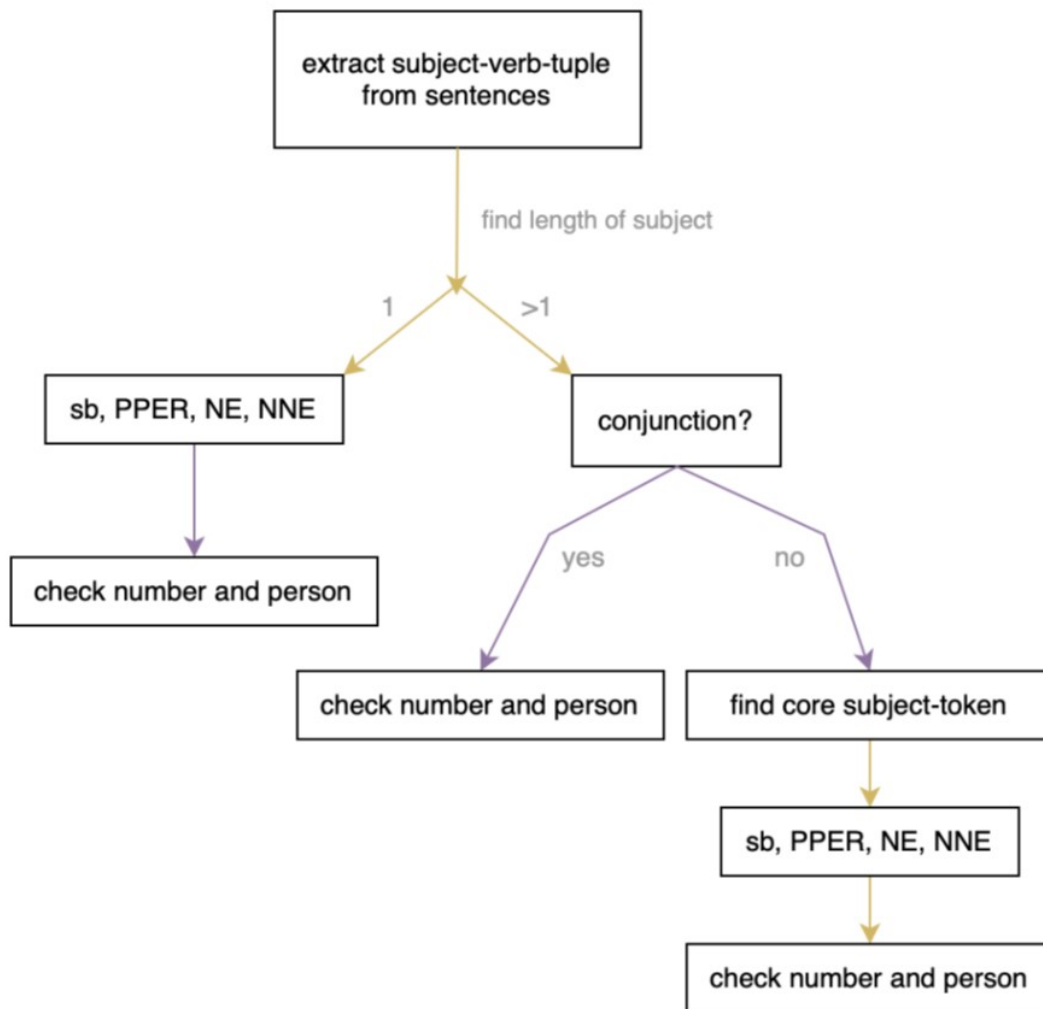
Figure 4: Flowchart of the subject-verb agreement function by Carpentieri [2022]

overview of the pipeline can be found in Figure 3.

Carpentieri [2022] extended the pipeline with two additional measurements for analysis. The two were subject-verb agreement, computed locally (a graphical overview is found in Figure 4) and the correct formation of plurals using a website to look them up (a graphical overview is found in Figure 5). The extended pipeline is shown in Figure 6.

At this stage, the tool consists of a linear pipeline called through the command line and works for speech in German.
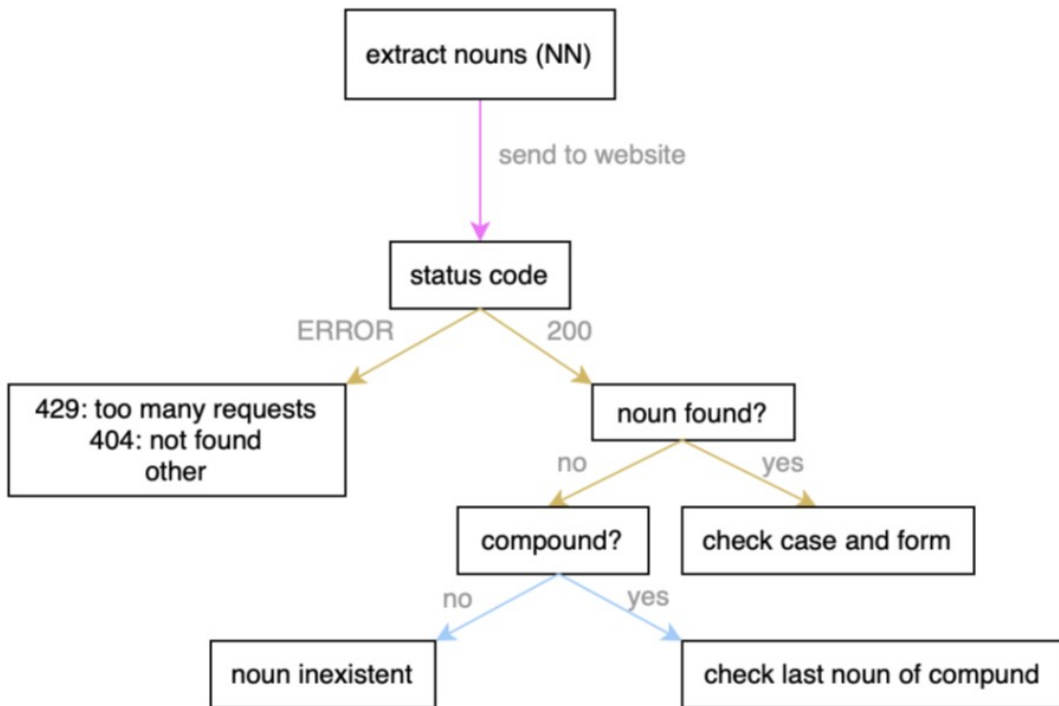
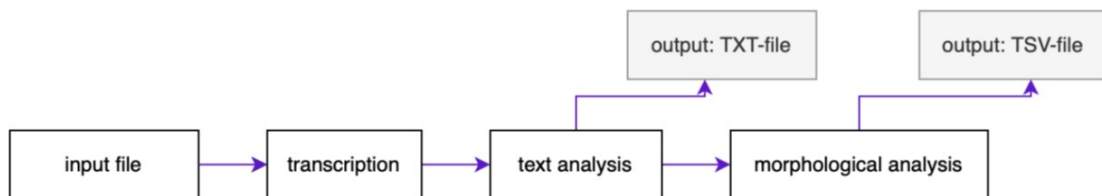Figure 5: Flowchart of the plural formation function by Carpentieri [2022]

Figure 6: An overview of the second pipeline by Carpentieri [2022]

3 Development of the Tool

3.1 Developing a Graphical User Interface

Before this thesis, the tool only existed as a backend architecture operated through the command line. This is not a problem for people used to programming but is a major hurdle for people not used to this. To make the tool more accessible for users a Graphical User Interface (GUI) is needed. This is also important for early engagement with the tool, which can only be further developed with feedback from end users. I developed a prototype of the tool in which I extended the functionality and replaced the linear approach with a more open one.

I decided to build the graphical part of this project in the Python library Tkinter[3]. As a standard library from Python, it is available on all computers which run Python. It is available for Linux as well as MacOS and Windows, which makes the tool usable on all platforms. Tkinter is recommended for starters in GUI development, as the first steps are easy to learn. However, mastering Tkinter to the level needed for such complex software as in this thesis has a steep learning curve.

In the first phase, I developed a GUI for the exact pipeline described in Carpentieri [2022], except for the functionality offered by Prodigy. A dialogue box guided the steps of the pipeline, telling the user to select the file, type in the login to IMS-speech[4] and so on step by step. I soon wanted to add the option to edit an already existing transcript or to annotate a finished transcript. This was not possible in the linear pipeline, which forces the user to go through every step every time it is started. I abandoned this approach for a more flexible one.

In the second phase, I built the tool for non-linear use.

---

[3]`https://docs.python.org/3/library/tkinter.html`

[4]`https://75474978-c3fa-43a5-aa6c-ee36f2513064.ma.bw-cloud-instance.org/ims-speech/`

The tool opens into a text editor, giving the user the choice of whether to transcribe an audio file or whether they want to edit an already existing transcript. The editor shows a txt file and gives the option to create a new file or change an existing file. The transcription is made by the web tool IMS-speech. I implemented a visual login, the tool logs the user in on the website of IMS-speech, uploads the chosen file and downloads it locally when the transcription is finished.

The Application Programming Interface (API) for Prodigy was barely compatible with the functioning of Tkinter. Through a parallel started program, I managed to implement the transcription function of Prodigy. This got redundant when I introduced the text editor within the tool and therefore I removed it. The label annotation of Prodigy used a Prodigy recipe created by Conrad [2021] and it was not possible to bring it in line with the tool.

I added an MP3 player to the tool. This makes it possible to listen to the audio file in parallel when displaying the transcript. For this, I used the music mixer and player from Pygame[5]. The linear pipeline calculated all measurements each time it was started, leaving the user without any choice. I added the option to select single or multiple analyses. The results are displayed in a dialogue box. For the measurements themselves, I used the code by Conrad [2021] and Carpentieri [2022]. This version was presented to the speech therapists in a presentation introducing the tool and giving visual representation for the first time. This made the tool and its possibilities, which were theoretical and abstract up to this point, more comprehensible.

After that, I started an iterative process to refine and extend the existing structure with my ideas as well as feedback from speech therapists. I introduced a more robust underlying data structure to keep track of opened files and all things relating to them (such as paths, corresponding audio files and annotated versions of the text, as well as the calculated measurements).

The transcription process can take a while, which left the tool irresponsive in the first iteration. To signify the running process to the user, I added a progress bar during the process.

The transcriptions coming from IMS-speech are encoded in utterances and contain time stamps. These falsify the linguistic analyses, so I added the option to clean the transcripts to raw text, removing the ims-specific time stamps.

The mixer from Pygame only supports MP3 files, so I added a conversion step for other audio formats to make them playable in the tool as well.

I changed the visual representation, naming labels and structures of the tool repet-

---

[5]https://www.pygame.org/news
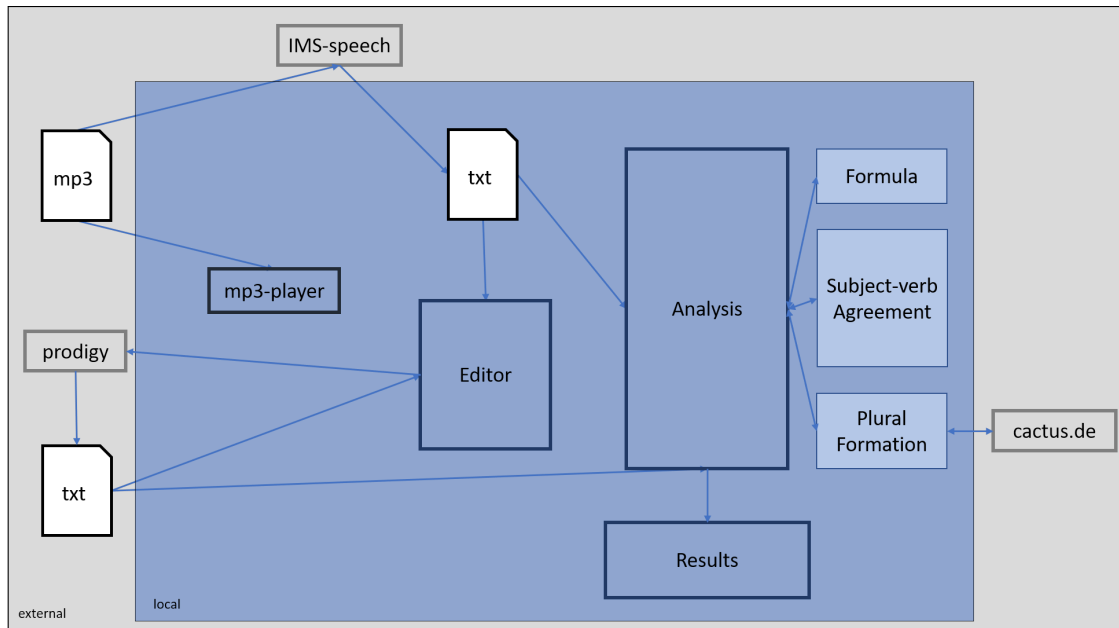
Figure 7: A visual representation of the recent iteration of the tool

itively to achieve a better understanding and easier operation of the tool.

