Research on the methodology of LSA with preschool children: a scoping review

Article *in* Clinical Archives of Communication Disorders · August 2023

DOI: 10.21849/cacd.2022.00738

Ulrike M. Lüdtke
Leibniz Universität Hannover
30 PUBLICATIONS **106** CITATIONS
SEE PROFILE

Hanna Ehlert
Leibniz Universität Hannover
35 PUBLICATIONS **58** CITATIONS
SEE PROFILE

Juan Bornman
Stellenbosch University
135 PUBLICATIONS **1,512** CITATIONS
SEE PROFILE

Original Article

http://e-cacd.org/ eISSN: 2508-5948

Clinical Archives of Communication Disorders / Vol. 8, No. 2:29-46 / August 2023

Open Access

https://doi.org/10.21849/cacd.2022.00738

Research on the methodology of LSA with preschool children: a scoping review

Ulrike Lüdtke[1], Hanna Ehlert[1], Dana Gaigulo[3], Juan Bornman[2]

[1]Leibniz Lab for Relational Communication Research, Leibniz University Hannover, Hannover, Germany; [2]Centre for Augmentative and Alternative Communication, University of Pretoria, Pretoria, South Africa; [3]Department of Education and Rehabilitation, Ludwig-Maximilians-University Munich, Munich, Germany

Purpose: Language Sample Analysis (LSA) is a prominent method in researching language development and is also used in clinical practice in the speech-language pathology (SLP) discipline. This scoping review aims to describe current contributions of research on LSA methodology, identify research gaps and explore areas of future advancement of LSA methodology related to its five components: determining the sample length/size, collecting, transcribing, coding and analyzing the sample.

Methods: A scoping review was conducted of studies on LSA methodology published between 2010–2020 that focused on preschool children. Relevant electronic databases and research platforms were searched using the PRISMA method for data identification, screening, selection and extraction.

Results: Of the 213 identified studies, 61 met the inclusion criteria, covering all aspects of the LSA process. Overall, a wide variability in study designs and research foci were found, reflecting the broad applicability of LSA. The two LSA aspects addressed most frequently are the first and last of the five LSA components: determining the length (or size) of the language sample and analyzing the sample. The methodological variability hinders the comparison of evidence and drawing implications which negatively impacts on research and clinical SLP practice.

Conclusions: Besides expanding research on LSA for multilingual children and establishing LSA guidelines for specific contexts, age groups and language backgrounds, it appears as if technological development, particularly in the (semi)automatic transcription, coding and analysis of child language, holds promise to improve LSA applicability and efficiency.

Keywords: Language sample analysis, Language disorder, Preschool children

CACD

Received: May 30, 2022
Revision:
Accepted: September 19, 2022

Correspondence:
Ulrike Lüdtke

Leibniz Lab for Relational Communication Research, Faculty of Humanities, Department of Special Education, Leibniz University Hanover, Schlosswender Str. 1, D-30159 Hanover, Germany
Tel: +49-511-762-17377
Fax: +49-511-762-17408
E-mail: ulrike.luedtke@ifs.uni-hannover.de

INTRODUCTION

Language Sample Analysis (LSA) has a long tradition in research focused on language development as well as in research and clinical practice within the speech-language pathology discipline where the focus is on the assessment of developmental language disorders in children [1], [2]. The advantages of LSA can be summarized in the fact that it enables both researchers and clinicians to obtain more realistic insights into children's functional language. This is particularly true for preschool children, because younger have greater difficulty in presenting their actual language abilities during formal testing (e.g., due to test-taking unfamiliarity or motivational reasons). This ecological validity paired with the breadth of data elicited - covering language aspects such as grammar, vocabulary, phonology, narrative, pragmatic and communicative abilities -

enhances its value. For example, Ebert and Pham ([3], p.43) stated: "The language sample can provide information on multiple language dimensions at the micro- and macrostructural levels." Eisenberg et al ([4], p.633). Used the clinical perspective to highlight that: "A diagnosis of language impairment in young children may be more accurately accomplished through the use of quantitative LSA measures than through standardized tests". This perspective is supported by researchers [5], [6] and clinicians [7], [5] who also argued that LSA of spontaneous language samples can be regarded as a culturally unbiased method that can help identify language impairment in multilingual children with higher accuracy than language tests. LSA has even been praised as the "gold standard procedure" when assessing the expressive language skills of both mono- and multilingual children ([8], p. 339) thus deeming it as useful for "complex language and cultural environments" ([9], p.499). Researchers are constantly attempting to advance the reliability, validity and practicability of LSA by producing a variety of different concepts, approaches and (digital) tools [10], and by making LSA applicable to diverse contexts and populations - thereby strengthening its usefulness as a method. However, LSA is still not routinely implemented in clinical practice [11].

The current study therefore attempts to conduct a scoping review of the current evidence on the methodology of LSA in order to investigate if that might hold a clue as to why it is not frequently implemented in clinical practice despite its obvious value and advantages. Particular attention is paid to language sampling contexts that represent spontaneous, non-imitative language (e.g., conversational or play-based) as these are regarded as being closest to everyday communication and do not tap into other cognitive abilities (e.g., story retelling/memory). Therefore, this review only focusses on studies that present original or archival data on the methodology of LSA guided by its chronological components, namely *determining* (i.e., size in number of words/utterances or length in minutes) *collecting, transcribing, coding* and *analyzing the language sample*. Unique questions are tied to each of these five components, but to our knowledge this is the first review on LSA methodology that attempts to provide an overview of all five components. Thereby, this review can map out the entire process of LSA, compare efforts across components, identify research gaps and provide a holistic procedural view. Thus we will refer to these five components in the methodology, results and discussion sections.

Determining language sample length/size

The discourse on this component revolves around using the shortest possible sample length that will yield reliable data (i.e., contain the targeted information to render a representative view of the child's language abilities). Sample length/size effects are calculated in the literature for different age groups [12], different language status (typical developing vs. developmental language disorders) [13], different sampling contexts [14] and different LSA measures [15].

Collecting

Research in this component aims to identify the specific contexts that elicit the most talking, with the most complex language repertoire possible to best reflect the child's functional communicative abilities. Advantages and limitations of specific elicitation conditions are discussed (e.g., narrative, play-based, expository), often with reference to their degree of structuredness [4]. Several individual variables have been considered when evaluating elicitation contexts, all of which may influence the collected sample. Among these are broader aspects such as setting (e.g., laboratory vs. home) [16], age and language status of the children [17], [18], communication partners [19], [20], as well as aspects regarding the communicative elicitation itself [21].

Transcribing

Research on this component mainly addresses questions regarding the efficiency and accuracy of translating oral language into text: the amount of time it takes to transcribe oral samples [22] as well as transcription accuracy/reliability [23].

Many transcription conventions have been developed in order to address the latter [24], [25].

Coding

Research related to the coding component is similar to that discussed in the transcribing component: the amount of time it takes to code samples [26] and coding accuracy/reliability [27]. Similarly, coding conventions have been developed [24], [25], such as those focused on specific LSA units (e.g., utterance units: [28]), measures (e.g., mean length of utterance [MLU]; [10]), or those focused on the level of detail (e.g., MLU in words, syllables or morphemes: [29], [30]).

Analyzing

Research on this core component of the LSA process constantly seeks to improve the methods' outcome in multiple language domains such as grammar, vocabulary and verbal fluency. New measures have been developed that share different levels of commonality, ranging from single-score general outcome measures to measures assessing separate components within a skill area in a more complex and detailed way [31]. Existing measures have been evaluated in terms of their diagnostic value for different languages [3] and also for mono- and bilingual children [32-35].

Research on how technology could assist and advance this component, has also grown in the 21st century. Several hardware and software tools have been developed to support the analysis of language samples, such as Computerized Language Analysis (CLAN) [24], Systematic Analysis of Language Transcripts (SALT) [25], Computerized Profiling (CP) [36], and Language Environment Analysis (LENA) [37]. Recent advances in machine learning and natural language processing have accelerated the research in this field even further [38].

As we approach this scoping review focused on the methodology of LSA, we aim to identify aspects that researchers have been addressing within these five components during the last decade. These are seen as direct answers to some of the most pressing needs for improving the validity, reliability and efficiency of LSA, thus promoting its applicability in research and clinical practice. In order to contribute and advance the field we will highlight areas that are well researched while revealing areas that are under researched and therefore require attention in the future. As such, the review will inform researchers and clinicians who are interested in LSA on which of these components consensus has been reached in terms of methodology, and where controversy still remains. Furthermore the review will specifically elaborate on current trends in LSA methodology with regard to automation as it is an important aspect of future development and might impact LSA fundamentally if expanded to all its components. The objectives of this scoping review are thus:

1. To examine and synthesize the current published research related to the five components of the LSA process: determining the length/size of the sample (I), collecting the sample (II), transcribing the sample (III), coding (IV) and analyzing the sample (V).
2. To analyze the contributions of recent research (2010–2020) regarding the methodological considerations related to these components of the LSA process.
3. To draw conclusions from these findings in terms of future research directions and needs for methodological advancement overarching the LSA process as a whole with a special focus on automation.

