Folia Phoniatrica et Logopaedica

Review Article

Folia Phoniatr Logop 2023;75:1–12
DOI: 10.1159/000527427

Received: June 10, 2022
Accepted: October 4, 2022
Published online: October 7, 2022

Multidisciplinary Perspectives on Automatic Analysis of Children's Language Samples: Where Do We Go from Here?

Ulrike Lüdtke[a] Juan Bornman[b] Febe de Wet[c] Ulrich Heid[d] Jörn Ostermann[a] Lars Rumberg[a] Jeannie van der Linde[e] Hanna Ehlert[a]

[a]Leibniz Lab for Relational Communication Research, Leibniz University Hannover, Hanover, Germany; [b]Centre for Augmentative and Alternative Communication, University of Pretoria, Pretoria, South Africa; [c]Department of Electrical, Electronic and Computer Engineering, North-West University, Potchefstroom, South Africa; [d]Institute for Information Science and Speech Technology, University of Hildesheim, Hildesheim, Germany; [e]Department of Speech-Language Pathology and Audiology, University of Pretoria, Pretoria, South Africa

Keywords
Language sample analysis · Child language · Automatic speech recognition · Assessment

Abstract
Background: Language sample analysis (LSA) is invaluable to describe and understand child language use and development for clinical purposes and research. Digital tools supporting LSA are available, but many of the LSA steps have not been automated. Nevertheless, programs that include automatic speech recognition (ASR), the first step of LSA, have already reached mainstream applicability. ***Summary:*** To better understand the complexity, challenges, and future needs of automatic LSA from a technological perspective, including the tasks of transcribing, annotating, and analysing natural child language samples, this article takes on a multidisciplinary view. Requirements of a fully automated LSA process are characterized, features of existing LSA software tools compared, and prior work from the disciplines of information science and computational linguistics reviewed. ***Key Messages:*** Existing tools vary in their extent of automation provided across the process of LSA. Advances in machine learning for speech recognition and processing have potential to facilitate LSA, but the specifics of child speech and language as well as the lack of child data complicate software design. A transdisciplinary approach is recommended as feasible to support future software development for LSA.

Introduction

Language sample analysis (LSA) has been an essential method in researching language development and use for half a century and is considered an important part of clinical language assessment [1, 2]. Spontaneous language samples are a valid and reliable means of gaining insights into a child's language ability and use in everyday communication settings [3]. LSA has developed greatly from handwritten transcriptions, manual annotation, and analysis [1] to computer programs and guidelines providing computational support with a variety of features for conversational LSA [3–7].

Nevertheless, despite the available software support, LSA remains a resource-intensive process. The amount of time to elicit/record, transcribe, and analyse child lan-

Karger@karger.com
www.karger.com/fpl

Correspondence to:
Hanna Ehlert, hanna.ehlert@ifs.uni-hannover.de

guage samples – especially spontaneous ones – limits the method's clinical applicability [2] and emphasizes the need for further technological advancement in LSA [3]. With growing popularity in mainstream applications, such as voice-enabled virtual assistants, the utility of automatic speech recognition (ASR) for clinical purposes in linguistics and speech language pathology (SLP) is also explored [8, 9]. However, the error rate of current ASR systems is still high when applied to children's speech [10] because of the latter's unique acoustic and linguistic characteristics. Furthermore, many of the current models and available data used to train them are designed for downstream tasks such as dialogue systems, where extracting the meaning of what is said is more important than accurately capturing what was said verbatim. This makes these models unsuitable for linguistic and LSA purposes, where a verbatim transcription of all utterances is necessary for analysing language use as well as developmental status including errors, neologisms, and deviations from adult speech.

Therefore, the purpose of this article is to overview automated LSA from a multidisciplinary perspective. By combining expertise from SLP, computational linguistics, and computer science, our aims are (i) to impart knowledge about what is required on a technical level to automate the transcription, coding, and analysis of language samples recorded in natural settings, (ii) to establish a status quo by illustrating which components of this process are already automated by existing LSA software and which are not, and (iii) to provide a realistic outlook on the prospects and pitfalls of technological support for SLP practice and to derive aspects of future research.

To address these aims, we examine digital language processing and analysis in detail by outlining the tasks that an "ideal" (most useful) digital LSA system would need to accomplish. We then provide a comparative overview of three contemporary digital tools in SLP practice, namely "Computerized Language ANalysis" (CLAN) [6], "Systematic Analysis of Language Transcripts" (SALT) [7], and "Language Environment Assessment" (LENA) [11] with reference to the previously sketched ideal system, and we identify needs in the LSA process not met by these tools. We then address prior work and challenges faced in the development of technology to support LSA of monolingual and multilingual children with different language abilities from an engineering and computational linguistics perspective. Finally, in the conclusions, we advocate for a transdisciplinary approach in combining efforts to advance existing technology.

The "Ideal" System

In this section, we briefly describe the components of an ideal system that would be able to support clinical practice in assessing mono- and multilingual children. The ideal system, illustrated in Figure 1, represents our vision for future development.

To be able to capture holistically the input (1) of verbal communication in vivo in natural environments (e.g., at home, at school), an ideal system must first separate and identify all acoustic components of the specific setting in a pre-processing/diarization (2) component. The step of diarization includes speech/non-speech classification, often referred to as voice activity detection (2a). Once speech and non-speech segments have been separated, additional information can be derived. This includes identifying different speakers and assigning each speech segment to one of the detected speakers. The languages spoken can also be identified and possible background media noise (which might include speech) (2b) can be detected. After pre-processing/diarization, the systems split into two different paths, leading to the analysis of the sample: a direct audio-based route (route I), relying on the recorded acoustic signal only and a text-based route (route II) requiring the intermediate steps of transcription and annotation. If transcription (3) is desired, then an orthographic (3a) and phonetic (3b) representation of all spoken communication needs to be compiled. This transcript should also be annotated automatically. At this point of the automated LSA process, "coding" or "annotation" (4) (the latter term is preferred by computational linguists, who usually define coding as programming) refers to the recognition and labelling of segments and phenomena in the transcribed sample (4a and 4b). Annotation adds information to the text that relates to individual tokens (e.g., words or sounds) or structures (e.g., sentences) and can be subsequently queried. Analysis (5) may be carried out according to the two different paths: audio-based (route I) or text-based (route II). Route I allows distributional analysis only, and route II beyond that an analysis of the sample's linguistic content. In summary, analysis (routes I and II) will compute a selection of several measures concerning speech (5a), all language(s) spoken (5b), the communicative interaction (5c), and the acoustic environment (5d). Consequently, the ideal system would span a wide range of outcome measures, such as the time the child has been exposed to electronic media during the day, developmental language profiles, and a detailed analysis of specific linguistic target structures or elements of the synchronicity of adult-child interaction (e.g., child/infant-directed speech).