A visual representation of the final tool can be seen in Figure 7. External text files (txt) or audio files (MP3) can be fed into the tool. Other audio files are converted before being fed to the MP3 player, as it can only process MP3 files. Audio files can be played in the MP3 player and an automatic transcript can be created using the service IMS-speech. Text files can be opened and edited in the text editor. An analysis can be done on an opened text file, calling the corresponding function in the script depending on the selected checkboxes. All mathematical measurements are calculated locally, as well as the subject-verb agreement, which calls a decision tree, returning all wrong forms. Plural formation looks up plural forms on a website. All results of the analysis are presented in the tool. All GUI elements which are part of the tool can be seen in Figure 8.

In this process, the tool was presented to the supervisors of DigiSpon regularly and their feedback was implemented, improving the tool step by step. Changes were made in naming different functions more clearly, changing the design of the layout and adding more functions, such as the MP3 player. In these meetings, ideas for further development were also discussed. The biggest request was the possibility of
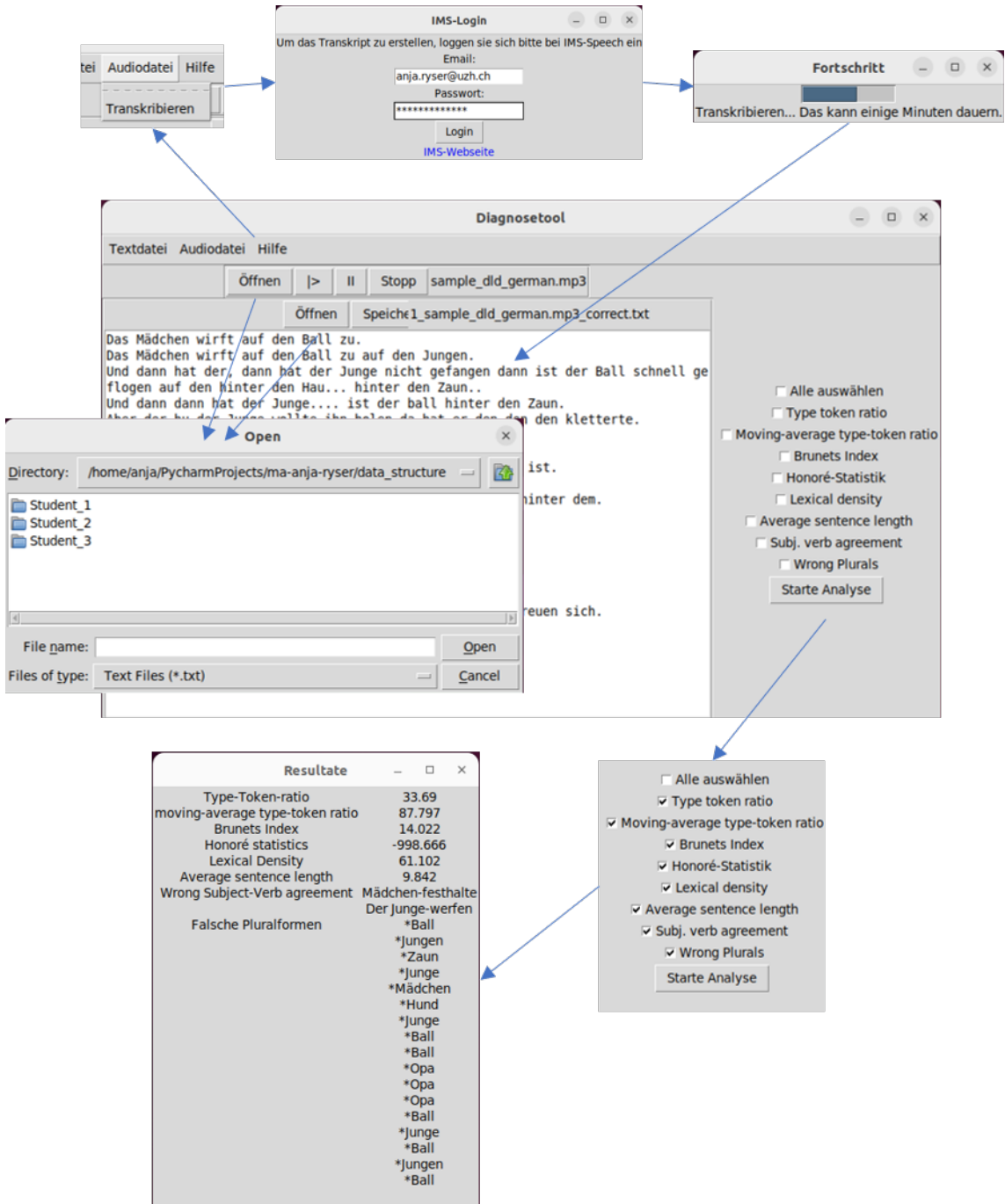
Figure 8: An overview over all graphical elements of the recent tool
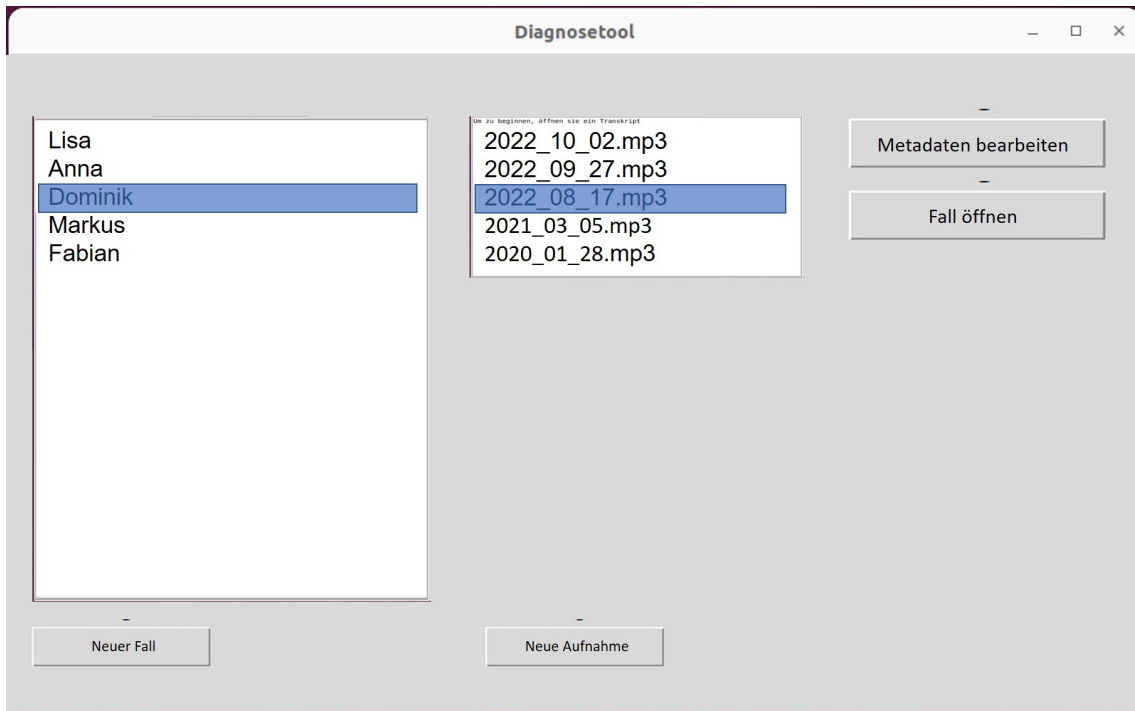
Figure 9: A first conceptualisation of a possible user interface for the case management function

managing cases. It was wished that the tool could sort files by individual children, allowing them to open more than one file at a time and show the development of measurements over time. I created a visual conceptualisation of how this could look (Figure 9). I did not implement it, as this would have gone beyond the scope of this work.

The code created in my thesis can be found at `https://gitlab.uzh.ch/anja.ryser/ma-anja-ryser/`.

3.2 First Tests for ASR

While working on this project, it became clear that the next important step for this tool is the Swiss German ASR to create transcriptions. To test a few existing Swiss German ASR systems, I selected a few example sentences. This gave insight into the performance of these tools on speech from children with DLD. The sentences were selected to be long structures with few errors which were of good audio quality.

Goldstandard: Und de hett sie gseh das Gewitter, schwarzi Wolke, denn sind die alte wäggange, hetter ihm sini Kappe ghebt nid so guet de de murkel, de isch abgheit, hett nid gmerkt

wav2vec-xslr-swiss-german (1): und de hensi bfrög dass s ge wit tüll pwachtsi wolkeregrend si auto wäg gängä he de in ginli kappe het nio gue do der muech kö öl sl au grabgeiche i glächkt

wav2vec-xlsr-swissgerman (2): um diehisung fün de weier wati morkenbegen i antegäng haten intei katepe mit so gut durh anu u desabgabekinge eingemerkt

stt4sg Demo: In dieser Hinsicht haben Sie gespürt, dass die Gewitter schwarze Wolken beeinflussen, Sie haben den Eindruck, es handle sich nicht so gut, dann muss ich sagen:

ims speech: Und daher sind solche Gebiete zwar zu meist verdrängt sie unterwachsen aber in den letzten Jahren nicht so gut dass der Muskel der seine Abgabe kälte nicht macht

Figure 10: An example of the performance of available Swiss German ASR on our data

I tested the Swiss German Kaldi system described in Povey et al. [2011]. The installation process was challenging and needed many different steps on different levels, some in the terminal and others in the folder structure itself. This showed that this system is not suited for use by people without expertise in programming.

I tested the Stt4sg system (speech-to-text for Swiss German) [Plüss et al., 2021]. By the time of this trial, it was not yet published, so I used the online demo version, which was limited to transcribing 15 seconds of speech.

Additionally, I tested two different versions of the Swiss German Wav2vec models published on Huggingface[67].

As a comparison, I also transcribed these sentences with IMS-speech, even though it is only trained on German data.

Looking at this minimal example, none of these systems met expectations and outputs are not usable for further processing. Only one system, one of the Wav2vec systems produced input worth of manual correction. The quality of all other systems would suggest that a manual transcription from scratch is more efficient. Using existing models will not suffice for a reliable result on speech from Swiss German-speaking children with DLD in this project.

The results of all tested ASR systems for a single sentence used as an example can

---

[6]`https://huggingface.co/scasutt/wav2vec2-large-xlsr-52_Swiss_German`

[7]`https://huggingface.co/Yves/wav2vec2-large-xlsr-53-swiss-german`

be seen in Figure 10

3.3 Considerations for Data Collection

As there is no dataset for ASR training with speech data of Swiss German-speaking children with DLD, this data has to be collected. Data collection requires thorough considerations regarding legal and ethical aspects. For this, researchers need to know many different regulations and need to find out what is binding for them. This is not a straightforward process. I researched the most important regulations and ethical guidelines and summarized what is in force for the data collection of DigiSpon and which ethical considerations are needed for the DigiSpon project. The results of this research can be found in Chapter 4.

3.4 Consideration to Metadata

An important part of data collection is to determine what raw data is collected. Additionally to the speech recording of children, it was discussed which metadata should be collected. The collected metadata should fulfil the needs of both, the machine learning process, and the further research by the speech therapists. It was agreed to collect the following data:

- Speaker (labelled with a random pseudonym)

- Age (in years and months)

- School year and class

- Gender

- Information about language development:

    - First language

    - Second/ third language

    - Age of onset for second/ third language

    - Spoken dialect

- Visits speech therapy (Yes/ No)

- Known language or developmental disorders

- Information about the recording context:

  - Recording date

  - Recording duration

  - Elicitation language (German or Swiss German)

  - Elicitation method

  - Pseudonym of eliciting person

- Pseudonym of transcriber

- Pseudonyms of other editors/ analyzers

4 Legal and Ethical Considerations of Data Collection and Research

When starting a new research project, researchers have to navigate the legal and ethical frame in which their work will happen. Navigating between federal, cantonal and university law, keeping within the guidelines and recommendations of guiding ethical instances such as ethical committees as well as navigating their personal ethical values is not always easy. Some aspects are regulated heavily under binding components such as laws and regulations, while others are mere recommendations and important guidelines are in between binding and non-binding. Which law is in force can change depending on the university, research area and funding partners. Researchers have to navigate the small line between ethics and law. Not always is following the law ethical, or following ethical principles legal. This can already be seen in the research of this chapter: I read almost 90 different papers, regulations, guidelines and standards, as well as countless websites from institutions as preparation.

Most regulations operate on very general levels and are written vaguely to make them conform to many different circumstances. When trying to apply them practically to a concrete project, this leaves researchers in the dark and they need to rely on their ethics committees and the university's services to make sure that all rules are fulfilled. This can lead to blindingly trusting and following established ways instead of thinking deeply about ethical influences on the study. This can for example be seen in consent forms. Their purpose should be to inform the participants fully about the research and its impact. While creating these forms, it can be a possibility for the researchers to think about the ethicality of their research beforehand. Unfortunately, because of the need to protect themselves legally and unclear guidelines leading to uncertainties, these forms are reduced to mere bureaucratic procedures formulated in hard-to-understand legal writing, which is not beneficial to participants or researchers.