METHODS

Search protocol

This scoping review is reported in accordance with the PRISMA-ScR statement (Tricco et al., 2018) as shown in Figure 1. In order to identify scientific publications that focus on recent research on LSA methodology a systematic search of the literature was conducted. It commenced with an electronic database search of the online research platforms EBSCOhost (APA psycinfo) and PubMed as they cover the databases most likely to index studies on LSA. Our search terms were generated on the basis of 100% consensus within the team after a preliminary review of the literature. Search terms using Boolean operators at either title or abstract level (with truncation) included the following: (AB child* AND AB “speech sampl*” AND AB analys*) OR (AB child* AND AB “language sampl*” AND AB analys*), respectively (child* [Title/Abstract] AND speech sampl* [Title/Abstract] AND analys* [Title/Abstract]) OR (child* [Title/Abstract] AND language sampl* [Title/Abstract] AND analys* [Title/Abstract]) for PubMed (finding all search terms as search mode). The electronic search was further supplemented by hand searching. After deleting duplicates, 213 papers remained for screening at the abstract level.

For the screening phase, inclusion and exclusion criteria were employed. We limited the publication date of the included papers to the last decade (2010–2020), to summarize recent research on the topic. The number of relevant publications in the databases showed a clear increase from 2010 onward, reflecting a growing interest. In addition, we only included studies that focused on preschool children (mean age of the participants ≤72 months) as LSA methodology differs considerably across age groups, with regards to elicitation contexts and measures. However, the review did not set any exclusion criteria regarding language status and included children with typical development, developmental language disorders, and those at risk for developmental language disorders. For participant descriptions, we relied on how the disorders were described in the original studies. Furthermore, we included both mono- and multilingual children. Irrelevant studies were excluded at title and abstract level in the first round conducted by two reviewers. We defined exclusion criteria as: (a) studies not published in English, German or French (i.e., wrong language studies); (b) studies analyzing the language only of school-age children, adolescents or

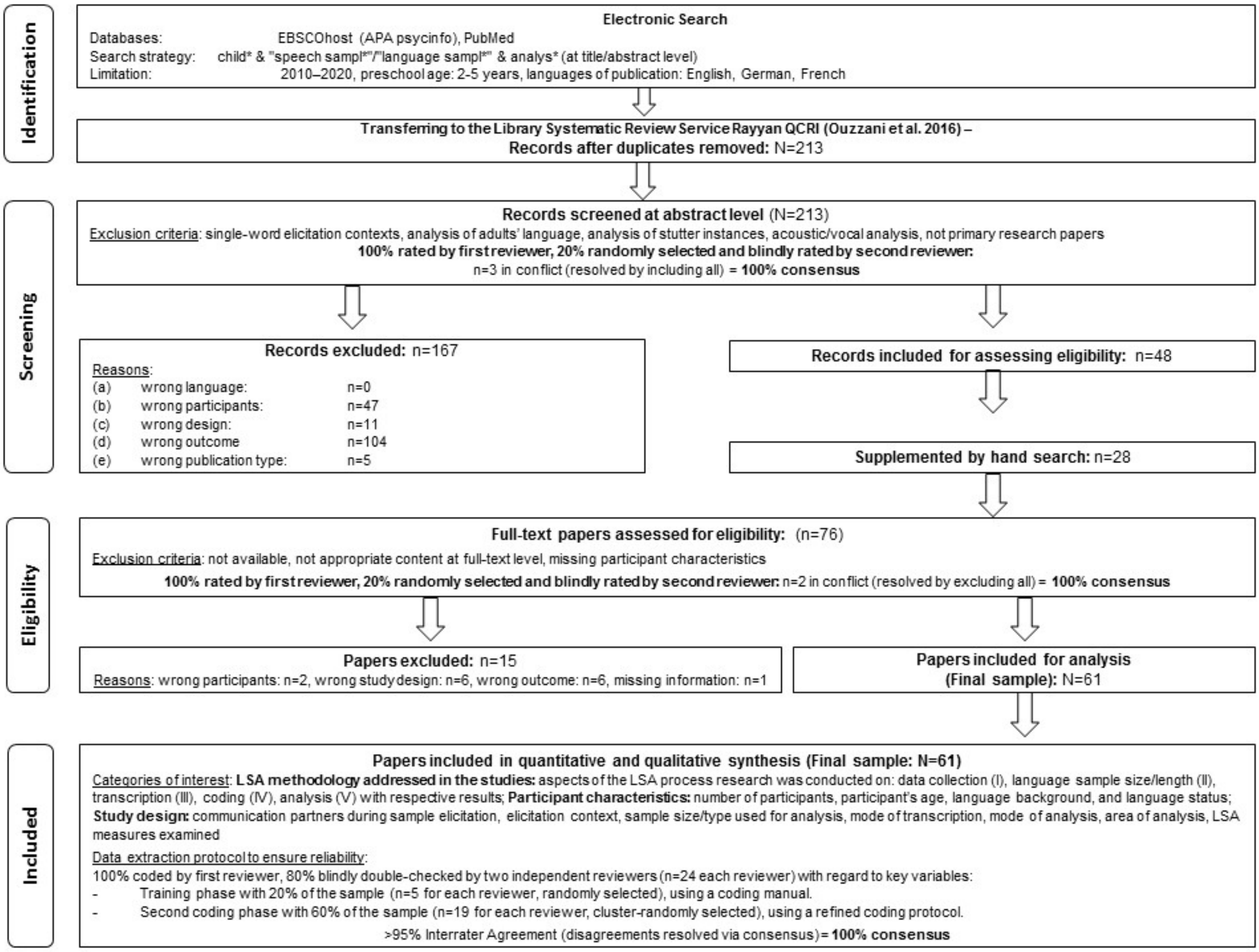
Identification
Electronic Search
Databases: EBSCOhost (APA psycinfo), PubMed
Search strategy: child* & "speech sampl*"/"language sampl*" & analys* (at title/abstract level)
Limitation: 2010–2020, preschool age: 2-5 years, languages of publication: English, German, French
Transferring to the Library Systematic Review Service Rayyan QCRI (Ouzzani et al. 2016) –
Records after duplicates removed: N=213
Screening
Records screened at abstract level (N=213)
Exclusion criteria: single-word elicitation contexts, analysis of adults' language, analysis of stutter instances, acoustic/vocal analysis, not primary research papers
100% rated by first reviewer, 20% randomly selected and blindly rated by second reviewer:
n=3 in conflict (resolved by including all) = 100% consensus
Records excluded: n=167
Reasons:
(a) wrong language: n=0
(b) wrong participants: n=47
(c) wrong design: n=11
(d) wrong outcome n=104
(e) wrong publication type: n=5
Records included for assessing eligibility: n=48
Supplemented by hand search: n=28
Eligibility
Full-text papers assessed for eligibility: (n=76)
Exclusion criteria: not available, not appropriate content at full-text level, missing participant characteristics
100% rated by first reviewer, 20% randomly selected and blindly rated by second reviewer: n=2 in conflict (resolved by excluding all) = 100% consensus
Papers excluded: n=15
Reasons: wrong participants: n=2, wrong study design: n=6, wrong outcome: n=6, missing information: n=1
Papers included for analysis
(Final sample): N=61
Included
Papers included in quantitative and qualitative synthesis (Final sample: N=61)
Categories of interest: LSA methodology addressed in the studies: aspects of the LSA process research was conducted on: data collection (I), language sample size/length (II), transcription (III), coding (IV), analysis (V) with respective results; Participant characteristics: number of participants, participant's age, language background, and language status; Study design: communication partners during sample elicitation, elicitation context, sample size/type used for analysis, mode of transcription, mode of analysis, area of analysis, LSA measures examined
Data extraction protocol to ensure reliability:
100% coded by first reviewer, 80% blindly double-checked by two independent reviewers (n=24 each reviewer) with regard to key variables:
- Training phase with 20% of the sample (n=5 for each reviewer, randomly selected), using a coding manual.
- Second coding phase with 60% of the sample (n=19 for each reviewer, cluster-randomly selected), using a refined coding protocol.
>95% Interrater Agreement (disagreements resolved via consensus) = 100% consensus

Figure 1. The PRISMA chart of the scoping review.

adults, studies analyzing LSA methodology in children with language disorders associated with biomedical conditions (i.e., wrong participant studies); (c) studies using single-word elicitation contexts or samples which cannot be seen as spontaneous, non-imitative language (including story retelling) (i.e., wrong design studies); (d) studies that employed LSA as a method, but that only presented results on the speech and language development of their participants and not on the LSA methodology itself, as well as studies concerned with acoustic, vocal or disfluency analyses, or studies only considering distributional measures of language analysis using LENA technology (which have already been reviewed by Ganek and Eriks-Brophy [40]) (i.e., wrong outcome studies); (e) studies that are background papers and that did not present original data on LSA methodology or that are not peer-reviewed such as grey literature (e.g., conference abstracts, presentations, tutorials, proceedings, government publications, reports, dissertations/theses, patents and policy documents) (i.e., wrong publication-type studies).

Of the 213 papers included in the screening phase, a total of 77.5% (n = 165) were excluded by mutual agreement between the two reviewers. The majority of these studies (n = 104) were excluded because they had the wrong outcome, mostly employing LSA as a method in their research without focusing on the methodology of LSA itself. In total, 48 papers from the electronic search plus 28 papers from the hand search remained after the screening phase for determining eligibility (n = 76).

During the eligibility phase, all 76 papers were reviewed at full-text level. 100% were reviewed for inclusion/exclusion by the first reviewer and 20% randomly selected and blindly rated by a second reviewer. After full-text reading, 16.4% (n = 15) of studies were excluded. The final sample of this review thus comprised 61 included papers.