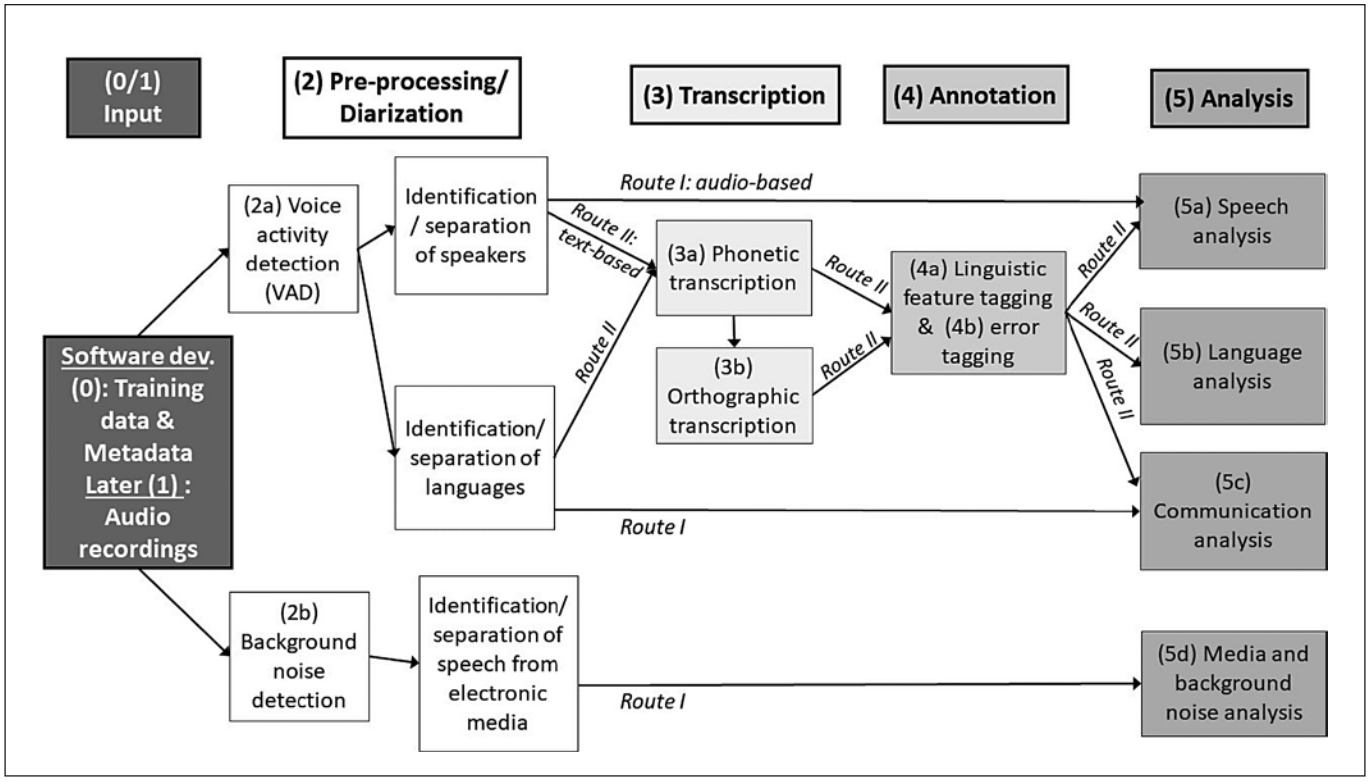

Fig. 1. Overview of the ideal system for automated LSA.

A component of Figure 1 that is not a part of the software but equally important for its development is the data needed to train the software and the associated metadata (e.g., date of recording, age of the recorded child, language(s) spoken in the recording) (0) allocated to every recording and transcript. Recordings of spontaneous child language can be converted into training data by manual labelling, which implies identifying and demarcating the occurrence of each acoustic event or linguistic content in a recording that should be detected or retrieved. This means that recordings accompanied by manual transcriptions and annotations are a prerequisite for software development and for automating the process of pre-processing/diarization, transcription, annotation, and analysis.

The Present: What Do We Have?

In this section, we outline which tools and solutions already exist for automatic analysis of children's language samples from a linguistics/SLP perspective. Three tools could be identified as cited most frequently in the past decade: CLAN, SALT, and LENA.

The open-source software CLAN was developed to search, manipulate, and analyse language data as part of the CHILDES ("CHild Language Data Exchange System") database for annotated media of child language acquisition (which was later merged into the larger TalkBank repository, a Clarin B-centre) [6]. SALT, on the other hand, was programmed to make computerized LSA available for SLP practice [12]. By contrast, the development of the LENA system was originally motivated by the preventative aim of furnishing parents of young children with an automatic feedback system for the speech occurring naturally in their home environment to encourage increased caregiver-child communication [13]. Other tools such as Computerized Profiling [5] or Sampling Utterances and Grammatical Analysis Revised (SUGAR) [3] also reported in the literature are either not very commonly used (any more) or not classified as specific LSA software, but rather utilize regular word processing software for the process (as in the case of SUGAR). While we recognize these approaches in their efforts to promote LSA, we do not include them in depth in this overview because we focus on more recent, dedicated computer programs. To illustrate the features and limitations of