As the use and ability of machine learning and artificial intelligence grows, it enters more and more research fields. This is a big opportunity, as machine learning algorithms can tackle research questions on big amounts of data, which would not be possible for a human to do manually, such as predicting disease spread through search queries on Google [Lin et al., 2020]. Machine learning algorithms have the potential to solve problems faster and more accurately than humans, which can be of advantage in research, such as pattern recognition in MRI [Morrow and Sormani, 2020]. Even the newest data protection and other laws from the last ten years have struggled to regulate machine learning and artificial intelligence [Gellert, 2022].

It becomes clear that these laws are not suited for this new technology and an adaptation is urgently needed. As this field is changing very fast and innovations such as Chat-GPT by OpenAI (a chatbot with never before seen performances, [OpenAI, 2023]) can change the market within weeks, it is difficult to fix legal requirements to regulate all existing and potential systems in the future. As law takes time to come into force, it happened that when regulations came into force, they were already outdated. To this point, there is no consensus on how to regulate machine learning legally in an ethical way to protect humans.

Projects such as DigiSpon are a good example to show the problems described above: It is a project involving multiple research organisations and research fields which follow different regulations. As data is collected from children, different laws will be taken into account.

It involves machine learning on potentially sensitive data, which is not sufficiently regulated. Navigating all the different laws, regulations, guidelines and recommendations, which this project falls under, is not an easy task, and researchers involved in it will have to invest time and thought in this topic. More about the specifics of DigiSpon and the legal and ethical complications will be described in Section 4.3.

In this chapter, I summarize the most important laws, regulations and guidelines for data collection with a focus on DigiSpon and explain these processes in more general matters. In the end, I adapt these general legal and ethical considerations to DigiSpon. As many of the ethical processes at universities are ruled internally, it is difficult to get information about other institutions. As a part of the University of Zurich, I have access to their processes and will mainly focus on their procedures, even though for this research, others are important as well.

4.1 Legal Considerations

4.1.1 Switzerland and The University of Zurich

In Switzerland, most universities are cantonal bodies (except for national research institutes such as ETH and EPFL) and thus are regulated under the respective canton's data protection laws. In research involving more than one university, data protection laws from all cantons participating in the research should be taken into consideration. To simplify this process, the standard is that the project coordinator's university and canton are guiding and ethics applications are submitted there [Diaz, 2022]. The University of Zurich is regulated by the cantonal data protection law of Zurich (*Gesetz über die Information und den Datenschutz*) (IDG) [Kanton Zürich, 2007]. According to these laws, the ethics committee of the university's faculties and, if necessary, the ethics committee of the canton are evaluating research projects.

The definition of data types is given in the Federal Data Protection Law (DSG) (Article 5, [DSG, 2022]) and applies to the cantonal IDG:

- Data: An umbrella term for all data types that do not concern persons (such as weather data). This type of data can be collected, used and published with no restrictions under data protection law.

- Personal data: All types of data concerning identified or identifiable persons, such as name, phone numbers, credit information etc. This data can only be used for the purpose for which they were collected. They cannot be published without consent and have to be anonymised or deleted after processing. They can only be published or processed for other purposes when they are anonymised sufficiently.

- Sensitive Data ("*Besonders schützenswerte Personendaten*"): All data of a person regarding sensitive topics, such as health, religion, political views and biometric data that can univocally identify a person. This data can harm a person when published or used immorally. This type of data always needs explicit free and informed consent to be collected and processed.

Free and informed consent is also needed when profiling is involved. Profiling describes the automatic processing of data to gain personal information about a person and is used in systems for hiring people, evaluating health and recognizing personal interests in advertisements [DSG, 2022].

In August 2022, the Swiss Federal Council decided on a fully revised version of the data protection law, which strengthens the protection and self-determination of personal data, adapts the law to technological advances and increases transparency in personal data collection [DSG, 2022]. This revision focuses on stricter regulations mostly for profiling and the processing of big data and mainly targets big companies. It will come into effect in September 2023 [DSG, 2022] after an extended period for the economy to adapt its processes to the new law. As the law was formulated a few years ago and the discussion about regulation of AI systems got public recently with the deployment of ChatGPT and GPT-4 [OpenAI, 2023], it remains to be seen whether the law can meet the requirements this new technology is requesting.

One difficulty for universities and researcher is that university falls under the same law as other public institutions and organisations and other federal bodies such as hospitals and local governments. In Zurich, the data protector (*Datenschutzbeauftragter*) publishes leaflets and guidelines for official bodies, which are binding for universities as well. This makes sense in administration, but for research, these leaflets are not suitable. In some of these leaflets, research is explicitly mentioned and has specific guidelines. The University of Zurich and its Legal Department and Data Privacy, in collaboration with the data protector, created internal leaflets for their researchers. In research, the ethics committee mostly is the decisive authority (more to this in section 4.2.1).

If research involves health-related data or if research in social studies or humanities involves data regarding human diseases, it falls under the regulation of the Human Research Act (HRA) (*Humanforschungsgesetz*). In this, all medical research including clinical trials is regulated and the rules are much stricter than in any other research. The definitions used in the HRA can be blurry and lead to uncertainties in research done in other disciplines involving personal data related to health, as to whether it falls under the HRA or not. For example, research regarding the structure, function and development of the human psyche, as done in psychology and educational research, are excluded from the HRA. In edge cases, the ethics committee are the deciding instance. Research falling under the regulations of the HRA always has to be approved by the cantonal ethics committee. Research under this regulation involving children is under even stricter regulation, as children fall into the category of particularly vulnerable people, as they are more sensitive to external influences and are not yet able to understand all the consequences of

their actions. For these reasons, research with children always needs the consent of their caregiver. The refusal of a child is always prioritised over the consent of their caregivers [Schweizer Bundesrat, 2011]. Additionally to the written form of consent and information to the caregivers, the participant needs to be informed age appropriately: Children under 14 are informed orally; children over 14 sign an additional consent form, which is adapted to the understanding of the participant. Research of this kind needs to have a direct use for the participants or result in insights that cannot be gained in other research not involving particularly vulnerable participants [Swissethics and Schweizerische Ethikkommissionen für die Forschung am Menschen, 2018].

Research involving children which does not fall under the regulations of the HRA needs to adhere to similar, but less strict guidelines. The HRA only allows research on children when the child has direct use of the study (such as new treatments or interventions). In research not falling under the HRA, solidarity is a legitimate reason to allow the participation of children in a study. This is possible, as children can think in solidarity with others from a young age. Consent is regulated the same as under the HRA. Research involving under-age participants cannot be financially compensated (other than reimbursement for travel costs etc.), to prevent unjustified incentives to the caregivers, as they need to give consent on behalf of their children [Kleist et al., 2017].

In 2021, in a workshop by the strategy lab of the Digital Society Initiative of the University of Zurich (DSI), members of the DSI and members of the Ad hoc Committee on Artificial Intelligence from the Europe Council (CAHAI) created a position paper requesting the Swiss Federal Council to create guidelines for AI in Switzerland, similar to the European regulations [Thouvenin et al., 2021]. In this, they call for a legal framework that leaves as much room as possible for the development and use of algorithmic systems, to use their benefits for individuals and society. At the same time, algorithmic systems need to be regulated to prevent disadvantages to all concerned, such as discrimination and manipulation of elections and votes. The paper states that concerns of privacy and data protection are sufficiently regulated by existing law. The authors demand better regulations in areas of recognizability and traceability of AI Systems (for example, clearly state when a person is talking to a conversational agent, also known as a chatbot, in customer service), rules against discrimination and manipulation by AI systems and clear liability, as well as data protection and security for training data, similar to the ones for autonomous vehicles and drones. They suggest that political instances need to think about bans

or moratoriums for specific, very dangerous technology, such as mass surveillance (for example facial recognition in public spaces), social credit scoring systems or autonomous surveillance systems with lethal force. The authors acknowledge the similarities to European rights but argue for an active approach of Switzerland to define their regulation in cooperation with research and experts in the field. [Thouvenin et al., 2021]

4.1.2 European Union - GDPR

In 2017, the General Data Protection Regulation (GDPR) came into force in the European Union (EU). It replaced the data protection directive of 1995. The GDPR is binding for all companies in the European Union and the European economic area (EEA), as well as for all companies if any personal data is processed by a service provider inside the EU or the EEA. This means the GDPR's impact spans world wide.

The GDPR regulates all processing of personal and sensitive data, which includes collection, storage and any processing of such data [EU-Parliament, 2016].

Personal data is defined as all data related to a natural person that is identified or identifiable directly or indirectly through this data. If personal data is sufficiently anonymised, it is no longer personal data and is out of the scope of the GDPR. In the GDPR, data is sufficiently anonymised if data is no longer re-identifiable or only with high effort and unlikely means. Personal data can only be processed when the person, to whom the data relates (data subject), gives free and informed consent. Exceptions allow for processing without consent when explicitly allowed by law or when processing ensures the "vital interests of the data subject" (Article 9, EU-Parliament [2016]).

The GDPR and its regulations can be summarized under the following data protection principles [Lagioia et al., 2020]:

- Fairness: The data subject needs to be informed of the data processing and its purpose. When consent is needed, it should communicated in a way that allows free and informed consent.

- Transparency: Information to the data processing is easily and understandably accessible to the public.

- Purpose limitation: Personal data needs to be collected for "specified, explicit and legitimate purposes and not further processed in a manner that is

incompatible with those purposes" (Article 5, EU-Parliament [2016])

- Data minimisation: Only personal data needed for the specified purpose is collected and not more.

- Accuracy: Collected personal data needs to be accurate and, where necessary, be kept up to date. When the collector knows about the inaccuracy of data it needs to be corrected or deleted.

- Storage limitation: Personal data should only be kept as long as its purpose lasts. When the purpose of named at collection is no longer fulfilled, personal data should be destroyed.

The GDPR is one of the strictest data protection laws and breaches of it are followed by devastating penalties for even the biggest companies worldwide. Breaches of the GDPR can, depending on which article is involved, lead to penalties from 10 million or 2% of the annual global turnover up to 20 million or 4% of the annual global turnover. [Gruschka et al., 2018]. Its relevance is much bigger worldwide than possibly assumed and it also heavily impacts data processing and protection in Switzerland. Many Swiss companies and researchers use services operating in the EU, such as cloud services and other data processing, which makes the GDPR binding even within Switzerland. Swiss authorities denied taking over the GDPR as it is, but the close relationship between Switzerland and the EU influenced the data protection law, which is very similar to the GDPR while still being an independent regulation.

As in all EU regulations, the GDPR itself and the definitions within it are formulated very broadly, as to make it applicable to all causes in the present and the future. Additionally, this gives the single member states the opportunity to add regulations to their national law and gives them room for interpretation. But through this, it is difficult to know how exactly it has to be applied to different areas concerned with it. For this, there are papers created by members of the parliament cooperatively with experts in different areas (such as artificial intelligence), which start to interpret and define clearer rules for their respective fields with an expert perspective, which the EU parliament cannot be expected to have. These papers do not have the binding factor such as the law itself; in practice, they become the guiding papers for all judges and lawsuits in the EU, which makes them highly influential. An example of this is Lagioia et al. [2020], where experts in the field of artificial intelligence reformulate the GDPR to suit the needs of this field, focusing mostly on the processing of big data and profiling.

Almost all research considering data protection, machine learning and related topics is done by researchers working under the GDPR. Most considerations can be applied more broadly, as they are general considerations surrounding these topics, however, the details are specific to the GDPR. Even though this thesis focuses on Swiss law, as there is no research to be found on specific Swiss law, these considerations are discussed nevertheless in this thesis.

Lagioia et al. [2020] try to set the GDPR in relationship to AI. AI is not explicitly mentioned in the GDPR, but many regulations are important for the use of it. Article 22 is most important to AI, as it generally prohibits automatic decision-making and profiling, "which produces legal effects concerning him or her or similarly significantly affects him or her" (p.59, Lagioia et al. [2020]). However, they allow a broad range of exceptions to this general prohibition. Automated decision-making is allowed when it is necessary for entering into a contract between the data subject and the data controller (Article 22, Paragraph 2a) when it is allowed by Union or member state law (Article 22, Paragraph 2b) or when there is explicit consent from the data subject (Article 22, Paragraph 2c). These broad exceptions allow automated decision-making in many different areas under specified conditions. When exactly these are met is heavily dependent on interpretation: When is automated decision-making in entering a contract "necessary"? In which form must consent be given? Article 22 makes it necessary that at the end of a decision-making process, the machine-generated output is always finalized by a human. Through the sheer number of decisions and trust in the automated system, it is questionable whether this step is always beneficial or becomes a mere formality. It is imaginable that there are decisions where the system outperforms humans, where it would be counter-productive to overwrite the machine's decision.

A very important question is how data inferred from personal data needs to be classified and whether the models themselves should be handled as personal data. Inference from personal data should be handled as a process of creating data, which means they create personal data that should underlie the same regulations as the training data. Models themselves only contain generalised knowledge, or knowledge about groups of persons with similar features and thus should not count as personal data.