Data extraction

In the final phase, data extraction was conducted on all 61 included papers using a custom-designed data extraction protocol. Variables focused on participant characteristics, study design as well as the five components of the LSA process. All information and aggregated data were extracted from the included papers. The collected information was then tabulated and analyzed both quantitatively and qualitatively at a descriptive level, noting patterns in methodology and gaps in knowledge. To better structure the amount of research in the analysis component (V) we differentiate between studies that focused on monolingual mainstream English-speaking children and children who speak any other languages/dialects (including multilingual children). Additionally, we also divided our data related to this component into several subcategories due to the variety of topics covered in the papers. First, (a) we summarize the development of new LSA measures (previously unpublished measures or measure modifications) or the adaptation of existing measures. Second, (b) we explore the value of LSA measures by describing the diagnostic value, growth value and value in terms of usefulness. Diagnostic value refers to the identification of children with (or at risk for) developmental language disorders, according to LSA measures (only studies computing statistics such as classification accuracy, ROC–analysis, sensitivity/specificity, or positive/negative likelihood ratio [LR+/-]). Growth value refers either to determining the stage or overall level of language development or change following intervention via LSA measures. Value in terms of usefulness refers to the exploration of feasibility, reliability and validity. Third (c) we synthesize data on automated LSA, by describing studies that compare digital vs. manual analysis, and examining the reliability of software support.

Reliability

A data extraction protocol, based on the following interrater ratios, was used: 100% coded by the second author, 80% double-blindly checked independently by the third and fourth authors (n=24 for each reviewer). A coding manual was used for categorizing variables in terms of LSA methodology. There was a training phase with 20% of papers (n=5) randomly selected for each of the three reviewers, using the coding manual. Training enhanced inter-rater reliability. The coding phase involved 60% of the sample (n=19) randomly selected for each of the reviewers, using a refined coding protocol. Interrater agreement of >95% could be reached for the final data extraction. Disagreements were resolved via consensus discussions. The PRISMA chart in Figure 1 summarizes all these steps.

RESULTS

Descriptive results regarding participant characteristics and study design of the 61papers included as the final sample are reported in supplement A. In the results section, we present summaries of the studies' results arranged according to the components of the LSA process, as per the aims of the review: determining the language sample length/size (I), collecting (II), transcribing (III), coding (IV) and analyzing (V) (see supplement B for the categorization criteria). Table 1 provides an overview of these components targeted in the included studies.

Determining the length or size/length of the language sample (I) is described in 17 of the 61 studies (28%); collecting the sample (II) in 11 studies (18%); transcribing (III) in two studies (3%); coding (IV) in six studies (10%), and analyzing the sample (V) in 48 studies (79%), making it the most frequently addressed component.

Determining the length/size of the language sample

Studies focused on the length/size of the language sample vary greatly e.g., according to the unit of comparison that is used (e.g., utterances, minutes, tokens, number of elicitation materials), the measures that are calculated (e.g., MLU, D, TTR, FVMC, TAPS, TNW, TDW), the sample sizes that are used in the comparisons (e.g., > 200 vs. 100 utterances, 100 vs. 50/25/12 utterances), the position of the subsamples (e.g., beginning/middle/end of the transcript, transcript split into two halves, even/uneven number of tokens), and the statistical method used to assess measurement reliability (e.g., relative vs. absolute reliability). This diversity complicates clustering and synthesis of results on this component of the LSA process. Nevertheless, seven studies reported a clinically relevant effect of sample size/length on the measures computed in their studies [41-47]. This effect was reported, for example in study [41], with regard to decreased diagnostic accuracy when comparing Finite Verb Morphology Composite (FVMC), and Tense and Agreement Productivity Score (TAPS) in samples of 50 utterances vs. samples of 100 utterance, and in study [42] in terms of the lower reliability of two global lexical measures (i.e., total number of words TNW/m and number of different words NDW/m) as well as mean length of utterances in morphemes (MLU_m) when comparing 3- and 7-minute consecu-

Table 1. Included studies and their research focus according to the five aspects of the LSA process

Author & year		Deciding Length / size	Collecting	Transcribing	Coding	V analyzing					
						Monolingual mainstream english			Other than monolingual mainstream english		
						MD	VM	AA	MD	VM	AA
1	Altenberg & Roberts (2016)							X			
2	Altenberg et al. (2018)				X						
3	Bedore et al. (2010)									U	
4	Bowles et al. (2020)					X					
5	Casby (2011)	X									
6	Eisenberg et al. (2012)					X					
7	Eisenberg & Guo (2013)						D				
8	Eisenberg & Guo (2015)	X									
9	Eisenberg & Guo (2018)	X				X	U				
10	Eisenberg et al. (2018)		X								
11	Gallagher & Hoover (2020)						G				
12	Gatt, Grech & Dodd (2014)						G, U				
13	Gladfelter & Leonard (2013)						D				
14	Guo & Eisenberg (2014)	X					D				
15	Guo & Eisenberg (2015)	X									
16	Hadley et al. (2014)						G				
17	Hadley et al. (2016)						D				
18	Heilmann et al. (2013)	X	X								
19	Hoffman (2013)	X			X	X					
20	Imgrund et al. (2019)						U				
21	Jalilevand et al. (2016)								X		
22	Jean-Baptiste et al. (2018)		X								
23	Justice et al. (2010)					X					
24	Kapantzoglou et al. (2017)		X							D	
25	Kazemi et al. (2015)									D	
26	Klein et al. (2010)		X								
27	Leonard et al. (2017)	X					G				
28	Lubetich & Sagae (2014)							X			
29	MacWhinney et al. (2020)							X			
30	Maillart et al. (2012)								X		X
31	Manning et al. (2020)		X								
32	McKenna & Hadley (2014)					X					
33	Miyata et al. (2013)	X							X	G, U	
34	Ooi & Wong (2012)									D, G, U	
35	Owens & Pavelko (2017)						U				
36	Owens et al. (2018)						G				
37	Pavelko & Owens (2017)		X	X		X	G				
38	Pavelko & Owens (2019)		X	X			D				
39	Pavelko et al. (2020)	X									
40	Peets & Bialystok (2015)		X				U			U	
41	Potapova et al. (2018)									G, U	
42	Qi et al. (2011)									U	
43	Roberts et al. (2020)							X			
44	Santos et al. (2015)									G, U	

(*Continued to the next page*)

Table 1. Continued

Author & year		Deciding Length / size	Collecting	Transcribing	Coding	V analyzing					
						Monolingual mainstream english			Other than monolingual mainstream english		
						MD	VM	AA	MD	VM	AA
45	Smith & Jackins (2014)	X					U				
46	Smolik & Málková (2010)		X						X	U	
47	Soleymani et al. (2016)	X							X		
48	Souto et al. (2014)						D				
49	Stockman (2010)				X						
50	Stockman et al. (2013)								X		
51	Stockman et al. (2016)									U	X
52	Thordardottir et al. (2011)									D	
53	Thordardottir (2016a)	X			X					G	
54	Tomas & Dorofeeva (2019)	X			X					G, U	
55	Tommerdahl & Kilpatrick (2013)	X									
56	van Severen et al. (2012)	X				X					
57	Washington et al. (2019)									U	
58	Wieczorek (2010)				X						
59	Wood et al. (2016)		X								
60	Wong et al. (2010)									D	
61	Zhang & Zhou (2020)								X	G	
	Sum	17	11	2	6	8	17	4	7	17	2

MD, measure development; VM, value of LSA measures (D = diagnostic value, G = growth value, U = use value); AA, automatic analysis.

tive segments to 10-minute consecutive segments or to the whole 22-minute samples. In contrast, five studies [13], [48-51] did not find sample size to have a clinically relevant effect. The shortest reliable sample lengths reported in these five studies were: $\mathrm{MLU}_{\mathrm{m}}$ in several 10 and 20-utterance subsamples of total samples comprising 100–150 utterances in children with developmental language disorder [13] as well as two of three SUGAR metrics ($\mathrm{MLU}_{\mathrm{SUGAR}}$, words per sentence [WPS], clauses per sentence [CPS]) with no statistically significant differences between the two conditions (i.e., 25 vs. 50 utterances) for $\mathrm{MLU}_{\mathrm{SUGAR}}$ or WPS in typically developing children [51].

A third group of studies specified their findings depending on the type of LSA measure with mixed results [53], [54], [55], [56]. Study [53] reported a greater influence of language sample length mostly on additive measures of diversity and productivity, such as number of different words (NDW), type/token ratio (TTR), tense marker total (TMT), and productivity score (PS). MLU or lexical diversity measured by *d*, for example, were less affected [53-55].

Collecting a language sample

Results regarding this component focus on advantages and limitations of specific elicitation conditions. Four subsections related to this component are described.