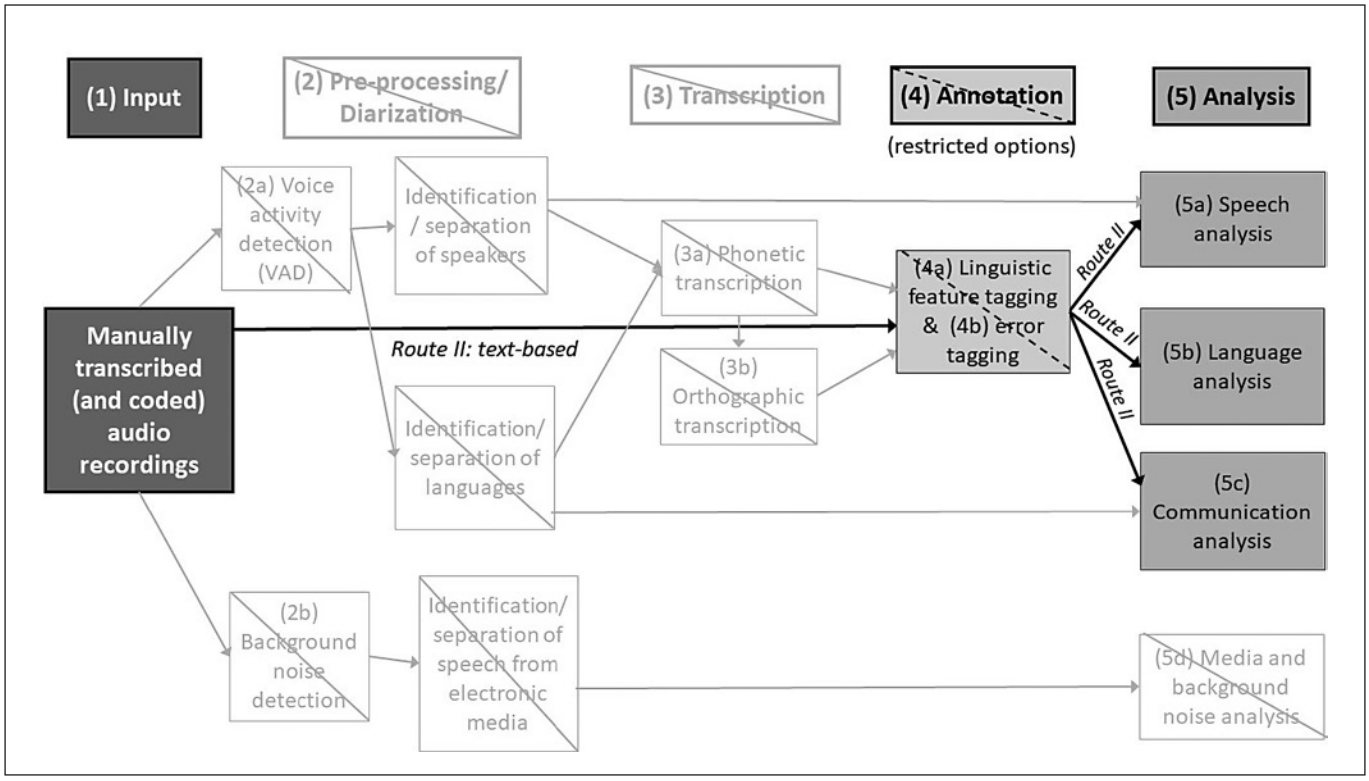

Fig. 2. Parts of the ideal system covered by CLAN and SALT (crossed out parts are not covered).

current LSA software tools and to derive necessary recommendations, they are compared to the components of our ideal system in the following sections.

CLAN and SALT

Figure 2 compares the features of CLAN and SALT against the components of our ideal system.

CLAN and SALT offer analysis (5) and to some extent annotation features (4) based on a manual transcript assembled according to the respective program-specific transcription conventions (SALT or CLAN/CHAT "Codes for the Human Analysis of Transcripts" conventions) (route II). Written conventions are typically drawn up for specific languages using the English conventions as a starting point.

CLAN and SALT offer some options for automatic speech and language annotation (4). For example, CLAN includes features for semiautomatic coding such as morphological parsing (MOR), part-of-speech tagging (POST), and grammatical dependency parsing (GRASP/MEGRASP), the latter being available for English and Japanese only. It is possible to use the MOR and POST programs to process bilingual transcripts automatically whenever MOR grammars exist for both languages and each sentence is identified for language. This includes English, French, German, Italian, Japanese, Mandarin-Cantonese, and Spanish [6]. Nevertheless, manual preannotation is required in both software programs for the subsequent analysis. Similar to manual transcription, CLAN/CHAT and SALT have their own coding conventions for manual annotation with overlapping yet differing coding options. Manual coding standards span a range of aspects from additional (meta)information provided with each transcript file, addition of dependent tiers/layers to the main transcript to utterance boundaries, unintelligible parts of the recording, or divergences of child forms from adult standards. With multilingual speakers, code switching can also be marked in CHAT and SALT at the word and/or utterance level [6, 14, 15].

CLAN offers the largest array of analysis options (5) with over 40 commands operating on CHAT files, enabling the exploration of conversational interaction, language development, and use or language disorders. In terms of linguistic areas, CLAN focuses on morphosyntactic analysis, but several vocabulary diversity scores can also be computed. CLAN was developed primarily for

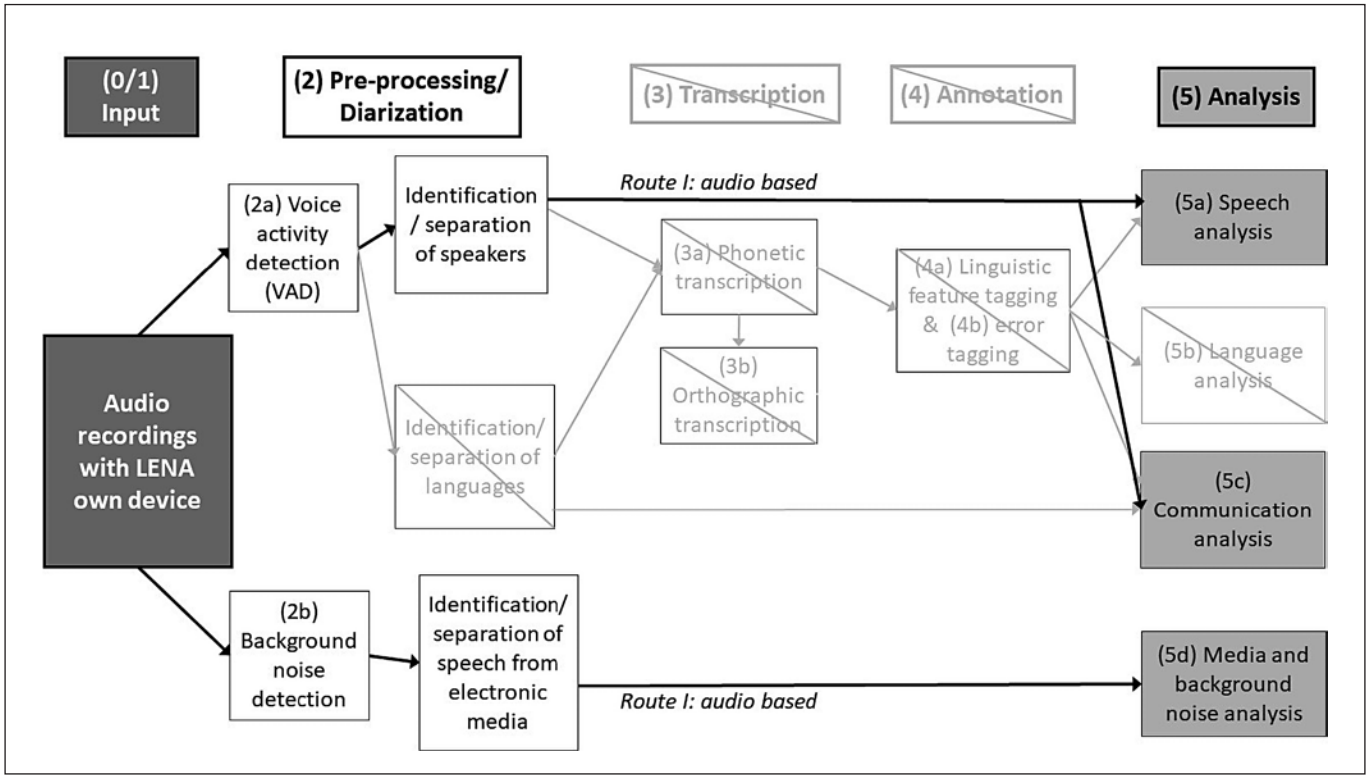

Fig. 3. Parts of the ideal system covered by LENA (crossed out parts are not covered).