The fundamental data protection principles mentioned above also apply to machine learning, but they can be understood in ways that allow AI with personal data and big data. This includes a more flexible interpretation of purpose limitations, as statistical purposes are generally allowed and most machine learning is statistical,

which allows machine learning on data as long as the purpose is compatible with the one originally used to collect the data. The limiting factor is the risk for the data subject; if it is unacceptable, the use of AI is not allowed. The interpretation of when risk is unacceptable is only very broadly covered in the GDPR. The principle of data minimisation can be interpreted as compatible with AI, as long as it is beneficial to data subjects and/or society. It may require pseudonymisation or anonymisation to allow the processing of data to reduce the ease of linking specific data to a person. Lagioia et al. [2020] argue that re-identification should not be a direct feature of the data itself, but should be defined as a process of creating personal data, which is ruled under the strict law of the GDPR. This would make the ones re-identifying data responsible for the processing of these data and would take pressure away from the original data processor. All in all, the GDPR allows the use of AI on personal data, when sufficient safeguards are in place and when its use is beneficial.

Rights on the side of the data subjects, which are part of the GDPR, including the right to access (Article 15), the right to erasure (Article 17), the right to portability (Article 19) and the right to objection (Article 21), are more difficult to adapt to machine learning. The right to access gives the data subject the right to request information about whether automated decision-making was used and about the logic involved. It is not clear, whether this means that the subject has the right to a detailed personal explanation or just more general information about the process. The right to erasure is the right that all data is deleted after the request of a data subject. It is not fully clear whether that only concerns the data itself or algorithms trained with that data as well. As knowledge in an algorithm is no longer personal data, the right to erasure only concerns the data itself. The erasure of used training data can risk the legitimacy of an algorithm, as the evaluation is dependent on the training data and can be distorted by deleting data points. The right to portability allows the data subject to request all of their data in a portable format and the right to transfer it to another data controller. It is not clear whether that right only covers data actively given to the controller, or whether it also applies to data automatically collected by the controller (such as necessary cookies on websites) and data inferred from other data (through machine learning). The right to object allows the data subject to request termination of the processing of their data. This right should lead data controllers to provide easily accessible and understandable ways to use this right.

One of the most important aspects in the handling of AI in the scope of the GDPR is the weighing of benefit and use against risk and harm. For example, AI can be used to gain insight into health-related issues in a group, such as tracking epidemics and help in contact tracing, as used in the COVID-19 pandemic or in anticipating

risk factors and future illnesses in single patients. When used in a medical setting, this can have major benefits as algorithms can outperform humans in some scenarios [Morrow and Sormani, 2020]. However, the same algorithms can be used by health insurance and employers to filter out weaker candidates, which would harm the same person in a major way. Generally forbidding certain techniques can prevent beneficial uses and generally allowing them can lead to harm. A suggested solution is that, in addition to the general regulations such as the GDPR, there need to be more specific, sectoral regulations that regulate specific sectors, such as medical diagnostics and recruitment.

Based on all these aspects, Lagioia et al. [2020] propose the following policies for the future:

- Specific regulations for AI and guidelines on how to apply general regulations

- Additional regulations for specific sectors

- Guidance through authorities (such as the data protection board and data protection officers) for data controllers as well as data subjects

- Room for discussion for more specific regulations with all parties involved (experts in the field, politicians, data protection authorities, data controllers, data subjects and academia)

- Encourage data controllers to integrate regulations and protection measurements in the design of AI systems itself (e.g. implementing sufficient safeguards or only processing anonymised data or necessary personal data)

Gellert [2022] argues that one of the main problems in the practical implementation of these laws is that the definitions of the basic terms do not match in the law and the field of machine learning. In the law, the term data was defined when humans were the main medium to communicate and process data. Data itself contained all information and thus was the only thing that needed consideration. Until recently, only simple processing of data was done digitally, such as storing, changing, saving, reading and outputting data. In machine learning, it looks different: the focus is not on the data itself but on inference from them. Machine learning generates new data or knowledge from existing data. Suddenly, the separation between data and knowledge is important. Additionally, the operations in machine learning are far more complex than before. Paradigms, such as "the less data is processed the better", are not able to picture the whole reality anymore. In a few steps, machine learning can do highly complex operations on data. Laws such as the GDPR focus solely on the data itself and use broad definitions, where they do not make the distinction between data and knowledge. Through that, whole areas of data processing are

not regulated under the law, for example, whether inferred data is seen as personal data or not. This comes from the lack of knowledge about the technology from the lawmakers, who are not able to sufficiently define the important terms and thus are not able to formulate important regulations. Specific regulation, such as how inferred data should be regulated is left to be interpreted by experts. This leads to documents such as Lagioia et al. [2020], which try to fill these holes but are not legally binding.

4.2 Ethical considerations

In addition to the legal framework a researcher has to navigate, they also have to navigate the ethical aspects of their research. Ethical aspects, contrary to legal aspects, are not necessarily formalised and are highly dependent on the topic of a project and on practical aspects. To make abstract concepts easier to apply practically to concrete areas of research, some organisations formalised their ethical principles as guidelines or charters. Others did explicitly not formalise them to prevent unforeseen restrictions of research. At many universities, the guiding force in ethics is the ethics committee. In law, it is regulated in which cases the approval of an ethics committee is compulsory.

Ethical considerations include balancing different principles. For example, research tries to move toward open access to data, through initiatives such as the FAIR principles (see Section 4.2.4), which is heavily supported by institutions such as the Swiss National Science Foundation (SNSF, see Section 4.2.3). Depending on the research and the data used this can pose new challenges if the protection of the data is contradicting the goal of open research. In this case, researchers, with the help of SNSF and the ethics committee, have to weigh the risks and benefits to find the best solution for research and the research participant.

The Swiss Centre of Expertise in the Social Sciences (FORS) published guides to survey methods and data management. Diaz [2019] defines three ethical principles, on which all research should be based:

- Free and informed consent: Data can only be collected from individuals who have given their consent. Consent should be given fully informed on how the data is collected and processed, as well as the risks and/or consequences of sharing the data. Consent needs to be given freely, without disproportionate

reward or through pressure. This should ensure that participation is voluntary and conscious.

- Respect for Privacy and Confidentiality: Data should, whenever possible, be anonymised to protect the participant and to prevent the information given to a researcher from being published publicly without consent.

- "Do no harm": Research should not cause anyone, especially the participant, any harm. This includes "extreme physical pain or death, but also involves such factors as psychological stress, personal embarrassment or humiliation, or myriad influences that may adversely affect the participants in a significant way" (p.14, Diaz [2019]). It requests risk assessment before doing research and a weighing of benefits and potential harm.

4.2.1 Ethics Committee

At the University of Zurich, the University of Zurich Ethics Commission is responsible for ethical considerations concerning the university as a whole. Its purpose is to "support members of the University in their perception of ethical responsibility in research and teaching" (para. 1, Center of Ethics, University of Zurich), to "promote ethical awareness within the University" (para. 1, Center of Ethics, University of Zurich) and to "represent ethical issues to the public at large" (para. 1, Center of Ethics, University of Zurich). One of its main tasks is the "development of curricula and the integration of ethical issues in teaching" (para. 2, Center of Ethics, University of Zurich).

Each faculty of the University of Zurich has its own ethics committee, which is responsible for ethically reviewing research projects and approving ethics applications if they comply with the guidelines of the respective ethics committee. For the DigiSpon project, only the ethics committee of the Faculty of Arts and Social Sciences is relevant, so I will only focus on this one. The ethics committee of the Faculty of Arts and Social Sciences consists of at least seven members from different fields or research. The majority of the members are professors of the faculty and at least one member comes from another faculty at the University of Zurich. If necessary, the committee may call in an internal or external advisor. Approval of the ethics committee is needed for research, where the participant is at risk of harm or disadvantages due to the study (sharing sensitive information, being exposed to stressful situation etc.), if a particularly vulnerable group is participating (such as minors or people with cognitive disabilities), if the participants are deceived about

the study (false information about the study, observational studies that require ignorance of the observation), or when a funding body for the research requires approval.

Research falling under the scope of the HRA needs to be reviewed and approved by the cantonal ethics committee. The ethics committee Zurich (*Ethikkommission Zürich*) is responsible for applications from the cantons of Glarus, Graubünden, Schaffhausen, Zurich and Liechtenstein. The cantonal ethics committees are part of Swissethics, an umbrella organisation that focuses on harmonising and coordinating the work processes of all members [Kantonale Ethik Komission des Kanton Zürichs].

4.2.2 Ethical Standards

Even though most ethical principles are not formalised or legally binding, for many organisations it makes sense to standardize ethical principles, either in specific research fields or in research in general. As ethics can be very specific to the field it is applied to, there are many different ethical guidelines. Almost all of them are based on broader, more general ones. In this section, I will present a selection of the most important ethical standards related to the project of this thesis.

In Switzerland, all research is unified under the Code of Conduct for scientific integrity [Swiss Academies of Arts & Sciences, 2021]. As it is used in many different fields of research, the codex is written very broadly and generally. It defines basic principles, how to implement these in different aspects of research and how to proceed in cases of violation. The basic principles of the Swiss Code of Conduct are based on the principles of the European Code of Conduct [Hiney, 2022]. The main principles are reliability, honesty, respect and accountability [Swiss Academies of Arts & Sciences, 2021].

The Helsinki Declaration is the ethical base for medical human research. It was created in 1964 and revised multiple times afterwards by the World Medical Association (WMA). It mainly addresses physicians, declaring the duty of physicians to safeguard the health, well-being and rights of patients (Article 4). At the same time, it recognizes that research is an integral part of medicine. In research, the primary goal should be to understand mechanics and develop and improve procedures (Article 6) while minimising the harm done to participants (Article 11) and never putting the collection of knowledge above the rights of an individual (Article

8) [World Medical Association, 1964]. The Helsinki Declaration is the basis of legal documents regarding medical research, including the HRA [Schweizer Bundesrat, 2011].

The American Psychological Association's (APA's) Ethical Principles of Psychologists and Code of Conduct [American Psychological Association, 2003] is another influential ethical code. It consists of an introduction, a preamble, five general principles and specific ethical standards. The five principles are Beneficence and Nonmaleficence, Fidelity and Responsibility, Integrity, Justice and Respect for People's Rights and Dignity. As psychology is a very broad field, including diagnostics and therapy as well as research, this code is kept very generally. Even though it was created for mental health practitioners and researchers in America, it has established itself as a basis and code around the world [American Psychological Association, 2003]. The Swiss ethical guidelines for psychologists adopt part of the APA's code directly or analogously. Other Swiss organisations that do not have a written form of their ethical principles often directly link to the APA, such as the Department of Special Education at the University of Fribourg.

Some organisations decide to formalise more specific rules for their field. One example is the ethics code of the Association of German Sociologists [DGS, 2017]. In it, the more general rules found in documents such as [Hiney, 2022] are written more specifically for the use of sociologists. In this codex, the following principles are set and discussed: integrity and objectivity, the rights of participants, rules for publications, assessments of work from others, professional interactions with students and colleagues and the composition and tasks of the ethics committee [Hiney, 2022]. Similar documents such as these can be found in many different research fields and from different organisations. Other organisations, that do not have their own written ethics code, take over these documents as non-binding guidelines.

4.2.3 Ethical Data Collection in Switzerland

The nature of ethical standards leaves them very general and without explicit suggestions on how to handle different topics. Almost all of them mention the processing of data in research, but their statements do not give concrete suggestions on how to handle data processing in practice. For example, in Swiss Academies of Arts & Sciences [2021], "data management" has its separate subsection. Data should be "stored appropriately" and "in compliance with the relevant regulations" (p. 19,

Swiss Academies of Arts & Sciences [2021]). In these general formulations, legal and ethical aspects blend together without any concrete ways of how to process data [Swiss Academies of Arts & Sciences, 2021]. Researchers have to find their own ways or rely heavily on the judgement of their ethics committee on how to handle this very important topic.

In Switzerland, the Swiss National Science Foundation (SNSF) is one of the most influential organisations in research in addition to their partners, such as higher education institutions (universities, etc.) and others. It was established in 1952 as a private organisation and, as of the end of 2021 funded 5700 projects with over 20,000 researchers in all research fields in Switzerland. Their goal is to allocate public research money in a way that "contributes to the high research standards in Switzerland" (para. 2, Swiss National Science Foundation [b]). In 2021, they allocated 882 million Swiss francs to their projects [Swiss National Science Foundation, b]. Through its influence, the SNSF does not only shape the project directly funded but also research throughout Switzerland. The SNSF requests a long-term Data Management Plan (DMP), for which they provide a leaflet[8] for all their funded projects. In it, researchers should plan the whole processing of the data, from collection and metadata over the processing within the research and short- and long-term storage of them. In the DMP leaflet of SNSF, researchers should answer questions on the 4 topics of "Data collection and documentation", "Ethics, legal and security issues", "Data storage" and "Preservation and Data sharing and Reuse". Even though the ethical standards and documents such as the DMP are not legally binding, as they are compulsory for applicants, they found their way into the whole research field of Switzerland as guiding standards [Swiss National Science Foundation, b].