Effect of different elicitation contexts

Five studies compared different elicitation contexts [4], [56], [18], [57], [58] and all highlight the impact of the elicitation context on LSA. Both study [4] and study [56], the latter for bilingual Spanish–English-speaking children, reported a significant influence of the elicitation context on the majority of the LSA measures included in their studies (DSS sentence point score, DSS pass–fail decisions, *D* for lexical diversity in typically developing children, $\mathrm{MLU}_{\mathrm{w}}$, and sentence complexity). No significant effect was found in the DSS overall score [4], or in the lexical diversity (*D* in children with developmental language disorders) or grammatical errors per communication unit [56]. In both studies, the children tended to score higher in the unstructured elicitation contexts (the former: play vs. picture description; the latter: story telling vs. retelling). This was especially true the younger the sampled children were [18]. Furthermore, when examining mean length of utterance in words ($\mathrm{MLU}_{\mathrm{w}}$) based on transcribed consecutive excerpts of LENA audio files (CEAFs: 50 utterances or 30 min) from monolingual and bilingual (English/Spanish) children, the $\mathrm{MLU}_{\mathrm{w}}$ in both groups of children were lower than expected on the published norms on traditional LSA procedures [68].

Effect of the linguistic content during elicitation

The three studies [53], [59], [60] concerned with the linguistic content of the elicitation (e.g., topics, question types), found contrasting results regarding its influence. Study [53] reported similar performance on LSA measures despite changes in topic (school-related activities vs. non-school activities), while study [60] found an effect of topic familiarity in their discourse tasks (explanation of home routine vs. function of a magnet), with children showing on average better performance in the LSA measures based on the familiar task. Study [59] reported that different types of questions (external state, procedural action, epistemic, causal) provided evidence of obligatory strength for the production of full-sentence responses. When the focus was narrowed to multi-verb (complex) sentences, open-ended external state questions (e.g., “What’s happening?”) elicited significantly more language than any other type of question.

Amount of time to collect samples

Only two studies [10], [26] reported on the amount of time to collect samples. Both indicated that on average, five to six minutes was needed to collect 50 utterance samples in conversational discourse using a standard language sampling protocol designed to promote production of complex language from children of different ages (36–95 months) and with different language status (typically developing vs. developmental language disordered). The average time did not differ according to age or language status.

Feasibility of samples collected via video

Only one study [61] investigated the feasibility, reliability and validity of language samples collected via telepractice. It reported no significant differences between the in-person and the video (telepractice) sample in terms of mean transcript reliability or the LSA measures calculated (percent intelligible child speech, MLU, NDW, TTR, language errors and omissions).

Transcribing

The two studies that addressed this component [10] and [26] both focused on the amount of time taken to transcribe language samples. Both studies showed that the mean time required to orthographically transcribe 50 intelligible child utterances and analyze a language sample for the four metrics they included (MLU_{SUGAR}, WPS, TNW, CPS) was 15.3 minutes. The amount of time did not differ markedly across age groups (36–95 months) or language status, although the children with developmental language disorder produced more unintelligible words.

Coding

Six studies included three subsections related to the coding component, namely:

Specification of coding with respect to individual LSA measures

Four of the six studies [62], [15], [45], [63] focused on LSA measures, mainly MLU. Study [62] presented a revised version of the index of productive syntax (IPSyn-R), which provides more detailed and less ambiguous guidelines for coding. Study [15] showed that the MLU_w and MLU_m in English and French correlated almost perfectly. MLU_w and MLU_m also correlated in a detailed analysis of two English-speaking children’s language trajectory (one typically developing, the other with developmental language disorder) despite some significant differences, thus suggesting that MLU_w and MLU_m reflect two separate types of linguistic development (lexicon vs. morphosyntactic development), which cannot substitute one another [65]. Acceptable agreement was found in Russian between MLUs and MLU_m with researchers concluding that MLU in syllables (with minor adjustments) is the easier-to-code alternative [45].

Identification of utterance boundaries

Only one study [27] investigated the identification of utterance boundaries by different listeners by exploring whether adult listeners could identify utterance boundaries in children’s natural spontaneous speech. Results showed significant differences in the number of utterances identified and in the utterance boundaries which were influenced by different speech characteristics, such as the length and grammatical complexity of the response, turns, pauses, sentence/clause boundaries, response turn shifts and a number of convergent features.

Real-time coding

Only one study [27] attempted to investigate real-time coding by rating 1-minute timed samples in a binary manner as either “acceptable” (structurally complete, grammatically correct, and contextually appropriate without syntactic or semantic errors) or as "incorrect". This coding was done in real time and then compared with traditional 100-utterance samples. Results revealed significant correlations in linguistic data across both forms of language sampling.

Analyzing

The component of LSA that deals with analyzing the data was divided into three subsections as described earlier: a) development of new LSA measures, b) value of LSA measures and c) automated LSA (c). Over 60 different measures were analyzed in the sampled studies (see supplement A). A total of 27 studies (56%) concentrated on monolingual English-speaking children. Of these eight introduced new LSA measures [64], [65], [49], [54], [66], [67], [19], [47], 17 focused on the value of LSA measures (use, growth and diagnostic value) [68], [49], [69], [70], [71], [41], [72], [73], [74], [52], [9], [75], [10], [26], [60], [51], [76], and four on the use of automated LSA [77-80]. A further 21 studies (44%) focused on other children (i.e., non-monolingual mainstream English-speaking children) as shown in Table 1. Of these studies, seven were concerned with new LSA measures or measure adaptation [81], [82], [43], [57], [45], [83], [84], 17 with the value of LSA measures (use, growth, and diagnostic value) [85], [56], [86], [43], [9], [60], [87], [88], [89], [57], [90], [55], [15], [45], [91], [92], [84], and two with automated LSA [82], [90].

Development of new measures

The previously unpublished measures for monolingual mainstream English-speaking children mostly target grammar. For example, the authors of study [65] aimed to develop normative expectations for the level of grammatical accuracy in 3-year-old children, by first introducing a measure of grammaticality, percentage of grammatical utterances including fragments (PGU), defined as C-units containing one or more errors in eliciting picture descriptions with four questions per picture. In Study [67] an approach is proposed for assessing sentence diversity in young children (30 and 36 months) by measuring the unique combinations of grammatical subjects and lexical verbs (USV). To be identified as a USV, a child's sentence is required to include an explicit noun or pronoun in the subject noun phrase (NP) position, to include a lexical verb, and to reflect a sufficiently different subject-verb combination.

Measures were also described for other languages and dialects. For example, for a Persian [83] and a Japanese [43] version of the DSS, a French version of the Language Assessment, Remediation and Screening Procedure (LARSP) [82] and a Czech version for analyzing MLU and mean number of words from certain grammatical categories per utterance [57]. Adaptations included either evaluating the appropriateness of the original English items, possible selection of language-specific equivalents of the original items, or gathering and grouping a completely new set of items based purely on observations of the child's development in the target language. Three additional studies introduced new sets of measures [44], [83], [84]. For example, the authors of study [83] expanded their own previously identified competencies for children speaking African American English to a minimal competence core of morphosyntax (MCC-MS) including MLUm, measures of sentence completeness, sentence clausal complexity, sentence variation (lexical, phrasal, clausal level), and sentence ellipsis.

Value of LSA measures

Table 1 shows that the studies captured different aspects of value (e.g., diagnostic, growth or use value) of existing LSA, with 17 studies each focusing on monolingual children and other children (i.e., non-monolingual English-speaking).

Results on *diagnostic value* regarding classification accuracy are summarized in Table 2. Overall diagnostic accuracy of the examined LSA measures consistently reached at least an acceptable level of >80% in almost all studies including monolingual mainstream English-speaking children. Diagnostic value in studies including other than monolingual mainstream English-speaking children was mixed, with three of the five studies reporting metrics not reaching adequate levels [9], [55], [92].

Study [69] and [52] focused on *growth value* by evaluating change following intervention, both in monolingual mainstream English-speaking children. Study [69] compared two general LSA measures (MLU, TTR) and two intervention-specific measures (percent accuracy, TAPS of the treated morpheme) and concluded that the measures more attuned to the intervention target are most appropriate for grammatical outcome measures from LSA. Study [52] showed that three measures (FVMC, TMT, Productivity Score), all provide unique and relevant information on language growth during intervention. The ability to detect developmental growth was also evaluated for DSS in Japanese [43], MLU for Portuguese [89], several vocabulary measures for Mandarin-Chinese [84] as well as for children in different age groups (e.g., young children: [70], [72], and older children: [10]) or children with different language status (e.g., typical language development:

Table 2. Results on diagnostic classification of LSA measures in the sampled studies