English but allows selected analysis steps for several other languages including Chinese, French, German, Italian, and Spanish. For clinical analysis of child language, for example, CLAN has developed a specific set of commands called KIDEVAL. This program computes lexical and grammatical measures such as mean length of utterance in morphemes or words, type-token ratio, lexical diversity (vocD), clause density, and number of major morphemes observed. For mainstream English, individual data can be compared to the KIDEVAL database for children under 6 years in an adult-child free-play setting. Additional grammatical profiles such as the Developmental Sentence Score (DSS) and the Index of Productive Syntax (IPSYn) can be calculated via CLAN commands for English-speaking children (and a few other languages).

The analysis options (5) of the SALT software focus on measures relevant to SLP assessment of speech and language samples including measures of syntax, lexicon, discourse, fluency, and speaking rate. For example, utterance intelligibility, mean length of utterance in words, lexical diversity, and mean length of turns in utterances from the child can be determined from the transcripts. In addition, SALT offers multiple reference databases for comparison to age- and grade-matched peers of English and English/Spanish speaking children in various elicitation settings (e.g., play, narrative, expository) [15]. The outcome measures of CLAN and SALT can be allocated to the subcomponents 5a, 5b, and 5c. Because of the text-based analysis route (II) of CLAN and SALT, measures evaluating the acoustic environment (5d) are not computed.

LENA

Figure 3 compares the features of LENA against the components of our ideal system. The LENA tool includes a recording device that enables audio recordings of several hours in natural settings. This feature sets it apart from the other two tools. In addition, of the three tools under consideration, LENA is the only one that incorporates automatic diarization and pre-processing features (2). The LENA software segments the audio recordings into key child, male adult, female adult, other child, overlapping speech, electronic media, noise, and silence [16]. Child age-specific modelling is used to distinguish speech-related vocalizations from cries and vegetative sounds [17], so the system may be used with children as young as

2 months. LENA developers report a sensitivity of 67% for child speech and 80% for adult speech. For detecting speech from television/media in recordings made with LENA devices, sensitivity and precision values of 61% and 34%, respectively, have been reported [18]. Finally, to prepare the recording for analysis of the speech of key child and adults, the LENA system eliminates overlapping speech, background noise, and child non-speech sounds the audio content [16]. A pre-processing feature not covered by LENA is the ability to process multilingual input (identification/separation of languages). However, its reliability has been validated in five languages: Chinese, English, French, Korean, and Spanish [19].

The measures provided as analysis options (5) by LENA differ from those measured by SALT or CLAN in that LENA analyses the audio recordings (route I) and not a transcript. Based on algorithmic estimation, the pre-processed audio is analysed directly by the LENA accompanying software: the step of creating an orthographic or phonetic transcript (3) and subsequent linguistic annotation (4) is thus avoided. Calculations can therefore be based on the acoustic signal itself only, which restricts the options for analysis to distributional measures such as conversational turn taking, number of vocalizations (e.g., canonical syllables), word counts, or identification of overlapping speech. A screening tool for autism spectrum disorder utilizing atypical vocalizations has also been developed by the LENA foundation [20]. All of these measures may be allocated to the speech and communication analysis components 5a and 5c of our proposed ideal system. Linguistic measures analysing language by drawing on grammar or vocabulary (5b) cannot be computed. However, LENA is the only program that analyses the acoustic environment (5d) in which samples are recorded (e.g., background noise, amount of time of language/sounds from electronic media in the child's environment) [18].

Summary

Overall, pre-processing/diarization (2) (regarding LENA) and analysis (5) (regarding all tools) are the components of the ideal system that existing software covers best. The specific abilities of each tool might be due to the purpose for which it was developed. While SALT was designed for SLP clinical practice, CLAN stems from a linguistic tradition of conversational analysis and LENA follows a preventative aim to promote language development. Differences between LENA and CLAN/SALT – apart from the LENA system containing a hardware component – can be identified mainly in the breadth and depth of their options for analysis. LENA offers the least qualitative insights as it calculates only measures that can be detected without the need to access the linguistic content of a sample (route I), but it is the only tool that enables analysis of the child's acoustic environment. By contrast, SALT and CLAN allow for text-based linguistic analysis (route II) predominantly in grammar, but they do not analyse the acoustic environment. Limitations concern mainly the transcription (3) and annotation (4) components as well as the ability to process multilingual input.

The Future: What Do We Need?

The one component of the ideal system not yet covered by any of the existing tools is automatic transcription (3). Transcription conventions such as SALT or CHAT offer standardized rules for transcribing samples elicited in different settings and many different types of utterances (e.g., babbling) and therefore enhance the accuracy and – to some extent – speed of manual transcriptions [21]. Nevertheless, manual transcription of language samples remains an extremely time- and resource-intensive process, taking significantly longer to compile than the actual duration of the sample. Different sources report orthographic transcription times of 5–50 times the duration of the audio recording (especially if time-aligned) [22, 23]. On the other hand, different measures or linguistic profiles call for different numbers of eligible utterances to be calculated (e.g., Developmental Sentence Score, IPSyn) [24], and sample length may also influence clinical applications of LSA, such as diagnostic accuracy for certain age groups [25]. Clinicians are therefore constantly bargaining for the shortest possible sample length that still provides a reasonable amount of accurate information [26]. This challenge is even greater in multilingual and multicultural low- and middle-income countries as LSA is the best and often the only measure to use to evaluate language [27]. From an SLP perspective, automatic transcription of child language could make the analysis of substantially longer and preferably more natural recordings possible and avoid compromises due to lack of resources.