4.2.4 Open Data and Data Privacy

Recently, the pressure on researchers grew to make their data available to other researchers for easier reusability. In Switzerland, since 2018 the SNSF requests for researchers: "to store the research data [...], to share these data with other researchers, unless they are bound by legal, ethical, copyright, confidentiality or other clauses and to deposit their data and metadata onto existing public repositories [...]" (para. 5, Swiss National Science Foundation [a]) for all its funded research.

---

[8]https://www.snf.ch/SiteCollectionDocuments/DMP_content_mySNF-form_de.pdf

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
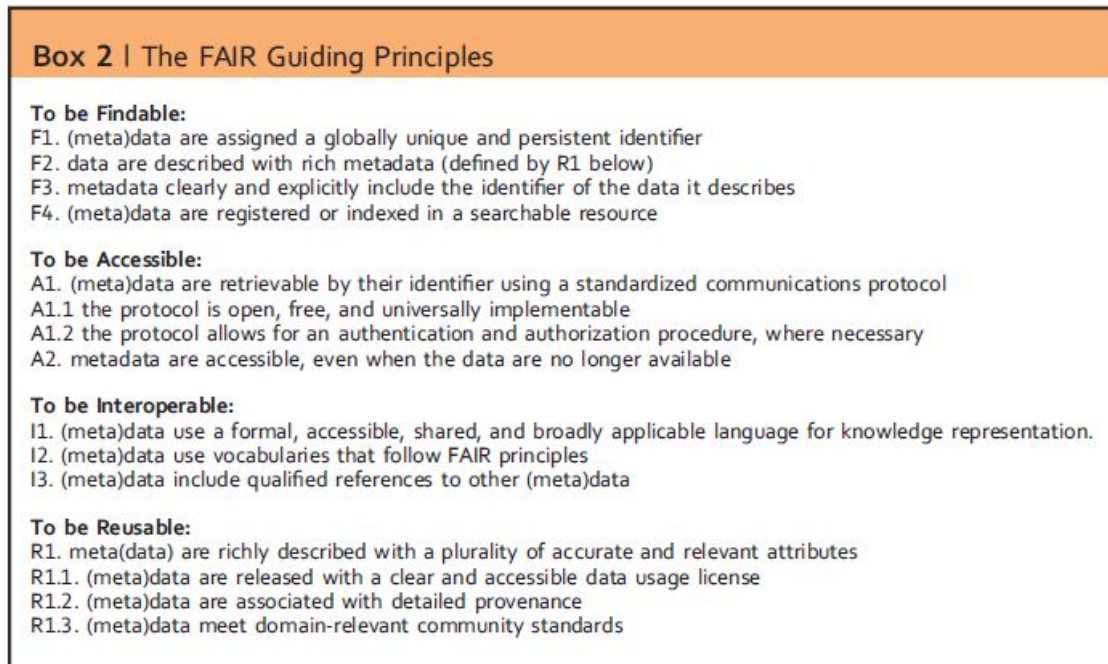R1.3. (meta)data meet domain-relevant community standards

Figure 11: An overview of the FAIR principles from Wilkinson et al. [2016]

At the same time, data protection becomes more important and legal frames for data processing, such as the DSG and the GDPR become stricter. Additionally, the ethical demands on data protection are growing, as more and more data is collected in all areas of life through the Internet and the Internet of Things. The risk of privacy breaches and the harm that they can cause grows. This needs to be taken into account in research, data protection and publishing of research data.

To improve the infrastructure of reusable data in research, make them readable for machines, as well as simplify the process of reusing data, members of academia, industry, funding agencies and publishers of scientific papers determined a set of principles on how to publish data. These are known as the FAIR principles [Wilkinson et al., 2016] and are today standard in many organisations, such as the SNSF and the Swiss code of conduct for scientific Integrity (Chapter 4.5) [Swiss Academies of Arts & Sciences, 2021]. The FAIR principles are targeted at the data itself, but also at the metadata and used algorithms to process the data. For a detailed overlook of the FAIR principles see picture Figure 11.

The SNSF recognizes the fact that not all data are shareable because of "legal,

ethical, copyright, confidentiality or other clauses" (para. 5, Swiss National Science Foundation [a]). If not the whole dataset can be published, data could be shared partially (for example only selected examples), conditionally (researchers need to apply for access to data), or not at all. When it is not possible to share data, mostly it is possible to at least publish the metadata of this research. If researchers are not able to publish their data, they need to explain specifically the circumstances, as to why they cannot share their data. Through this, researchers have the opportunity to reflect on their data and the research practices surrounding them. Through all factors influencing the state of research data, Open Data should be considered as a spectrum rather than a hard category, whereas it should be the goal to publish as much data as possible within legal and ethical considerations [Diaz, 2019].

4.2.4.1 Anonymisation

Many of the legal aspects of data collection can be simplified through the anonymisation of data. As the DSG, as well as the GDPR and the HRA, are only binding for personal data, complete anonymisation frees researchers and other data handlers from their duties under these laws. Additionally, many ethical aspects of data processing are easier when data is anonymised, as there is less harm to the data givers when data cannot be linked to a real person. In data anonymisation, one has to separate between anonymisation and pseudonymisation.

After Stam and Kleiner [2020], anonymisation is defined as "a process by which the elements allowing the identification of a person are definitively removed from the data and the related documentation, such that an individual cannot be identified without significant effort" (p. 3). This corresponds to the legal definition. The two important aspects are that anonymisation is a permanent, irreversible process and that it holds a strong protection threshold.
Pseudonymisation on the other hand is defined by Stam and Kleiner [2020] as "the removal or replacement of identifiers with pseudonyms or codes, where the identifiers are retained separately and secured by technical and organisational measures" (p. 4). As long as the key is kept, pseudonymised data is personal data and underlies data protection law. When the key is destroyed or data is shared without the key, the data counts as anonymised.

The GDPR excludes data that "has been effectively anonymised so that it has lost its connection to particular individuals" (Article 4(1)) from its scope. In the HRA,

anonymised data is defined as "data which cannot (without disproportionate effort) be traced to a specific person" (Article 3i). None of these laws defines when exactly data is sufficiently anonymised or when the effort to deanonymise data is disproportionate. In the last years, many cases have proven that data, which was thought to be anonymous, was re-identifiable, such as the re-identification of Netflix user data with IMDB ratings [Narayanan and Shmatikov, 2008] and the study of De Montjoye et al. [2015], in which they showed that from an anonymised dataset with credit card data, four data points with a known time and place of the transactions are enough to uniquely identify 90% of individuals in the dataset. With more and more datasets available, the risk of cross-identifications such as seen with the Netflix data [Narayanan and Shmatikov, 2008] becomes bigger and a dataset, which counts as anonymous today could be re-identified with another dataset published tomorrow, without the publisher of any of these knowing about it and the risk it bears.

Some data types are easier to anonymise, mostly datasets that can be stored in table format. However, as shown before, even a small amount of data can be re-identified. Other types of data, such as audio and video data are even harder to anonymise due to their structure, such as interview transcripts, which hold information in natural language and video and audio recordings [Bowen [2022], Saunders et al. [2015], Dubagunta et al. [2022]].

Siegert et al. [2020] look at real-world data collection of multi-modal data and how to pseudonymise or anonymise them. Especially in audio and video data, anonymisation is difficult. In audio recordings, one has information about a person on two different levels: First, the voice speaking in a recording is itself identifiable biometric information, which categorizes it as personal data. Second, the spoken content of the recording can contain personal and/or sensitive data. This means, that for sufficient anonymisation, both levels of information need to be anonymised separately. Depending on the project, both or one of the levels contain information needed for the experiment. This means, that anonymising the data leads to a loss of relevant information. Anonymising the content of the recording needs to be done manually, as the information is contained in natural language and thus is not easily and reliably processed by automatic processes. Anonymisation should be irreversible, which is difficult in voice data and only a few techniques are yet developed for that. As this article looks at data collection under the GDPR, it focuses on big private companies using speech data, for example for voice-controlled virtual assistant software. In these scenarios, processing within data protection regulations can be achieved through local processing of the voice data, anonymising them on the mobile device

and only sending anonymised data to the data controller, or through mobile devices sending dummy requests to the provider to add noise to the data.

Bowen [2022] proposes different ways to anonymise a census dataset for different purposes. Ways to anonymise data that he proposes are: suppression (remove data points most likely to be identified completely from the dataset), top and bottom coding (summarize outliers in general groups to protect them) and rounding (summarize different absolute values in categories), add noise (add artificial data points), sample the data (only use or publish a subset) or swap a certain amount of data points. With methods of machine learning, there are examples where a system was trained on the dataset and then generated an augmented dataset, which contains the properties of the dataset without having the data of real persons. All of these methods have their advantages and disadvantages and are not suited for all experiments. Choosing the right anonymisation technique can reduce the risk of re-identification or data breaches heavily.

Saunders et al. [2015] explore ways to anonymise a highly sensitive dataset of interviews with family members of people in a vegetative state. This dataset poses unique challenges, as the community concerned in this topic is very small and interrelated and many professionals know all these families and the families know each other well. Through the limited group and the interlinking, a very small set of information is enough to re-identify an interviewee. As very sensitive topics are discussed (for example families wishing that their relative would die) and many answers are related to shame or even legal consequences it is important to not be able to re-identify the persons involved. Such data points could not be destroyed, as they were highly relevant to the study. When data is unstructured, such as interviews in natural language, the anonymisation process is very time-consuming, as it needs to be done manually. The research team formulated rules, on how to anonymise the data without losing relevant context. They replaced identifying data, such as relationship to the patient, job, religion and location with placeholders, where they were not relevant. When they were relevant, they let them in. To prevent cross-identification through pseudonyms, they changed them in each interview and split up parts of the interviews, which were not relevant to each other but could lead to cross-identification. All in all, their approach was manual anonymisation and they needed to decide on how to proceed on every single data point. When anonymisation was not possible, they asked for consent to share bits of the interview without anonymisation. In interviews with minorities, anonymisation would have erased important details explaining certain experiences. In these cases, the identifying detail,

such as race or religion, was kept. The researchers stated that open communication with the people involved, a flexible option to opt-in, opt-out and different levels of anonymisation for different topics, as well as open communication during and after research and discussion with participants when circumstances changed, were the best ways. This is a very expensive way to anonymise data, however, they open up opportunities with rare and very valuable datasets that could never be published without this process.

In most research, anonymisation is seen as an ethical imperative. However, there are cases where the act of removing personal information about a statement reduces its verifiability or its possible impact, for example in history research, where the name and personal details of a statement giver are a very important factor in their importance in the research, as they deliver context on how the person lived and how they experienced the situation [Diaz, 2019].

All in all, anonymisation is a very important and complex topic. There is no general way to anonymise datasets, which is suited for all circumstances. Anonymisation should be specific and individual for each dataset and should be an important and integral part of considerations around data collection, processing and publishing.

4.2.5 Data Protection in Machine Learning

Looking at research with machine learning, one finds itself with ethical challenges specific to the research area. On one hand, one has to think about the handling of training data, on the other hand, one has to think about the ethical implications of the machine learning model itself. In addition to ethical considerations surrounding machine learning in the present, it makes sense to consider the future. As this field is quickly changing and developing, ethical considerations need to include systems that are not yet developed but are probable to emerge in the next few years.

While the collection and handling of training data is regulated legally, one has to consider ethical aspects in their work. First, when using big data, it is not always possible to get free and informed consent from the users. How can such data be used ethically? In most research projects, data is destroyed when no longer used. This can be harmful to research in machine learning, as these datasets can be important bases for later improvement of the system and are integral parts of the evaluation and quality assurance of systems [Lagioia et al., 2020]. Researchers have to weigh up

what is more important and what produces less risk of harm. Additionally, the reuse of collected data for training other systems is not fully regulated. Repurposing data is allowed when it aligns with the participant's given consent. It is not always easy to decide whether another system is in the scope of the same purpose and whether data is ethically usable for other projects [Future of Privacy Forum, 2018].

Machine learning systems tend to reinforce existing biases from data. This happens when the annotation of the data itself is biased, (for example racial bias in training data for a system assessing the risk of re-offending [Angwin et al.]), which is reinforced and appears in the prediction of systems. Bias can also be created with data that is not itself biased through under-representation. For example, a review of available datasets with images of skin, as well as skin cancer showed that the majority of images were images of white skin and darker skin types were underrepresented [Wen et al., 2022] This lack of examples can bias systems for skin cancer recognition, as they are not able to reliably predict cancer on darker skin types, as they did not see that enough in training.

Some systems produce very sensitive data from personal or even non-personal data. For example, machine learning systems used in medicine can recognize anomalies in MRI scans [Morrow and Sormani, 2020]. These systems generate knowledge from personal data. Highly personal data can be inferred from non-personal data, such as recognizing a teenage pregnancy through buying patterns in Target stores [Hill, 2012] or recognition of depression through language patterns in tweets [Amanat et al., 2022]. The GDPR tries to regulate this data, but there are many edge cases where the legal frame fails and the decisions on how to handle such data are ethical: how big is the risk of misuse of such data and what harm can be done with them and how to prevent this?