Study #	Author & year	Sample language & age	Results
7	Eisenberg & Guo (2013)	English 36-47 months	Percentage grammatical utterances (PGU): sensitivity: 100%, specificity: 88% DSS sentence point (PSP): sensitivity: 100%, specificity: 82% Percentage verb tense usage (PVT): sensitivity: 100%, specificity: 82%
13	Gladfelter & Leonard (2013)	English younger: 48-54 months; older: 60-66 months)	Finite verb morphology composite (FVMC): sensitivity: 92.31%/100%, specificity 93.33%/100% Tense marker total (TMT): sensitivity: 76.92%/83.33%, specificity: 86.67%/80.00% Productivity score: sensitivity: 66.67%/84.62%, specificity: 86.67%/80.00% TMT + Productivity Score combined: sensitivity: 83.33%/84.62%, specificity: 86.67%/80.00%
14	Guo & Eisenberg (2014)	English 36-47 months	FVMC: sensitivity: 83.0%, specificity: 89.0% Tense and agreement productivity Score (TAPS): sensitivity: 89.0%, specificity: 78.0%
24	Kapantzoglou et al. (2017)	Spanish/English 55 months (mean)	Best combination story retelling: Grammatical errors per communication unit (GE/CU) & lexical diversity (*D*), classification accuracy: 87.5%, sensitivity: 90%, specificity: 85%; LR+: 8.5, LR-: 0.17 Best combination story telling: GE/CU & subordination index (SI), classification accuracy: 82.5%, sensitivity: 85%, specificity: 80%; LR+: 4.25, LR-: 0.19 MLUw did not contribute to the predictive utility of the classification model, once other measures were accounted for
25	Kazemi et al. (2015)	Persian 36-60 months	Measures with good/acceptable sensitivity and specificity: Grammaticality (Sens: 98%, Spec: 84%), MLUw-exc (Sens: 82%, Spec: 98%) and semantic errors (Sens: 92%, Spec: 96%); Measures with LR+ point estimates of 10 or greater: MLUw-exc (45.92), total errors (41.44), MLUm-exc (36.96), semantic errors (24.75), and wrong responses (16.87); Measures with LR– point estimates of less than 0.20: semantic errors (0.09), MLUm and MLUw (0.011), and MLUw-exc (0.18)
34	Ooi & Wong (2012)	Chinese/English 44-81 months	Only IPSyn was found to successfully differentiate children by clinical status (classification accuracy: 77.8%), MLU and *D* did not.
38	Pavelko & Owens (2019)	English 36-95 months	$\text{MLU}_{\text{SUGAR}}$ + clauses per sentence CPS combined: sensitivity: 97.22%, specificity: 82.96%; LR+: 5.71, LR-: 0.03
48	Souto et al. (2014)	English 48-70 month (2 groups: four-year-olds and five-year-olds)	DSS mean tense/agreement developmental score: no group difference (4-year-olds); sensitivity: 64%/82% , specificity: 80% (5-year-olds) Mean of the five highest DSS tense/agreement developmental scores: sensitivity: 71%/14%, specificity: 69%/94% (4-year-olds); sensitivity: 73%/82%, specificity 87% (5-year-olds) FVMC: sensitivity: 93%, specificity: 94%/100% (4-year-olds); sensitivity: 91%/82%, specificity: 93% (5-year-olds) Mean DSS sentence point: sensitivity: 93%/100%, specificity: 94%/100% (4-year-olds); sensitivity & specificity: 100% (5-year-olds) Overall DSS score: sensitivity: 79%/93%, specificity: 94% (4-year-olds); sensitivity: 72%/82%, specificity: 87% (5-year-olds)
52	Thordardottir et al. (2011)	French 49-71 months	MLU (words): sensitivity: 71%, specificity: 71% MLU (morphemes): sensitivity: 71%, specificity: 68%
60	Wong et al. (2010)	Chinese 49-60 months	MLU + *D* + age: Sensitivity: 73.3%, specificity: 57.1%; LR+: 1.71 (95% CI 0.87–3.37), LR-: 0.47

[75], and various language abilities or developmental language disorder: [88], [15]). Most studies (irrespective of their focus on monolingual or other children) reported a correlation between their respective LSA measures and age (vocabulary: [70]; TAP: [72]; MLU_{SUGA}R: [10]; DSS Japanese version: [43]; TAP and TMT: [87]; MLU_{m} and NDW: [88]; MLU_{w}: [89]; TNW, NDW, MLU and morphological diversity: [15]; average number of unique grammatical forms (AvUniqF): [45]; and vocD: [84]). However, some measures did not correlate with age or predict age in non-monolingual mainstream English-speaking children, e.g., MLU [9], T/A accuracy [87], NDW and TTR [84] and were worse in demonstrating growth in older children (e.g., MLU for Russian children older than 3;0 years [46].

In evaluating the validity, feasibility and applicability (*use value*) of LSA, measures were either correlated with standardized assessments, or with other LSA measures; examined for their ability to show group differences in children with various language abilities; or their general appropriateness for other than monolingual mainstream English-speaking children. Two studies addressing the latter aimed to examine whether measures were independent of English dialect or language use ([91]: IPSyn, $\mathrm{MLU_m}$, NDW for English/Jamaican Creole bilingual children), English exposure ([9]: MLU, *D*, IPSyn for English/Chinese bilingual children), or maternal education ([9], [91]), both reporting no effect on the listed possible influences. Several studies reported significant correlations (of various value) between LSA measures and standardized assessment of general language skills ([74], [88], [89]), grammar ([49], [10], [57]), or vocabulary ([70], [57]). Some authors pointed out that although LSA measures are correlated with standardized assessment, the often low numeric values indicate that language sampling measures and standardized assessment tap into relatively different aspects of the same underlying abilities (e.g., expressive language) [88]. Correlation of LSA measures within the same linguistic area were often strong in the published studies: in the Japanese DSS version and $\mathrm{MLU_m}$ [43], in MCC-MS and IPSyn in AAE children [90], in $\mathrm{MLU_m}$ and AvUniqF in Russian [44]. Nevertheless, some authors reported mixed results or low correlation between LSA measures mostly including other than monolingual mainstream English-speaking children (e.g., T/A accuracy, TAP, TMT, MLU, NDW) for Spanish/English bilingual children [89] as well as composite discourse score for English/various languages speaking children [60]. Two studies calculated cross-linguistic correlations to specify connection between measures for bilingual language contexts with positive [91] and mixed results [85].

Automated LSA

Four of the six studies that used automated LSA focused on monolingual mainstream English-speaking children with one focusing on French [82] and one focusing on dialects (e.g., African American English) [90]. IPSyn was the most targeted language measure in five studies [62], [78], [79], [80], [90]. Different software approaches to automatically derive IPSyn scores were proposed, which either involved encoding IPSyn target structures in a language-specific inventory as complex patterns over parse trees [62], or were fully-data driven and used only language-independent feature templates applied to syntactic dependency trees [78]. Results and conclusions derived from studies on the accuracy of machine scoring differ, ranging those that report on acceptable reliability when compared to manual scoring (i.e., a mean point-to-point agreement of around 95%, e.g., IPSyn: [79], [91]) and the French version of the LARSP [82], to those studies that still recommend caution in the clinical application of automated LSA, with point-to-point agreement below 85% [62], [80].

DISCUSSION

The purpose of this review was to collate research produced in the past decade on five components of the LSA process, namely determining the length/size of the sample (I), collecting (II), transcribing (III), coding (IV), and analyzing (V) the language sample with the aim of informing researchers and clinicians about recent methodological advancements and to identify future research needs and potential areas for improvement of LSA methodology. In this discussion we firstly address each of the components separately and reconnect it to the current state of the research evidence outlined in the introduction. Thereafter we identify topics overarching all aspects of the LSA process, and focus on the possibilities of technological support of LSA in the future.

The five components of LSA

Determining the size/length of the language sample (Component I) is a source of variability in LSA. Measures are often calculated as a proportion of the total transcript, or as means, or by calculating the frequency of specific linguistic structures in spontaneous oral language. Therefore, these measures are potentially influenced by the amount of text analyzed. The results in this review illustrate that general consensus on the prerequisite sample size/length in LSA may not be achievable, but that it might be more appropriate to be specified for different types of measures. Evidence is growing that some measures, such as MLU, may be less influenced by sample length than measures that evaluate the linguistic content of the sample in more detail [52]. Nevertheless, the size/length of the language sample is an aspect researchers and clinicians have to be cognizant of when employing LSA as a method.

In terms of *collecting the language sample* (Component II), this review expanded the field by highlighting the effect of elicitation contexts on specific LSA measures, the effect of the language during elicitation (topic, question type), as well as the interaction between these two constructs. This evidence does not resolve the structured vs. unstructured controversy but rather specifies the matching of elicitation contexts and questions asked during elicitation to specific target language/structures aimed to elicit from children during language sampling. The following picture arises: On the one hand preschool children tend to produce longer and more complex sentences and even score better on certain LSA measures (e.g., DSS) if the elicitation context is less structured [4], [56], [18]. This phenomenon is even more noticeable in younger preschool children [18]. On the other hand, the question type that elicits the most complex sentence structures (external state questions), occurred more frequently in the higher structured activities (e.g., barrier games and storytelling) [59]. Therefore, including open-ended "What happened?" type questions into play-based elicitation contexts, might be a way to combine both advantages.

Transcribing the sample (Component III) focused on the amount of time needed to transcribe child language samples, and was only described in two of the included 61 papers, highlighting a paucity of research. Results offer a perspective of improved efficiency of LSA regarding clinical implementation [10], [26]. Furthermore, the singular focus on this aspect related to transcription might be interpreted as a general indicator of satisfaction with existing transcription conventions (e.g., SALT or CLAN software).

However, in contrast to the transcribing component where there appears to be consensus on the procedures and rules, the same is not true for *coding* (Component IV). The results related to coding emphasized the continuing challenge of coding phenomena of spoken language, such as utterances or specific LSA measures (e.g., IPSyn-R), and the high variability of coding specific measures across languages (e.g., MLU). This component of the LSA process is of core importance for the reliability of LSA, because it addresses how basic measures such as MLU are calculated. The results presented in this review illustrate that final consensus related to coding is not reached yet in terms of how to segment utterances in either English, or across other languages when attempting to do cross-linguistic comparisons. In coding it becomes most obvious that spoken language differs from written language and although the text is being analyzed in LSA (i.e., transcribed recordings), the ultimate aim is to evaluate the oral language (spoken language) by means of the written sample. Hence, strategies to enhance and optimize this component should be considered and prioritized. Additionally, researchers and clinicians should be explicit in reporting how they coded their sample and/or specific LSA measures in order to allow comparison or norm use.