Annotation/coding (4) is the other component that has to be done manually in the existing tools, except for some basic coding features in CLAN and SALT [6]. Here, the time-consuming aspect of locating and labelling all relevant linguistic structures in the transcripts manually applies as well, but also the heterogeneity of options for

segmenting the sample adds to the complexity of the task. The utterance unit as the basic unit of observation, for example, is defined in several different ways based on phonological, grammatical, or prosodic criteria that are derived from either written or spoken language [28]. Furthermore, complexity in annotation is increased when considering multilingual contexts. An ideal system should be able to handle and interpret multilingual input, especially because disparities in language tests exist within these populations making LSA the gold standard in assessing these children [29]. This need applies to the step of pre-processing/diarization (2) as well where, besides different speakers in the audio recording, the different languages should be identified and separated to provide transcripts for the analysis of multilingual recordings.

Technical Perspective on Development of LSA Software

In this section, we scrutinize the challenges and progress in establishing the components of the ideal digital LSA system outlined earlier. For this purpose, we draw on the research of two disciplines concerned primarily with the development of tools for the automatic processing of speech and text: information science and computational linguistics. While various applications of speech technology are well established for adult speech, many are yet to be implemented for children's speech and language. The following sections briefly introduce available data sets as well as existing technology for the five components of the ideal system. In each section, prior work is discussed, challenges specifically related to child speech and language are presented, and future tasks in software development are highlighted.

Training Data (0)

Speech recognition technology that could be implemented in the ideal system proposed herein is nowadays based on machine learning techniques. As is the case for all applications of machine learning, suitable training data from the target domain (in this case spoken language) are a prerequisite for model development. Once appropriate models have been developed using the training data, they can be applied to process new, unlabelled data from the same or similar target domains. Training data for adult voices are already widely available in some languages, but not for children. The acoustic properties of the training data should be representative of the target data. In a typical dialogue scenario, this would entail the voice of an adult, a child, and some background noise associated with the setting, for example, picking up elicitation objects and putting them down or moving objects around. The properties of longer recordings are usually much more complex because they aim to capture all the acoustic events in the child's day (e.g., TV, PC, mobile phone, animals, sports, traffic). Recordings can be converted into training data by labelling them manually. An appropriate label from a predefined set should be assigned to each event, and transcriptions should be provided for segments that contain speech. In the dialogue scenario, basic labelling would involve adding time stamps to the recording to indicate speaker turns as well as other acoustic events (e.g., laughter, coughing, and toy falling). Longer recordings might require more voices, different languages, and different features of the acoustic environments (e.g., various types of noise) to be labelled. Software for annotation purposes (e.g., part-of-speech taggers) needs labelled training data as well. Creating accurate manual labels is an extremely time-consuming task, but without appropriate training data, automatic diarization, transcription, annotation, and analysis cannot be implemented.

To train an ASR system for adult speech and language, typically hundreds to thousands of hours of annotated speech are used [30–32]. The amount of data required to implement accurate ASR for children is expected to be even higher because of the wide interspeaker and intraspeaker variability in the acoustic and linguistic characteristics of children's speech [33, 34]. Some children's speech corpora, usable for training of speech recognition systems, are already publicly available. For example, the HomeBank repository [35] contains day-long child-centred recordings in natural environments: many of the recordings were recorded using the LENA device [17]. The level of detail in the metadata differs between recordings, with most including automated or human-generated speaker labels, but manual orthographic transcription is provided for only a few recordings. Other corpora include the OGI Kids' Speech corpus [36], the CMU Kids Corpus [37], and the UltraSuit corpus [38]. A more complete list of existing corpora can be found in [39].

Viewed from an SLP perspective, none of these data sets were recorded with the specific aim of supporting SLP clinical practice. Data sets such as large corpora of annotated therapy sessions and HomeBank-like recordings with detailed transcriptions in different languages are required to enable future development of the relevant automatic systems.

Pre-Processing/Diarization (2)

Diarization techniques including voice activity detection and speaker tracking have been implemented successfully for adult voices [40, 41]. Authors of [9, 42] reported similar results for recordings of therapy sessions with one child and one adult, that is, a diarization error rate of around 10%. Much higher error rates are associated with HomeBank-like recordings, where the recording conditions are less structured and more background noise is present. For example, Xie et al. [43] reported an error rate of above 30% for the BabyTrain subset [44] of the HomeBank repository. As mentioned earlier, the LENA tool achieves a sensitivity of 67% for child speech. Overlapping speech may exist in these recordings, which has to be separated for downstream tasks such as ASR and automatic analysis. Wang et al. [45] explored diarization and overlapping speech separation using the HomeBank repository, emphasizing the associated challenges. Diarization and overlapping speech separation using the HomeBank repository are explored in preliminary emphasizing its challenge in the results. In multilingual societies, automatic language identification can be used to assign a language label to each speech segment. Some techniques use information derived from the non-speech segments in a recording, for example, detected background noise can be used to enhance the quality of the speech [46]. Non-speech segments can also be used to determine to how much "speech from media" children are exposed to as the LENA system does [18].

Challenges that should still be addressed to improve pre-processing for child speech include robust segmentation techniques for longer, more natural recordings. Language identification and speaker verification systems should also be benchmarked for children of different ages.

Transcription (3)

The aim of ASR is often not the transcription itself; rather transferring the speech signal into graphic symbols is a required intermediate step for downstream tasks such as human computer interfaces in dialogue systems (e.g., voice-enabled virtual assistants) where the transcription is used to extract the meaning of the speech. For automated LSA, the analogue downstream task would be analysing of the spoken language(s) for clinical purposes, which is the topic of this article. In the following paragraphs, a short review is given of some recent work on child speech and language recognition, independent of the downstream task. Liao et al. [31] presented a large vocabulary speech recognition system trained on a large proprietary data set extracted from Google Voice Search traffic. They reported a word error rate of around 10% for utterances produced by children, which was close to adult speech recognition performance at the time. While obtaining a similar amount of training data is not feasible for most other researchers, the results show that accurate child speech recognition can be achieved with a sufficiently large data set. Kennedy et al. [47] evaluated multiple ASR systems in the setting of child-robot interaction, while Yeung and Alwan [10] showed that child speech recognition is especially difficult for children in kindergarten age and younger; they also showed that an age difference of even a few years between the children in the training and testing data reduces the performance drastically. Wu et al. [34] argued that the problems with child speech recognition are similar to those with adult speech recognition for low-resource languages; by applying a model that works well for low-resource languages, they managed to improve the performance for child speech recognition.

Much of the recent research on ASR for children has been focused on how data on adult speech can be used during training to improve recognition for children. Authors of [48, 49] investigated how to fine-tune models trained on adult speech recognition with child data. Fine-tuning is the process of first training a machine learning model on one domain with large amounts of data (in this case adult speech) and then retraining either parts of the model or the whole model on the target domain (here children's speech). When similar features exist in both domains, the model has already learned them, thereby leveraging the large amount of data of the source domain before training on the target domain. An alternative to fine-tuning is multitask learning, where the model is trained on both domains simultaneously. Tong et al [33] explored multitask learning for child and adult speech recognition, and Rumberg et al. [50] constrained the model to learn features that are independent of speaker age, leading to better transfer between the domains. Other work augmented adult speech by making it more similar to child speech; this involved simulating phenomena such as vowel prolongation that are typically associated with speech produced by children [51].