When training machine learning systems in areas such as medicine, highly sensitive data, is compulsory for training models. Is it responsible to use this data or should this data be protected and all use should be forbidden? Is a good cause, such as cancer recognition, justification enough to train a system with this data [Datatylsinet, 2018] Weighing between the benefits and the risks of harm and misuse of systems are important considerations that need to be done by researchers. As systems in the medicinal context can produce inferences that can have life-altering consequences, in this field specifically, the users (e.g. the doctors) must understand what the system is doing and be capable users of their systems. This raises the demand for

explainability of the systems and the decisions based on them [Vayena et al., 2018]. In medicine, the doctor-patient relationship is a very important factor in treatment and healing. When using machine learning techniques, developers should make sure that their systems are helpful and efficient without risking the trust and the relationship of patients with their doctors [Aloufi et al., 2021].

The risk of using data is high when training systems on biometric data, such as voice (speech recognition, assistant programs) or faces (facial recognition, emotion detection). This data is particularly vulnerable to data breaches, as they cannot be replaced when leaked, unlike passwords or credit cards. Additionally, biometric data gains importance as an identifying factor in many relevant areas of life, such as access to bank accounts over the phone using voice recognition [Aloufi et al., 2021].

When a system is deployed, more data is needed to improve the model. In a commercial system, data can be gained through real-world users of the system. A user can permit the collection of their data, but in some cases (such as voice assistants on mobile phones) the system will also record people who did not permit the collection of their data (such as conversations in the background) [Aloufi et al., 2021]. The risk also poses itself for users who have given permission to record their data, when the system activates itself accidentally and records sensitive conversations or other data without the user being aware of this. Data recorded from real-world users can be highly valuable, as these are real-life data that are specifically beneficial for this exact system. However, they pose their very own ethical problems of use and risk of harm to users.

4.2.5.1 Approaches to More Ethical Data Collection for Machine Learning

Concerns about privacy with personal training data, as well as concerns of consent on big data and repurposing, can be alleviated by using synthesized data to train the systems [Future of Privacy Forum, 2018]. This technique was for example used with a Bayesian Network to generate artificial VKontakte profiles which produced usable augmented data [Deeva et al., 2020].

To mitigate the risks of training highly sensitive data, systems can be trained with anonymised data. This technique is used for example in training automatic speech recognition. This method leads to a small loss of information, but if trained from the beginning with such data, a system can compensate for this and reach almost the

same performance as systems trained on not anonymised data [Tomashenko et al., 2020]. Through this technique, mobile devices can send collected data back to the data collectors without breaching data privacy. To add additional layers of protection, for example in voice assistant systems, data sent back to the data collector can be cut into small, independent pieces, to not gain access to full private conversations. Another approach is to add noise to the data for example through dummy requests which are randomly generated between real-world data [Siegert et al., 2020].

Using federated learning in systems trained with highly sensitive data can be a solution if sending data to a central provider is not an option or bears big risks. In federated learning, a trained system is distributed to all user devices. After that, these models are further trained locally with specific data. This was used in hospitals for systems which are trained on confidential patient data [Xu et al., 2021]. To further improve the base model, all locally trained models are sent back to the provider of the system punctually. Through this, only generalised information (e.g. the weights of a model) is shared with the provider, whereas confidential patient information is kept locally. The provider updates the base model, whose improved version is sent back to the users, where they are again locally trained. This could be an approach to more secure voice assistant systems on mobile devices, as no personal voice data is shared with the provider [Cui et al., 2021].

4.3 Practical Considerations

After examining the legal and ethical aspects of data collection and its use in machine learning, this section describes the regulations and standards that apply to the data collection of DigiSpon, as well as ethical considerations specifically for our data and the project in general.

4.3.1 Regulations

DigiSpon is a collaboration of the Department of Computational Linguistics at the University of Zurich, the University of Teacher Education in Special Needs (*Interkantonale Hochschule für Heilpädagogik*, HfH) in Zurich, the Department of Special Education at the University of Fribourg, as well as the University of Teacher Education in Berne (*Pädagogische Hochschule Bern*, PHB). Through its collaborative character involving more than one canton, the research falls under the regulations of the project coordinators' canton and institute [Diaz, 2022]. The coordinator is

from the HFH Zürich. This makes the Data Protection Law of Zurich (IDG) binding. Per definition in the IDG and the DSG, voice recordings are biometric, thus personal data, which demands free and informed consent from the participants at collection [DSG, 2022]. As the participants are minors (children from three to six years old), written consent is needed from the caregivers. The children should be informed orally and age-appropriately. If a child shows signs of refusal, this is to be placed above the consent of the caregivers and counts as a veto right of the participant [Kleist et al. [2017], Swissethics and Schweizerische Ethikkommissionen für die Forschung am Menschen [2018]].

Research in medical settings, as well as studies in the humanities and social sciences dealing with human diseases [Bislimi et al., 2009], fall under the scope of the HRA. Whether our research meets this criteria is not clearly defined. Additionally, we are collecting metadata about known speech disorder diagnoses of the children recorded. As Bislimi et al. [2009] acknowledge, it is not always easy to decide whether research is in the scope of the HRA or not. DigiSpon does not test direct intervention on children with speech disorders, but in the first phase only collects speech recordings, which are used for the development of a speech recognition model included in a tool supporting diagnostics in this group. Through this, this research falls in the grey area of the HRA and no guideline can be relied upon for the decision, as this case is nowhere described. It is dependent on the decision of approving instances whether an ethics application by the cantonal ethics committee is needed or not.

The GDPR is not binding, as long as data is not processed by a service in the EU [EU-Parliament, 2016]. At the submission of this work, the transcriptions are made with IMS-Speech from the University of Stuttgart [Denisov and Vu, 2019] and plural forms of nouns are looked up at `https://www.cactus2000.de`. As both of these services are in Germany, the GDPR is binding for all data undergoing this process. IMS-Speech is only able to transcribe files in German but not Swiss German and it is planned to replace it with our own ASR system as soon as possible. As the website to look up the plural forms is not reliable, the possibility is high that both of these tools will no longer be used in the next phase of the project. As this aspect of the project is still in development, one has to keep in mind the implications of using external services and the use of such should be re-evaluated regularly.

If this research were coordinated by the University of Zurich, it would need an application to the ethics committee, as all research involving children at the University

of Zurich needs to be approved by the ethics committee. The HFH Zurich does not have an ethics committee. It is not common to create an application for similar research at the HFH. As the HFH researches children regularly, their projects could be covered by a general ethics application which covers all research at the HFH. In this case, an ethics application is only necessary if the research differs in its procedures from the research covered under the general application.

At the Department for Special Education at the University of Fribourg, it is possible to submit a project to an internal research commission, which evaluates projects and points out potential ethical difficulties. Approvals such as this can later help in peer-reviewed publishing and make sure, that research is compliant with data protection laws. As the recording of audio files can be critical, it is helpful to create a DMP, in which the handling and processing of data is planned in detail and the procedures are documented in written form.

The collection of the data will be carried out in the cantons of Zurich, Fribourg and Berne. Additionally to the differing data protection laws, in all of these cantons data collection in speech therapy and schools is regulated differently. Depending on the canton, such a project needs to be approved by the school board, a cantonal association of speech therapists or similar institutions. This will be done by the corresponding coordinator of data collection at the different institutions before the data collection starts.

Depending on the funding partners of this project, other guidelines could be binding, such as a mandatory DMP, or a request for an ethics application independent of what the research is about. These are not yet known by the time this thesis was written.

As research with children is working with particularly vulnerable participants and thus can bear a higher risk of harm, Swissethics has published thorough guidelines about research with children. They published a general guideline[9], a guideline on how to inform the participants[10] and guidelines on how the HRA is to be applied to children[11].

---

[9]`https://swissethics.ch/assets/pos_papiere_leitfaden/forschung_an_gesunden_minderjaehrigen.pdf`

[10]`https://swissethics.ch/assets/kinder_notfall/leitfaden_pi_kinder_d.pdf`

[11]`https://swissethics.ch/assets/kinder_notfall/kinder_checkliste_d.pdf`

4.3.2 Concrete Ethical Considerations

The fact that the participants in this project are children does not only have legal but also ethical implications. Children are protected more by regulations than adults. However, only following the law does not cover all ethical aspects of research with children. Children's participation is decided by their caregiver, which can be a problem if the wish of the child does not correspond to the decision of the caregiver. The case, in which the caregiver wants the child to participate and the child does not want to, is regulated, as the refusal of the child vetoes the decision of the caregiver [Kleist et al., 2017]. There is no explicit regulation for when a child wants to participate, but the caregiver does not consent. As the consent needs to come from the caregiver, it makes participation impossible. This can be beneficial when the child cannot yet fully understand the consequences and risks of their participation, but it can be disadvantageous when a child wants to participate, for example as a participant of a control group out of solidarity with an ill classmate, but the caregiver does not allow this.

When collecting health data of children, one has to know that possible leakages or misuse of data can have a much bigger impact and harm than in adults. The effect of data breach data related to an impairment or illness could last long into the future and in the worst case influence job applications, insurance evaluations or similar important processes, and could be devastating.

In this research, the majority of the participants have a disordered speech development, which can impact reception and understanding of language. Through this, it is not guaranteed that participants have the same language ability as children with a typical development of the same age. This makes informing the participants appropriately even more challenging and should be kept in mind by the data collectors. It can influence the communication during a recording session, for example when a participant has difficulties communicating that they want to stop their participation.

The recordings collected in this project have personal data on two levels: First, they contain voice data, which is biometric data and thus counts as personal data. Second, they contain spoken conversations with children. As one elicitation method is asking about their family, or what they did the day before, children will tell personal details in these conversations, about themselves and other persons, such as parents, teachers or friends. If a child tells something sensitive, the collectors as well as the data coordinator need to decide which conversations they can keep and which conversations bear too much personal or sensitive information to integrate

them into a dataset. Anonymising recordings after collection through bleeping sensitive information could make the integration of such data in the dataset possible. Less sensitive data would be elicited when using techniques such as retelling a story, or inventing stories through pictures, as the subject of the conversation is not the child and their family itself.

The lack of clarity on whether the HRA is in force or not for this research elicits further ethical questions. It is not always easy to decide whether one should go with the harsher regulations, or whether the less strict regulation is sufficient. Researchers should base their decisions on well-considered ethical questions. If researchers need help or want their decision to be revised, ethics committees and other ethical organisations could help in the decision process, as they have more expertise than most researchers. In this project, we tend towards seeing our data as health-related, which would request ethical approval from the cantonal ethics committee.

The dataset created in this project will be unique, as there is no comparable dataset to this point. We will collect spontaneous speech from children with DLD in Swiss German. This could mean that the general interest in such a dataset could be high. Due to its nature, it is a highly sensitive dataset and cannot be published in its raw form. When preparing such a dataset for publication, one has to invest thorough considerations in how to anonymise such a dataset. Possible approaches could be to anonymise or pseudonymise the voice of the participants, such as suggested by Kai et al. [2022] and Dubagunta et al. [2022]. Other ways to anonymise voice could be the publication of the transcripts only, as done in Saunders et al. [2015]. To protect the information within conversations, a subset of manually selected conversations or only snippets of a few seconds instead of the whole conversation could be published. The most secure but also the most expensive way would be to check every recording manually and anonymise relevant details individually for each recording, such as in Saunders et al. [2015]. If up-front publishing is not possible, restricted access could be thought of, where each access needs to be approved by the creators of the dataset.

Later in the project, we will train an automatic speech recognition model on the collected data. Before training, one should ensure that the dataset is representative and contains enough data on different potential patients diagnosed with this system. For this, the dataset needs to contain enough data of, in the best case, all dialects in Switzerland, children who speak Standard German, children with or without accents, children who do not have Swiss German or German as their first language,

children with many different sociolects as well as a balanced ratio of female and male speakers. Only with diverse data and enough data from minorities can the system reliably produce predictions in all use cases. This is difficult, as we collect atypical data and are limited in resources to collect them. The short-term goal is to collect enough data to train a first baseline system. Questions about representativeness need to be postponed to further phases of this project, but should not be forgotten.

When researching in the area of machine learning, one should always consider the benefits, the harms and the risks of potential misuse of their system. The benefit of such a system would be that practitioners could be supported in their processes, such as transcribing and analysing spontaneous speech of their patients more efficiently and steps in the process, such as transcription, would take significantly less time. It is important that such a tool is not seen as a threat to professionals and it should never take the position of such. It should be a mere gain in efficiency and support in pointing out important factors for a decision. The final decision and diagnoses should always be made by the professional, without influencing it through the system in a non-beneficial way. This should already be ingrained in the concept and design of the software itself.

To make the tool and its improvement more attractive in medical use cases, such as hospitals, approaches of distributed learning, such as silo learning [Cui et al., 2021], could be considered in the future.

5 Conclusion

In this thesis I developed a GUI, bringing the development of an LSA tool for Swiss German-speaking children with DLD a step further. In close collaboration with speech therapists, the tool was improved for a better user experience. I brought the linear pipeline in a more open form, allowing non-linear processing of recordings and transcripts. In the next step data needed for the training of a Swiss German ASR system for children with DLD will be collected.