The results of this review regarding *analysis* (Component V) expand the breadth and depth of LSA by developing measures for specific purposes, focusing mainly on grammar. In addition, LSA measures are increasingly adapted to target populations other than monolingual mainstream English-speaking children emphasizing the specificity of these populations when utilizing LSA just as in formal assessment with language tests. All these efforts extend the applicability of LSA for example to younger children [65], [67] or to languages other than English [81], [82]. It also attempts to improve the applicability of LSA by refining the guidelines [62] and by developing alternative measures [10]. This review also presented further evidence on the predictive and discriminant validity of LSA measures in several languages/dialects, thereby strengthening the general value of LSA without avoiding its limitations such as demonstrating its growth value after intervention [69] or its challenges such as sampling bias [47]. Results on diagnostic accuracy of LSA measures calculated from spontaneous language samples for preschool children are promising and are not lagging behind those of standardized language tests [93] confirming the utility of LSA for clinical practice. Finally, the two approaches in machine scoring of LSA measures were shown: on the one hand those that continue to use traditional language-specific approaches of computational linguistics [62], and on the other hand those that have adopted recent approaches of purely data-driven potentially language-unspecific machine learning approaches [68]. Improvement in this area is closely related to general advancements in machine learning and to the associated paradigm change from rule-based to data-based applications in automated speech recognition and processing [94]. A fully automated LSA process from transcription to coding and linguistic analysis of the transcript is not offered by any of the existing software or tools yet. Controversy arises in this field on how to compare machine and manual scoring [80].

The overarching topic of automation

In this section we focus on the *possibilities and requirements of technological advancement* in supporting LSA. Other needs that can be clearly derived from the results of this review as well, such as an increase of research on LSA methodology regarding multilingual preschool children (almost 90% of the studies in this review focused on monolingual mainstream English-speaking children or monolingual children speaking other languages or dialects) or the establishment of guidelines tailoring LSA to specific purposes, age groups, language status and language/dialect backgrounds to inform best practice are not addressed in detail (but see e.g., [5]).

Technological support (hardware and software) can be associated to every one of the five components of the LSA process that framed this review, but the central most impacting development in the future will be automatic speech recognition applications that are able to process spontaneous child language leading to automated transcription and coding. Observing the distributed areas of interest in the studies included in this review (Table 1), these parts of the LSA process have yielded the least research in the last decade. Future expansion of automation might change this. Currently, software support is only available for the last component – analyzing the language samples (with the exception of CLAN offering some features for coding) [24], [25]. Tied to the expansion of transcribing and coding automation is the ability to record truly spontaneous language samples in natural settings and analyze their linguistic content. With software support the two most time-consuming and resource intensive steps of the LSA process determining the sample size will become unnecessary. What has to be kept in mind however, is that spontaneous conversation produces a different type of data than samples elicited under structured conditions [57]. This will lead to the development of new LSA measures and require the establishment of specific norms for parameters derived from everyday communication [58]. In the still mostly "manual" present, the following dilemma arises: The more natural (unstructured) the elicitation context the more representative the sampled language, but also the longer it takes to collect (and then transcribe and code) a desired amount of language to calculate LSA measures, because the frequency of targeted structures (e.g., complex language) may be much lower in natural communication [65], [46], [47]. Automatic transcription (orthographic and possibly even phonetic) would reduce the time-consuming aspect of LSA substantially, improve its efficiency and therefore avoid this dilemma. Additionally, automated transcription and coding would strengthen the applicability of LSA and if long-dated, it could even be expanded to fully develop its potential in providing insights into multilingual language learning and how it is used within communicative interaction in natural settings, including children and their caregivers, peers, teachers and other communication partners. In terms of advancements related to the coding and analyzing components of LSA, machine learning opens new possibilities as not only the text data but also audio data (and possibly even video data) will become accessible for exploration. This might lead to an even further expanded applicability of using speech and language samples for diagnostic and research purposes combining parameters derived from transcripts and audio/video recordings [95]. At present we only have tools relying either on the (manually transcribed) text data (CLAN: [24], SALT [25]) or the processed audio data (LENA: [36]). Finally, a fully automated LSA process might enable the evaluation of SLP intervention outcome according to change in communicative participation in everyday situations as ultimately aspired by the conceptual framework of the International Classification of Functioning, Disability, and Health, (ICF) proposed by the World Health Organization (WHO) [96]. To develop these tools, interdisciplinary collaboration (e.g., engineers, SLPs, computer linguists) is necessary. The contribution of SLP is firstly, to inform machine learning models with knowledge on typical and atypical language development on all linguistic levels. Secondly to design and supervise the collection of appropriate data for model training, because as has been stated earlier the shift in machine-learning development to data-based, so called "deep learning", requires large amount of adequate data to train and test the applications for the intended purpose. Lastly, to address ethical implications of automation relevant to the context of assessing the developing communicative abilities of children for clinical and research purposes.

Limitations

Some limitations of this current study must be considered. The first deals with our search and study selection strategy. The present scoping review included only studies that were accessible via the listed databases. Due to the very specific focus of this review (research on LSA methodology) on the one hand and the broad range of topics covered by this focus on the other hand (the five targeted components of the LSA process), deciding on appropriate search terms was challenging. Not all of the studies included "language/speech sample" in their key words, but rather broad descriptors such as "language assessment", which would have significantly enlarged the electronic search, resulting in a large number of excluded studies. Therefore, we decided to rather augment the electronic search with thorough hand searching [97]. Furthermore, we labeled picture naming, imitation of target words and story retelling as elicitation contexts which do not reflect spontaneous language close enough, resulting in the exclusion of several studies. However, it is acknowledged that this decision (especially with regard to story retelling) may not be subject to consensus among the entire scientific community, and the results of these studies might also contain information contributing to research on LSA methodology [33], [98]. The focus of this review is also only on preschool children, therefore, implications drawn might not be applicable across other age groups (i.e., school-age and adolescence). In addition, the exclusion of dissertations and theses that may have evaluated components of LSA methodology could introduce publication bias into the review. Furthermore, the generalizability of the results is negatively impacted by the fact that within some of the five LSA components building our main variables reviewed, there were only a few papers included (e.g., for the transcribing component). Overrepresentation of IPSyn studies in the section on automated LSA may be due to the fact that other studies on automated LSA had technical aims and therefore more mixed groups of participants or lacked information on participant characteristics and thus did not meet the participant inclusion criterion e.g. [99], [100]. Finally, not all studies provided information on all aspects included in the data extraction (resulting in missing data), or reported it in the same way (e.g., software use in different parts of the LSA process). Although it may limit the validity of our results, it also draws attention to the absence of guidelines in how to report the methods and results in LSA studies, limiting opportunities for conducting meta-analysis of data [101].

CONCLUSIONS

This review shares data from 61 studies on LSA methodology published between 2010 and 2020, by synthesizing existing evidence on five components of LSA: determining the sample length/size, collecting, transcribing, coding and analyzing the sample. The review highlighted the current debates in the field and described the breadth of topics covered by the extant literature that reflects the wide range of potential LSA applicability. At the same time, this variability expressed in how studies were conducted and how results were reported, makes it challenging to compare studies, to draw clinical conclusions, and to show implications for future research [102], [10], [80]. Therefore, it is recommended that future research might establish refined guidelines to optimize the potential of LSA for diverse purposes in mono- and multilingual children of different ages and language status. The greatest singular potential in future LSA advancement is seen in technological support expanding to the transcription and coding of child language data, enabling the recording of complex linguistic data in real-life contexts, thus fully redeeming the promise of ecological validity.

SUPPLEMENTS

Sample set results on participant characteristics and study design (supplement A), Criteria for study categorization (supplement B).

CONFLICT OF INTERESTS

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

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

1. Brown R. A first language: The early stages. Cambridge, MA: Harvard University Press; 1973.
2. Miller JF. Assessing language production in children. Baltimore, MD: University Park Press; 1981.
3. Ebert KD, Pham G. Synthesizing information from language samples and standardized tests in school-age bilingual assessment. Language, Speech, and Hearing Services in Schools. 2017;48:42-55.
4. Eisenberg SL, Guo LY, Mucchetti E. Eliciting the language sample for developmental sentence scoring: A comparison of play with toys and elicited picture description. American Journal of Speech-Language Pathology. 2018;27:633-646.
5. Ebert KD. Language sample analysis with bilingual children: translating research into practice. Topics in Language Disorders. 2020; 40:182-201.
6. Gutiérrez-Clellen VF, Simon-Cereijido G. Using language sampling in clinical assessments with bilingual children: Challenges and future directions. Seminars in Speech and Language. 2009;30:234-245.
7. American Speech-Language-Hearing Association. Preferred practice patterns for the profession of speech-language pathology. 2004.Retrieved from https://www.asha.org/policy/PP2004-00191/
8. Heilmann J, Rojas R, Iglesias A, Miller JF. Clinical impact of wordless picture storybooks on bilingual narrative language production: A comparison of the 'Frog' stories. International Journal of Language & Communication Disorders. 2016;51:339-345.
9. Ooi CCW, Wong AMY. Assessing bilingual Chinese-English young children in Malaysia using language sample measures. Interna-

tional Journal of Speech-Language Pathology. 2012;14:499-508.