However, the task remains challenging from an SLP perspective: child language cannot be treated the same as adult language. Capturing children's speech and language – which is still developing – in written form is difficult because it always involves interpretation, whether attempted by a human or by a computer [6]. Researchers will have to find ways to address this issue without simply

relying on huge volumes of data, because recording and annotating child speech and language is a challenge in itself. The resources that are available for children will not grow and become available at the same rate as those that are already available for adult speech and language.

Annotation (4)

Annotation of a transcript can be seen as another intermediate step in the LSA process. Subsequent linguistic analysis is enabled by querying and combining the information added previously via annotation. Both statistical measures and (interactive) query tools are per se language-independent [6]. Language dependency is relevant for linguistic annotation (where part-of-speech tag sets, manually annotated data from child language and methods for normalization are needed) and for the details of (interactive) exploration, where users have to determine for which constructs they intend to search. Computational linguistics typically uses automatic corpus annotation (for the text-based route II) at the level of word forms, which means that word class labels and base forms/lemmata are assigned to individual words. For example, for some analyses, part-of-speech labels and grammatical categories (e.g., word class, tense, number, case) could be assigned to the individual words in a sentence, such as in the annotation options offered by CLAN [6].

For many languages, tools that provide such annotations are available. Some rely on lexicons and statistics, while the latest ones are based on neural networks as in transcription [52]. While the former type is language-specific (because of the lexicon), neural systems often do not require a lexical resource but again an adequate amount of training data. A major issue for both approaches is the fact that children's speech does not always conform to the standard variety on which most of these tools are trained. Typical non-standard phenomena include morphological variation (often due to pronunciation variation) and word contractions. Such phenomena have recently been described in word class annotation tag sets (for German: refer [53]). Glaznieks et al. [54] explored spontaneous interaction with adults, for example. However, a distinction should be made between non-standard language – typically used in spoken discourse in natural settings or as part of mono- and multilingual language development – and language by children with speech and/or language disorders. CLAN's MOR programme achieves accuracies over 99% in part-of-speech-tagging for productions from adult native English speakers and over 95% for adult speakers of French, German, Japanese, Mandarin-Cantonese, and Spanish in TalkBank repositories such as CHILDES. However, the reliable determination of accuracy in tagging for child utterances is seen as more difficult by the authors and hence not reported [6].

Neural part-of-speech taggers have the advantages of being trainable and also being able to recognize data that deviate from the typical training examples and/or variable input. There are tools that "translate" between deviating forms and normalized forms, and both statistical and neural network-based machine translation approaches can be used to this end. Lexicon-based taggers perform relatively poor on non-standard speech, such as child speech. Neural-network-based taggers would have to be trained on large corpora of normal child language to produce meaningful analysis of child speech. This approach does not seem to have been investigated yet. While learner language corpora containing material from advanced learners can typically also be annotated at a syntactic level (e.g., phrases, valency structures), the quality of such analyses on child language (and in particular of those with language disorders) remains to be verified.

Analysis (5)

As pointed out earlier, automatic analysis of speech for diagnostic purposes can be done using an annotated transcription (as described in the previous section) or by analysing the audio/speech signal directly. In this section, we first discuss the case of analysis without transcription (route I), followed by analysis using automatic transcription (route II).

Route I

Parameters that can be estimated without a transcription (route I) include word and utterance counts using high-level speech features and environment analysis (e.g., classifying background noise). For example, the LENA tool provides multiple analytic measures for children during early language development (2–48 months) without using transcriptions. The LENA Automatic Vocalization Assessment [55] estimates the expressiveness of the child's language by using simple regression models on an intermediate representation of the child's speech provided by an ASR system for adults.

In recent work, authors of [8, 9] differentiated between disordered and typically developed speech on the basis of audio recordings without a transcript. The former uses paralinguistic features as an intermediate representation, while the latter applied a speaker recognition framework. They showed that the features used to distinguish between different speakers are suitable for identifying child speech at risk for speech sound disorders.

Route II

Text-based language analysis of automatic transcriptions of speech recordings allows a more detailed analysis of utterance content. Language analysis comprises both qualitative and quantitative measures. Many quantitative measures – such as word counts, (mean) length of utterances, and most lexical richness scores – simply rely on counts of word forms. These measures essentially compute lexical statistics and have mostly been integrated into workbenches or tool collections such as CLAN or SALT [6, 12]. If an analysis at word class and/or lemma level is available, then more sophisticated counts are possible, as well as qualitative analyses, for example, type-token ratios, counts of lexical repetition, and analyses of lexical richness with respect to single word classes. Even if no syntactic analysis is available, pattern-based search on word class and lemma annotations can provide insight into the use of certain grammatical constructions (e.g., complex tenses, passive, questions).

Corpus linguistics provides a whole set of query and inspection tools for text corpora. These typically allow for regular expression-based search over any combination of annotations, for example, word forms, parts of speech, and lemmas. Furthermore, such tools are aware of metadata, thereby allowing searches combining any annotated property with available metadata (e.g., the Open Corpus WorkBench [56]). A well-known tool is ANNIS [57], which allows for search in multiply annotated corpora (e.g., a speech-based transcript, a normalized one, as well as an arbitrary number of annotations on different annotation layers). An advantage of such a multi-layered query architecture is that the process of corpus exploration can, in principle, be used to feed analysis results back into the corpora, possibly as a (perhaps only partly populated) new layer of annotation. There are also tools to query joint speech/language annotated corpora; an example is EXMARaLDA [58], a tool underlying large-scale analysis and exploration of spoken corpora. The accuracy of automatic analysis techniques depends both on the task at hand (e.g., reading or speaking spontaneously) and on the age of the participant. Researchers will have to agree on acceptable accuracy levels in different scenarios and for different age groups to guide further technology development in this domain. Large corpora of normative data will be required for each language in which corpus analysis is to be used for automatic analysis. Such corpora will need to be stratified such that normal language usage and development including developmental errors and deviations from adult language for different age groups are adequately represented.