When collecting data, researchers have to follow many different legal regulations and ethical guidelines. Data will be collected, namely audio recordings of Swiss German-speaking children with and without DLD, which are in the first phase manually transcribed. As this project is intercantonal and interdisciplinary, the project coordinator's location is relevant. For this research, the data protection law of the canton of Zurich is binding. Personal data is collected, which needs free and informed consent from the participants. As they are children, the consent of parents is needed as well, and the children are to be informed age-appropriately. It is not fully clear whether this research falls under the HRA, in this case, the decision on the research application is awaited. The tool includes services located in the EU, which means the stricter GDPR is in force. Before beginning data collection, the coordinators in the cantons where data is collected need to be allowed to do so by cantonal bodies such as school boards and similar.

Children are an especially vulnerable group, which leads to ethical consideration regarding consent, assessing consequences and risks. Possible misuse of data can have long-lasting effects. As the participants are children with DLD, one has to pay close attention that the participants understand the study. Voice recordings have two different levels of personal information, the voice itself as well as the content of the utterances. Both levels need to be protected. As this dataset is unique, its publishing could benefit many researchers. Different anonymisation techniques were discussed, which would make this possible. When doing research in machine learning, one has to consider the representativeness and fairness of a system. This

can be made by the thoughtful creation of a dataset, ensuring that it is diverse and minorities are represented sufficiently. Additionally, when doing research, one has to weigh potential benefits and risks for all parties involved.

6 Future Research

DigiSpon is in its starting phase, leaving many open ways for further development.

The imminent next step in this project is the preparation for data collection. A functionality for recording speech, as well as a way to collect metadata and an organized way to manage files, would make the tool more attractive to users in the first phase of the project. To make sure that the tool is usable by professionals without coding skills for the intended purpose, the tool needs to go through a process of user testing. Professionals, who will be collecting the data, are not familiar with the details of the data collection and its requirements, the creation of clear and easy instructions regarding the tool and the data collection is important.

Data collection takes time, during which further development of the tool can be done simultaneously. Getting access to more training data can improve the future fine-tuning of the ASR tool. Potential data that can be used is: the kidsTALC database [TALC, c], data collected by EmTik [Isler, 2022] as well as Swiss German adult speech data such as the Swiss Parliament Corpus [Plüss et al., 2020].

The tool is constructed for German data. All functionalities need to be replaced with Swiss German models. For this, the NLP tool kit [Hollenstein and Aepli, 2015] and the dependency parser [Aepli, 2018] by Noemi Aepli can be used.

Extending the tool to be able to make annotations, such as labelling different parts of the transcripts and creating grammatical structure representations (such as dependency trees) could broaden the potential use of the tool. Most of these analyses can be done automatically, which leads to the need for an easy way to correct the automated output.

Further, the linguistic analysis can be expanded. Right now, the analysis only

consists of basic measurements. Other studies similar to this project suggested a wide range of analyses that proved to help diagnose DLD [Gabani et al. [2011], Hassanali et al. [2012]]. These include:

- Number of unique words in 100 words

- Numbers of interventions from the eliciting person

- Measuring mazes (including repetitions, corrections and restarts of sentences and filler words)

- Likelihood of a sentence given a trained language model

- Readability measures such as the Flesh-Kincaid score

- Measurements of coherence

- Analysis of the narrative structure

- Measurements for the quality of text

To raise the willingness of the intended user to use the tool, it could be made more appealing. This includes the launch of it as a web application, which has no need for local installation and makes the user interface easier to use. Another possibility is a visual representation of the analyses, including colour coding POS tags and error types, automatically highlighting errors in speech and providing standard scores for analyses to compare results to.

If the first phase of data collection is concluded, a Swiss German ASR system will be trained. Using the same data, the Swiss German NLP tools could be fine-tuned to the specific data of Swiss German children with DLD, which will contain atypical language structures and many errors, which causes difficulties for existing methods of POS-tagging and dependency-parsing.

In the far future, the tool could be enhanced by automated speaker recognition and noise identification during transcription, adding analyses of the audio file, such as pause rate, speaking tempo and prosodic features such as the range of the fundamental frequency, and adding analysis tool over multiple files from the same child, making tracking of progress over time possible.

Using the data collected with all the additional analytics could also be used to train a text classification system, that automatically recognises features of DLD in children.

The more data is collected, the more robust the model will be for different dialects and ages of children, increasing performance overall. In the next step, analysis for other impairments with linguistic markers, such as ADHD and ASD could be added.

References

Schweizerisches Datenschutzgesetz. *Neues Datenschutzrecht ab 1. September 2023*, Aug 2022. URL `https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-90134.html`.

N. Aepli. *Parsing approaches for swiss german.* PhD thesis, University of Zurich, 2018.

R. Aloufi, H. Haddadi, and D. Boyle. Configurable privacy-preserving automatic speech recognition. *arXiv preprint arXiv:2104.00766*, 2021.

A. Amanat, M. Rizwan, A. R. Javed, M. Abdelhaq, R. Alsaqour, S. Pandya, and M. Uddin. Deep learning for depression detection from textual data. *Electronics*, 11(5):676, 2022.

American Psychological Association. Ethical principles of psychologists and code of conduct, june 2003. URL `https://www.apa.org/ethics/code`. last access 22.05.2023.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications.

Y. Arabskyy, A. Agarwal, S. Dey, and O. Koller. Dialectal speech recognition and translation of swiss german speech to standard german text: Microsoft's submission to swisstext 2021. *arXiv preprint arXiv:2106.08126*, 2021.

A. AWMF. Leitlinien der Deutschen Gesellschaft für Phoniatrie und Pädaudiologie - Sprachentwicklungsstörungen bei Kindern, october 2010. URL `https://register.awmf.org/assets/guidelines/049-006_S1_Sprachentwicklungsstoerung_lang_09-2008_09-2013.pdf`. last access 22.05.2023.

G. Barthel, D. Djundja, M. Meinzer, B. Rockstroh, and C. Eulitz. Aachener Sprachanalyse (ASPA): evaluation bei Patienten mit chronischer Aphasie. *Sprache· Stimme· Gehör*, 30(03):103–110, 2006.

Berufsverband Deutscher Psychologinnen und Psychologen. Interdisziplinäre S2k-Leitlinie - Diagnostik von Sprachentwicklungsstörungen (SES), unter Berücksichtigung umschriebener Sprachentwicklungsstörungen (USES), december 2011. URL `https://www.dgpp.de/cms/media/download_gallery/S2k-LL-SES.pdf`. last access 22.05.2023.

D. V. Bishop. Grammatical errors in specific language impairment: Competence or performance limitations? *Applied Psycholinguistics*, 15(4):507–550, 1994.

D. V. Bishop, B. Clark, G. Conti-Ramsden, C. F. Norbury, and M. J. Snowling. Ralli: An internet campaign for raising awareness of language learning impairments. *Child Language Teaching and Therapy*, 28(3):259–262, 2012.

D. V. Bishop, M. J. Snowling, P. A. Thompson, T. Greenhalgh, and C. Consortium. Catalise: A multinational and multidisciplinary delphi consensus study. identifying language impairments in children. *PLOS one*, 11(7): e0158753, 2016.

D. V. Bishop, M. J. Snowling, P. A. Thompson, T. Greenhalgh, C.-. Consortium, C. Adams, L. Archibald, G. Baird, A. Bauer, J. Bellair, et al. Phase 2 of catalise: A multinational and multidisciplinary delphi consensus study of problems with language development: Terminology. *Journal of Child Psychology and Psychiatry*, 58(10):1068–1080, 2017.

R. Bislimi, I. Bischofberger, J. Drewe, R. L. Galeazzi, A. Kesselring, C. Kind, C. Rehmann-Sutter, M. Salathé, and D. Sprumont. Forschung mit Menschen: ein Leitfaden für die Praxis, 2009.

C. M. Bowen. The art of data privacy. *Significance*, 19(1):14–19, 2022.

S. Carpentieri. Computer-assisted diagnostics of developmental language disorder in german-speaking children. 2022.

Center of Ethics, University of Zurich. University of zurich ethics commission. URL `https://www.ethik.uzh.ch/en/ethikkommission.html`. last access 22.05.2023.

H. Clahsen, S. Bartke, and S. Göllner. Formal features in impaired grammars: A comparison of english and german sli children. *Journal of Neurolinguistics*, 10 (2-3):151–171, 1997.

J. Clegg, C. Hollis, L. Mawhood, and M. Rutter. Developmental language disorders–a follow-up in later adult life. cognitive, language and psychosocial outcomes. *Journal of child psychology and psychiatry*, 46(2):128–149, 2005.

S. Conrad. Toward automatic diagnosis of medical conditions involving speech impairment. 2021.

G. Conti-Ramsden, P. L. Mok, A. Pickles, and K. Durkin. Adolescents with a history of specific language impairment (sli): Strengths and difficulties in social, emotional and behavioral functioning. *Research in developmental disabilities*, 34 (11):4161–4169, 2013.

X. Cui, S. Lu, and B. Kingsbury. Federated acoustic modeling for automatic speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6748–6752. IEEE, 2021.

Datatylsinet. Artificial intelligence and privacy. online, last access 22.05.2023, january 2018. URL `https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf`.

Y.-A. De Montjoye, L. Radaelli, V. K. Singh, and A. Pentland. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347 (6221):536–539, 2015.

I. Deeva, P. D. Andriushchenko, A. V. Kalyuzhnaya, and A. V. Boukhanovsky. Bayesian networks-based personal data synthesis. In *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good*, pages 6–11, 2020.

P. Deevy and L. B. Leonard. The comprehension of wh-questions in children with specific language impairment. 2004.

P. Denisov and N. T. Vu. Ims-speech: A speech to text tool. *arXiv preprint arXiv:1908.04743*, 2019.

D. DGS. Ethik-kodex der Deutschen Gesellschaft für Soziologie (DGS) und des Berufsverbandes Deutscher Soziologen (BDS), june 2017. URL `https://soziologie.de/dgs/ethik/ethik-kodex`. last access 22.05.2023.

P. Diaz. Ethics in the era of open research data: some points of reference. *FORS Guide*, 3, 2019.

P. Diaz. Data protection: legal considerations for research in switzerland. *FORS Guide*, 17, 2022.

S. P. Dubagunta, R. J. van Son, and M. M. Doss. Adjustable deterministic pseudonymization of speech. *Computer Speech & Language*, 72:101284, 2022.

H. Ehlert, E. Beaulac, M. Wallbaum, C. Gebauer, L. Rumberg, J. Ostermann, and U. Lüdtke. Collecting and annotating natural child speech data–challenges and interdisciplinary perspectives. In *Konferenz Elektronische Sprachsignalverarbeitung*, pages 72–78. TUDpress, Dresden, 2023.

EU-Parliament. European general data protection regulation, may 2016. URL `https://gdpr-info.eu/`.

S. Fricke, C. Bowyer-Crane, A. J. Haley, C. Hulme, and M. J. Snowling. Efficacy of language intervention in the early years. *Journal of Child Psychology and Psychiatry*, 54(3):280–290, 2013.

Future of Privacy Forum. The privacy expert's guide to artificial intelligence and machine learning. online, last access 22.05.2023, october 2018. URL `https://iapp.org/media/pdf/resource_center/FPF_Artificial_Intelligence_Digital.pdf`.

K. Gabani, M. Sherman, T. Solorio, Y. Liu, L. Bedore, and E. Pena. A corpus-based approach for the prediction of language impairment in monolingual english and spanish-english bilingual children. In *Proceedings of human language technologies: the 2009 annual conference of the North American chapter of the Association for Computational Linguistics*, pages 46–55, 2009.

K. Gabani, T. Solorio, Y. Liu, K.-n. Hassanali, and C. A. Dollaghan. Exploring a corpus-based approach for detecting language impairment in monolingual english-speaking children. *Artificial Intelligence in Medicine*, 53(3):161–170, 2011.

C. Gebauer, L. Rumberg, and J. Ostermann. Pronunciation modeling for children's speech. In *Conference on Electronic Speech Signal Processing*, Munich, 2023.

R. Gellert. Comparing definitions of data and information in data protection law and machine learning: A useful way forward to meaningfully regulate algorithms? *Regulation & governance*, 16(1):156–176, 2022.

J. Gilkerson and J. A. Richards. The lenatm developmental snapshot. *LENA Foundation: Boulder, CO, USA*, 2008.

H. Grimm. Störungen der Sprachentwicklung (2., überarbeitete Auflage). *Göttingen: Hogrefe*, 2003.

N. Gruschka, V. Mavroeidis, K. Vishi, and M. Jensen. Privacy issues and data protection in big data: a case study analysis under gdpr. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 5027–5033. IEEE, 2018.

K.-n. Hassanali, Y. Liu, and T. Solorio. Evaluating nlp features for automatic prediction of language impairment using child speech transcripts. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

J. J. Heilmann, J. F. Miller, and A. Nockerts. Using language sample databases. 2010.

K. Hill. How target figured out a teen girl was pregnant before her father did. *Forbes, Inc*, 7:4–1, 2012.

M. Hiney. Der Europäische Verhaltenskodex für Integrität in der Forschung. *Katharina Miller, Milena Valeva, Julia Prieß-Buchheit (Hrsg.)*, page 47, 2022.