10. Pavelko SL, Owens RE. Sampling Utterances and Grammatical Analysis Revised (SUGAR): New normative values for language sample analysis measures. Language, Speech, and Hearing Services in Schools. 2017;48:197-215.
11. Pavelko SL, Owens RE, Ireland M, Hahs-Vaughn DL. Use of language sample analysis by school-based SLPs: Results of a nationwide survey. Language, Speech, and Hearing Services in Schools. 2016;47:246-258.
12. Tilstra J, McMaster K. Productivity, fluency, and grammaticality measures from narratives. Communication Disorders Quarterly. 2007; 29:43-53.
13. Casby MW. An examination of the relationship of sample size and mean length of utterance for children with developmental language impairment. Child Language Teaching and Therapy. 2011; 27: 286-293.
14. Heilmann J, Nockerts A, Miller JF. Language sampling: Does the length of the transcript matter? Language, Speech, and Hearing Services in Schools. 2010;41:393-404.
15. Thordardottir E. Grammatical morphology is not a sensitive marker of language impairment in Icelandic in children aged 4-14 years. Journal of Communication Disorders. 2016a;62:82-100.
16. Bornstein MH, Haynes OM, Painter KM, Genevro JL. Child language with mother and with stranger at home and in the laboratory: A methodological study. Journal of Child Language. 2000;27:407-420. 33.Allen SEM, Dench C. Calculating mean length of utterance for eastern Canadian Inuktitut. First Language. 2015;35:377-406.
17. Duinmeijer I, Jong J, Scheper A. Narrative abilities, memory and attention in children with a specific language impairment. International Journal of Language & Communication Disorders. 2012;47: 542-555.
18. Klein HB, Moses N, Jean-Baptiste R. nfluence of context on the production of complex sentences by typically developing children. Language, Speech, and Hearing Services in Schools. 2010;41:289-302.
19. Bornstein M, Painter K, Park J. Naturalistic language sampling in typically developing children. Journal of Child Language. 2002;29: 687-699.
20. Hoff E. Context effects on young children's language use: The influence of conversational setting and partner. First Language. 2010; 30:461-472.
21. Southwood F, Russell AF. Comparison of conversation, freeplay, and story generation as methods of language sample elicitation. Journal of Speech, Language, and Hearing Research, 2004;47:366-376.
22. Roy B, Roy D. Fast transcription of unstructured audio recordings. In International Speech Communication Association (Chair), INTERSPEECH. Symposium conducted at the meeting of the International Speech Communication Association, Brighton. 2009. Retrieved from https://dspace.mit.edu/handle/1721.1/67363
23. Seifert M, Morgan L, Gibbin S, Wren Y. An alternative approach to measuring reliability of transcription in children's speech samples: Extending the concept of near functional equivalence. Folia Phoniatrica Et Logopaedica. 2020;72:suppl;84-91.
24. MacWhinney B. The CHILDES Project: Tools for analyzing talk (3rd. ed.). Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
25. Miller JF, Andriacchi K, Nockerts A. Assessing language production using SALT software. A clinician's guide to language sample analysis. 3rd ed.. SALT Software LLC; 2019.
26. Pavelko SL, Owens RE. Diagnostic accuracy of the Sampling Utterances and Grammatical Analysis Revised (SUGAR) measures for identifying children with language impairment. Language, Speech, and Hearing Services in Schools. 2019;50:211-223.
27. Stockman I. Listener reliability in assigning utterance boundaries in children's spontaneous speech. Applied Psycholinguistics. 2010; 31:363-395.
28. Foster P, Tonkyn A, Wigglesworth G. Measuring spoken language: A unit for all reasons. Applied Linguistics. 2000;21:354-375.
29. Ezeizabarrena MJ, Fernandez IG. Length of utterance, in morphemes or in words?: MLU3-w, a reliable measure of language development in early Basque. Frontiers in Psychology. 2018;8:1-17.
30. Parker MD, Brorson K. A comparative study between mean length of utterance in morphemes (MLUm) and mean length of utterance in words (MLUw). First Language. 2005;25:365-376.
31. Durán P, Malvern D, Richards B, Chipere N. Developmental trends in lexical diversity. Applied Linguistics. 2004;25:220-242.
32. Allen SEM, Dench C. Calculating mean length of utterance for eastern Canadian Inuktitut. First Language. 2015;35:377-406.
33. Kapantzoglou M, Fergadiotis G, Auza BA. Psychometric evaluation of lexical diversity indices in Spanish narrative samples from children with and without developmental language disorder. Journal of Speech, Language, and Hearing Research. 2019;62:70-83.
34. Klee T, Stokes SF, Wong AMY, Fletcher P, Gavin WJ. Utterance length and lexical diversity in Cantonese-speaking Children with and without specific language impairment. Journal of Speech, Language, and Hearing Research. 2004;47:1396-1410.
35. Thordardottir E. Long versus short language samples: A clinical procedure for French language assessment. Canadian Journal of Speech-Language Pathology and Audiology (CJSLPA). 2016b;40: 176-197. Record ID: 1202
36. Long SH, Fey ME, Channell RW. Computerized Profiling (Version 9.0.3-9.2.7) [Computer software]. Cleveland, OH: Department of Communication Sciences, Case; 1996-2000.
37. Gilkerson J, Richards JA. The power of talk. LENA Foundation Technical Report ITR-01-2. Boulder, CO: LENA Foundation. 2009.
38. McKechnie J, Ahmed B, Gutierrez-Osuna R, Monroe P, McCabe P, Ballard KJ. Automated speech analysis tools for children's speech production: A systematic literature review. International Journal of Speech-Language Pathology. 2018;20:583-598.
39. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMAScR): Checklist and explanation. Annals of Internal Medicine. 2018;169:467-473.

40. Ganek H, Eriks-Brophy A. Language ENvironment analysis (LENA) system investigation of day long recordings in children: A literature review. Journal of Communication Disorders, 2018;72:77-85.
41. Guo LY, Eisenberg SL. The diagnostic accuracy of two tense measures for identifying 3-year-olds with language impairment. American Journal of Speech-Language Pathology. 2014;23:203-212.
42. Guo LY, Eisenberg SL. Sample length affects the reliability of language sample measures in 3-year-olds: Evidence from parent-elicited conversational samples. Language, Speech, and Hearing Services in Schools. 2015;46;141-153.
43. Miyata S, MacWhinney B, Otomo K, Sirai H, Oshima-Takane Y, Hirakawa M, et al. Developmental sentence scoring for Japanese. First Language. 2013;33:200-216.
44. Soleymani Z, Nematzadeh S, Tehrani LG, Rahgozar M, Schneider P. Language sample analysis: Development of a valid language assessment tool and determining the reliability of outcome measures for Farsi-speaking children. European Journal of Developmental Psychology. 2016;13:275-291.
45. Tomas E, Dorofeeva S. Mean length of utterance and other quantitative measures of spontaneous Speech in Russian-speaking children. Journal of Speech, Language, and Hearing Research. 2019; 62:4483-4496.
46. Tommerdahl J, Kilpatrick C. The reliability of morphological analyses in language samples. Language Testing. 2013;31:3-18.
47. Van Severen L, van den Berg R, Molemans I, Gillis S. Consonant inventories in the spontaneous speech of young children: A bootstrapping procedure. Clinical Linguistics & Phonetics. 2012;26: 164-187.
48. Eisenberg SL, Guo LY. Sample size for measuring grammaticality in preschool children from picture-elicited language samples. Language, Speech, and Hearing Services in Schools. 2015;46:81-93.
49. Eisenberg SL, Guo LY. Percent grammatical responses as a general outcome measure: Initial validity. Language, Speech, and Hearing Services in Schools. 2018;49:98-107.
50. Pavelko SL, Price LR, Owens RE. Revisiting reliability: Using Sampling Utterances and Grammatical Analysis Revised (SUGAR) to compare 25- and 50-utterance language samples. Language, Speech, and Hearing Services in Schools. 2020;51:778-794.
51. Smith AB, Jackins M. Relationship between longest utterances and later MLU in late talkers. Clinical Linguistics & Phonetics. 2014;28: 143-152.
52. Leonard LB, Haebig E, Deevy P, Brown B. Tracking the growth of tense and agreement in children with specific language impairment: Differences between measures of accuracy, diversity, and productivity. Journal of Speech, Language, and Hearing Research. 2017;60:3590-3600.
53. Heilmann J, DeBrock L, Riley-Tillman TC. Stability of measures from children's interviews: The effects of time, sample length, and topic. American Journal of Speech-Language Pathology. 2013;22:463-475.
54. Hoffman LM. An exploratory study of clinician real-time morphosyntactic judgements with pre-school children. International Journal of Speech-Language Pathology. 2013;15:198-208.
55. Thordardottir E. et al. Sensitivity and specificity of French language and processing measures for the identification of primary language impairment at age 5. Journal of Speech, Language, and Hearing Research. 2011;54:580-597.
56. Kapantzoglou M, Fergadiotis G, Restrepo MA. Language sample analysis and elicitation technique effects in bilingual children with and without language impairment. Journal of Speech, Language, and Hearing Research. 2017;60:2852-2864.
57. Smolík F, Málková G. Validity of language sample measures taken from structured procedures in Czech. Československá Psychologie. 2011;55:448-458.
58. Wood C, Diehm EA, Callender MF. An investigation of language environment analysis measures for Spanish-English bilingual preschoolers from migrant low-socioeconomic-status backgrounds. Language, Speech, and Hearing Services in Schools. 2016;47:123-134.
59. Jean-Baptiste R, Klein HB, Brates D, Moses N. What's happening? And other questions obligating complete sentences as responses. Child Language Teaching and Therapy. 2017;34:191-202.
60. Peets KF, Bialystok E. Academic discourse: Dissociating standardized and conversational measures of language proficiency in bilingual kindergarteners. Applied Psycholinguistics. 2015;36:437-461.
61. Manning BL, Harpole A, Harriott EM, Postolowicz K, Norton ES. Taking language samples home: Feasibility, reliability, and validity of child language samples conducted remotely with video chat versus in-person. Journal of Speech, Language, and Hearing Research. 2020;63:3982-3990.
62. Altenberg EP, Roberts JA, Scarborough HS. Young children's structure production: A revision of the Index of Productive Syntax. Language, Speech, and Hearing Services in Schools. 2018;49:995-1008.
63. Wieczorek R. Using MLU to study early language development in English. Psychology of Language and Communication. 2010;14: 59-69.
64. Bowles RP, Justice LM, Khan KS, Piasta SB, Skibbe LE, Fostera TD. Development of the Narrative Assessment Protocol-2: A Tool for Examining Young Children's Narrative Skill. Language, Speech, and Hearing Services in Schools. 2020;51:390-404.
65. Eisenberg SL, Guo LY, Germezia M. How grammatical are 3-year-olds? Language, Speech, and Hearing Services in Schools. 2012; 43:36-52.
66. Justice LM, Bowles RP, Pence K, Gosse C. A scalable tool for assessing children's language abilities within a narrative context: The NAP (Narrative Assessment Protocol). Early Childhood Research Quarterly. 2010;25:218-234. 78.Altenberg EP, Roberts JA. Promises and pitfalls of machine scoring of the Index of Productive Syntax. Clinical Linguistics & Phonetics. 2016;30:433-448.
67. McKenna MM, Hadley PA. Assessing sentence diversity in toddlers at-risk for language disorders. Perspectives on Language