Conclusion

In this paper, we have outlined current tools for software-based LSA (CLAN, SALT, and LENA), recent advances in ASR technology, and corpus analysis tools for supporting automated LSA. By highlighting the components of an ideal system, we attempted to create a framework to analyse the abilities and limitations of three frequently used tools in linguistic/SLP research and SLP clinical practice. We also presented the technical challenges and preliminary solutions for required software development from a multidisciplinary view. A highly desirable goal for future research is the development of digital solutions that enable research on monolingual and multilingual children's speech and language development and use and that support clinical assessment on the basis of representative – and thus longer – natural language samples recorded in vivo. While machine learning has accelerated progress in ASR and digital corpus analysis in the past decade, applying these approaches to children's variable, non-standard, and developing speech and language remains difficult, at least when linguistic and SLP purposes are considered. The first and last steps of the ideal system (pre-processing/diarization and analysis) are the ones covered best by existing software. Several other relevant tasks required for a fully automated LSA process – such as speaker tracking and spoken language identification – have been explored for adults and for structured simplified acoustic contexts. The transfer to settings with natural (multilingual) communicative interaction of children in acoustically complex and unstructured everyday contexts remains a future challenge. What becomes apparent is the need for appropriate spontaneous language data to develop ASR software for children. Machine learning techniques are state of the art for this task, and the quality of their programming relies heavily on how well the training data match the intended purpose. Besides that, refined training models could help compensate for the child data gap.

To ensure future automation of LSA, a truly transdisciplinary approach is needed. The disciplines of engineering, information science, linguistics, computer linguistics, and SLP each hold different, but equally relevant knowledge that can inform technology development in unique ways. However, this knowledge and existing data cannot be accessed or adopted if disciplinary siloed approaches remain. For instance, CHILDES data are edited or provided in a way not to allow software training but solely from an applied linguistics/SLP research perspective. Collaboration is required as well to create appropriately annotated data sets of child language that not only

capture target phenomena but can also be used as training data for machine learning. Furthermore, SLP/linguistic researchers and clinicians should guide technology development to ensure that systems are optimized to generate both accurate and useful results.

Conflict of Interest Statement

The authors declare that no competing interests existed at the time of publication.

Author Contributions

Ulrike Lüdtke and Juan Bornman: conceptualized the manuscript and formulated overarching goals of the manuscript, participated in writing the first draft of manuscript, and critically reviewed and revised the manuscript. Febe de Wet, Ulrich Heid, Jörn Ostermann, and Lars Rumberg: conceptualized the manuscript, participated in writing the first draft of manuscript, and critically reviewed and revised the manuscript. Jeannie van der Linde and Hanna Ehlert: conceptualized the manuscript and formulated overarching goals of the manuscript, participated in writing the first draft of manuscript, and critically reviewed and revised the manuscript.

Funding Sources

The authors declare that they did not receive funding for this work.

References

1 Brown R. A first language. Boston: Harvard University Press; 1973.
2 Pavelko SL, Owens RE, Ireland M, Hahs-Vaughn DL. Use of language sample analysis by school-based slps: results of a nationwide survey. Lang Speech Hear Serv Sch. 2016;47(3):246–58.
3 Pavelko SL, Owens RE. Sampling utterances and grammatical analysis revised (sugar): new normative values for language sample analysis measures. Lang Speech Hear Serv Sch. 2017;48(3):197–215.
4 Gilkerson J, Richards JA. The LENA natural language study (Tech. Rep. No. LTR-02-2). Boulder, Connecticut: LENA Foundation; 2008.
5 Long FMSH, Channell R. Computerized profiling, versions 9.0.3-9.2.7 (ms-dos) [computer program] [Computer software manual]. Cleveland, OH: 1996–2000.
6 MacWhinney B. The CHILDES project: tools for analyzing talk. 3rd ed. 2000.
7 Miller JF, Chapman R. Systematic analysis of language transcripts [Computer software manual]. Madison, WI; 1985.
8 Kothalkar PV, Rudolph J, Dollaghan C, McGlothlin J, Campbell TF, Hansen JHL. Automatic screening to detect "at risk" child speech samples using a clinical group verification framework. 40th Annu Int Conf IEEE Eng Med Biol Soc (EMBC). 2018;2018:4909–13.
9 Shahin M, Zafar U, Ahmed B. The automatic detection of speech disorders in children: challenges, opportunities, and preliminary results. IEEE J Sel Top Signal Process. 2020;14(2):400–12.
10 Yeung G, Alwan A. On the difficulties of automatic speech recognition for kindergarten-aged children. Proceedings INTERSPEECH 2018. 2018. p. 1661–1665.
11 Gilkerson J, Richards JA. The LENA™ developmental snapshot (Tech. Rep. No. LTR-07-2). Boulder, Connecticut: LENA Foundation; 2008.
12 Miller JF, Freiberg C, Holland MB, Reeves MA. Implementing computerized language sample analysis in the public school. Top Lang Disord. 1992;12(2):69–82.
13 Greenwood CR, Schnitz AG, Irvin D, Tsai SF, Carta JJ. Automated language environment analysis: a research synthesis. Am J Speech Lang Pathol. 2018;27(2):853–67.
14 Miller JF, Andriacchi K, Nockerts A. Assessing language production using SALT software: a clinician's guide to language sample analysis [computer software manual]. Middleton, WI; 2015.
15 Miller JF, Andriacchi K, Nockerts A. Using language sample analysis to assess spoken language production in adolescents. Lang Speech Hear Serv Sch. 2016;47(2):99–112.
16 Xu D, Yapanel U, Gray S. Reliability of the LENA language environment analysis system in young children's natural home environment (Tech. Rep. No. LTR-05-2). Boulder, Connecticut: LENA Foundation; 2009.
17 Xu D, Yapanel U, Gray S. The LENA™ language environment analysis system: the Interpreted Time Segments (ITS) File (Tech. Rep. No. LTR-04-2). Boulder, Connecticut: LENA Foundation; 2008.
18 Gilkerson J, Richards JA. A guide to understanding the design and purpose of the LENA System (Tech. Rep. No. LTR-12-1). Boulder, Connecticut: LENA Foundation; 2020.
19 Wang Y, Hartman M, Aziz NAA, Tunison E, Arora S, Shi L, et al. A systematic review of the use of LENA technology. Am Ann Deaf. 2017;162(3):295–311.
20 Richards JA, Xu D, Gilkerson J. Development and performance of the LENA automatic autism screen (Tech. Rep. No. LTR-10-1). Boulder, Connecticut: LENA Foundation; 2010.
21 Overton S, Wren Y. Outcome measurement using naturalistic language samples: a feasibility pilot study using language transcription software and speech and language therapy assistants. Child Lang Teach Ther. 2014;30(2):221–9.
22 Heilmann JJ. Myths and realities of language sample analysis. Perspect Lang Learn Educ. 2010;17(1):4–8.
23 Roy BC, Roy D. Fast transcription of unstructured audio recordings. In Proceedings INTERSPEECH 2009. 2009. p. 6–10.
24 Garbarino J, Ratner NB, MacWhinney B. Use of computerized language analysis to assess child language. Lang Speech Hear Serv Sch. 2020;51(2):504–6.
25 Guo LY, Eisenberg S. The diagnostic accuracy of two tense measures for identifying 3-year-olds with language impairment. Am J Speech Lang Pathol. 2014;23(2):203–12.
26 Wren Y, Titterington J, White P. How many words make a sample? Determining the minimum number of word tokens needed in connected speech samples for child speech assessment. Clin Linguist Phon. 2021;35(8):761–78.
27 Southwood F, van Dulm O. The challenge of linguistic and cultural diversity: does length of experience affect South African speech-language therapists' management of children with language impairment? S Afr J Commun Disord. 2015;62(1):1–14.
28 Foster P, Tonkyn A, Wigglesworth G. Measuring spoken language: a unit for all reasons. Appl Linguist. 2000;21(3):354–75.
29 Heilmann JJ, Rojas R, Iglesias A, Miller JF. Clinical impact of wordless picture storybooks on bilingual narrative language production: a comparison of the "frog" stories. Int J Lang Commun Disord. 2016;51(3):339–45.