N. Hollenstein and N. Aepli. A resource for natural language processing of swiss german dialects. 2015.

D. Isler. Erwerbsunterstützung mündlicher Textfähigkeiten im Kindergarten(EmTiK) - Projektbeschrieb, 2022. URL `https://www.phtg.ch/forschung/organisation/forschungsabteilung/aktuelle-projekte/emtik/`. last access 29.05.2023.

H. Kai, S. Takamichi, S. Shiota, and H. Kiya. Lightweight and irreversible speech pseudonymization based on data-driven optimization of cascaded voice modification modules. *Computer Speech & Language*, 72:101315, 2022.

S. Kannengieser. *Schweizer Zeitschrift für Psychiatrie & Neurologie*, 2013.

Kanton Zürich. Gesetz über die Information und den Datenschutz (IDG), february 2007. URL `http://www2.zhlex.zh.ch/appl/zhlex_r.nsf/WebView/A6BC9ADB1514D7C1C12588C300321D76/$File/170.4_12.2.07_118.pdf`.

Kantonale Ethik Komission des Kanton Zürichs. Zuständigkeit der Kantonalen Ethikkommission. URL `https://www.zh.ch/de/gesundheit/ethik-humanforschung/zustaendigkeit-kantonale-ethikkommission.html`. last access 22.05.2023.

C. Kauschke, C. Lüke, A. Dohmen, A. Haid, C. Leitinger, C. Männel, et al. Delphi-Studie zur Definition und Terminologie von

Sprachentwicklungsstörungen.–eine interdisziplinäre Neubestimmung für den deutschsprachigen Raum. *Logos*, 31(1):2–20, 2023.

T. Kew, I. Nigmatulina, L. Nagele, and T. Samardzic. Uzh tilt: A kaldi recipe for swiss german speech to standard german text. In *SwissText/KONVENS*, 2020.

C. Kiese-Himmel. Sprachentwicklungsstörung, spezifische (umschriebene) im Dorsch Lexikon der Psychologie, 2022a. URL `https://dorsch.hogrefe.com/stichwort/sprachentwicklungsstoerung-spezifische-umschriebene`. [Online; Stand 25.05.2023].

C. Kiese-Himmel. Früherkennung primärer Sprachentwicklungsstörungen–zunehmende Relevanz durch Änderung der Diagnosekriterien? *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz*, 65(9):909–916, 2022b.

P. Kleist, S. Driessen, and P. Gervasoni. Leitlinie zur Forschung mit gesunden Kindern und Jugendlichen. Leitlinien von Swissethics, 2017. Last access 22.05.2023.

B. Kolonko and T. Seglias. Ältere Kinder und Jugendliche mit Spracherwerbsstörungen. *Forschungsbericht. HfH, Zürich*, 2004.

F. Lagioia et al. The impact of the general data protection regulation (gdpr) on artificial intelligence. 2020.

M. Lahey and J. Edwards. Naming errors of children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, 42(1):195–205, 1999.

H. Liao, G. Pundak, O. Siohan, M. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani. Large vocabulary automatic speech recognition for children. 2015.

Y.-H. Lin, C.-H. Liu, and Y.-C. Chiu. Google searches for the keywords of "wash hands" predict the speed of national spread of covid-19 outbreak among 21 countries. *Brain, behavior, and immunity*, 87:30–32, 2020.

U. Lüdtke, J. Bornman, F. de Wet, U. Heid, J. Ostermann, L. Rumberg, J. Van der Linde, and H. Ehlert. Multidisciplinary perspectives on automatic analysis of children's language samples: Where do we go from here? *Folia Phoniatrica et Logopaedica*, 75(1):1–12, 2023.

B. MacWhinney. *The CHILDES Project: Tools for analyzing talk. transcription format and programs*, volume 1. Psychology Press, 2000.

J. Miller, R. Chapman, et al. Systematic analysis of language transcripts. *Madison, WI: Language Analysis Laboratory*, 1985.

A. P. Monaco. Multivariate linkage analysis of specific language impairment (sli). *Annals of Human Genetics*, 71(5):660–673, 2007.

J. M. Morrow and M. P. Sormani. Machine learning outperforms human experts in mri pattern analysis of muscular dystrophies, 2020.

A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

I. Nigmatulina, T. Kew, and T. Samardzic. 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, 2020.

OpenAI. Gpt-4 technical report, 2023.

J. Paradis, M. Crago, F. Genesee, and M. Rice. French-english bilingual children with sli: How do they compare with their monolingual peers?(vol 46, pg 122, 2003). *Journal of Speech Language and Hearing Research*, 46(2):404–404, 2003.

Z. Penner. Plädoyer für eine präventive Frühintervention bei Kindern mit Spracherwerbsstörungen. *Therapie von Sprachentwicklungsstörungen. Anspruch und Realität, Kohlhammer, Stuttgart*, 2002.

M. Plüss, L. Neukom, C. Scheller, and M. Vogel. Swiss parliaments corpus, an automatically aligned swiss german speech to standard german text corpus. *arXiv preprint arXiv:2010.02810*, 2020.

M. Plüss, L. Neukom, and M. Vogel. Swisstext 2021 task 3: Swiss german speech to standard german text. In *Proceedings of the Swiss Text Analytics Conference*, volume 2021, 2021.

D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society, 2011.

Raising awareness of developmental language disorders.
Sprachentwicklungsstörung (Ses) – Die Fakten, september 2020. URL `https://radld.org/wp-content/uploads/2020/09/DLD-Fact-Sheet-German.pdf`.
[Online; Stand 22.05.2023].

S. M. Redmond. Conversational profiles of children with adhd, sli and typical development. *Clinical linguistics & phonetics*, 18(2):107–125, 2004.

J. Reilly, M. Losh, U. Bellugi, and B. Wulfeck. "frog, where are you?" narratives in children with specific language impairment, early focal brain injury, and williams syndrome. *Brain and language*, 88(2):229–247, 2004.

L. Rumberg, H. Ehlert, U. Lüdtke, and J. Ostermann. Age-invariant training for end-to-end child speech recognition using adversarial multi-task learning. In *Interspeech*, pages 3850–3854, 2021.

L. Rumberg, C. Gebauer, H. Ehlert, U. Lüdtke, and J. Ostermann. Improving phonetic transcriptions of children's speech by pronunciation modelling with constrained ctc-decoding. In *Proc. Interspeech*, pages 1357–1361, 2022.

SALT. Salt software - products. URL `https://www.saltsoftware.com/products`. last access 29.05.2023.

B. Saunders, J. Kitzinger, and C. Kitzinger. Anonymising interview data: Challenges and compromise in practice. *Qualitative research*, 15(5):616–632, 2015.

F. Schiel, C. Draxler, and J. Harrington. Phonemic segmentation and labelling using the maus technique. 2011.

T. Schlett, P. Mäder, A. Frank, and T. Günther. Vergleich von verschiedenen Varianten der Spontansprachanalyse bei der Diagnostik von Kindern mit Aussprachestörungen und/oder Dysgrammatismus. *Sprache· Stimme· Gehör*, pages 37–41, 2013.

Y. Schraner, C. Scheller, M. Plüss, and M. Vogel. Swiss german speech to text system evaluation. *arXiv preprint arXiv:2207.00412*, 2022.

Schweizer Bundesrat. Bundesgesetz über die Forschung am Menschen, Humanforschungsgesetz, september 2011. URL `https://www.fedlex.admin.ch/eli/cc/2013/617/de`. last access 22.5.2023.

I. Siegert, V. S. Varod, N. Carmi, and P. Kamocki. Personal data protection and academia: Gdpr issues and multi-modal data-collections. *Online Journal of Applied Knowledge Management (OJAKM)*, 8(1):16–31, 2020.

J. Siegmüller, H. Bartels, and L. Höppe. *Leitfaden Sprache Sprechen Stimme Schlucken*. Elsevier Health Sciences, 2022.

D. V. Smith, A. Sneddon, L. Ward, A. Duenser, J. Freyne, D. Silvera-Tawil, and A. Morgan. Improving child speech disorder assessment by incorporating out-of-domain adult speech. In *Interspeech*, pages 2690–2694, 2017.

G. Snider. Comparison of language sample analysis procedures. Poster. URL `https://www.eiu.edu/studentresearch/posters/snider.pdf`. last access: 29.05.2023.

T. Solorio. 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, 2013.

T. Solorio and Y. Liu. Using language models to identify language impairment in spanish-english bilingual children. In *Proceedings of the workshop on current trends in biomedical natural language processing*, pages 116–117, 2008.

M. C. St Clair, A. Pickles, K. Durkin, and G. Conti-Ramsden. A longitudinal study of behavioral, emotional and social difficulties in individuals with a history of specific language impairment (sli). *Journal of communication disorders*, 44(2): 186–199, 2011.

A. Stam and B. Kleiner. Data anonymisation: legal, ethical, and strategic considerations, 2020.

Swiss Academies of Arts & Sciences. Swiss code of conduct for scientific integrity, 2021. URL `https://api.swiss-academies.ch/site/assets/files/25607/kodex_layout_en_web-1.pdf`. last access 22.05.2023.

Swiss National Science Foundation. Open research data, a. URL `https://www.snf.ch/en/dMILj9t4LNk8NwyR/topic/open-research-data`. last access 22.05.2023.

Swiss National Science Foundation. Profile of the snsf, b. URL `https://www.snf.ch/en/GrjwOKMdGiigVhgY/page/theSNSF/profile`. last access 22.05.2023.

Swissethics and Schweizerische Ethikkommissionen für die Forschung am Menschen. Forschung an und mit Kindern und Jugendlichen ¡ 18 Jahren - Leitfaden zur Studieninformation. Leitfaden von Swissethics online, april 2018. Last access 22.05.2023.

TALC. Talc-lsa (language sample analysis), a. URL
https://www.leibnizlab-communication.uni-hannover.de/de/forschung/
projekte/talc/projektbeschreibung. last access 29.05.2023.

TALC. Talc-dira (data integrated reading assessment), b. URL
https://www.leibnizlab-communication.uni-hannover.de/de/forschung/
projekte/talc/projektbeschreibung-talc-dira. last access 29.05.2023.

TALC. KIDSTALC DATENBANK, c. URL
https://www.leibnizlab-communication.uni-hannover.de/de/forschung/
projekte/talc/kidstalc-datenbank. last access 29.05.2023.

TALC. Talc - tools for analyzing language and communication, d. URL
https://www.leibnizlab-communication.uni-hannover.de/de/forschung/
projekte/talc. last access 29.05.2023.

F. Thouvenin, M. Christen, A. Bernstein, N. Braun Binder, T. Burri, K. Donnay,
L. A. Jäger, M. Jaffé, M. Krauthammer, M. Lohmann, et al. Positionspapier:
Ein Rechtsrahmen für Künstliche Intelligenz. *Workshop of the DSI (Digital
Society Initiative) Strategy Lab, Balsthal, 26 August 2021 - 28 August 2021*,
2021.

C. Till, J. Winkes, and E. Hartmann. Diagnostik des Satzverständnisses bei
Deutschschweizer Kindern mit und ohne Sprachentwicklungsstörung. *Forschung
Sprache E-journal*, 2017.

N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch,
J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, et al.
Post-evaluation analysis for the voice privacy 2020 challenge: Using anonymized
speech data to train attack models and asr, 2020.

J. B. Tomblin, N. L. Records, P. Buckwalter, X. Zhang, E. Smith, and M. O'Brien.
Prevalence of specific language impairment in kindergarten children. *Journal of
speech, language, and hearing research*, 40(6):1245–1260, 1997.

E. Vayena, A. Blasimme, and I. G. Cohen. Machine learning in medicine:
addressing ethical challenges. *PLoS medicine*, 15(11):e1002689, 2018.

W. Von Suchodoletz. *Sprech-und Sprachstörungen*. Hogrefe Verlag GmbH &
Company KG, 2013.

D. Wen, S. M. Khan, A. J. Xu, H. Ibrahim, L. Smith, J. Caballero, L. Zepeda,
C. de Blas Perez, A. K. Denniston, X. Liu, et al. Characteristics of publicly

available skin cancer image datasets: a systematic review. *The Lancet Digital Health*, 4(1):e64–e74, 2022.

M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

World Medical Association. Declaration of helsinki – ethical principles for medical research involving human subjects, june 1964. URL `https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-hu` last access 22.05.2023.

J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5:1–19, 2021.

S. Zauke and S. Neumann. Die kommunikative Partizipation von Kindern im Vorschulalter mit Sprachentwicklungsstörungen (S) SES-Erste Ergebnisse anhand des FOCUS©-G. *logopädieschweiz*, pages 15–25, 2019.

Lebenslauf

Persönliche Angaben

Anja Ryser

Kirchstrasse 17

8214 Gächlingen

anja.ryser@uzh.ch

Schulbildung

2012-2016	Kantonsschule Schaffhausen, Altsprachliches Profil
2017-2020	Bachelor-Studium Computerlinguistik und Sprachtechnologie an der Universität Zürich
seit 2020	Master-Studium Computerlinguistik und Sprachtechnologie an der Universität Zürich