Learning and Education. 2014;21:159-172.

68. Eisenberg SL, Guo LY. Differentiating children with and without language impairment based on grammaticality. Language, Speech, and Hearing Services in Schools. 2013;44:20-31.

69. Gallagher JF, Hoover JR. Measure what you treat: Using language sample analysis for grammatical outcome measures in children with developmental language disorder. Perspectives of the ASHA Special Interest Groups. 2020;5:350-363.

70. Gatt D, Grech H, Dodd B. Early expressive vocabulary skills: A multi-method approach to measurement. First Language. 2014;34:136-154.

71. Gladfelter A, Leonard LB. Alternative tense and agreement morpheme measures for assessing grammatical deficits during the preschool period. Journal of Speech, Language, and Hearing Research. 2013;56:542-552.

72. Hadley PA, Rispoli M, Holt JK, Fitzgerald C, Bahnsen A. Growth of finiteness in the third year of life: Replication and predictive validity. Journal of Speech, Language, and Hearing Research. 2014;57: 887-900.

73. Hadley PA, Rispoli M, Hsu N. Toddlers' verb lexicon diversity and grammatical outcomes. Language, Speech, and Hearing Services in Schools. 2016;47:44-58.

74. Imgrund CM, Loeb DF, Barlow SM. Expressive language in preschoolers born preterm: Results of language sample analysis and standardized assessment. Journal of Speech, Language, 884 and Hearing Research. 2019;62:884-895.

75. Owens RE, Pavelko SL, Bambinelli D. Moving beyond mean length of utterance: Analyzing language samples to identify intervention targets. Perspectives of the ASHA Special Interest Groups. 2018;3: 5-22.

76. Souto SM, Leonard LB, Deevy P. Identifying risk for specific language impairment with narrow and global measures of grammar. Clinical Linguistics & Phonetics. 2014;28:741-756.

77. Altenberg EP, Roberts JA. Promises and pitfalls of machine scoring of the Index of Productive Syntax. Clinical Linguistics & Phonetics. 2016;30:433-448.

78. Lubetich S, Sagae, K. Data-driven measurement of child language development with simple syntactic templates. In Proceedings of COLING: the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland. 2014. Retrieved from https://www.aclweb.org/anthology/C14-1203.pdf

79. MacWhinney B, Roberts JA, Altenberg EP, Hunter Madison. Improving automatic IPSyn coding. Language, Speech, and Hearing Services in Schools. 2020;51:1187-1189.

80. Roberts JA, Altenberg EP, Hunter M. Machine-scored syntax: Comparison of the CLAN automatic scoring program to manual scoring. Language, Speech, and Hearing Services in Schools. 2020;51: 479-493.

81. Jalilevand N, Kamali M, Modarresi Y, Kazemi Y. The Persian developmental sentence scoring as a clinical measure of morphosyntax in children. Medical Journal of the Islamic Republic of Iran. 2016;30: 435.

82. Maillart C, Parisse C, Tommerdahl J. F-LARSP 1.0: An adaptation of the LARSP language profile for French. Clinical Linguistics & Phonetics. 2012;26:188-198.

83. Stockman I, Guillory B, Seibert M, Boult J. Toward validation of a minimal competence core of morphosyntax for African American children. American Journal of Speech-Language Pathology. 2013;22:40-56.

84. Zhang Y, Zhou J. Building a norm-referenced dataset for vocabulary assessment based on Chinese vocD and word classes. Journal of Chinese Writing Systems. 2020;4:5-17.

85. Bedore LM, Peña ED, Gillam RB, Ho TH. Language sample measures and language ability in Spanish-English bilingual kindergarteners. Journal of Communication Disorders. 2010;43:498-510.

86. Kazemi Y, Klee T, Stinger H. Diagnostic accuracy of language sample measures with Persian-speaking preschool children. Clinical Linguistics & Phonetics. 2015;29:304-318.

87. Potapova I, Kelly S, Combiths PN, Pruitt-Lord SL. Evaluating English morpheme accuracy, diversity, and productivity measures in language samples of developing bilinguals. Language, Speech, and Hearing Services in Schools. 2018;49:260-276.

88. Qi CH, Kaiser AP, Marley SC, Milan S. Performance of African American preschool children from low-income families on expressive language measures. Topics in Early Childhood Special Education. 2012;32:175-184.

89. Santos ME, Lynce S, Carvalho S, Cacela M, Mineiro A. Mean length of utterance-words in children with typical language development aged 4 to 5 years. Revista CEFAC. 2015;17:1143-1151.

90. Stockman IJ, Newkirk-Turner BL, Swartzlander E, Morris LR. Comparison of African American children's performances on a minimal competence core for morphosyntax and the Index of Productive Syntax. American Journal of Speech-Language Pathology. 2016;25:80-96.

91. Washington KN, Fritz K, Crowe K, Kelly B, Wright Karem R. Bilingual preschoolers' spontaneous productions: Considering Jamaican Creole and English. Language, Speech, and Hearing Services in Schools. 2019;50:179-195.

92. Wong AMY, Klee T, Stokes SF, Fletcher P, Leonard LB. Differentiating Cantonese-Speaking preschool children with and without SLI using MLU and lexical diversity (D). Journal of Speech, Language, and Hearing Research. 2011;53:794-799.

93. Spaulding T, Plante E, Farinella KA. Eligibility criteria for language impairment. Language, Speech, and Hearing Services in Schools. 2006;37:61-72.

94. Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, et al. Deep Speech: Scaling up end-to-end speech recognition. 2014.

95. Richards JA, Xu D, Gilkerson J. Development and performance of the LENA automatic autism screen; LTR-10-1. Hardware Model: LR-0121 Software Version: V3.1.0; Lena Foundation: Boulder, CO, USA, 2010. Available online: https://www.lena.org/wp-content/uploads/2016/07/LTR-10-1-LENA-Automatic-Autism-Screen.pdf

96. Cunningham BJ, Washington KN, Binns A, Rolfe K, Robertson B, Rosenbaum P. Current Methods of Evaluating Speech-Language Outcomes for Preschoolers With Communication Disorders: A Scoping Review Using the ICF-CY. Journal of Speech, Language,

and Hearing Research. 2017;60:447-464.

97. Richards D. Handsearching still a valuable element of the systematic review. Evidence-Based Dentistry. 2008;9:85.

98. Westerveld MF, Vidler K. The use of the Renfrew Bus Story with 5-8-year-old Australian children. International Journal of Speech-Language Pathology. 2015;17:304-313.

99. 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 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2018;4909-4913.

100. Morley E, Hallin AE, Roark B. Challenges in automating maze detection. In Association for Computational Linguistic (Chair), Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Real. Maryland, USA: Symposium conducted at the meeting of Association for Computational Linguistics, Baltimore; 2014.

101. Finestack LH, Payesteh B, Disher JR, Julien HM. Reporting child language sampling procedures. Journal of Speech, Language, and Hearing Research. 2014;57:2274-2279.

102. Heilmann J, DeBrock L, Riley-Tillman TC. Stability of measures from children's interviews: The effects of time, sample length, and topic. American Journal of Speech-Language Pathology. 2013;22:463-475.