30 Hinton G, Deng L, Yu D, Dahl GE, Mohamed Ar, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag. 2012;29(6):82–97.
31 Liao H, Pundak G, Siohan O, Carroll MK, Coccaro N, Jiang Q-M, et al. Large vocabulary automatic speech recognition for children. In Proceedings INTERSPEECH 2015. 2015. p. 1611–5.
32 Panayotov V, Chen G, Povey D, Khudanpur S. Librispeech: an asr corpus based on public domain audio books. Proceedings ICASSP'15. 2015. p. 5206–10.
33 Tong R, Wang L, Ma B. Transfer learning for children's speech recognition. International Conference on Asian Language Processing (IALP). 2017. p. 36–9.
34 Wu F, García-Perera LP, Povey D, Khudanpur S. Advances in automatic speech recognition for child speech using factored time delay neural network. Proceedings INTERSPEECH 2019. 2019. p. 1–5.
35 MacWhinney B, Warlaumont A, Bergelson E, Cristia A, Soderstrom M, De Palma P, et al. Homebank: an online repository of daylong child-centered audio recordings. Semin Speech Lang. 2016;37(2):128–42.
36 Shobaki K, Hosom JP, Cole RA. The OGI kids' speech corpus and recognizers. Sixth international conference on spoken language processing (ICSLP); 2000.
37 Eskenanzi M, Mostow J, Gradd D. The CMU kids corpus LDC97S63. Linguistic Data Consortium; 1997.
38 Eshky A, Ribeiro MS, Cleland J, Richmond K, Roxburgh Z, Scobbie J, et al. Ultrasuite: a repository of ultrasound and acoustic data from child speech therapy sessions. Proceedings INTERSPEECH 2018. 2018. p. 1888–92.
39 Ramteke PB, Supanekar S, Hegde P, Nelson H, Aithal V, Koolagudi SG. Nitk kids' speech corpus. Proceedings INTERSPEECH 2019. 2019. p. 331–5.
40 Snyder D, Garcia-Romero D, Sell G, McCree A, Povey D, Khudanpur S. Speaker recognition for multi-speaker conversations using x-vectors. Proceedings ICASSP'19. 2019. p. 5796–800.
41 Van Leeuwen DA, Huijbregts M. The AMI speaker diarization system for NIST RT06s meeting data. International workshop on machine learning for multimodal interaction. 2006. p. 371–84.
42 Koluguri NR, Kumar M, Kim SH, Lord C, Narayanan C. Meta-learning for robust child-adult classification from speech. In Proceedings icassp'20. 2020. p. 8094–8.
43 Xie J, Sia S, Garcia P, Povey D, Khudanpur S. Mixture of speaker-type pldas for children's speech diarization. 2020. arXiv.
44 Garcia P, Villalba J, Bredin H, Du J, Castan D, Cristia A, et al. Speaker detection in the wild: lessons learned from JSALT 2019. 2019. arXiv.
45 Wang X, Du J, Cristia A, Sun L, Lee C-H. A study of child speech extraction using joint speech enhancement and separation in realistic conditions. In Proceedings ICASSP'20. 2020. p. 7304–8.
46 Sun M, Li Y, Gemmeke JF, Zhang X. Speech enhancement under low snr conditions via noise estimation using sparse and low-rank nmf with kullback–leibler divergence. IEEE/ACM Trans Audio Speech Lang Process. 2015;23(7):1233–42.
47 Kennedy J, Lemaignan S, Montassier C, Lavalade P, Irfan B, Papadopoulos F, et al. Child speech recognition in human-robot interaction: evaluations and recommendations. Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. 2017. p. 82–90.
48 Smith D, Sneddon A, Ward L, Duenser A, Freyne J, Silvera-Tawil D, et al. Improving child speech disorder assessment by incorporating out-of-domain adult speech. In Proceedings INTERSPEECH 2017. 2017. p. 2690–4.
49 Shivakumar PG, Georgiou P. Transfer learning from adult to children for speech recognition: evaluation, analysis and recommendations. Comput Speech Lang. 2020;63:101077.
50 Rumberg L, Ehlert H, Lüdtke U, Ostermann J. Age-invariant training for end-to-end child speech recognition using adversarial multitask learning. In Proceedings INTERSPEECH 2021. 2021. p. 3850–4.
51 Nagano T, Fukuda T, Suzuki M, Karuta G. Data augmentation based on vowel stretch for improving children's speech recognition. IEEE Automatic Speech Recognition and Understanding workshop (ASRU). 2019. p. 502–8.
52 Schmid H. Deep learning-based morphological taggers and lemmatizers for annotating historical texts. DATeCH2019 Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage. 2019. p. 133–7
53 Westpfahl S. Pos-tagging für Transkripte gesprochener Sprache. Tübingen: Gunter Narr Verlag; 2020.
54 Glaznieks A, Frey JC, Nicolas L, Stopfner M, Zanasi L, Nicolas L. Leonide: A longitudinal trilingual corpus of young learners of italian, german and english. Int J Learn Corpus Res. 2022;8(1):97–120.
55 Richards JA, Gilkerson J, Paul T, Xu D. The LENA automatic vocalization assessment (Tech. Rep. No. LTR-08-1). Boulder, Connecticut: LENA Foundation; 2008.
56 Evert S, Hardie A. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. Proceedings of the Corpus Linguistics 2011 Conference. University of Birmingham; 2011. p. 1–21.
57 Krause T, Zeldes A. ANNIS3: a new architecture for generic corpus query and visualization. Digit Scholarsh Humanit. 2016;31(1):118–39.
58 Schmidt T, Wörner K. Exmaralda. Handbook on corpus phonology. Oxford University Press; 2014. p. 402–19.